MANAGEMENT SCIENCE
Vol. 29, No. 4, April 1983
Printed in U.S.A.

UNTOLD HORRORS OF THE WAITING ROOM: WHAT THE EQUILIBRIUM DISTRIBUTION WILL NEVER TELL ABOUT THE QUEUE-LENGTH PROCESS*

WARD WHITT†

This paper cautions against using only the equilibrium distribution to describe the behavior of a queue. It is suggested that fluctuations in the queue-length process should also be described, for example, by various first-passage-time distributions. The range of possible fluctuations associated with a given equilibrium queue-length distribution is described for the $GI/M/1$ queue. The theory of complete Tchebycheff systems in Karlin and Studden [11] is applied to construct appropriate extremal distributions, i.e., interarrival-time distributions having the given equilibrium queue-length distribution and maximum or minimum values of fluctuation measures such as the relaxation time.
(QUEUES; STOCHASTIC MODELS; APPLIED PROBABILITY)

1. Introduction

It is common to describe the behavior of a queue by giving the mean equilibrium queue-length or, equivalently (by Little's formula, $L = \lambda W$), the mean equilibrium delay. The variance and/or various tail probabilities of these equilibrium distributions are also sometimes given, but not as often as they should be considering that the mean is a notoriously limited description of an entire probability distribution. The primary purpose of this paper is to point out that even the entire equilibrium distribution may not be enough. This is trivially true, of course, if equilibrium is never attained, as is the case with a nonstationary arrival process. However, even when equilibrium is attained, in general there can be many stochastic processes having the same equilibrium distribution. In addition to the equilibrium distribution, it is useful to describe the fluctuations or transient behavior of the stochastic process. This can be done, for example, with various first-passage times such as the busy period or with the serial correlations which are revealingly described through the spectral density. The rate at which the stochastic process approaches steady-state also can be described by the relaxation time; see Cohen [4]. The transient behavior is important because it can contribute to the costs and benefits of operating a system. For example, when buffers are allocated in real time by a central processor, the equilibrium distribution of buffer content may be used to determine the required number of buffers, but the fluctuations will determine the load on the central processor for buffer allocation.

We illustrate the limitations of the equilibrium distribution in a simple context: the $GI/M/1$ queue. We assume that the equilibrium queue-length distribution is known, from which we can determine the traffic intensity and thus the arrival rate and service rate, up to an arbitrary choice of the measurement (time) scale. From the structure of the $GI/M/1$ queue, it is then easy to characterize all renewal arrival processes with this arrival rate and associated equilibrium queue-length distribution. We obtain an interesting description of this class of interarrival-time distributions by identifying two extremal distributions. Roughly speaking, one extremal distribution maximizes long-run fluctuations and the other maximizes short-run fluctuations. We use these extremal interarrival-time distributions to describe the range of possible values for several different fluctuations measures.

*Accepted by John P. Lehoczky; received April 9, 1981. This paper has been with the author 6 months for 2 revisions.
†Bell Laboratories, Holmdel, N.J. 07733.

0025-1909/83/2904/0395$01.25

We obtain the extremal distributions by applying the theory of complete Tchebycheff systems as described in Karlin and Studden [11]. This theory was applied to provide bounds for partially specified queueing systems by Rolski [20]–[22], Holtzman [8], Eckberg [6], Bergman et al. [1], and Whitt [27], [28].

The problem of finding all arrival processes or all queue-length processes associated with a given equilibrium distribution is an inverse problem. An introduction to inverse problems is contained in Keller [16]. The general problem of characterizing the class of Markov transition matrices with a given equilibrium distribution is treated by Karr [12]. Other stochastic inverse problems are treated in Karr and Pittenger [14], Karr [13], and references there. Of course, the special structure of the queue restricts the range of possible transition kernels for the natural Markov chain associated with the $GI/M/1$ queue, namely, the sequence of queue-lengths at arrival epochs. It would be interesting to compare different stochastic models to see how much each restricts the inverse map.

Before considering the inverse problem for the $GI/M/1$ queue, it is appropriate to mention the contrasting situation with the $M/G/1$ queue. Unlike the $GI/M/1$ queue and more general models, the equilibrium queue-length and delay distributions of the $M/G/1$ queue each completely characterizes the arrival and service processes, of course again up to a scale factor. Moreover, the first k moments of the equilibrium delay determine the first $k + 1$ moments of the service-time distribution and vice versa. From the equilibrium distribution it is possible to reconstruct the entire model and then calculate any desired fluctuations in the queue-length process.

This work is closely related to Whitt [25] in which we described an indirect or inverse method to approximate an unknown arrival process to a queue by using a partial characterization of the equilibrium queue-length distribution. The present paper obviously indicates limitations in that approach. However, if the arrival process is assumed to satisfy additional regularity conditions, then a much smaller class of arrival processes is possible. The impact of various additional constraints on the interarrival-time distribution in a $GI/M/1$ queue, such as shape constraints (e.g., unimodality and log-convexity), is studied in Klincewicz and Whitt [18] and Whitt [28]. That work indicates that if the interarrival-time distribution is not irregular then the range of fluctuations should be reasonably narrow and such two-moment approximation schemes should work reasonably well. In §3.8 here we examine the possible arrival processes and possible fluctuations in the queue length process of a $GI/M/1$ queue with hyperexponential interarrival-time distributions and a given equilibrium distribution. Specifying the equilibrium distribution does not pin down the variance of the interarrival times, so the range of possible behavior is greater than in Klincewicz and Whitt [18] where the mean and variance are specified in addition to the shape.

In closing this introduction, it is appropriate to point out that there is considerable literature bearing on the theme of this paper. There are many descriptions of time-dependent behavior and many examples showing the limitations of equilibrium distributions. A good illustration is a Jackson network of queues. The equilibrium distribution of queue lengths has the product form as if the service facilities were independent and all arrival processes were Poisson processes, but the facilities and the associated queue length processes are not independent and the arrival processes within the network are in general not Poisson; see Melamed [19]. This is one way in which the equilibrium distribution does not capture the time-dependent behavior.

The rest of this paper is organized as follows. In §2, to put our work on the $GI/M/1$ queue in perspective, we briefly review the results for finite Markov chains. In §3 we treat the $GI/M/1$ queue. We show that the extremal distributions reveal the range of possible values for various random quantities describing the fluctuations of the $GI/M/1$ queue.

