MANAGEMENT SCIENCE
Vol. 37, No. 3, March 1991
Printed in U.S.A.

THE POINTWISE STATIONARY APPROXIMATION FOR $M_t/M_t/s$ QUEUES IS ASYMPTOTICALLY CORRECT AS THE RATES INCREASE*

WARD WHITT
AT&T Bell Laboratories, Murray Hill, New Jersey 07974-2070

Green, Kolesar and Svoronos (in press) and Green and Kolesar (in press) use numerical methods to investigate the behavior of multiserver Markov queues with a Poisson arrival process having a sinusoidal arrival rate. For this model they propose an approximation for long-run average performance measures called the pointwise stationary approximation (PSA), which consists of an appropriate weighted average of the performance measure that would result at each point in time if the system were stationary with the arrival rate that applies at that point in time. In this paper we verify their conjecture that PSA is asymptotically correct as the service and arrival rates increase with the instantaneous traffic intensity held fixed (corresponding to long arrival rate cycles). We actually establish both pointwise and average versions of this result for general time-dependent birth-and-death processes.

(QUEUES; NONSTATIONARY QUEUES; BIRTH-AND-DEATH PROCESSES; LIMIT THEOREMS; APPROXIMATIONS)

1. Introduction

Green, Kolesar and Svoronos (in press) and Green and Kolesar (in press) investigate the behavior of $M_t/M/s$ queues with periodic sinusoidal arrival rates. They (abbreviated GKS) propose an approximation for average performance measures called the *pointwise stationary approximation* (PSA) which consists of an appropriate weighted average of the performance measure that would result at each point in time if the system were stationary with the arrival rate that applies at that point in time. (Obviously this approximation breaks down for unbounded performance measures if the instantaneous traffic intensity is allowed to exceed 1, but that case is ruled out.) They empirically observe that the quality of the approximation improves as the individual service rate increases with the instantaneous traffic intensities remaining unchanged (less than 1), and conjecture, quite naturally, that the approximation is asymptotically correct in the limit. In this paper we verify that conjecture. We actually establish a stronger result for a more general model: we show that both a pointwise version of PSA and the integrated (average) version of PSA considered by GKS are asymptotically correct for general time-dependent birth-and-death processes (TDBD's). In TDBD's, the transition rates can depend on both the state and time.

The limit for TDBD's is obtained by multiplying the transition rates by n and letting $n \to \infty$. The limiting behavior for this model (and more general models) should be anticipated because the process approaches steady state (locally) more quickly as the rates increase. Indeed, the limit corresponds exactly to the uniform acceleration used by Massey (1985) to analyze the $M_t/M_t/1$ queue. (Related work was done by Newell 1968, 1982 and McClish 1979.) Massey goes beyond the relatively simple limit we consider to develop refinements. Indeed, the 0th order term in Massey's Theorem 1 is just the pointwise version of PSA for the $M_t/M_t/1$ queue. Formula (9.1) of Newell (1982, p. 263) also expresses a variant of the pointwise version of PSA. The upper bound in Rolski (1986) is a variant of PSA as well. (A deterministic periodic arrival rate can be expressed

* Accepted by Linda Green; received October 29, 1989. This paper has been with the author 2 weeks for 1 revision.

0025-1909/91/3703/0307$01.25

as the limit of a sequence of Markov intensity processes.) Indeed, PSA goes back to Palm (1943, 1988). Nevertheless, a relatively simple proof of the limit for TDBD's seems to be of interest. It is also of interest to see how PSA performs, as shown by Green and Kolesar (in press) for the $M_t/M/s$ model with sinusoidal arrival rates.

The rest of this paper is organized as follows. In §2 we define the TDBD model. In §3 we state the limit theorem and two corollaries for this model, and in §4 we prove them. We conclude in §5 with some discussion; e.g., there we propose a new approximation based on a stationary model, in which the arrival rate is determined by averaging the instantaneous arrival rate over an interval prior to the time of interest. In particular, we suggest making the length of the interval proportional to the mean service time. This *average stationary approximation* (ASA) may be any attractive alternative to the two relatively extreme schemes discussed by GKS, namely, using the instantaneous arrival rate (PSA) or using the long-run average arrival rate. With the averaging, we can treat short periods of instability; only the average traffic intensities must be less than one. The averaging also produces a lag in the performance measure behind the arrival rate that is observed in examples.

