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

On Approximations for Queues, III: Mixtures of Exponential Distributions

By W. WHITT*

(Manuscript received May 3, 1983)

To evaluate queueing approximations based on a few parameters (e.g., the first two moments) of the interarrival-time and service-time distributions, we examine the set of all possible values of the mean queue length given this partial information. In general, the range of possible values given such partial information can be large, but if in addition shape constraints are imposed on the distributions, then the range can be significantly reduced. The effect of shape constraints on the interarrival-time distribution in a GI/M/1 queue was investigated in Part II (see "On Approximations for Queues, II: Shape Constraints," this issue) by restricting attention to discrete probability distributions with probability on a fixed finite set of points and then solving nonlinear programs. In this paper we show how one kind of shape constraint—assuming that the distribution is a mixture of exponential distributions—can be examined analytically. By considering GI/G/1 queues in which both the interarrival-time and service-time distributions are mixtures of exponential distributions with specified first two moments, we show that additional information about the distributions is more important for the interarrival time than for the service time.

I. INTRODUCTION AND SUMMARY

Many approximations for the mean steady-state queue length in the GI/G/1 queue are based on the first two moments of the general interarrival-time and service-time distributions. To evaluate these approximations, it is natural to compare the approximations with the set of possible values of the mean queue length given this limited

* AT&T Bell Laboratories.

moment information. For several special cases, the minimum and maximum values of the mean queue length are attained by simple two-point extremal distributions. In Part I the extremal distributions were used to calculate the extreme values of the mean queue length in the GI/M/1 queue and show how they depend on the traffic intensity, the second moment of the interarrival-time distribution, and an upper bound on the distribution.[1] Extremal distributions were also used to compare different parameters for approximations.

Unfortunately, the range of possible values of the mean queue length in the GI/M/1 queue given this limited moment information can be very wide. However, since the extremal interarrival-time distributions are quite unusual, this still leaves the possibility that the range would not be too wide for typical distributions. Part II showed that the range of possible values for the mean queue length in the GI/M/1 queue can indeed be reduced dramatically by imposing shape constraints such as unimodality and log-convexity on the interarrival-time distributions with given first two moments.[2] This was done by restricting attention to discrete distributions with all mass on a fixed finite set of points and solving nonlinear programs.

Unlike Part I, the approach in Part II was computational, based on nonlinear programs. However, the extremal distributions obtained from the nonlinear programs exhibit regularity that suggests the possibility of an analytic treatment similar to Part I. This paper sets out to treat analytically one kind of shape constraint. We show that the theory underlying Part I also applies to mixtures of exponential distributions. Within this class of distributions there are extremal distributions with respect to the same partial orderings based on Laplace transforms used in Part I. The extremal distributions in this class of mixtures are obtained by using the extremal distributions of Part I as the mixing distributions. These extremal distributions yield the minimum and maximum mean queue length as interarrival-time distributions in the GI/M/I queue and as service-time distributions in the K_2/G/1 queue (with interarrival-time distributions having a rational Laplace-Stieltjes transform with a denominator of degree 2, see Section V of Part I and Section VII here).

The rest of this paper is organized as follows. In Section II, we briefly review the theory yielding distributions that minimize or maximize the Laplace-Stieltjes transforms for all arguments. In Section III, we show that this theory applies to mixtures of exponential distributions, and in Section IV, we apply the results to the H/M/1 queue, having interarrival-time distributions that are mixtures of exponential distributions. In Section V, we examine the case of H_2 interarrival-time distributions (mixtures of two exponential distributions) in more detail. In Section VI, we indicate how some of the

results for H/M/1 queues extend to GI/G/1 queues with interarrival-time having increasing mean residual life (the service-time distribution is general instead of exponential and the interarrival-time distribution can be more general than a mixture of exponentials). Finally, in Section VII, we indicate how the ordering of transforms can be applied to compare different service-time distributions in K_2/H/1 queues. There, Table III gives a good picture of the way the mean queue length can vary in a large class of GI/G/1 queues with the first two moments of the interarrival time and the service time specified.