2. Finite Markov Chains

Consider a discrete-time Markov chain with state space $\{1,2,\ldots,n\}$, transition matrix P, and unique equilibrium distribution (probability vector) α. To understand the limitations of equilibrium distributions, we are led to ask: (1) What is the set of all transition matrices P having a given equilibrium distribution α? and (2) How much can the transient behavior vary over this class of transition matrices? Karr [12] studied the first question; we review some of what is known about the second.

2.1. *Decomposability and Periodicity*

An instructive example is the uniform equilibrium distribution: $\alpha = (1/n, \ldots, 1/n)$. It is well known that P has a uniform equilibrium distribution if and only if P is doubly stochastic, i.e., if all the column sums as well as the row sums are one. The obvious extreme cases (in their behavior and in the convex set of transition matrices) are the permutation matrices, having a single entry 1 in every row. The permutation matrices exhibit two kinds of unusual behavior: decomposability (no transition away from a subset of states) and periodicity. Neither of these phenomena is revealed by the equilibrium distribution, which is a serious shortcoming. Even if we assume that P is indecomposable (ergodic) and aperiodic, which occurs if and only if P^n has no zero entries for some n, there can be bizarre behavior primarily because P may be nearly decomposable or nearly periodic.

EXAMPLE 1. Consider the two-state chain with $P_{12} = P_{21} = p$, $0 \leqslant p \leqslant 1$. For all p, $(1/2, 1/2)$ is an equilibrium distribution. For $p = 0$, P is decomposable, i.e., $P = I$; for $p = 1$, P is periodic. As p varies from 0 to 1, the correlation between $X(n)$ and $X(n+1)$ decreases from 1 to -1, with the correlation being 0 in the independent case: $p = 1/2$. For p near 0, P is nearly decomposable and, for p near 1, P is nearly periodic.

2.2. *The Spectral Representation*

A convenient representation for the k-step transition matrix P^k can be obtained if P is diagonizable,, i.e., if there exists a complete linearly independent family of left (or right) eigenvectors x associated with the eigenvalues λ of P, satisfying $xP = \lambda x$. Then P can be expressed as $P = B\Lambda B^{-1}$ where Λ is the diagonal matrix of eigenvalues and $P^k = B\Lambda^k B^{-1}$. The multiplicity of the eigenvalue 1 indicates the number of ergodic subsets; the complex roots of unity indicate the periodic structure; and the eigenvalue with the largest modulus less than one indicates the rate of convergence. The time-dependent behavior is relatively easy to describe in this manner when the eigenvectors and eigenvalues are real-valued, which is always the case for time-reversible chains; see Keilson [15]. This spectral representation also leads to a spectral density associated with the autocorrelation of the stationary Markov chain; see Keilson [15, pp. 34 and 118].

Consider Example 1. The eigenvalues are 1 and $1 - 2p$. The spectral representation yields

$$P^k = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}(1-2p)^k, \qquad k \geqslant 0.$$

The associated autocorrelation function of the stationary version of the chain is $\rho_k = (1-2p)^k$, $k \geqslant 0$. The approach to steady-state is described by $|1-2p|$, which is very slow when p is near 0 or 1, i.e., in the nearly decomposable and nearly periodic cases.

2.3. *The Fundamental Matrix*

A natural way to look at the time-dependent behavior of a chain with transition matrix P and equilibrium distribution α is to compare the time-dependent behavior with some convenient reference chain which also has equilibrium distribution α. The independent case provides such a convenient reference chain; it has transition matrix A with each row of A being α. For each pair of states (i, j), we can compare the expected number of visits in k steps to j starting in i using P and A. For P this expected value is $\sum_{m=1}^{k} P_{ij}^{m}$. For A, this expected value is $k\alpha_j$. Obviously both expected values diverge to infinity as $k \to \infty$, but it turns out that the difference always converges to a finite limit. In matrix form, this limit is $Z - A$ where Z is the fundamental matrix introduced by Kemeny and Snell [17, p. 75]:

$$Z = (I - P + A)^{-1} = \sum_{k=0}^{\infty} (P - A)^k.$$

The fundamental matrix Z is the basis for simple formulas for the mean and the variance of the first passage times; see §§4.4 and 4.5 of Kemeny and Snell. The mean time to return to state j from state j is $1/\alpha_j$. Since it depends only on the equilibrium distribution, it does not help describe the time-dependent behavior. However, the variance of the return time to j, T_{jj}, is

$$\operatorname{Var}(T_{jj}) = (2Z_{jj} - \alpha_j - 1)/\alpha_j^2.$$

In Example 1,

$$Z = \begin{bmatrix} \frac{1}{2} + \frac{1}{4p} & \frac{1}{2} - \frac{1}{4p} \\ \frac{1}{2} - \frac{1}{4p} & \frac{1}{2} + \frac{1}{4p} \end{bmatrix} \quad \text{and} \quad \operatorname{Var}(T_{11}) = \operatorname{Var}(T_{22}) = \frac{2}{p} - 2.$$

Note that the return time variance is near 0 in the nearly periodic case (p near 1) and near ∞ in the nearly decomposable case (p near 0). In the independent case ($p = 1/2$) this variance is 2.

The moment formulas for the first passage times from i to j, $i \neq j$, are also available. In fact, from all the mean first passage times it is possible to reconstruct the transition matrix P; Theorem 4.4.12(c) of Kemeny and Snell. A useful summary description of the time-dependent behavior is the mean first passage time to j starting from equilibrium, which equals Z_{jj}/α_j. In Example 1, this equilibrium mean first passage time is $2Z_{jj} = 1 + 1/2p$, which cannot be too small, but can be arbitrarily large as p approaches 0. The average mean first passage time to j starting in equilibrium, weighting j by a_j, which is just $\sum_{j=1}^{n} Z_{jj}$, the trace of Z, is another possible summary measure; see Corollary 4.3.6 of Kemeny and Snell.

3. The $GI/M/1$ Queue

Consider a $GI/M/1$ queue characterized by a nonlattice interarrival-time cdf F and mean service time μ^{-1}. Let λ^{-1} be the mean of F and let ϕ be its Laplace–Stieltjes transform, i.e.,

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

Given μ, the equilibrium behavior of this system depends on the interarrival-time cdf F through two real parameters. The first is the mean λ^{-1} or, equivalently, the traffic intensity $\rho = \lambda/\mu$, which we assume is less than one; the second is the root σ, with $0 < \sigma < 1$, of the equation

$$s = \phi(\mu(1-s)). \tag{2}$$

The equilibrium distribution of the continuous-time queue-length process $Q(t)$ is well defined (since F is nonlattice) and has the distribution

$$P(Q=k) = \begin{cases} 1-\rho, & k=0, \\ \rho(1-\sigma)\sigma^{k-1}, & k \geqslant 1, \end{cases} \tag{3}$$

with mean $EQ = \rho(1-\sigma)^{-1}$ and variance $\operatorname{Var} Q = \rho(1-\rho+\sigma)(1-\sigma)^{-2}$; Cohen [4, §II.3]. From the equilibrium distribution of the discrete-time process $\{Q_k\}$ obtained by looking at $Q(t)$ at arrival epochs, we would obtain the root σ but not ρ. Given only σ, much more extreme time-dependent behavior is possible; given σ and μ, λ could still vary over a wide range.