2. Time-Dependent Birth-and-Death Processes

For each integer $k \geq 0$, let $\lambda(t, k)$ and $\mu(t, k)$ be nonnegative bounded real-valued functions of $t \geq 0$ with left and right limits everywhere. Moreover, assume that there exist constants λ, μ and k^* such that

$$\lambda(t, k) \leq \lambda < \mu \qquad \text{for all } t \text{ and } k, \text{ and} \tag{1}$$

$$\mu(t, k) \geq \mu \qquad \text{for all } t \text{ and} \qquad k \geq k^*. \tag{2}$$

We regard $\lambda(t, k)$ and $\mu(t, k)$ as time-dependent birth and death rates in state k for a time-dependent birth-and-death process (TDBD) on the nonnegative integers. The special case in which $\lambda(t, k) = \lambda(t)$ (i.e., $\lambda(t, k)$ is independent of k) and $\mu(t, k) = \mu \min \{k, s\}$ for some constant μ is the $M_t/M/s$ queue. If, in addition, $\lambda(t)$ is sinusoidal, then we have the model considered by GKS.

The TDBD can be constructed using independent Poisson processes, with one Poisson process associated with each state k. Let $\mu(k)$ be an upper bound for $\mu(t, k)$, which exists by assumption. Let the Poisson process associated with state k have intensity $\alpha(k) = \lambda + \mu(k)$. Start the TDBD with some proper initial distribution on the nonnegative integers. If the process starts in state k, then let the Poisson process in state k generate the first potential transition, and suppose that it occurs at time t. Let this potential transition be a real jump up to $k + 1$ with probability $\lambda(t, k)/\alpha(k)$; let this transition be a real jump down to $k - 1$ with probability $\mu(t, k)/\alpha(k)$; and let this potential transition not correspond to a real transition with probability

$$[\alpha(k) - \lambda(t, k) - \mu(t, k)]/\alpha(k).$$

After time t generate the next potential transition using the Poisson process associated with the resulting state, and continue inductively. By assumptions (1) and (2), there can be only finitely many transitions in finite time. (This is a consequence of comparison results in §4.)

3. The Limit Theorem

We now consider a family of TDBD's indexed by n. For $n \geq 1$, let the birth and death rate functions $\lambda_n(t, k)$ and $\mu_n(t, k)$ be defined in terms of the birth and death rate functions in §2 simply by increasing the rates, i.e., by

$$\lambda_n(t, k) = n\lambda(t, k) \qquad \text{and} \qquad \mu_n(t, k) = n\mu(t, k). \tag{3}$$

Let the initial state be given by the proper nonnegative integer-valued random variable $X(0)$. For each $n \geq 1$, let $\{X_n(t): t \geq 0\}$ be the resulting TDBD with rate functions in (3) and $X_n(0) = X(0)$.

For each $s > 0$, let $\{Y_s(t): t \geq 0\}$ be a time-homogeneous birth-and-death process (BD) with birth and death rate functions $\lambda_s(k)$ and $\mu_s(k)$ defined by

$$\lambda_s(k) = \lambda(s, k) \qquad \text{and} \qquad \mu_s(k) = \mu(s, k). \tag{4}$$

Let $\Rightarrow$ denote convergence in distribution; see Billingsley (1968). From the previous assumptions, it is easy to see that, for each $s > 0$, there exists a proper random variable $Y_s(\infty)$ such that $Y_s(t) \Rightarrow Y_s(\infty)$ as $t \rightarrow \infty$. The random variable $Y_t(\infty)$ is the *pointwise stationary approximation* (PSA) to $X_n(t)$ for each n. To see this, note that the stationary distribution of $\{Y_s(t): t \geq 0\}$ is unchanged by multiplying both $\lambda_s(k)$ and $\mu_s(k)$ by n for all k.

Our main result is that PSA is asymptotically correct (in a pointwise sense instead of an average sense).

THEOREM 1. *If $\lambda(t, k)$ and $\mu(t, k)$ satisfying (1) and (2) are strictly positive and left-continuous at t for all k, then*

$$X_n(t) \Rightarrow Y_t(\infty) \qquad \text{as } n \rightarrow \infty.$$

The following result implies that PSA is asymptotically optimal in the integral or average sense of GKS.

COROLLARY 1. *If, under the assumptions of Theorem 1, $f(t)$ and $w(t)$ are bounded measurable real-valued functions; then*

$$\lim_{n \rightarrow \infty} \int_a^b Ef[X_n(t)]w(t)dt = \int_a^b Ef[Y_t(\infty)]w(t)dt$$

for $0 \leq a < b < \infty$.

COROLLARY 2. *If, in addition to the assumptions of Theorem 1 and Corollary 1, $E[X(0)^k] < \infty$, then*

$$\lim_{n \rightarrow \infty} \int_a^b E[X_n(t)^j]w(t)dt = \int_a^b E[Y_t(\infty)^j]w(t)dt$$

for all $j \leq k$ and $0 \leq a < b < \infty$.