II. EXTREME VALUES OF THE LAPLACE-STIELTJES TRANSFORM

As in Eckberg[3] and references there, we obtain the extremal distributions for queues with specified moments for the interarrival times and service times from extremal distributions for the Laplace-Stieltjes transform. For the transform, the object is to find a cdf (cumulative distribution function) F with support on the interval $[0, bm_1]$, $b \leq \infty$, to minimize or maximize the transform $\phi(s)$, defined by

$$\phi(s) = \int_0^\infty e^{-st} dF(t), \qquad s \geq 0, \tag{1}$$

subject to moment constraints

$$m_j = \int_0^\infty t^j dF(t), \tag{2}$$

for $j = 1, 2, \cdots, n$. The key idea is to apply the theory of Tchebycheff systems in Karlin and Studden,[4] which implies that the optimization problem involving (1) and (2) has a very nice solution. First, the minimizing and maximizing cdf's are independent of the variable s in the transform $\phi(s)$. Second, the extremal distributions are discrete distributions with positive mass on at most $(n + 2)/2$ mass points. Finally, the points with positive mass and the associated probability masses are obtained simply by solving a system of linear equations. (See Section 2 of Eckberg[3] and Section II of Part I[1] for more discussion.)

III. MIXTURES OF EXPONENTIAL DISTRIBUTIONS

Now we consider the optimization problem in Section II for distributions that are mixtures of exponential distributions. It turns out that the theory of Tchebycheff systems can be applied again because the extremal distributions in this class of mixtures can be obtained by using extremal mixing distributions.

A cdf F is a mixture of exponential distributions if it satisfies

$$1 - F(x) = \int_0^\infty e^{-x/t} dG(t), \qquad x \geq 0, \tag{3}$$

for some mixing cdf G. Densities of mixtures of exponential distributions are also called completely monotone; see Section 5.4 of Keilson.[5] A density f has this property if and only if it has derivatives $f^{(n)}$ of all orders n and $(-1)^n f^{(n)}(x) \geq 0$ for all x and n. Mixtures of exponentials are log-convex (see Part II) and thus are DFR (have decreasing failure rate).

It turns out that the moments and transform of F are easily expressed via G:

$$m_k(F) = \int_0^\infty t^k dF(t) = k! \int_0^\infty t^k dG(t) = k! m_k(G) \tag{4}$$

and

$$\phi(s) = \int_0^\infty e^{-sx} dF(x) = \int_0^\infty (1 + st)^{-1} dG(t). \tag{5}$$

Moreover, the functions $1, t, 2t^2, \cdots, (k!)t^k, (1 + st)^{-1}$ form a complete Tchebycheff system, so extremal distributions F within the class of mixtures are obtained by using the associated extremal mixing cdf's G. If the first n moments of F are specified as $m_1, m_2, \cdots, m_n$, then the first n moments of G are $m_1, m_2/2, \cdots, m_n/n!$

First suppose that the two moments of F are specified as m_1 and m_2. Let c^2 be the squared coefficient of variation of F, i.e., $c^2 = (m_2 - m_1^2)/m_1^2$. Also require that the mixing cdf G has support on the interval $[0, bm_1]$, $b < \infty$. Then the extremal distributions are:

1. Upper bound—the two-point mixture with mass $(c^2 - 1)/(c^2 + 1)$ on 0 and mass $2/(c^2 + 1)$ on the exponential distribution with mean $m_1(1 + c^2)/2$, which has cdf

$$F_u(x) = 1 - [2/(1 + c^2)]e^{-2x/m_1(1+c^2)}, \qquad x \geq 0, \tag{6}$$

and

2. Lower bound—the mixture of two exponential distributions, one having mean bm_1 with probability $(c^2 - 1)/(c^2 - 1 + 2(b - 1)^2)$ and the other having mean $m_1[1 - (c^2 - 1)/2(b - 1)]$ with probability $2(b - 1)^2/(c^2 - 1 + 2(b - 1)^2)$; the cdf is

$$\begin{aligned} F_\ell(x) = 1 - [c^2 - 1 + 2(b - 1)^2]^{-1}\{&(c^2 - 1)e^{-bm_1 x} \\ &+ 2(b - 1)^2 e^{-m_1[1 - (c^2 - 1)/2(b - 1)]x}\}, \quad x \geq 0. \end{aligned} \tag{7}$$