Assume that the equilibrium distribution (3) is known, from which we extract ρ and σ. Suppose μ is given, by fixing the measurement scale. Then the cdf F is partially characterized by knowing its mean λ^{-1} and the root σ of (2).

3.1. *Extremal Two-Point Distributions*

We now define two special interarrival-time distributions having mean λ^{-1} and satisfying (2) with given σ. These are two-point distributions that tend to exhibit extreme fluctuation behavior.

Let $\mathscr{F} \equiv \mathscr{F}(m_1; s_0, \theta; b)$ be the set of all cdf's on $[0,b]$, $b \leqslant \infty$, having mean m_1 and the transform $\theta(s)$ assuming the value θ at $s = s_0$. We now define a partial order relation on the set $\mathscr{F}$. This partial order relation is apparently not standard, but it arises naturally in the theory of complete Tchebycheff systems. Let ϕ_i be the transform associated with F_i.

DEFINITION 1. $F_1 \leqslant_t F_2$ in $\mathscr{F}(m_1; s_0, \theta; b)$ if $\phi_1(s) \leqslant \phi_2(s)$ for $s \leqslant s_0$ and $\phi_1(s) \geqslant \phi_2(s)$ for $s \geqslant s_0$.

The theory of complete Tchebycheff systems yields extremal cdf's in the space $(\mathscr{F}, \leqslant_t)$.

DEFINITION 2. (LOWER BOUND). Let F_L be the cdf of the two-point distribution with mass m_1/x on x and mass $1-(m_1/x)$ on 0, where x is the unique positive root to the equation $x = m_1[1-e^{-xs_0}]/(1-\theta)$.

(UPPER BOUND, $b < \infty$). Let F_U be the cdf of the two-point distribution with mass $p = (b-m_1)/(b-x)$ on x and mass $1-p$ on b, where x is the unique root to the equation $pe^{-xs_0} + (1-p)e^{-bs_0} = \theta$.

(UPPER BOUND, $b = \infty$). Let $\hat{F}_U$ be the cdf of the one-point distribution with unit mass on $(-\log\theta)/s_0$.

From (5) and (15) of Eckberg [6], we obtain the following result.

THEOREM 1. (i) *For $b < \infty$, F_L and F_U belong to $\mathscr{F}$ and $F_L \leqslant_t F \leqslant_t F_U$ for all $F \in \mathscr{F}$.*

(ii) *For $b = \infty$, F_L belongs to $\mathscr{F}$ (but not $\hat{F}_U$) and $F_L \leqslant_t F \leqslant_t \hat{F}_U$ for all $F \in \mathscr{F}$.*

By constructing other complete Tchebycheff systems, we also have

THEOREM 2. *For all $F \in \mathscr{F}$ (with $b < \infty$ for the second inequality),*

$$\int_0^\infty x^k\, dF_L(x) \leqslant \int_0^\infty x^k\, dF(x) \leqslant \int_0^\infty x^k\, dF_U(x) \tag{4}$$

for all $k \geqslant 2$.

REMARKS. (i) Theorem 2 might lead one to conjecture a stronger stochastic ordering, namely,

$$\int_0^\infty g(x)\,dF_L(x) \leqslant \int_0^\infty g(x)\,dF(x) \leqslant \int_0^\infty g(x)\,dF_U(x) \tag{5}$$

for all $F \in \mathscr{F}$ and all nondecreasing convex real-valued functions g on $[0, b]$, see Whitt [24] and references there; but to see that (5) fails note that (5) would also hold for all convex functions since the means are equal, but this is not possible here because $\phi_L(s) \geqslant \phi(s) \geqslant \phi_U(s)$ for $s > s_0$.

(ii) As $b \to \infty$, the lower bound remains unchanged while the upper bound converges to $\hat{F}_U$. This limit is not in $\mathscr{F}(m_1; s_0, \theta; \infty)$ because the small mass at b has disappeared, causing the mean to cease being m_1.

(iii) The lower bound (possibly also the upper bound) is not quite proper for the queue because it is a lattice distribution, but a minor perturbation would make it nonlattice.

3.2. *Extremal GI/M/1 Queues*

Associated with the extremal distributions F_L and F_U are $GI/M/1$ queueing systems. When the interarrival-time distribution is F_L, the queue is a $D^N/M/1$ system, i.e., there are geometrically-distributed batch arrivals at fixed intervals. When the interarrival-time distribution is F_U, the queue alternates between two regimes. For a geometrically-distributed number of interarrival times, it behaves like a $D/M/1$ system; then there is an interval of length b in which there are no arrivals at all. Intuitively, it is clear that F_L causes greater short-run fluctuations (probability of many arrivals in a short interval), whereas F_U causes greater long-run fluctuations.

EXAMPLE 2. In order to dramatize the range of fluctuations in $GI/M/1$ systems with a common equilibrium distribution, we consider a specific example. To have a basis for comparison, we start with the first prototype interarrival-time distribution used in Klincewicz and Whitt [18]. This distribution is a mixture of two geometric distributions on the nonnegative integers, truncated at 20, chosen to have mean $m_1 = 2$ and squared coefficient of variation $c^2 = 2$ (variance $= 8$). We let $\rho = 0.9$ ($\mu = 5/9$). In this case, the root is $\sigma = 0.93238$. We use the prototype distribution to generate a reasonable root σ. We shall not restrict attention to distributions on the integers or require that 20 be the upper bound of the support. In fact, we shall let the upper bound be $b = 100$. Hence, we consider the class of all interarrival times in $\mathscr{F}(m_1; \mu(1-\sigma), \sigma; b)$, here being $\mathscr{F}(2; 0.03757, 0.93238; 100)$.