4. Proofs

The essential idea in the proof of Theorem 1 is that the scaling in (3) makes the processes approach steady-state (locally) faster as n increases. Equivalently, the scaling in (3) is tantamount to considering TDBD's with the original rate functions in a different time scale; i.e., the nth process has the same rate functions stretched by a factor of n. With this interpretation, the rate functions become asymptotically locally constant as $n \rightarrow \infty$. In the context of GKS, this means long cycles for given amplitude.

We actually prove Theorem 1 by bounding the TDBDs $X_n(t)$ above and below by BD's in the neighborhood of t and showing that these BD's converge. Since the BD's can be made arbitrarily close, the process $X_n(t)$ sandwiched between them must converge as well.

For this purpose, we need a comparison result for TDBD's. The following is a corollary to Theorem 9 of Whitt (1981), but we give a direct proof. We say that one random variable X_1 is stochastically less than or equal to another X_2, and write $X_1 \leq_{st} X_2$, if $Ef(X_1) \leq Ef(X_2)$ for all nondecreasing bounded real-valued functions f.

LEMMA 1. *For $i = 1, 2$, let $\{X_i(t): t \geq 0\}$ be TDBDs with rate functions $\lambda_i(t, k)$ and $\mu_i(t, k)$ such that each TDBD has only finitely many transitions in finite time w.p.1. If $X_1(0) \leq_{st} X_2(0)$, $\lambda_1(t, k) \leq \lambda_2(t, k)$ and $\mu_1(t, k) \geq \mu_2(t, k)$ for all t and k, then it is possible to construct TDBDs $\{\tilde{X}_i:(t): t \geq 0\}$ on the same sample space such that $\{\tilde{X}_i(t): t \geq 0\}$ has the same finite-dimensional distributions as $\{X_i(t): t \geq 0\}$ for each i and*

$$P(\tilde{X}_1(t) \leq \tilde{X}_2(t) \textit{ for all } t) = 1,$$

so that

$$X_1(t) \leq_{st} X_2(t) \qquad \textit{for all } t. \tag{5}$$

PROOF. First, construct $\tilde{X}_1(0) \leq \tilde{X}_2(0)$ in the usual way; i.e., let F_i be the *cdf* of $X_i(0)$ and let U be a uniform random variable on the interval $[0, 1]$; then let $\tilde{X}_i(0) = F_i^{-1}(U)$. Next, construct the two TDBD's as described in §2 using the same Poisson processes (independent of the uniform random variable above). When the two processes are in the same state, the next potential transition in both processes occurs at the same time. Generate the actual transitions at this time using a common uniform random variable (using a different independent uniform variable for each potential transition). Use the condition on the rates to guarantee that process 2 has a jump up whenever process 1 does (and possibly when process 1 does not) and that process 1 has a jump down whenever process 2 does (and possibly when process 2 does not). Hence, the conclusion follows by induction on the potential transition epochs when the two processes are in the same state. By assumption, the number of transitions when the two processes are in same state is finite in finite time, so that the construction covers all time. □

Let $\{Q_n(t): t \geq 0\}$ represent the queue length process (number in system) in an $M/M/1$ queue with arrival rate $n\lambda$ and service rate $n\mu$, with $\rho \equiv \lambda/\mu < 1$. Then $\{Q_n(t): t \geq 0\}$ is a BD with $Q_n(t) \Rightarrow Q(\infty)$ as $t \to \infty$, where

$$P(Q(\infty) = k) = (1 - \rho)\rho^k, \qquad k \geq 0.$$

(Note that the same limit holds for all n.) Let $Q_n(t; X)$ represent $Q_n(t)$ with $Q_n(0) = X$, where X is a random variable independent of $\{Q(t) - Q(0): t \geq 0\}$.

LEMMA 2. *Under the assumptions of Theorem* 1,

$$\begin{aligned} X_n(t) &\leq_{st} k^* + Q_n(t; X(0)) \\ &\leq_{st} k^* + X(0) + Q_n(t; 0) \\ &\leq_{st} k^* + X(0) + Q(\infty) \end{aligned}$$

for all n and t, where k^* is defined in (2).