As $b \to \infty$, the lower bound approaches (converges in law) to

3. Limiting lower bound—the exponential distribution with mean m_1, having cdf $F_\ell(x) = 1 - e^{-m_1 x}$, $x \geq 0$.

The upper bound cdf F_u may not be considered a mixture of exponential distributions because of the atom at 0, but the atom at 0 can be thought of as an exponential distribution having mean 0. Alternatively, F_u can be realized as the limit in distribution of mixtures of two exponential distributions having means λ_1^{-1} and λ_2^{-2} and proper moments where $\lambda_1^{-1} \to 0$ and $\lambda_2^{-1} \to m_1(1 + c^2)/2$.

Let $\phi_{\hat{\ell}}(s)$, $\phi_\ell(s)$, and $\phi_u(s)$ be the transforms of the extremal cdf's $F_{\hat{\ell}}$, F_ℓ, and F_u, respectively. The theory of Tchebycheff systems implies that

$$\phi_{\hat{\ell}}(s) \leq \phi_\ell(s) \leq \phi(s) \leq \phi_u(s) \tag{8}$$

for all s and the transforms ϕ of cdf's F of the form (3) having first two moments m_1 and m_2.

Remark: It is no doubt possible to study extremal distributions for other kinds of mixtures, but we have not. Mixtures of exponentials seem particularly appropriate for the queueing application.

IV. THE H/M/1 QUEUE

The results of Section III apply immediately to GI/M/1 queues in which the interarrival-time distribution is a mixture of exponential distributions; see Section II of Part I. Since the mixture of k exponential distributions is called hyperexponential and is denoted by H_k, we use H to refer to interarrival-time distributions that are general mixtures of exponentials.

Note that the upper bound cdf F_u in (6) as an interarrival-time distribution corresponds to a batch Poisson arrival process with geometrically distributed batches having mean $m_B = (1 + c^2)/2$ and squared coefficient of variation $c_B^2 = (m_B - 1)/m_B$. Let M^B represent a batch Poisson arrival process. Of course, the limiting lower bound corresponds to a Poisson arrival process with intensity $1/m_1$. What we obtain is the ordering

$$\text{M/M/1} \leq \text{H/M/1} \leq \text{M}^\text{B}\text{/M/1}, \tag{9}$$

which means that the mean queue lengths (expected number in the system, including any in service) are ordered and in fact the entire steady-state queue-length distributions are stochastically ordered as in (9), provided the traffic intensity ρ and the squared coefficient of variation of the interarrival-time distribution, c^2, are fixed. We obtain these orderings because in the case of exponential service-time distributions the entire steady-state queue-length distribution depends only on the traffic intensity ρ, which is fixed, and the root σ in the interval (0, 1) of the equation

$$\phi[\mu(1 - \sigma)] = \sigma. \tag{10}$$

It is easy to see that the queue-length distributions $P(Q_i \leq k)$ are stochastically ordered, i.e.,

$$P(Q_1 \geq k) \leq P(Q_2 \geq k) \quad \text{for all} \quad k \geq 0 \tag{11}$$

if the roots satisfy $\sigma_1 < \sigma_2$. Moreover, it is easy to see that the roots are ordered if the transforms are ordered in the sense (8).

Let σ_w and L_w be the probability of delay and mean queue length in the H/M/1 queue with interarrival-time distribution F_w, and similarly for F_ℓ and $F_{\hat{\ell}}$. Here are the main results:

Theorem 1: For an H/M/1 queue with traffic intensity ρ and interarrival-time squared coefficient of variation c^2,

$$\sigma_{\hat{\ell}} = \rho \quad \textit{and} \quad \sigma_w = 1 - 2(1 - \rho)/(1 + c^2), \tag{12}$$

so that

$$L_{\hat{\ell}} = \rho/(1 - \rho), \qquad L_w = L_{\hat{\ell}}(1 + c^2)/2 \tag{13}$$

and the maximum relative error (MRE) is

$$MRE \equiv (L_w - L_{\hat{\ell}})/L_{\hat{\ell}} = (\sigma_w - \sigma_{\hat{\ell}})/(1 - \sigma_w) = (c^2 - 1)/2. \tag{14}$$

Proof: Since $\sigma = \rho$ for an M/M/1 queue, $\sigma_{\hat{\ell}} = \rho$. For σ_w, follow the proof of Theorem 2 in Part I, making the change of variables $(1 - \sigma_{\hat{\ell}}) = (1 - \sigma_w)(1 + c^2)/2$.