For these parameter values, F_L has mass 0.35 on 5.715 and mass 0.65 on 0. The first three moments of F_L are: $m_{L1} = 2$, $m_{L2} = 11.43$, and $m_{L3} = 65.33$; the variance and squared coefficient of variation are thus $\sigma_L^2 = 7.43$ and $c_L^2 = 1.86$. The geometrically distributed batch of arrivals in the $D^N/M/1$ system has mean 2.86 and variance 2.45.

For these parameter values, F_U has mass 0.998113 on 1.8147 and mass 0.001887 on 100. The first three moments of F_U are: $m_{U1} = 2$, $m_{U2} = 22.16$, and $m_{U3} = 1893$; the variance and squared coefficient of variation are thus $\sigma_U^2 = 18.16$ and $c_U^2 = 4.54$. The geometrically-distributed $D/M/1$ "on" phase has mean 529. The parameters of the $D/M/1$ system are $\rho = 0.9919$ and $\sigma = 0.9838$. (The overall ρ and σ are unchanged; these larger values are obtained using the conditional interarrival-time distribution given that the larger value in the two-point distribution does not occur. In other words, this is the deterministic interarrival-time distribution obtained by conditioning on being in an "on" phase.) The mean equilibrium queue length of the $D/M/1$ regime is thus $\rho/(1-\sigma) = 61.3$. Of course, equilibrium is usually not attained even in the approximately 500 interarrival times, but this gives some idea of the fluctuations. Note that the average "on" phase of 529 arrivals is of length 960 time units, about 10 times the deterministic "off" phase of length 100.

The limiting upper bound cdf $\hat{F}_U$ has unit mass on 1.8637. The resulting $D/M/1$ queue has traffic intensity $\rho = 0.9658$.

Of course, we have chosen $\mathscr{F}$ so that $GI/M/1$ queues with F_L, F_U, and $\hat{F}_U$ as interarrival-time cdf's all have the same equilibrium distribution, namely, that determined by $\rho = 0.9$ and $\sigma = 0.93238$. In particular, $EQ = 13.3$. Note that the equilibrium mean in the $D/M/1$ "on" phase associated with F_U is 4.6 times this overall equilibrium mean. This can be somewhat misleading, however, since equilibrium is usually not attained in an on-phase. The equilibrium during an on-phase does help describe what is going on though if properly interpreted, because the number of customers served during an on-phase is geometric with a large mean and a very large variance. Hence, most on-phases will be relatively short or only moderately long, e.g., 500 customers, leading to congestion no greater than described by the overall equilibrium distribution. (This can be verified by using the heavy-traffic approximation for the time-dependent distribution with $\rho = 1$; see Whitt [23].) However, there will also be occasional very long on-phases when many customers experience unusual congestion. This unusual congestion is precisely the on-phase equilibrium distribution described above.

To see how different F_L and F_U are, it is instructive (and easy) to look at typical sample paths of the renewal arrival processes associated with F_L and F_U. For $GI/M/1$ queues with the service rate fixed, a description of the time-dependent behavior of the arrival process is similar to a description of the time-dependent behavior of the queue-length process. (The queue-length process will be somewhat smoother due to the exponential service times.) Good summary descriptions of these different arrival processes are provided by the number of arrivals in an interval of length t for various values of t. The distributions of these numbers are given in Table 1 for $t = 10$, 100, 1,000, and 10,000. For F_L, the number of arrivals in $[0, t]$ is just the sum of $1 + [t/5.715]$ iid geometric random variables, where $[x]$ is the integer part of x; for $t \geqslant 100$, this is approximately normally distributed by the central limit theorem. For F_U, the number of arrivals is deterministic except for off-phases. For example, for $t = 10$, there are $1 + [t/1.8147] = 6$ possible arrivals. There are no off-phases with probability $(0.998113)^6 = 0.9887$. If an off-phase is initiated, it is of length 100 and so extends to the end of the interval.

Notice that for $t = 10$, the number associated with F_L is much more variable than the number associated with F_U, but this is reversed for $t = 10,000$. Notice that the distribution of the number associated with F_L rapidly becomes "regular," i.e., it approaches the normal distribution, being very close for $t = 100$, but the distribution of the number associated with F_U is not yet close at $t = 1000$.

To compare the different arrival processes, it is also instructive to remove the waiting room and compute the blocking probabilities. This will obviously favor the process with the least short-run fluctuations, which is the arrival process associated with the upper bound F_U. Recall that the blocking probability in a $GI/M/1$ loss system is just $\phi(\mu)$, which is also known as the peakedness; see Eckberg [6] and references there. As an immediate consequence of Theorem 1, we have $\phi_L(\mu) \geqslant \phi(\mu) \geqslant \phi_U(\mu) \geqslant \phi_{\hat{U}}(\mu)$ for all $F \in \mathscr{F}(m_1; \mu(1-\sigma), \sigma; b)$. For Example 2, $\phi_L(\mu) = 0.66$, $\phi_U(\mu) = 0.36$, and $\phi_{\hat{U}}(\mu) = 0.35$. Even the equilibrium blocking probability of the $D/M/1$ "on" phase associated with F_U is only 0.365. Obviously F_L and F_U are dramatically different in this view.

TABLE 1

Approximate Probability Distribution of the Number of Arrivals in an Interval Beginning at an Arrival Epoch: Comparing F_L and F_U for Various Interval Lengths

Length of Interval	Distribution of Number of Arrivals based on F_L	Distribution of Number of Arrivals Based on F_U
10	2 w.p. 0.123 3 w.p. 0.159 4 w.p. 0.155 5 w.p. 0.134 $\geqslant$ 6 w.p. 0.429 (mean 5.72, variance 4.90; obtained by convoluting two geometric distributions)	6 w.p. 0.9887 1, . . . , 5 w.p.$\approx$ 0.002 each
100	$\approx$ normal (51.5, 44.1)	56 w.p. 0.8996 12, . . . , 55 w.p.$\approx$ 0.002 each
1,000	$\approx$ normal (501, 429)	552 w.p. 0.353 496 w.p. 0.367 440 w.p. 0.150 others w. total p. 0.130
10,000	$\approx$ normal (5,000, 4,288)	$\approx$ normal (5,000, 22,700)

Notes:
(1) In normal (a, b), a is the mean and b is the variance.
(2) w.p. means "with probability."
(3) $\approx$ means "approximately equal to."

