STATIONARY-PROCESS APPROXIMATIONS FOR THE NONSTATIONARY ERLANG LOSS MODEL

WILLIAM A. MASSEY

Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey

WARD WHITT

AT&T Labs, Murray Hill, New Jersey

(Received November 1992; revision received August 1994; accepted September 1994)

In this paper we consider the $M_t/G/s/0$ model, which has s servers in parallel, no extra waiting space, and i.i.d. service times that are independent of a nonhomogeneous Poisson arrival process. Arrivals finding all servers busy are blocked (lost). We consider approximations for the average blocking probabilities over subintervals (e.g., an hour when the expected service time is five minutes) obtained by replacing the nonstationary arrival process over that subinterval by a stationary arrival process. The stationary-Poisson approximation, using a Poisson (M) process with the average rate, tends to significantly underestimate the blocking probability. We obtain much better approximations by using a non-Poisson stationary (G) arrival process with higher stochastic variability to capture the effect of the time-varying deterministic arrival rate. In particular, we propose a specific approximation based on the heavy-traffic peakedness formula, which is easy to apply with either known arrival-rate functions or data from system measurements. We compare these approximations to exact numerical results for the $M_t/M/s/0$ model with linear arrival rate.

Most real queueing systems reveal significant time variation in the arrival rates, e.g., see Hall (1991). However, queueing models with nonstationary arrival processes are relatively difficult to analyze. Here we propose a way to approximate a queueing model with a nonstationary arrival process over a fixed time interval by a queueing model with a stationary arrival process. *The main idea is to introduce extra stochastic variability to approximately capture the fluctuations over time in the deterministic arrival rate function.*

The specific model we consider is the *nonstationary Erlang loss model* or $M_t/G/s/0$ queue, which has s servers in parallel, no extra waiting spaces, and i.i.d. service times with a general distribution that are independent of a nonstationary Poisson arrival process. The arrival process has a deterministic arrival-rate function $\lambda(t)$ defined over the time interval $[0, T]$. Our goal is to predict the average blocking probability over the interval.

A common approach to this problem is to compute the average arrival rate over this time interval, $\bar{\lambda} \equiv T^{-1} \int_0^T \lambda(t)\, dt$, and approximate the nonstationary $M_t/G/s/0$ model by the associated stationary $M/G/s/0$ model, obtained by replacing the nonstationary Poisson process with a stationary Poisson arrival process having rate $\bar{\lambda}$. Here we develop an alternative procedure that can do much better in predicting the average blocking probability over the interval $[0, T]$.

Our approach is to act as if the M_t arrival process were a stationary G arrival process, and then try to approximately characterize the stochastic variability. For this we use the concept of *peakedness*; see Eckberg (1983), Whitt (1984), and Chapter 7 of Wolff (1989). This approach has the advantage that it can be applied to general G_t arrival processes, without actually knowing how much of the variability is due to fluctuations in a time-varying deterministic arrival rate or non-Poisson stochastic variability. Over appropriate subintervals we simply act as if the arrival process were actually stationary and try to assess the stochastic variability under this assumption. From that perspective, our approach can be regarded as part of current engineering practice (when indeed an attempt is made to estimate peakedness). From that point of view, we are investigating the quality of the approximations that are being done.

We evaluate our approximations for the special case of exponential service times by making comparisons with exact numerical results obtained from a discrete-time Markov chain (DTMC) algorithm, which is described in Section 5 of Davis et al. (1995). Runge-Kutta methods for solving ordinary differential equations could also have been used, as in Green et al. (1991) and Taaffe and Ong (1987). The DTMC algorithm also applies to the $PH_t/PH_t/s/r$ model (with more required computation); indeed it is used in Davis et al. to study the influence of the service-time distribution on the time-dependent blocking in the $M_t/PH/s/0$ model. (The impact of the service-time distribution can be substantial.) It is significant that our approach to the $M_t/G/s/0$ model here, replacing it by a $G/GI/s/0$ model, can capture the effect of this service-time distribution; see Section 4.2 of Davis et al.

Variations of our approximation methods also can be applied to queueing models with waiting rooms. However, the time-dependent behavior of a loss model is easier to analyze because the system has less memory, i.e., high arrival rates in the past can at most make all servers be busy

Subject classifications: Queues, nonstationary: stationary-process approximations. Queues, approximations: the nonstationary Erlang loss model
Area of review. Stochastic Processes and Their Applications

0030-364X/96/4406-0976 $01.25

initially. For loss models, it is more reasonable to do the kind of "local" analysis we do.

Here is how the rest of this paper is organized. In Section 1 we discuss initial conditions and specify the performance measures of interest. In Section 2 we introduce the stationary-process approximations. In Section 3 we make a connection between our approximation method and a method for estimating peakedness in a stationary model, due to Holtzman and Jagerman (1979). In Section 4 we examine numerical examples, and in Section 5 we state our conclusions.

1. INITIAL CONDITIONS AND PERFORMANCE MEASURES

We do not specify the initial conditions; we hope to obtain reasonable approximations without explicitly considering the initial conditions. In our numerical examples we will consider three cases: starting with all servers idle, starting with all servers busy, and starting with the stationary distribution associated with the initial arrival rate $\lambda(0)$. We consider the stationary distribution with the initial arrival rate as the principal case, because it seems to be a reasonable approximation. By considering all three cases, we investigate the impact of the initial conditions.