PROOF. To establish the first inequality, apply Lemma 1. Note that we can regard $k^* + Q_n(t; X(0))$ as a TDBD with $\mu(t, k) = 0$ for $k \leq k^*$. By (1) and (2), the assumptions of Lemma 1 are satisfied. The dominating process is well known to have finitely many transitions in finite time. The second inequality is elementary. The third inequality follows because a birth-and-death process is stochastically monotone, see Keilson (1979, Chapter 9), so that $Q_n(t; 0)$ increases stochastically toward steady-state limit $Q(\infty)$. □

REMARK 1. Lemma 2 implies that, under the assumptions of Theorem 1, $\{X_n(t): n \geq 1\}$ is tight, so that any subsequence has a subsubsequence converging in distribution; see Billingsley (1968, pp. 9, 37). It only remains to show that $Y_t(\infty)$ must be the limit of any subsequence converging in distribution; see Billingsley (1968, p. 16). □

PROOF OF THEOREM 1. Fix $\epsilon > 0$. For all $n \geq 1$, we bound the TDBD $X_n(t)$ in the interval $[t - \epsilon, t]$ above and below by BD's. Let the upper bound BD $\{X_{\epsilon n}^+(t): t \geq 0\}$ have rate functions

$$\lambda^+_{\epsilon n}(k) = \sup_{t-\epsilon \leq s \leq t} \{n\lambda(s, k)\} \quad \text{and} \quad \mu^+_{\epsilon n}(k) = \inf_{t-\epsilon \leq s \leq t} \{n\mu(s, k)\}, \quad k \geq 0, \quad (6)$$

let the lower bound BD $\{X^-_{\epsilon n}(t): t \geq 0\}$ have rate functions

$$\lambda^-_{\epsilon n}(k) = \inf_{t-\epsilon \leq s \leq t} \{n\lambda(s, k)\} \quad \text{and} \quad \mu^-_{\epsilon n}(k) = \sup_{t-\epsilon \leq s \leq t} \{n\mu(s, k)\}, \quad k \geq 0. \quad (7)$$

Let $X^+_{\epsilon n}(t - \epsilon) = k^* + X(0) + Q(\infty)$, the stochastic upper bound established in Lemma 2; let $X^-_{\epsilon n}(t - \epsilon) = 0$, a trivial lower bound. By Lemma 1,

$$X^-_{\epsilon n}(s) \leq_{st} X_n(s) \leq_{st} X^+_{\epsilon n}(s), \qquad t - \epsilon \leq s \leq t. \qquad (8)$$

(Stronger sample path comparisons could also be made.) By an elementary change in time scale, $\{X^+_{\epsilon n}(t): t \geq 0\}$ is distributed the same as $\{X^+_{\epsilon 1}(nt): t \geq 0\}$, and similarly for $\{X^+_{\epsilon n}(t): t \geq 0\}$. By Lemma 2, $X^+_{\epsilon n}(t)$ is stochastically bounded above. Hence, $X^+_{\epsilon 1}(nt) \Rightarrow X^+_{\epsilon 1}(\infty)$ and $X^-_{\epsilon 1}(nt) \Rightarrow X^-_{\epsilon 1}(\infty)$ as $n \to \infty$. As noted in Remark 1 above, Lemma 2 implies that $\{X_n(t): n \geq 1\}$ is tight. Let $X_\infty(t)$ be the limit of some convergent subsequence of $\{X_n(t)\}$ as $n \to \infty$. Since convergence in distribution preserves stochastic order, we can combine this with (8) to obtain

$$X^-_{\epsilon 1}(\infty) \leq_{st} X_\infty(t) \leq_{st} X^+_{\epsilon 1}(\infty). \qquad (9)$$

Finally, we will show that $X^+_{\epsilon 1}(\infty) \Rightarrow Y_t(\infty)$ and $X^-_{\epsilon 1}(\infty) \Rightarrow Y_t(\infty)$ as $\epsilon \to 0$. This implies that $X_\infty(t) \overset{d}{=} Y_t(\infty)$, where $\overset{d}{=}$ denotes has the same distribution, so that indeed $X_n(t) \Rightarrow Y_t(\infty)$ as $n \to \infty$.