3.3. *The Busy Period*

Let B be the cdf of the busy period in a $GI/M/1$ queue, i.e., the first-passage time from state i to state i for any $i \geqslant 0$, beginning just before an arrival. Let $\phi_B(s)$ be the Laplace–Stieltjes transform of B in (1). From Cohen [4, p. 226],

$$\phi_B(s) = \frac{1 - \eta}{\mu^{-1}s + 1 - n}, \qquad s \geqslant 0, \tag{6}$$

where $\eta \equiv \eta(s)$ is the unique root with $0 < \eta(s) < 1$ of the equation

$$\eta = \phi(s + \mu(1 - \eta)), \tag{7}$$

with ϕ being the transform (1) of the interarrival time cdf F. By differentiating (6), we obtain the first two moments of B:

$$m_{B1} = 1/\mu(1 - \sigma) \quad \text{and} \tag{8}$$

$$m_{B2} = 2\big(\mu^{-1} - \eta'(\mu(1 - \sigma))\big)/\mu(1 - \sigma)^2. \tag{9}$$

The variance can be expressed as

$$\sigma_B^2 = \big[1 - \mu\phi'(\mu(1 - \sigma))\big]/\big[1 + \mu\phi'(\mu(1 - \sigma))\big]\,\mu^2(1 - \sigma)^2. \tag{10}$$

Note that the mean m_{B1} is fixed for a given equilibrium distribution, just as with the return time in §2.3.

Let B_L and B_U be the busy-period cdf's associated with F_L and F_U. We can apply Theorem 1 to obtain the following comparison result.

THEOREM 3. *Let all interarrival times be in* $\mathscr{F}\ (m_1;\ \mu(1-\sigma),\ \sigma;\ b)$. *Then*

(i) $$\phi_{B_L}(s) \leqslant \phi_B(s) \leqslant \phi_{B_U}(s), \qquad s \geqslant 0, \quad \textit{and}$$

(ii) $$\sigma^2_{B_L} \leqslant \sigma^2_B \leqslant \sigma^2_{B_U} \leqslant \sigma^2_{B_{\hat{U}}} = (1+\gamma)/(1-\gamma)\,\mu^2(1-\sigma),$$

where $\gamma = -\sigma(\log\sigma)/(1-\sigma)$.

PROOF. From (7), first note that $\eta(s) \leqslant \sigma$ for all $s \geqslant 0$ and then note that $\eta_1(s) \geqslant \eta_2(s)$ for each s if $F_1 \leqslant_t F_2$ in $\mathscr{F}$, from which (i) follows easily from (6). Finally, (ii) follows from (10) and Theorem 1 because $\phi_1'(\mu(1-\sigma)) \geqslant \phi_2'(\mu(1-\sigma))$ if $F_1 \leqslant_t F_2$ in $\mathscr{F}$. It is easy to see the upper bound is attained at $F_{\hat{U}}$ as $b \to \infty$.

Consider Example 2. There $m_{B1} = 26.6$. Here are the key transform derivatives and the variances:

$$\begin{aligned} \phi_L'(\mu(1-\sigma)) &= -1.6138, & \sigma^2_{B_L} &= 12,996, \\ \phi_U'(\mu(1-\sigma)) &= -1.6963, & \sigma^2_{B_U} &= 23,894, \\ \phi_{\hat{U}}'(\mu(1-\sigma)) &= -1.7370, & \sigma^2_{B_{\hat{U}}} &= 39,783. \end{aligned} \tag{11}$$

To see the range of possibilities, note that $\sigma^2_{B_U}/\sigma^2_{B_L} = 1.84$ and $\sigma^2_{B_{\hat{U}}}/\sigma^2_{B_L} = 3.06$. To put these values in perspective, for an $M/M/1$ queue with $\rho = 0.93238$, the mean and variance of the busy period are $1/\mu(1-\rho) = 26.6$ and $(1+\rho)/\mu^2(1-\rho)^3 = 20,250$ respectively.

3.4. *A Step in the Embedded Random Walk*

The discrete-time sequence of queue-lengths just prior to arrivals in a $GI/M/1$ queue is a random walk with an impenetrable barrier at zero. One way to describe the fluctuations is to describe the distribution of the individual steps. The step is of course just one minus the number of potential services generated by a Poisson process with rate μ in an interarrival time. It obviously suffices to focus on the number of potential services in an interarrival time, which is just the cdf D with transform

$$\phi_D(s) = \phi(\mu(1-e^{-s})). \tag{12}$$

It is easy to check that

$$m_{D1} = \rho^{-1} \quad \text{and} \quad m_{D2} = \mu^2 m_2 + \mu m_1. \tag{13}$$

Moreover, it follows immediately from Theorem 1 that $D \in \mathscr{F}(\rho^{-1}; -\log\sigma, \sigma; \infty)$ if D is associated with $F \in \mathscr{F}(m_1;\ \mu(1-\sigma), \sigma; \infty)$. Hence, if $F_1 \leqslant_t F_2$ in $\mathscr{F}(m_1;\ \mu(1-\sigma), \sigma; b)$, then $D_1 \leqslant_t D_2$ in $\mathscr{F}(\rho^{-1}, -\log\sigma, \sigma; \infty)$ for cdf's D_i associated with F_i. Hence,

$$D_L \leqslant_t D \leqslant_t D_U \tag{14}$$

for all $F \in \mathscr{F}(m_1;\ \mu(1-\sigma), \sigma; b)$ if D, D_L and D_U are associated with F, F_L and F_U, respectively. Note, however, that the upper bound D_U is not the upper bound over all cdf's in $\mathscr{F}(\rho^{-1}, -\log\sigma, \sigma; \infty)$ because only step cdf's D associated with interarrival-time cdf's F are considered.

In Example 2, $\sigma^2_{D_L} = 3.01$ and $\sigma^2_{D_U} = 6.22$. Of course, $\sigma^2_{D_U} \to \infty$ as $b \to \infty$.

3.5. *Covariance Structure*

Let Q_k be the queue length just prior to the kth arrival, with a superscript L or U to denote the interarrival-time cdf's F_L and F_U. We first describe the one-step correlation $\text{Cor}(Q_k, Q_{k+1})$ in light and heavy traffic. We assume the queue is in equilibrium.

As $\rho \to 0$, the root σ of (2) converges to $F(0)$, the probability mass at 0. With the moment formulas of Cohen [4, p. 210], it is then easy to show that $\mathrm{Cor}(Q_k, Q_{k+1})$ converges to $F(0)$ as $\rho \to 0$. Since F_L attaches the maximum mass to 0 among all interarrival-time cdf's if $\mathscr{F}$, F_L yields the maximum one-step correlation as $\rho \to 0$. Of course, $(\mathrm{Cor}\, Q_k^U, Q_{k+1}^U) \to 0$ as $\rho \to 0$.