Throughout this paper we assume that the individual mean service time is 1. In order to be able to ignore the initial conditions, we assume that T is not too small, e.g., $T \geq 5$. In our numerical examples we let $T = 12$.

Let $Q(t)$ represent the number of busy servers at time t. The time-dependent blocking probability is

$$\beta(t) = P(Q(t) = s) . \tag{1}$$

Note that $\beta(t)$ is the probability that the system is full at time t, which would be the probability of blocking if there were an arrival at time t (an event of probability zero). Since the arrival process is Poisson, it has independent increments. Hence $\beta(t)$ also represents the conditional probability of blocking at time t given that there is an arrival at time t. (Conditioning on an arrival at time t does not alter the distribution of the remaining arrival process.)

We consider two "average blocking" performance measures. First, we consider the expected proportion of time during $[0, T]$ that the system is full,

$$\bar{\beta}_t = \frac{1}{T}\int_0^T \beta(t)\, dt , \tag{2}$$

and second, we consider the ratio of the expected number of lost customers to the expected number of arrivals. Letting $B(t)$ be the number of blocked calls in the interval $[0, t]$ and $A(t)$ the number of arrivals in $[0, t]$, this ratio is

$$\bar{\beta}_c = \frac{EB(T)}{EA(T)} = \frac{\int_0^T \lambda(t)\beta(t)\, dt}{\int_0^T \lambda(t)\, dt} , \tag{3}$$

where the final display depends on having an M_t arrival process.

We call $\bar{\beta}_t$ in (2) the *time congestion* and $\bar{\beta}_c$ in (3) the *call congestion*. This is consistent with usage for stationary non-Poisson arrivals. Since our goal is usually to satisfy customer demand, we usually are primarily interested in the call congestion.

Often the time congestion and call congestion do not differ greatly, but this is not always the case. For example, having all arrivals in $[0, T]$ occur at the single instant T clearly maximizes call congestion and minimizes time congestion, yielding $\bar{\beta}_t = 0$.

Note that (3) is the ratio of the expected number blocked to the expected number of arrivals; which is *not* necessarily the expected proportion of arrivals blocked, where 0/0 is defined to be 0. (In general, we do not have $E(X/Y) = EX/EY$.) To better understand (3), suppose that we simulate n independent replications of the $M_t/M/s/0$ model over $[0, T]$ and calculated the numbers A_i and B_i of arrivals and blocked arrivals in run i, $1 \leq i \leq n$, and their averages $\bar{A}_n = n^{-1}\sum_{i=1}^n A_i$ and $\bar{B}_n = n^{-1}\sum_{i=1}^n B_i$. Then, by the law of large numbers, the estimate $\bar{B}_n/\bar{A}_n$ converges as $n \to \infty$ to $\bar{\beta}_c$. However, in general we do *not* have $E(\bar{B}_n/\bar{A}_n) = \bar{\beta}_c$. Continuing the example above further, suppose that all arrivals in $[0, T]$ occur at the instant T and that the number of arrivals is equally likely to be s or $3s$. Then $\bar{\beta}_c = 1/2$, while the expected proportion of blocked customers is 1/3.

We conclude this section by pointing out that $\bar{\beta}_c$ in (3) is *nearly* the expected proportion of blocked calls, due to Proposition 10.1(c) of Massey and Whitt (1994). Surprisingly, for an M_t arrival process it turns out that the call congestion equals the expected conditional proportion of calls blocked, given that there is at least one arrival, i.e.,

$$\bar{\beta}_c = E\left[\frac{\int_0^T \beta(t)\, dA(t)}{A(T)}\middle| A(T) > 0\right] ,$$

where the integration is the sum associated with each sample path.

2. STATIONARY-PROCESS APPROXIMATIONS

We now consider approximations for the time congestion $\bar{\beta}_t$ in (2) and the call congestion $\bar{\beta}_c$ in (3) in the $M_t/G/s/0$ model. An important consideration is the length T of the relevant subinterval. The length T should be neither too small nor too large. The relevant time scale is the mean service time, which here has been fixed at 1. In order to predict the blocking with precision (i.e., not to be doing too much averaging in (2) and (3)), we would usually prefer to have T as small as possible. However, to have a reasonable approximation, the length T should be long enough that for most points t, e.g., for $T/4 \leq t < T$, most of the relevant past to determine the blocking at time t is included in the interval $[0, T]$. On the other hand, T should not be so large that the blocking in many subintervals (such as $((k-1)T/10, kT/10]$, $1 \leq k \leq 10$) are nearly independent. Then there is needless averaging that reduces the precision of our performance measure.

Reasonable values of T might be between 6 and 20. We anticipate that $T = 1$ is too small, while $T = 100$ is too large. For example, consider telephone calls, with an average length of five minutes. Consistent with engineering practice, an appropriate length of the averaging interval might be one hour, which corresponds to $T = 12$. Indeed, in our numerical examples we let $T = 12$. With shorter or longer call holding times, the interval might range from fifteen minutes to two hours.