From Corollary 1 of Part I and Theorem 1, we see that the shape constraint reduces the maximum relative error from c^2 to $(c^2 - 1)/2$. If c^2 is near its lower limit 1 for mixtures of exponentials, then of course the MRE is very small.

Given the first two moments, the upper bound is hard and the lower bound is soft: The upper bound depends on c^2; the lower bound does not. The upper bound is not improved by specifying the third moment; the lower bound is. From Section IV of Part I, we see that the extremal distributions given three moments are two-point mixtures of exponentials:

Theorem 2: For H/M/1 queues, specifying the third moment of the interarrival-time distribution in addition to the first two does not change the upper bound cdf F_w and makes the lower bound cdf $F_{\hat{\ell}}$ the unique H_2 distribution (two-point mixing distribution) specified by these three parameters.

The formula for calculating H_2 parameters given the first three moments is given in (3.5) and (3.6) of Ref. 6.

Example 1: Consider an interarrival-time distribution with moments $m_1 = 2.00$, $m_2 = 12.00$, and $m_3 = 119.01$, which are the moments of Prototype Distribution I in Part II. With mixtures of exponential distributions, the upper bond cdf is

$$F_u(x) = 1 - 0.6667e^{-.3333x}, \qquad x \geq 0,$$

and the lower bound cdf is

$$F_\ell(x) = 1 - 0.5146e^{-0.2964x} + 0.4854e^{-1.8386x}, \qquad x \geq 0.$$

From Theorem 1, given just the first two moments, $\sigma_\ell = 0.6667$ and $\sigma_u = 0.7778$ for $\rho = 0.6667$ and $\sigma_\ell = 0.9000$ and $\sigma_u = 0.9333$ for $\rho = 0.9000$. From Theorem 2, also specifying the third moment changes the lower bound to $\sigma_\ell = 0.76705$ for $\rho = 0.6667$ and $\sigma_\ell = 0.93259$ for $\rho = 0.9000$. To get these, we solved the appropriate $H_2/M/1$ queue. Imposing the shape constraint in addition to the first two moments reduced the MRE from $c^2 = 2.0$ to $(c^2 - 1)/2 = 0.50$. Also specifying the third moment further reduces the MRE to 0.048 when $\rho = 2/3$ and 0.011 when $\rho = 9/10$.

V. MIXTURES OF TWO EXPONENTIALS: H_2 DISTRIBUTIONS

Mixtures of two exponential distributions, i.e., H_2 distributions, play a key role in many approximations. This is a three-parameter distribution with density

$$f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}, \qquad x > 0, \tag{15}$$

where $p_2 = 1 - p_1$. Instead of the three parameters p_1, λ_1, and λ_2, one may choose to work with the first three moments m_1, m_2, and m_3 or the mean m_1, the squared coefficient of variation c^2, and the proportion of the total mean in the component with the smaller mean r, defined by

$$r = \frac{p_1/\lambda_1}{(p_1/\lambda_1) + (p_2/\lambda_2)}, \tag{16}$$

where $\lambda_1 > \lambda_2$. Given the parameters p_1, λ_1, and λ_2, it is easy to calculate any of the other parameters. The formulas for p_1, λ_1, and λ_2 given the first three moments appear in (3.5) and (3.6) of Ref. 6. Given m_1, c^2, and r, $m_2 = m_1^2(c^2 + 1)$, $p_1 = rm_1\lambda_1$, $\lambda_2 = (1 - rm_1\lambda_1)/(1 - r)m_1$ and

$$\lambda_1 = (-B + \sqrt{B^2 - 4AC})/2A, \tag{17}$$

where $A = rm_1m_2/2$, $-B = (m_2/2) + (rm_1)^2 - (1 - r)^2m_1^2$, and $C = rm_1$.