On the other hand, $\sigma \to 1$ and $\mathrm{Cor}(Q_k, Q_{k+1}) \to 1$ as $\rho \to 1$ for all interarrival-time distributions. We conjecture that F_L maximizes and F_U minimizes the one-step correlation for all values of ρ, but we have not proved it. We also conjecture that F_L minimizes and F_U maximizes the m-step correlations for very large m. These conjectures are supported but not established by §3.4. It is known that $\mathrm{Cor}(Q_k, Q_{k+m})$ is decreasing in m for all F; see Bergmann and Stoyan [2] and references there.

A useful summary measure of dependence is the sum $S = \sum_{l=0} \mathrm{Cor}(Q_k, Q_{k+l})$. It is related to the cumulative process $\{\sum_{k=1}^{k=n} Q_k, n \geqslant 1\}$ associated with the stationary process $\{Q_n, n \geqslant 0\}$ by

$$\lim_{n \to \infty} n^{-1} \mathrm{Var} \sum_{k=1}^{n} Q_k = S \,\mathrm{Var}\, Q_k .$$

This asymptotic average variance is also the normalizing constant in the central limit theorem, i.e.,

$$\left[\sum_{k=1}^{n} Q_k - nEQ_1 \right] \Big/ \left[nS \,\mathrm{Var}(Q_1) \right]^{1/2} \Rightarrow N(0,1),$$

where $\Rightarrow$ denotes convergence in distribution and $N(a,b)$ is the normal distribution with mean a and variance b; see Iglehart [9] and Billingsley [3, p. 174].

The sum S has been calculated by Daley [5]. From the discussion above, it follows that S is maximized by F_L for very small values of ρ. In fact,

$$\lim_{\rho \to 0} S(\rho) = 1 + \sum_{k=1}^{\infty} kF(0)^k = 1 + (1 - F(0))^{-2}.$$

In Example 2, this limit is 9.16 for F_L and 1 for F_U.

Daley has also given an expansion for $S(\rho)$ as $\rho \to 1$ that shows it is maximized by F_U:

$$\lim_{\rho \to 1} (1-\rho)^2 S(\rho) = 1 + m_2 / 2m_1^2,$$

where m_i is the ith moment of F. Recall that m_1 is fixed but $m_{U2} \geqslant m_2 \geqslant m_{L2}$ for all $F \in \mathscr{F}$. In Example 2, $m_{L2} = 11.43$ and $m_{U2} = 22.16$.

3.6. *The Relaxation Time*

The rate at which the queue-length process approaches equilibrium can be described by the relaxation time; see Cohen [4, pp. 181 and 589]. For $GI/G/1$ systems, if $Q(t)$ is the queue-length at time t with any given initial conditions, then $P(Q(t) > 0)$ and $EQ(t)$ differ from their limits by constant multiples (which depend on the initial conditions) of a term $t^{-3/2} e^{-\eta t}$ (under minor regularity conditions). The crucial exponential decay rate η is found by solving the equation

$$\psi(\xi)\phi(s - \xi) = 1, \tag{15}$$

where $\psi(s)$ is the Laplace–Stieltjes transform of the service-time cdf; here $\psi(s) = \mu/(\mu + s)$. The parameter $-\eta$ is the unique value of s such that the two real roots $\xi(s)$ of (15) coincide. The reciprocal η^{-1} is the relaxation time.

THEOREM 4. *For interarrival times in* $\mathscr{F}(m_1;\ \mu(1-\sigma),\sigma;b)$,

$$\eta_L^{-1} \leqslant \eta \leqslant \eta_U^{-1}. \tag{16}$$

PROOF. From Cohen [4, p. 590], we obtain η by solving

$$\log \Psi(\xi) = \log(1/\phi(s-\xi)) = -\log(\phi(s-\xi)) \tag{17}$$

and the corresponding equation after taking derivatives of both sides of (17) with respect to ξ, i.e.,

$$\frac{\Psi'(s)}{\Psi(s)} = \frac{\phi'(s-\xi)}{\phi(s-\xi)} \tag{18}$$

for (s,ξ); then $\eta = -s$. As noted by Cohen, $\log \Psi(\xi)$ is decreasing and convex and $-\log(\phi(s-\xi))$ is decreasing and concave within the relevant interval. It is thus easy to see that a solution (s,ξ) always satisfies $\xi < s < 0$; see Figure 1.

Let ϕ correspond to F in $\mathscr{F}$. By Definition 1, $\phi_L(s) \leqslant \phi(s) \leqslant \phi_U(s)$ for $0 \leqslant s \leqslant s_0 = \mu(1-\sigma)$. Consequently,

$$-\log \phi_U(s-\xi) \leqslant -\log \phi(s-\xi) \leqslant -\log \phi_L(s-\xi)$$

for $0 \leqslant s - \xi \leqslant \mu(1-\sigma)$. Hence, it suffices to show that the solution (s_U, ξ_U) to (17) and (18) for ϕ_U satisfies $s_U - \xi_U \leqslant \mu(1-\sigma)$. See Figure 1. In other words, as s is increased form $-\infty$, the curve $-\log \phi_U(s-\xi)$ will hit the curve $\log \Psi(\xi)$ before the curves $-\log \phi(s-\xi)$ and $-\log \phi_L(s-\xi)$ for $\xi \leqslant s - \mu(1-\sigma)$. Hence, if the curve $-\log \phi_U(s-\xi)$ hits the curve $\log \Psi(\xi)$ first as s increases at (s_U, ξ_U) with $\xi_U \geqslant s_U - \mu(1-\sigma)$, then necessarily $s_L < s < s_U$, which is equivalent to (16).

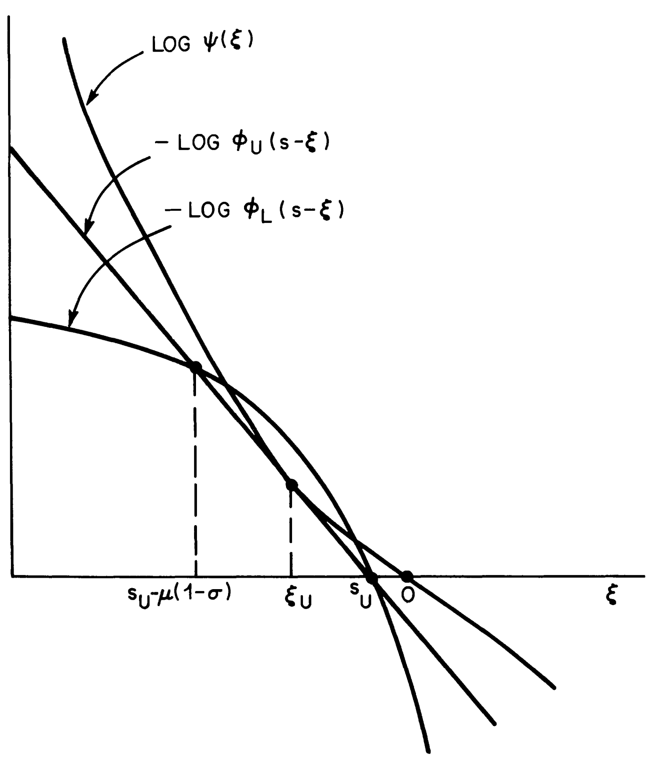

FIGURE 1. The Relaxation-Time Equations: Displaying the Solution for $\eta_U = |s_U|$.