At the outset we assume that we have a reasonable value of T such as $T = 12$. Our approximation strategy is to use a stationary arrival process with the average arrival rate

$$\bar{\lambda} = \frac{1}{T}\int_0^T \lambda(t)\,dt\,. \tag{4}$$

The *stationary-Poisson approximation* is a stationary-Poisson (M) process with arrival rate $\bar{\lambda}$ in (4). The approximate time congestion and call congestion (which agree for M) are then given by the *Erlang blocking formula*

$$B(s, \bar{\lambda}) = (\bar{\lambda}^s/s!)\Big/\sum_{k=0}^{s}(\bar{\lambda}^k/k!)\,. \tag{5}$$

As we might well anticipate, the stationary-Poisson approximation typically underestimates the call congestion and the time congestion because it ignores the time fluctuations of the arrival rate. A second stationary approximation uses a non-Poisson stationary point process to approximate the nonhomogeneous Poisson arrival process. We introduce extra stochastic variability into the stationary point process to represent the fluctuations over time. Obviously there are several ways to do this. We describe two.

Our *first approach* is to act as if the arrival process is a stationary point process (i.e., has stationary increments). Indeed, we *make* the nonstationary Poisson process a stationary point process by, first, considering the periodic extension in which independent copies of the original Poisson process on $(0, T]$ are placed on $(kT, (k + 1)T]$ for all integers k and, second, moving the origin to a point uniformly distributed in the interval $(0, T]$. (It is easy to verify that the new process is a stationary point process.)

Let $N \equiv \{N(t) : t \geq 0\}$ be the stationary point process so constructed. We partially characterize the variability of the stationary point process N using the *index of dispersion for counts* (IDC), i.e.,

$$I(t) \equiv \frac{\operatorname{Var} N(t)}{EN(t)} = \frac{\operatorname{Var} N(t)}{\bar{\lambda}t}\,, \tag{6}$$

as in Cox and Lewis (1966) and Fendick and Whitt (1989).

By our construction, the number of arrivals in any interval of length T is Poisson with mean $\bar{\lambda}T$. Hence $I(kT) = 1$ for all $k \geq 1$. Moreover, it is not difficult to show that $I(t) \to 1$ as $t \to 0$. However, $I(t)$ is typically greater than 1 for $0 < t < T$.

Given that the arrival process is a nonhomogeneous Poisson process with arrival-rate function $\lambda(t)$, it is not difficult to calculate the IDC $I(t)$. Note that

$$I(t) \equiv \frac{\operatorname{Var} N(t)}{EN(t)} = \frac{E[N(t)^2] - (\bar{\lambda}t)^2}{\bar{\lambda}t}, \quad \text{where} \tag{7}$$

$$E[N(t)^2] = \frac{1}{T}\int_0^T [\Lambda_t(s) + \Lambda_t(s)^2]\,ds, \quad \text{with} \tag{8}$$

$$\Lambda_t(s) = \begin{cases} \displaystyle\int_s^{s+t} \lambda(u)\,du, & 0 \leq s \leq T - t, \\ \displaystyle\int_s^T \lambda(u)\,du + \int_0^{s-T+t} \lambda(u)\,du, & T - t \leq s \leq T. \end{cases} \tag{9}$$

Since the mean service time in the loss model is 1, it is natural to use

$$c^2 \equiv I(1)\,, \tag{10}$$

as an approximate variability parameter. We suggest (10) because the mean service time indicates the time scale we are interested in. The parameter c^2 can be thought of as the squared coefficient of variation (variance divided by the square of the mean) in a renewal process approximation for the stationary process N; see Fendick and Whitt.

Formulas (7)–(10) are convenient for obtaining the variability parameter c^2 from an explicit arrival rate function $\lambda(t)$. However, we could also use another variant of this approach with data. We could apply any procedure for estimating $I(1)$ to the stationary point process obtained from assuming the periodic independent extension with origin uniformly distributed over $[0, T]$; see Cox and Lewis and Section III.B of Fendick et al. (1991). The arrival process need not be Poisson and the arrival rate function need not be given explicitly.

However, it is often reasonable to assume that the arrival process is a nonhomogeneous Poisson process. In that case, it is often natural to assume that the arrival rate is linear when the subintervals are not too long. Estimation procedures for that case are studied in Massey et al. (1996). Hence, in this paper, we will consider the special case in which $\lambda(t) = a + bt$, $0 \leq t \leq T$. We can of course compute $I(1)$ directly from (7)–(9). For simplicity, we introduce the approximation

$$\Lambda_t(s) \approx (a + bs)t, \quad 0 \leq s \leq T\,. \tag{11}$$

We obtain

$$c^2 \equiv I(1) \approx 1 + \frac{b^2T^2}{6(2a + bT)}\,, \tag{12}$$

from (7)–(11).

Our *second approach* is motivated by what we might do with system measurements. (See Section 3 for further discussion on this point.) As with the first approach, we act as if the arrival process is stationary. We might then estimate the variability by looking at the counts in n disjoint subintervals of length T/n. As an approximation, we assume that the numbers of points in different subintervals are independent. For this independence approximation, the intervals should not be too short. Our approximating discrete-time arrival process is thus a process with stationary independent increments.

Hence, our second approach to the M_t arrival process is to divide the interval $(0, T]$ into n subintervals $((k-1)T/n, kT/n]$, $1 \leq k \leq n$, and perform a random permutation on them. We act as if what we see in any one subinterval is a mixture of the resulting n Poisson distributions.