For two-moment approximations based on H_2 distribution, one of the three parameters is often eliminated by setting $r = 1/2$; see Section 3.1 of Ref. 6. The range of all possible values given the first two moments is indicated in Section IV since both the upper and lower bounds are H_2 distributions. Since this range is pretty wide, it is natural to ask how the distribution and the GI/M/1 queue characteristics vary with the third parameter—either r or m_3. For what values of r is the approximation by $r = 1/2$ reasonable?

In order to answer this question, we have calculated the third moment m_3 and the queue characteristics σ and L for two values of c^2 (2 and 12), three values of ρ (0.3, 0.7, and 0.9), and thirteen values of r (ranging from 0.001 to 0.999). The results appear in Tables I and II.

For $c^2 = 2.0$, the approximation by $r = 1/2$ appears quite robust. For r in the interval [0.2, 0.8], the maximum relative error is 15.8

Table I—The possible third parameters and queue characteristics for an $H_2/M/1$ queue given $c^2 = 2.0$ with $\rho = 0.3$, 0.7, and 0.9

Proportion of Total Mean in Component With Smaller Mean, r	Skewness, Third Moment m_3/m_1^3	Key Root, Probability of Delay σ			Mean Queue Length, L		
		$\rho = 0.3$	$\rho = 0.7$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.7$	$\rho = 0.9$
Upper bound	13.5	0.5333	0.8000	0.9333	0.643	3.500	13.500
0.001	13.5	0.5323	0.7999	0.9333	0.641	3.499	13.499
0.01	13.6	0.5230	0.7992	0.9333	0.629	3.486	13.486
0.10	14.6	0.4627	0.7933	0.9327	0.558	3.386	13.381
0.20	15.4	0.4280	0.7885	0.9323	0.525	3.309	13.291
0.30	16.2	0.4059	0.7842	0.9319	0.505	3.244	13.210
0.40	17.1	0.3894	0.7801	0.9314	0.491	3.183	13.127
0.50	18.0	0.3757	0.7757	0.9310	0.481	3.121	13.036
0.60	19.2	0.3633	0.7707	0.9304	0.471	3.053	12.924
0.70	20.9	0.3512	0.7643	0.9295	0.462	2.970	12.771
0.80	23.9	0.3382	0.7552	0.9281	0.453	2.860	12.522
0.90	32.1	0.3226	0.7394	0.9248	0.443	2.686	11.966
0.99	167.9	0.3029	0.7065	0.9074	0.430	2.385	9.715
0.999	1518.0	0.3003	0.7007	0.9009	0.429	2.339	9.080
Lower bound	∞	0.3000	0.7000	0.9000	0.429	2.333	9.000

Table II—The possible third parameters and queue characteristics for an $H_2/M/1$ queue given $c^2 = 12.0$ with $\rho = 0.3$, 0.7, and 0.9

Proportion of Total Mean in Component With Smaller Mean, r	Skewness, Third Moment m_3/m_1^3	Key Root, Probability of Delay σ			Mean Queue Length, L		
		$\rho = 0.3$	$\rho = 0.7$	$\rho = 0.9$	$\rho = 0.3$	$\rho = 0.7$	$\rho = 0.9$
Upper bound	253.5	0.8923	0.9539	0.9846	2.789	15.17	58.50
0.001	253.8	0.8921	0.9538	0.9846	2.779	15.16	58.49
0.01	256.1	0.8897	0.9536	0.9846	2.721	15.10	58.43
0.10	280.7	0.8590	0.9516	0.9844	2.128	14.48	57.80
0.20	312.6	0.8006	0.9488	0.9842	1.505	13.68	56.99
0.30	351.6	0.7114	0.9451	0.9839	1.040	12.74	56.01
0.40	401.2	0.6142	0.9396	0.9836	0.778	11.59	54.76
0.50	468.0	0.5311	0.9311	0.9831	0.640	10.16	53.10
0.60	565.1	0.4650	0.9163	0.9823	0.561	8.36	50.73
0.70	722.7	0.4120	0.8881	0.9808	0.510	6.25	47.00
0.80	1031.7	0.3686	0.8380	0.9776	0.475	4.32	40.20
0.90	1946.0	0.3319	0.7708	0.9646	0.449	3.05	25.39
0.99	18,287.	0.3030	0.7070	0.9089	0.430	2.39	9.88
0.999	181,638.	0.3003	0.7007	0.9009	0.429	2.34	9.08
Lower bound	∞	0.3000	0.7000	0.9000	0.429	2.33	9.00

percent, 15.7 percent, and 6.1 percent for $\rho = 0.3$, 0.7, and 0.9. Very large values of r greatly extend the range.