We now show that $s_U - \xi_U \leqslant \mu(1-\sigma)$. We solve (17) and (18) with $\log \Psi(\xi) = \log \mu - \log(\mu + \xi)$ and $-\log \phi_U(s-\xi) = (s-\xi)z$ for $z = -(\log \sigma)/\mu(1-\sigma)$; see Definition 2. From (17), we have

$$\log \mu - \log(\mu + \xi) = (s - \xi)z. \tag{19}$$

After taking derivative in (19) with respect to ξ, we have $-(\mu+\xi)^{-1} = -z$ or

$$\xi = z^{-1} - \mu. \tag{20}$$

Putting (20) into (19), we obtain

$$s - \xi = \frac{\log \mu - \log(z^{-1})}{z}.$$

Hence, it suffices to show that

$$\log \mu + \log z \leqslant \mu(1-\sigma)z \quad \text{or}$$

$$\log \mu + \log(-\log \sigma) - \log \mu - \log(1-\sigma) \leqslant -\log \sigma \quad \text{or}$$

$$\sigma^{-1} \leqslant e^{(1-\sigma)/\sigma},$$

which is easily seen to hold for $0 < \sigma < 1$ (compare the derivatives).

3.7. *Maximum Queue Length*

Extreme values are a natural way to describe sample path variation. Let $M(s,t) = \max\{Q(u) : s \leqslant u \leqslant s+t\}$ for $s, t \geqslant 0$. As $t \to \infty$, $M(s,t) \to \infty$ even though $\rho < 1$. There are extreme value limit theorems showing that $M(s,t)$ grows like $\log t$ as $t \to \infty$; see Cohen [4, p. 615], Heyde [7], Iglehart [10] and references there. What is important here is that the constants in the limits depend on the queue-length process only through the equilibrium distribution. For regenerative processes, therefore, the extremes over very long time intervals only describe the tails of the equilibrium distribution.

However, if we look at the maximum over n busy cycles, i.e., $M(s, C_1 + \cdots + C_n)$ where C_i is the ith busy cycle, then the variation in the busy period (§3.3) causes the time-dependent behavior to matter; Cohen [4, p. 617]: As $k \to \infty$,

$$(-\log \sigma) M(s, C_1 + \cdots + C_n) - \log(n\gamma)$$

fails to converge in distribution as $n \to \infty$, but the distributions have proper upper and lower bounds as $n \to \infty$, with $\gamma = \mu\phi'(\mu(1-\sigma)) + 1$. As shown in the proof of Theorem 3, $\gamma_L \leqslant \gamma \leqslant \gamma_U$ for all $F \in \mathscr{F}$. The derivatives are given for Example 2 in (11).

More striking differences in extreme values are seen if we fix t in $M(s,t)$ and consider $\hat{M}(s,t) = \max\{M(u,t) : 0 \leqslant u \leqslant s\}$. For any fixed t, $\hat{M}_L(s,t) \to \infty$ as $s \to \infty$, but $\hat{M}_U(s,t) \to 1 + [t/x]$, where x is the root in Definition 2 and $[x]$ is the integer part of x. This is another way to look at Table 1.

3.8. *Hyperexponential Interarrival-Time Distributions*

It is interesting to see what happens if, in addition to fixing the parameters σ and ρ in the $GI/M/1$ queue, we stipulate that the interarrival-time distribution is hyperexponential, i.e., the mixture of two exponential distributions, having density

$$h(x) = q\lambda_1 e^{-\lambda_1 x} + (1-q)e^{-\lambda_2 x}, \qquad x \geqslant 0, \tag{21}$$

for $0 \leqslant q \leqslant 1$ and $\lambda_1 \geqslant \lambda \geqslant \lambda_2 \geqslant 0$. This distribution is often used in approximations;

see Whitt [25], [26] and references there. Since this distribution is unimodal with a mode at 0 and has a log-convex density, this distributional assumption is very strong and can be expected to strongly restrict the possible fluctuation behavior.

Knowledge of ρ, σ and a service rate μ gives us λ and the constraints

$$\frac{q}{\lambda_1} + \frac{1-q}{\lambda_2} = \frac{1}{\lambda} \quad \text{and} \tag{22}$$

$$\frac{q\lambda_1}{\mu(1-\sigma)+\lambda_1} + \frac{(1-q)\lambda_2}{\mu(1-\sigma)+\lambda_2} = \sigma. \tag{23}$$

Since we have three parameters and two constraints, we have a one-parameter family of hyperexponential distributions. The distributions with q near 0 and 1 obviously indicate the range of possible fluctuations.

As $q \to 0$, $\lambda_1 \to \infty$ and $\lambda_2 \to \lambda$. The extremal arrival process is approximately a Poisson process at rate λ with geometrically distributed batches of size only rarely more than one. This is the analog of F_L; instead of $D^N/M/1$ queue, we obtain an $M^N/M/1$ queue which is essentially the same as an $M/M/1$ queue with intensity ρ.

On the other hand, as $q \to 1$, $\lambda_1 \to \mu\sigma$ and $\lambda_2 \to 0$. The extremal arrival process alternates between two regimes; for a large geometrically distributed number of customers there is an "on" phase during which the arrival process is a Poisson process at rate $\mu\sigma$; and there is a rare exponentially distributed "off" phase during which there are no arrivals at all. This is the analog of F_U; instead of a $D/M/1$ queue during the "on" phase, we obtain an $M/M/1$ queue with intensity σ. Since $\sigma > \rho$ for hyperexponential distributions, the equilibrium mean queue length during the "on" phase associated with $q \to 1$ is strictly greater than the equilibrium mean associated with $q \to 0$.