To describe the distribution in a randomly selected interval, let

$$\lambda_k = \int_{(k-1)T/n}^{kT/n} \lambda(u)\, du, \quad 1 \leq k \leq n . \tag{13}$$

Then the number of arrivals in any one (random) subinterval has mean

$$\bar{\lambda}_n = \frac{1}{n} \sum_{k=1}^{n} \lambda_k = \frac{\bar{\lambda} T}{n}, \tag{14}$$

and variance

$$\sigma_n^2 = \bar{\lambda}_n + \frac{1}{n} \sum_{k=1}^{n} (\lambda_k - \bar{\lambda}_n)^2 . \tag{15}$$

(Use the fact that the second moment is the mixture of the second moments.) The number of arrivals in $(0, T]$ thus has mean $n\bar{\lambda}_n = \bar{\lambda} T$. Assuming stationary and independent increments for the discrete-time process, the number of arrivals in $[0, T]$ has variance

$$n\sigma_n^2 = \bar{\lambda} T + \sum_{k=1}^{n} (\lambda_k - \bar{\lambda}_n)^2 . \tag{16}$$

Based on this analysis, we approximate the original M_t arrival process by a stationary point process $N \equiv \{N(t) : t \geq 0\}$ partially characterized by its intensity $E[N(1)] = \bar{\lambda}$ in (4) and a variability parameter c^2 corresponding to the variance to mean ratio for counts in $[0, T]$, i.e.,

$$c^2 \equiv \frac{\mathrm{Var}[N(T)]}{E[N(T)]} = 1 + \frac{1}{\bar{\lambda} T} \sum_{k=1}^{n} (\lambda_k - \bar{\lambda}_n)^2 . \tag{17}$$

We now consider simplifications of Formula (17). First, assuming that λ is constant over each subinterval $((k-1)T/n, kT/n]$, we can write

$$c^2 \approx 1 + \frac{1}{\bar{\lambda} T} \left(\frac{T}{n}\right) \int_0^T (\lambda(s) - \bar{\lambda})^2\, ds . \tag{18}$$

From (18) it is clear that the formula for c^2 depends critically on n. Indeed, the formula for the variance in (16) depends critically on n: For $n = 1$, $n\sigma_n^2 = \bar{\lambda} T$; while $n\sigma_n^2 \rightarrow \bar{\lambda} T$ as $n \rightarrow \infty$. Intermediate values of n capture the additional variability due to the fluctuations in the arrival rate. It seems reasonable to let the length of a measurement interval be of the order of one mean service time, because that is the relevant time scale in the loss model. Hence, as a further simplification, in (18) we let $n \approx T$ and obtain

$$c^2 \approx 1 + \frac{1}{\bar{\lambda} T} \int_0^T (\lambda(s) - \bar{\lambda})^2\, ds . \tag{19}$$

Formula (19) is a convenient simplification for working with explicit arrival-rate functions, while (17) is convenient for measurements. Note that (17) can capture both stochastic variability and deterministic time-variations in the arrival rate function in G_t arrival processes.

Before proceeding, we want to exclude a pathological situation. We want to exclude extremely rapid oscillations in $\lambda(t)$ from c^2 in (18) and (19). It is known that if we have an M_t process with arrival rate function $\lambda(t) = a + b \sin(\gamma t)$ with very high frequency γ that, from the point of view of the queueing models, the M_t process is actually approximately equivalent to an M process with constant arrival rate function a; e.g., see Theorem 4.5 of Eick et al. (1993b). Another example is $\lambda(t) = a + (-1)^{\lfloor nt \rfloor} b$ for $0 < b < a$ and very large n. For each n such that nT is an even integer, $\int_0^T (\lambda(s) - \bar{\lambda})^2\, ds = b^2 T$, but as $n \rightarrow \infty$ the process approaches a Poisson process with constant rate a. Hence, we assume that before (18) or (19) is applied the arrival-rate function is appropriately smooth over short time scales. For example, if necessary, we might replace $\lambda(t)$ by its average over a single mean service time, i.e., $\bar{\lambda}(t) = \int_{t-1}^{t} \lambda(s)\, ds$, and then calculate c^2 in (18) or (19). Note that extra smoothing is not necessary with (17).

For the special case of a linear arrival rate function, i.e., when $\lambda(t) = a + bt$, $0 \leq t \leq T$, (19) becomes

$$c^2 = 1 + \frac{b^2 T^2}{6(2a + bT)}, \tag{20}$$

Note that (20) agrees with (12).

Finally, we approximate the distribution of $Q(t)$ in the $M_t/G/s/0$ model over the interval $(0, T]$ by the behavior of the stationary $G/G/s/0$ model with this arrival process N partially characterized by the parameters $\bar{\lambda}$ and c^2. To analyze the $G/G/s/0$ model, we use further approximations; for background, see Eckberg, Whitt, and Chapter 7 of Wolff. In particular, below we use a heavy-traffic peakedness approximation. The *peakedness* is defined as the ratio of the variance to the mean of the steady-state number of busy servers in an associated infinite-server model with the same service-time distribution and the same arrival process. By Little's law, the steady-state mean number of busy servers in the infinite-server model is always the arrival rate divided by the individual service rate. To obtain more tractable formulas for the peakedness, we consider the limiting behavior of the peakedness as the arrival rate grows. The heavy-traffic peakedness is

$$z = 1 + (c^2 - 1) \frac{1}{E[S]} \int_0^\infty [1 - G(t)]^2\, dt , \tag{21}$$

where G is the service-time cdf and c^2 comes from (10), (12), (19), or (20); see Eckberg and p. 692 of Whitt. Combining (19) and (21), we obtain the expression

$$z = 1 + \left(\frac{1}{\bar{\lambda} T}\int_0^T (\lambda(s) - \bar{\lambda})^2\, ds\right) \cdot \left(\frac{1}{E[S]}\int_0^\infty [1 - G(t)]^2\, dt\right) \geqslant 1 . \qquad (22)$$