On the other hand, for very large values of c^2 such as 12, the approximation by $r = 1/2$ is not robust: two moments do not pin down the H_2 distribution well. Using $r = 1/2$ as an approximation works better as ρ increases and c^2 decreases. Of course, by Theorem 1, ρ plays no role in the MRE over all r, but if we bound r, then ρ plays a role. We interpret these results as providing support for H_2 approximation based on $r = 1/2$, but large values of m_2 or m_3 are clear danger signals.

Example 2: Example 1 was based on Prototype Distribution I from Part II. Since Prototype I is a discrete probability mass function it is not a mixture of exponential distributions, and is thus not entirely satisfactory. Suppose we use the H_2 density with balanced means ($r = 0.5$) as a prototype instead. With $m_1 = 1$ and $c^2 = 2.0$, the prototype H_2 density is

$$f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}, \qquad x \geq 0,$$

where

$$p_1 = 0.78867, \ \lambda_1 = 1.577, \text{ and } \lambda_2 = 0.42265.$$

Given the first two moments with $\rho = 0.7$ and 0.9, σ_{ω} can be obtained from Table II of Part I and $\sigma_{\hat{\ell}}$ can be obtained from Theorem 2 there. The values are $\sigma_{\hat{\ell}} = 0.466$ and $\sigma_{\hat{\omega}} = 0.822$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.808$ and $\sigma_{\hat{\omega}} = 0.936$ for $\rho = 0.9$. The corresponding extremal characteristics among H_2 densities are $\sigma_{\hat{\ell}} = 0.700$ and $\sigma_{\omega} = 0.800$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.900$ and $\sigma_{\omega} = 0.933$ for $\rho = 0.9$.

The third moment 18.0 (see Table I) pins down the H_2 distribution, but among all H densities it is a lower bound. Among H densities with $m_3 = 18.0$, $\sigma_{\ell} = 0.7757$ for $\rho = 0.7$ and $\sigma_{\hat{\ell}} = 0.9310$ for $\rho = 0.9$. The MRE given only two moments is 200 percent for $\rho = 0.7$ and 0.9. Working with mixtures of exponentials reduces the MRE to 50 percent. Specifying the third moment too reduces the MRE to 12 percent for $\rho = 0.7$ and 3 percent for $\rho = 0.9$.

VI. THE H/G/1 QUEUE

The assumption of exponential service-time distributions played a crucial role in Section IV. With exponential service-time distributions, the mean queue length L depends on the transform of the interarrival-time distribution, so that we can apply the ordering in (8). However, it turns out that the ordering in (9) also applies for the mean queue length with general service-time distributions, i.e., we have

$$\mathrm{M/G/1} \leq \mathrm{H/G/1} \leq \mathrm{M^B/G/1}, \tag{18}$$

by which we mean that $L_{\hat{\ell}} \leq L \leq L_{\hat{u}}$ (but not the more general stochastic order) for all systems with common service-time distribution and given first two moments of the interarrival-time distribution.

To obtain (18), it suffices to observe that known formulas for L in the $M^B/G/1$ and $M/G/1$ systems agree with previously established lower and upper bounds for L in $GI/G/1$ queues having interarrival-time distributions with increasing mean residual life and with the first two moments of the interarrival times and service times specified. (See Ref. 7 for more details.) This result dramatically demonstrates that these papers have applicability beyond the special case of the $GI/M/1$ model.

VII. THE $K_2/H/1$ QUEUE

Whenever the interarrival-time distribution or the service-time distribution in a $GI/G/1$ queue has a Laplace-Stieltjes transform that is a rational function, then the steady-state distribution can be characterized in terms of the roots of an equation involving the transforms of the interarrival-time and service-time distributions; see II.5.10,11 of Cohen.[8] When the interarrival-time distribution has a rational transform with a denominator of degree 2, denoted by K_2, the mean queue length and the probability of delay depend on the service-time distribution only through its first two moments and a single root of an equation involving the transforms of the interarrival-time and service-time distributions; see p. 330 of Cohen,[8] Section V of Part I,[1] and Ref. 9.