It is not difficult to show that the distributions associated with extremal q are actually extremal in the ordering $\leqslant_t$. As in Whitt [28], we can apply the theory of complete Tchebycheff systems to the mixing distributions to obtain

THEOREM 5. *The mixture of two exponential distributions with* $\lambda_1 = b$ *and* q *small* ($\lambda_2 = a$ *and* q *large*) *determined by* (22) *and* (23) *is minimal* (*maximal*) *in the ordering* $\leqslant_t$ *among all mixtures of exponential distributions with parameters* λ_i *in the interval* $[a, b]$ *having a given mean and root* σ.

EXAMPLE 3. As in Example 2, consider arrival processes partially specified by having mean $m_1 = 2$, traffic intensity $\rho = 0.9$ ($\mu = 5/9$) and the key root $\sigma = 0.93238$. However, here restrict attention to mixtures of exponential distributions. Suppose the exponential parameter λ is restricted to the interval $[0.01, 100]$, so that exponential distributions with mean λ^{-1} ranging from 0.01 to 100 are allowed in the mixtures. Of course, the actual distributions are unbounded above so this case is not directly comparable to Example 2.

With these parameter values, the lower bound cdf has $q = 0.99275$, $\lambda_1 = 0.778$ and $\lambda_2 = 0.01$; the variance of the associated busy period is 3493 and the blocking probability of the associated $GI/M/1$ loss system if $\phi(\mu) = 0.579$. The upper bound cdf has $q = 0.01735$, $\lambda_1 = 100$ and $\lambda_2 = 0.4914$; the variance of the associated busy period is 33,685 and the blocking probability of the associated $GI/M/1$ loss system is $\phi(\mu) = 0.478$. To put the busy period variances in perspective, the corresponding busy period variances for $M/M/1$ queue with $\rho = 0.9$ and 0.93238 are 6150 and 20,250, respectively.

[1]The author is grateful to Shlomo Halfin, Moshe Segal, and the referees for their helpful suggestions.

References

1. Bergmann, R., Daley, D. J., Rolski, T. and Stoyan, D., "Bounds for Cumulants of Waiting-times in $GI/G/1$ Queues," *Math. Oper. Statist. Ser. Optim.*, Vol. 10 (1979), pp. 257–263.
2. ——— and Stoyan, D., "Monotonicity Properties of Second Order Characteristics of Stochastically Monotone Markov Chains," *Math. Nachr.*, Vol. 82 (1978), pp. 99–102.
3. Billingsley, P., *Convergence of Probability Measures*, John Wiley and Sons, New York, 1968.
4. Cohen, J. W., *The Single Server Queue*, North-Holland, Amsterdam, 1969.
5. Daley, D. J., "Monte Carlo Estimation of the Mean Queue Size in a Stationary $GI/M/1$ Queue," *Oper. Res.*, Vol. 16 (1968), pp. 1002–1005.
6. Eckberg, A. E., Jr., "Sharp Bounds on Laplace–Stieltjes Transforms, with Applications to Various Queueing Problems," *Math. Oper. Res.*, Vol. 2 (1977), pp. 135–142.
7. Heyde, C. C., "On the Growth of the Maximum Queue Length in a Stable Queue," *Oper. Res.*, Vol. 19 (1971), pp. 447–452.
8. Holtzman, J. M., "The Accuracy of the Equivalent Random Method with Renewal Inputs," *Bell System Tech. J.*, Vol. 52 (1973), pp. 1673–1679.
9. Iglehart, D. L., "Functional Limit Theorems for the Queue $GI/G/1$ in Light Traffic," *Adv. in Appl. Probab.*, Vol. 3 (1971), pp. 269–281.
10. ———, "Extreme Values in the $GI/G/1$ Queue," *Ann. Math. Statist.*, Vol. 43 (1972), pp. 627–635.
11. Karlin, S. and Studden, W. J., *Tchebycheff Systems: With Applications in Analysis and Statistics*, John Wiley and Sons, New York, 1966.
12. Karr, A. F., "Markov Chains and Processes with a Prescribed Invariant Measure," *Stochastic Process Appl.*, Vol. 7 (1978), pp. 277–290.
13. ———, "Extreme Points of Certain Sets of Probability Measures, with Applications," *Math. Oper. Res.*, Vol. 8 (1983) (to appear).
14. ——— and Pittenger, A. O., "The Inverse Balayage Problem for Markov Chains," *Stochastic Processes Appl.*, Vol. 7 (1978), pp. 165–178.
15. Keilson, J., *Markov Chain Models—Rarity and Exponentiality*, Springer-Verlag, New York, 1979.
16. Keller, J. B. "Inverse Problems," *Amer. Math. Monthly*, Vol. 83 (1976), pp. 107–118.
17. Kemeny, J. G. and Snell, J. L., *Finite Markov Chains*, Van Nostrand, Princeton, N.J., 1960.
18. Klincewicz, J. G. and Whitt, W., "On the Quality of Two-Moment Approximations for Queues, II: Shape Constraints," (1982).
19. Melamed, B., "Characterizations of Poisson Traffic Streams in Jackson Queueing Networks," *Adv. in Appl. Probab.*, Vol. 11 (1979), pp. 422–438.
20. Rolski, T., "Some Inequalities for $GI/M/n$ Queues," *Zast. Mat.*, Vol. 13 (1972), pp. 43–47.
21. ———, "Some Inequalities in Queueing Theory," *Colloq. Math. Soc. János Bolyai* (1974), pp. 653–659.
22. ———, "Order Relations in the Set of Probability Distribution Functions and Their Applications in Queueing Theory," *Dissertationes Math.*, Polish Scientific Publishers, Warsaw, 1976.
23. Whitt, W., "Complements to Heavy Traffic Limit Theorems for the $GI/G/1$ Queue," *J. Appl. Probab.*, Vol. 9 (1972), pp. 185–191.
24. ———, "The Effect of Variability on the $GI/G/s$ Queue," *J. Appl. Probab.*, Vol. 17 (1980), pp. 1062–1071.
25. ———, "Approximating a Point Process by a Renewal Process: The View Through a Queue, an Indirect Approach," *Management Sci.*, Vol. 27 (1981), pp. 619–636.
26. ———, "Approximating a Point Process by a Renewal Process: Two Basic Methods," *Oper. Res.*, Vol. 30 (1982), pp. 125–147.
27. ———, "On the Quality of Two-Moment Approximations for Queues, I: Extremal Distributions," (1982).
28. ———, "On the Quality of Two-Moment Approximations for Queues, III: Mixtures of Exponential Distributions," (1982).