For the case of exponential service times primarily considered in this paper,

$$z = 1 + \frac{c^2 - 1}{2} = 1 + \frac{1}{2\bar{\lambda} T}\int_0^T (\lambda(s) - \bar{\lambda})^2\, ds . \qquad (23)$$

For the special case of exponential service times, we can approximate the time and call congestion in a $G/M/s/0$ system where the arrival process is partially specified by its rate and peakedness using the *equivalent random method*; see Section 7.5 of Wolff. The idea is to replace the given parameter triple ($\bar{\lambda}$, s, z) with the parameter triple ($\bar{\lambda}$, $\bar{s} + s$, 1) which produces approximately the same overflow rate. An algorithm for the equivalent random method is in Jagerman (1984). We call this approximation the *stationary-peakedness (PK) approximation.*

Alternatively, for the $G/G/s/0$ model, we could use the *Hayward approximation*; i.e., the blocking probability $B(s, \bar{\lambda}, z)$ is approximated by $B(s/z, \bar{\lambda}/z, 1) \equiv B(s/z, \bar{\lambda}/z)$, where $B(s, a)$ is the Erlang blocking formula for s servers and offered load a in (5) extended to nonintegral s; see Fredericks (1980), Jagerman (1984), Whitt, and Wolff.

It is important to note that our stationary-process approximations are invariant under time reversal; i.e., the approximations for $\lambda(t)$, $0 \leqslant t \leqslant T$, and $\lambda(T - t)$, $0 \leqslant t \leqslant T$, are the same. Moreover, our approximations do not depend on the initial conditions. It is natural to consider refinements to our stationary approximations that are sensitive to order and initial conditions, and we do consider an elementary one in Section 4 below. However, the stationary-process approximations here are appealing because of their simplicity.

3. ESTIMATING PEAKEDNESS FROM MEASUREMENTS

It is significant that the stationary-peakedness approximations above are closely related to what occurs if we act as if the M_t arrival process were a stationary G process and estimate the arrival rate and peakedness from a sample; see Cox and Lewis and Holtzman and Jagerman (1979). In particular, following Holtzman and Jagerman, suppose that we use (21) as the definition of peakedness with c^2 defined by $\mathrm{Var}[N(T)]/E[N(T)]$ as in (17). Suppose that we can observe the total number X_k of arrivals in the subinterval $((k - 1)T/n, kT/n]$ for each k, $1 \leqslant k \leqslant n$. Then, assuming that

$$\frac{\mathrm{Var}\, N(T)}{EN(T)} \approx \frac{\mathrm{Var}\, N(T/n)}{EN(T/n)} , \qquad (24)$$

a reasonable tractable estimate of c^2 is the ratio $S_n^2/\bar{X}_n$ where S_n^2 is the sample variance and $\bar{X}_n$ is the sample mean of the n observations from the n subintervals of $(0, T]$, i.e.,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 . \qquad (25)$$

Indeed, we propose $\bar{X}_n$ and $S_n^2/\bar{X}_n$ as direct estimates of $\bar{\lambda}$ and c^2 in (4) and (17).

We remark that (24) above would hold as an *equality* if the process had stationary and *independent* increments. The idea in (24) is that both T and T/n are sufficiently large that the ratios are reasonable approximations for the limit as $T \to \infty$.

Following Holtzman and Jagerman, assume that the expected value of the ratio is approximately equal to the ratio of the expectations, so that we can approximate c^2 by $c_e^2 = E[S_n^2]/E[\bar{X}_n]$, which leads to a minor variant of (17). In particular, when this analysis is applied to a nonhomogeneous Poisson process, we obtain

$$c_e^2 = 1 + \frac{1}{\bar{\lambda}(n - 1)}\sum_{k=1}^{n} (\lambda_k - \bar{\lambda}_n)^2 . \qquad (26)$$

To compute (26), the expectations ES_n^2 and $E\bar{X}_n$ are computed with respect to the original M_t process. Note that (26) and (17) yield the same result when $T = n - 1$, which supports our choice of $n \approx T$.

Holtzman and Jagerman calculated (26) to show that inflated values of peakedness would result from nonstationarity when the arrival process is nonhomogeneous Poisson. Indeed, since the distribution of $Q(t)$ in an $M_t/G/\infty$ model is Poisson for all t (for appropriate initial conditions), we always have the *time-dependent peakedness* of a nonhomogeneous Poisson process being

$$z(t) \equiv \frac{\mathrm{Var}[Q(t)]}{E[Q(t)]} = 1 \quad \text{for all } t .$$

Nevertheless, we believe that the larger peakedness in (21) with (26), (17), or (19) can be useful to represent the fluctuations over time in approximations.