Hence, for $K_2/G/1$ queues it is possible to find extremal service-time distributions using the ordering of transforms in (8). Let GE_2 denote the convolution of two exponential distributions (an Erlang, E_2, is a special case), which is K_2. An H_2 distribution is also K_2. Paralleling Section V of Part I, we obtain from the analysis in Ref. 9 that

$$GE_2/M^B/1 \leq GE_2/H/1 \leq GE_2/\hat{M}/1 \tag{19}$$

and

$$H_2/\hat{M}/1 \leq H_2/H/1 \leq H_2/M^B/1, \tag{20}$$

by which we mean that the mean queue lengths are ordered as indicated. A significant feature of (19) and (20) is that the maximizing distributions are different for the different K_2 interarrival-time distributions. (This is explained in Ref. 9.) By $\hat{M}$, we mean the extremal service-time distribution $F_{\hat{\ell}}$ for large b. As $b \to \infty$, the distribution approaches the exponential distribution, but the fixed variance of $F_{\hat{\ell}}$ is lost in the limit. As $b \to \infty$, the key root in the equation for the $K_2/G/1$ queue approaches the root for the $K_2/M/1$ queue, but the

mean queue length also depends on the variance of $F_{\nearrow}$. The mean queue length in the $K_2/\hat{M}/1$ system is the limit as $b \to \infty$ of the mean queue length in the $K_2/G/1$ system with service-time distributions $F_{\nearrow}$. This limiting mean queue length can be computed by using the fixed service-time variance together with the root for the $K_2/M/1$ system.[8,9]

As in Section V, if we specify three service-time moments instead of two, the M^B bound is unchanged, but the $\hat{M}$ bound is replaced by the H_2 distribution uniquely determined by the three moments, i.e., with the interarrival-time distribution and *three moments* of the service time specified, we get

$$GE_2/M^B/1 \leq GE_2/H/1 \leq GE_2/H_2/1 \tag{21}$$

and

$$H_2/H_2/1 \leq H_2/H/1 \leq H_2/M^B/1. \tag{22}$$

We conclude by exhibiting the mean queue length, L, for several $K_2/H_2/1$ queues. We consider five different H_2 service-time distributions with a common mean 0.7 and a common squared coefficient of variation $c_s^2 = 2.0$. (We use subscripts "s" and "a" to indicate that parameters are associated with the service-time distribution or the interarrival-time distribution.) As in Section V, the H_2 distributions are characterized by the parameter r_s. We consider distributions close to the two extremal distributions M^B ($r_s = 0.01$) and $\hat{M}$ ($r_s = 0.99$), as well as the intermediate values $r_s = 0.1$, 0.5, and 0.9. The case $r_s = 1.0$ differs from the exponential distribution because the small mass at a large value, necessary to have $c^2 = 2.0$ instead of 1.0, still has an effect. (This is not the case for the H_2 interarrival-time distributions.)

We consider six interarrival-time distributions: the same five H_2 distributions and the Erlang (E_2) distribution. All the interarrival-time distributions have mean 1.0, so that the traffic intensity is always $\rho = 0.7$. As with the service-time distributions, the H_2 interarrival-time distributions have squared coefficient of variation $c_a^2 = 2.0$.

The results for the 30 cases are displayed in Table III. For the extremal H interarrival-time distributions, M^B and M, the mean queue length, L, does not depend on r_s because L depends on the service-time distribution only through its first two moments.[7] The range of L values over r_s increases for H_2 interarrival-time distributions as r_a moves away from the endpoints 0.0 and 1.0. The range is bigger for $c_a^2 = 2.0$ (H_2) than for $c_a^2 = 0.5$ (E_2) when $r_a = 0.5$, but obviously not for all r_a.