4. NUMERICAL EXAMPLES

To see how the stationary-process approximations perform, we now examine some examples. To focus on the invariance of our approximations under time reversal of the arrival-rate function, we consider time-reversed versions of all our arrival-rate functions. Motivated by telephone applications in which average holding times are about five minutes and a time interval of interest is one hour, we consider $M_t/M/s/0$ models having mean service time 1 over the time interval [0, 12]. Moreover, motivated by what seems to be a simple natural approximation, each example has λ linear over the time interval in question.

Example 1. In our first example the increasing linear arrival rate function is $\lambda(t) = 10 + 0.833t$ and the decreasing time-reversed arrival rate function is $\lambda(t) = 20 - 0.833t$, $0 \leqslant t \leqslant 12$. For both arrival rate functions, the average arrival rate is $\bar{\lambda} = 15$ and the peakedness via (23) and (20) is $z = 1.278$. (If we had used (21) and (26) instead for $n = 12$ corresponding to five-minute summary measurements, then we would have $z = 1.30$.)

Four different numbers of servers are considered: 20, 25, 30, and 35. Exact performance measures are calculated by the DTMC algorithm in Section 5 of Davis et al. for three different initial conditions: starting out empty, starting out full (all servers busy), and starting out with the stationary distribution associated with $\lambda(0)$. The exact call congestion in (3) and time congestion (2) are displayed in each case in Table I. For the case of increasing λ, the maximum time congestion is also displayed. In addition, the number of servers required to achieve (be less than) 0.01 and 0.001 blocking criteria are indicated (under the assumption of stationary initial condition).

We calculated the approximations using D. L. Jagerman's program TRAFCALC based on Jagerman (1984). From Table I we see that the blocking probability with the stationary-Poisson approximation is consistently too low (call and time congestion agree for a homogeneous Poisson arrival process), but that the stationary-peakedness (PK) approximation is not too bad. For each number of servers considered, the PK value is within the interval of possible exact values spanned by the six cases (up and down with three initial conditions). However, the range of exact values for each s is rather wide. From Table I, we see that the relative range increases as the number of servers increases (and the blocking probability decreases).

Even restricting attention to stationary initial conditions, we see that the blocking probability is significantly larger going down than going up (e.g., 0.017 versus 0.011 at $s = 25$). Upon reflection, this is consistent with intuition. This intuition can be strengthened by looking at the explicit $M_t/G/\infty$ results; e.g., Theorem 2.4, (7) and (8) of Eick et al. (1993a).

The PK approximation is quite good when $\lambda(t)$ is increasing and the blocking probability is not too small (e.g., above 0.001), but PK significantly underestimates the time and call congestion when $\lambda(t)$ is decreasing. A refinement that might be used is to compute the parameters $\bar{\lambda}$ and c^2 using $\lambda(t)$ in the subinterval $[-1, T]$ instead of $[0, T]$. Since this refinement helps here only when $\lambda(t)$ is decreasing, the refinement could be confined to that case. For example, this refinement when $\lambda(t)$ is decreasing increases $(\bar{\lambda}, z)$ from (15, 1.28) to (15.42, 1.35). The most significant effect is on $\bar{\lambda}$. The new call (time) congestion with $(\bar{\lambda}, z) = (15.42, 1.35)$ and 20, 25, 30, and 35 servers becomes 0.074, 0.0143, 0.0014, and 0.00007, (0.061, 0.0115, 0.0011, and 0.00006), respectively, which indeed is an improvement.

While the exact blocking probabilities with λ increasing and decreasing (with stationary initial conditions) are quite different, the number of servers required to meet 0.01 and 0.001 blocking criteria in Table I are not very different. Indeed, they usually differ by only 1. The PK approximation performs pretty well from this perspective too, with a maximum error of 2 servers. However, the stationary-Poisson approximation produces a significant error at the 0.001 level.

Table I

A Comparison of Approximations with Exact Call and Time Congestion (the Averages in (3) and (2)) in the $M_t/M/s/0$ Model with Mean Service Time 1 and Arrival Rate $\lambda(t)$, $0 \leqslant t \leqslant 12$, in Example 1. The Number of Servers Needed to Satisfy 0.01 and 0.001 Blocking Criteria Are Also Given.

		Exact up $\lambda(t) = 10 + 0.833t$			Exact down $\lambda(t) = 20 - 0.833t$		Equivalent random method (PK) $\bar{\lambda} = 15$ $z = 1.28$		
Number of servers	Initial condition	Call	Time	Max	Call	Time	Call	Time	Stationary Poisson approx. $\bar{\lambda} = 15$
	full	0.066	0.060		0.084	0.072			
20	stationary	0.059	0.051	0.151	0.073	0.064	0.062	0.053	0.046
	empty	0.059	0.050		0.048	0.044			
	full	0.0146	0.0143		0.0295	0.0248			
25	stationary	0.0110	0.0090	0.043	0.0172	0.0144	0.0100	0.0084	0.0050
	empty	0.0110	0.0090		0.0072	0.0065			
	full	0.0034	0.0044		0.0123	0.0095			
30	stationary	0.0011	0.00085	0.0058	0.0023	0.0018	0.00079	0.00065	0.00022
	empty	0.0011	0.00085		0.00051	0.00046			
	full	0.00178	0.00262		0.0070	0.0053			
35	stationary	0.000052	0.000031	0.0004	0.00015	0.00012	0.00003	0.00003	0.00000
	empty	0.000052	0.000031		0.000017	0.000015			
0.01 blocking level	stationary	26	25	29	27	26	25	25	24
0.001 blocking level	stationary	31	30	34	32	32	30	30	28

Table II
A Comparison of Approximations with Exact Call and Time Congestion (the Averages in (3) and (2)) in the $M_t/M/s/0$ Model with Mean Service Time 1 and Arrival Rate $\lambda(t)$, $0 \leq t \leq 12$, in Example 2. The Initial Number of Busy Servers Has the Stationary Distribution with $\lambda(0)$ in Each Case. The Number of Servers Needed to Satisfy 0.01 and 0.001 Blocking Criteria Are Also Given.

Number of servers	Exact up $\lambda(t) = 1.667t$			Exact down $\lambda(t) = 20 - 1.667t$		Equivalent random method (PK) $\bar{\lambda} = 10$ $z = 2.67$		Stationary Poisson approx. $\bar{\lambda} = 10$
	Call	Time	Max	Call	Time	Call	Time	
15	0.133	0.080	0.321	0.163	0.103	0.133	0.064	0.037
20	0.038	0.021	0.141	0.059	0.035	0.038	0.017	0.0019
25	0.0060	0.0033	0.0341	0.0145	0.0082	0.0071	0.0030	0.00003
30	0.00047	0.00025	0.0038	0.00198	0.00109	0.00087	0.00035	0.00000
0.01 blocking level	24		28	27		25		18
0.001 blocking level	29		32	32		31		21

Example 2. A similar but more extreme example is obtained by considering $\lambda(t) = 1.667t$ and $\lambda(t) = 20 - 1.667t$ for $0 \leq t \leq 12$ as in Table II. Here the average arrival rate is $\bar{\lambda} = 10$ and the peakedness from (23) and (20) is 2.67. (It would be 2.81 from (26) with $n = 12$.) As with Example 1, the stationary-Poisson approximation seriously underestimates the call and time congestion, while *PK* is a pretty good approximation for $\lambda(t)$ increasing, but underestimates the call and time congestion when $\lambda(t)$ is decreasing. Unlike Table I, PK actually overestimates the call congestion when $\lambda(t)$ is increasing for small blocking probabilities. As before, the exact blocking *probability* for decreasing λ *is* significantly greater than for increasing λ, with the ratio increasing as the number of servers increases.

Example 3. To have an example with more servers and less extreme (relative) slope, we consider $\lambda(t) = 98 + 4t$ and $\lambda(t) = 146 - 4t$, $0 \leq t \leq 12$, in Table III. Here the peakedness from (23) and (20) is 1.79. (It would be 1.85 from (26) with $n = 12$.) The results in this case are similar to those in Examples 1 and 2. The range of arrival rates in this example from 98 to 146 is large, but not nearly as large (relatively) as in Examples 1 and 2. Hence, this example may be considered more realistic (and less demanding). In many applications, the differences between the congestion measures with $\lambda(t)$ increasing and decreasing shown in Table III might not be considered extraordinarily great, in view of other uncertainties. With such modest standards, the stationary-peakedness approximation might be considered very suitable.

Table III
A Comparison of Approximations with Exact Call and Time Congestion (the Averages in (3) and (2)) in the $M_t/M/s/0$ Model with Mean Service Time 1 and Arrival Rate $\lambda(t)$, $0 \leq t \leq 12$, in Example 3. The Number of Servers Needed to Satisfy 0.01 and 0.001 Blocking Criteria Are Also Given.

Number of servers	Initial condition	Exact up $\lambda(t) = 98 + 4t$			Exact down $\lambda(t) = 146 - 4t$		Equivalent random method (PK) $\bar{\lambda} = 122$ $z = 1.79$		Stationary Poisson approx. $\bar{\lambda} = 122$
		Call	Time	Max	Call	Time	Call	Time	
135	full	0.0358	0.0326	0.1118	0.0448	0.0403	0.0374	0.0274	0.0198
	stationary	0.0336	0.0300		0.0416	0.0377			
	empty	0.0333	0.0297		0.0206	0.0195			
145	full	0.0150	0.0138	0.061	0.0235	0.0207	0.0143	0.0104	0.0044
	stationary	0.0135	0.0119		0.0195	0.0173			
	empty	0.0135	0.0119		0.0063	0.0059			
155	full	0.0051	0.0049	0.031	0.0117	0.0102	0.0040	0.0029	0.00053
	stationary	0.0039	0.0034		0.0072	0.0064			
	empty	0.0039	0.0034		0.0012	0.0011			
165	full	0.00171	0.00184	0.0063	0.00625	0.00532	0.00077	0.00055	0.00003
	stationary	0.00075	0.00064		0.00193	0.00168			
	empty	0.00075	0.00064		0.00014	0.00013			
0.01 blocking level	stationary	146			149		147		140
0.001 blocking level	stationary	160			167		165		153

For this model, if we estimate $(\bar{\lambda}, z)$ based on the interval $[-1, T]$ instead of $[0, T]$, then we obtain the parameter pairs (120, 1.93) and (124, 191) when $\lambda(t)$ is increasing and decreasing, respectively, instead of (122, 1.79). The heuristic refined *PK* approximations for call congestion with 135, 145, and 155 servers are 0.034, 0.0129, 0.0036 when $\lambda(t)$ is increasing and 0.046, 0.0197, 0.0064 when $\lambda(t)$ is decreasing. The corresponding time congestion values with 135, 145, and 155 servers are 0.024, 0.0089, 0.0025 when $\lambda(t)$ is increasing and 0.033, 0.0139, 0.0044 when $\lambda(t)$ is decreasing. As in Example 1, this heuristic refinement to PK seems to be effective. Indeed, it is reasonable for both increasing $\lambda(t)$ and decreasing $\lambda(t)$.