Table III gives an indication of the quality of two-moment approximations for GI/G/1 queues when $c_a^2 = c_s^2 = 2.0$ and $\rho = 0.7$. A natural two-moment approximation would be based on the $H_2/H_2/1$ queue

Table III—The mean queue length, L, in several $K_2/H_2/1$ systems with traffic intensity $\rho = 0.7$

Interarrival-time distribution			Service-Time Distribution				
			(M^B)	Hyperexponential (H_2)			($\hat{M}$)
			$r_s = 0.01$	$r_s = 0.1$	$r_s = 0.5$	$r_s = 0.9$	$r_s = 0.99$
		E_2	2.61	2.62	2.63	2.63	2.63
	(M)	$r_a = 1.0$	3.15	3.15	3.15	3.15	3.15
	Hyperexponential (H_2)	$r_a = 0.9$	3.60	3.60	3.59	3.56	3.52
		$r_a = 0.5$	4.02	4.01	3.99	3.96	3.94
		$r_a = 0.1$	4.23	4.23	4.21	4.21	4.20
	(M^B)	$r_a = 0.0$	4.32	4.32	4.32	4.32	4.32

Notes: 1. The hyperexponential (H_2) distributions all have squared coefficient of variation $c^2 = 2.00$. 2. The Erlang (E_2) distribution has squared coefficient of variation $c^2 = 0.5$. 3. The $\hat{M}$ service-time distribution differs from an exponential distribution because of the small mass at a very large value. This causes the $H_2/\hat{M}/1$ values of L to differ from the $H_2/M/1$ values of L in Table I.

with $c_a^2 = c_s^2 = 2.0$ and $r_a = r_s = 0.5$. The range of $H_2/H_2/1$ values as r_a and/or r_s varies indicates the possible deviations from the approximations when the distributions are required to be mixtures of exponential distributions. The maximum relative error is (4.32-3.15)/3.15 or 37 percent, but would be much less if we restricted r_a and r_s to some reasonable interval, e.g., [0.2, 0.8].

Table III enables us to compare the effect of additional information about the interarrival-time and service-time distributions. Table III shows that, given two moments, other properties of the distribution are much more important for the interarrival-time distribution than for the service-time distribution in determining the mean queue length. This phenomenon was previously noted by Sahin and Perrakis.[10]

The program for calculating the mean queue length and the probability of delay in a $K_2/G/1$ queue used to obtain Table III is being used as part of a three-parameter procedure for approximating general G/G/1 queues with bursty, possibly nonrenewal arrival processes.[11] The general bursty arrival process is approximated by a renewal process with an H_2 interarrival-time distribution, which is character-

ized completely by the first three moments of the renewal interval.[6] Then the expected queue length and probability of delay are calculated exactly for the resulting $H_2/G/1$ model. Additional descriptions of the $H_2/G/1$ queue, such as an entire waiting-time distribution, are obtained using approximations similar to the ones in the software package QNA (see Section 5.1 of Ref. 12). This approach is part of a new three-parameter algorithm for QNA.

VIII. ACKNOWLEDGMENT

This research was motivated by the results in Part II obtained jointly with John Klincewicz. I am grateful to him for his assistance. I am also grateful to Richard Stubing for writing programs to obtain the data in Tables I through III.

REFERENCES

1. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," AT&T Bell Lab. Tech. J., this issue.
2. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," AT&T Bell Lab. Tech. J., this issue.
3. 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.
4. S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, New York: John Wiley and Sons, 1966.
5. J. Keilson, *Markov Chain Models—Rarity and Exponentiality*, New York: Springer-Verlag, 1979.
6. W. Whitt, "Approximating a point process by a renewal process, I: two basic methods," Oper. Res., *30,* No. 1 (January–February 1982), pp. 125–47.
7. W. Whitt, "The Marshall and Stoyan bounds for IMRL/G/1 queues are tight," Oper. Res. Letters, *1,* No. 6 (December 1982), pp. 209–13.
8. J. W. Cohen, *The Single Server Queue*, Amsterdam: North-Holland, 1969.
9. W. Whitt, "Minimizing delays in the GI/G/1 queue," Oper. Res., to be published.
10. I. Sahin and S. Perrakis, "Moment inequalities for a class of single server queues," INFOR, *14,* No. 2 (June 1976), pp. 144–52.
11. W. Whitt, unpublished work.
12. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62,* No. 9 (November 1983), pp. 2779–815.

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.