In contrast, the adjusted stationary-Poisson approximations for call and time congestion with rate 120(124) and 135, 145, and 155 servers are 0.015, 0.0029, and 0.00030 (0.025, 0.0063, and 0.00029), respectively. When $\lambda(t)$ is increasing, this modification obviously just makes a poor approximation worse. When $\lambda(t)$ is decreasing, this modification is an improvement, but the estimates are still way too low.

5. CONCLUSIONS

We have proposed approximating average blocking probabilities over a subinterval in a nonstationary loss model by approximations of the corresponding steady-state blocking probabilities in an associated stationary loss model. The underlying idea is that appropriate stochastic variability in the stationary arrival process can approximately capture the effect of the fluctuations over time in the deterministic arrival-rate function in the nonstationary model.

We have investigated the stationary-Poisson approximation and the stationary-peakedness approximation for the time and call congestion in (2) and (3) in the $M_t/M/s/0$ model. When the arrival rate function has significant time variation, as in Examples 1–3 in Section 3, the stationary-Poisson approximation performs poorly, as anticipated. The stationary-peakedness approximation performs significantly better, but the rather large differences between the congestion measures when $\lambda(t)$ is increasing and decreasing dramatically reveal limitations of the stationary-peakedness approximation. These differences motivate considering refinements and alternative methods that are not invariant under time reversal of $\lambda(t)$. We suggested one such refinement in our stationary-process framework, in particular, computing $\bar{\lambda}$ and c^2 based on the interval $[-1, T]$ instead of $[0, T]$ (assuming that the mean service time is 1). This refinement seems to help consistently when $\lambda(t)$ is decreasing, but not when $\lambda(t)$ is increasing. Overall, the stationary-peakedness approximation seems reasonably good, especially when time variations in the arrival rate function such as are described in Examples 1–3 here can be regarded as relatively extreme (highly variable) cases. To the extent the PK approximation and refinements are not adequate, the examples serve as motivation for alternative methods which more directly address the time-dependence.

ACKNOWLEDGMENT

The authors thank Stephen G. Eick and the referees for assistance.

REFERENCES

COX, D. R. AND P. A. W. LEWIS. 1966. *The Statistical Analysis of Series of Events*. Methuen, London.

DALEY, D. J. AND D. VERE-JONES. 1988. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.

DAVIS, J. L., W. A. MASSEY, AND W. WHITT. 1995. Sensitivity to the Service-Time Distribution in the Nonstationary Erlang Loss Model. *Mgmt. Sci.* **41,** 1107–1116.

ECKBERG, A. E. 1983. Generalized Peakedness of Teletraffic Processes. *Proc. Tenth Int. Teletraffic Congress*. Montreal, Canada, 4.4.6.3.

EICK, S. G., W. A. MASSEY, AND W. WHITT. 1993a. The Physics of the $M_t/G/\infty$ Queue. *Opns. Res.* **41,** 731–742.

EICK, S. G., W. A. MASSEY, AND W. WHITT. 1993b. $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Mgmt. Sci.* **39,** 241–252.

FENDICK, K. W., V. R. SAKSENA, AND W. WHITT. 1991. Investigating Dependence in Packet Queues with the Index of Dispersion for Work. *IEEE Trans. Comm.* **39,** 1231–1244.

FENDICK, K. W. AND W. WHITT. 1989. Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue. *Proc. IEEE,* **77,** 171–194.

FREDERICKS, A. A. 1980. Congestion in Blocking Systems—A Simple Approximation Technique. *Bell System Tech. J.* **59,** 805–827.

GREEN, L., P. KOLESAR, AND A. SVORONOS. 1991. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Opns. Res.* **39,** 502–511.

HALL, R. W. 1991. *Queueing Methods for Services and Manufacturing.* Prentice Hall, Englewood Cliffs, NJ.

HOLTZMAN, J. M. AND D. L. JAGERMAN. 1979. Estimating Peakedness from Arrival Counts. *Proceedings Ninth Int. Teletraffic Congress.* Torremolinos, Spain.

JAGERMAN, D. L. 1984. Methods in Traffic Calculations. *AT&T Bell Lab. Tech. J.* **63,** 1283–1303.

MASSEY, W. A., G. A. PARKER, AND W. WHITT. 1996. Estimating the Parameters of a Nonhomogeneous Poisson Process with Linear Rate. *Telecommunication Systems* **5,** 361–388.

MASSEY, W. A. AND W. WHITT. 1994. A Stochastic Model to Capture Space and Time Dynamics in Wireless Communication Systems. *Prob. Eng. Inf. Sci.* **8,** 541–569.

TAAFFE, M. R. AND K. L. ONG. 1987. Approximating Ph(t)/M(t)/S/C Queueing Systems. *Anns. Opns. Res.* **8,** 103–116.

WHITT, W. 1984. Heavy-Traffic Approximations for Service Systems with Blocking. *AT&T Bell. Lab. Tech. J.* **63,** 163–175.

WOLFF, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.