We now show that $X^+_{\epsilon 1}(\infty) \Rightarrow Y_t(\infty)$ and $X^-_{\epsilon 1}(\infty) \Rightarrow Y_t(\infty)$ as $\epsilon \to 0$. This step is relatively easy because all processes are BD's. In this step, we use the strict positivity of $\lambda(t, k)$ and $\mu(t, k)$ to ensure that the BD $\{Y_t(u): u \geq 0\}$ is irreducible. Since $\{Y_t(u): u \geq 0\}$ is stochastically bounded, it has a proper limiting distribution, which is

$$\pi_{tk} \equiv P(Y_t(\infty) = k) = \beta_{tk} \Big/ \sum_{j=0}^{\infty} \beta_{tj}, \qquad (10)$$

where $\beta_{t0} = 1$, $\beta_{tk} = \prod_{j=1}^{k} \rho_{tj}$, $k \geq 1$, and $\rho_{tj} = \lambda(t, j - 1)/\mu(t, j) > 0$. Let

$$\pi^+_{\epsilon k} \equiv P(X^+_{\epsilon 1}(\infty) = k) = \beta^+_{\epsilon k} \Big/ \sum_{j=0}^{\infty} \beta^+_{\epsilon j}, \qquad (11)$$

where $\beta^+_{\epsilon 0} = 1$, $\beta^+_{\epsilon k} = \prod_{j=1}^{k} \rho^+_{\epsilon j}$, $k \geq 1$, and $\rho^+_{\epsilon j} = \lambda^+_{\epsilon 1}(j - 1)/\mu^+_{\epsilon 1}(j)$, and similarly, for $\pi^-_{\epsilon k}$. By (1) and (2),

$$\rho^-_{\epsilon j} \leq \rho_{tj} \leq \rho^+_{\epsilon j} \leq \lambda/\mu < 1, \qquad j \geq k^*. \qquad (12)$$

Since $\rho^+_{\epsilon j} > \rho_{tj} > 0$ for all j and (11) holds, for sufficiently small ϵ, $0 < \rho^+_{\epsilon j} < \infty$ for all j, so that (11) is indeed valid. (Since $X^+_{\epsilon 1}(t)$ is stochastically bounded, the sum in (11) must converge.) For any $\epsilon > 0$, we may have $\lambda^-_{\epsilon 1}(j) = 0$ for some j, but the analog of (11) for $\pi^-_{\epsilon k}$ is also valid because of the initial condition $X^-_{\epsilon n}(0) = 0$. (We have $\pi^-_{\epsilon k} = 0$ for $k \geq j + 1$ if $\lambda^-_{\epsilon 1}(j) = 0$.)

Finally, by the left continuity of $\lambda(t, k)$ and $\mu(t, k)$, we have $\rho^+_{\epsilon j} \to \rho_{tj}$ and $\rho^-_{\epsilon j} \to \rho_{tj}$ as $\epsilon \to 0$ for each j. To establish $\pi^+_{\epsilon k} \to \pi_{tj}$ and $\pi^-_{\epsilon k} \to \pi_{tj}$, we also use the uniformity in the tails provided by (12); i.e.,

$$\sum_{j=0}^{\infty} \beta^+_{\epsilon j} \to \sum_{j=0}^{\infty} \beta_{tj} \quad \text{and} \quad \sum_{j=0}^{\infty} \beta^-_{\epsilon j} \to \sum_{j=0}^{\infty} \beta_{tj}. \quad \square$$

REMARK 2. The final step in the proof involved a continuity result for BD's, as in Karr (1975) and Whitt (1980). For example, we could also have applied Lemma 1 of Whitt (1980). □

PROOF OF COROLLARY 1. The function f can be regarded as a continuous function on the nonnegative integers. Hence, by the convergence in distribution (see Billingsley 1968, p. 12), Theorem 1 implies that $\mathrm{E}f[X_n(t)] \to \mathrm{E}f[Y_t(\infty)]$ for all t which are left-continuity points of $\lambda(t, k)$ and $\mu(t, k)$ for all k. Since $\lambda(t, k)$ and $\mu(t, k)$ have left and right limits, there are only countably many points of discontinuity for each k (see Billingsley 1968, p. 110) and, thus, for all k. Hence, the convergence holds almost everywhere with respect to Lebesgue measure. The stated result then holds by the bounded convergence theorem. It should also be noted that the integrals are well defined. First, $\mathrm{E}f[X_n(t)]$ has limits from the left and right, and so is measurable, because $X_n(t)$ has limits from the left and right in probability. Similarly $\mathrm{E}f[Y_t(\infty)]$ has limits from the left and right because the distribution of $Y_t(\infty)$ is continuous in the rates. In particular, we can use the argument at the end of the proof of Theorem 1 again, noting that

$$X_{\epsilon 1}^{-}(\infty) \leq_{st} Y_{t\pm\delta}(\infty) \leq_{st} X_{\epsilon 1}^{+}(\infty) \tag{13}$$

whenever $0 < \delta < \epsilon$, paralleling (9). □

PROOF OF COROLLARY 2. The assumption together with Lemma 2 implies that $\{X_n(t)^j: n \geq 1\}$ is uniformly integrable for $j \leq k$; see Billingsley (1968, p. 32). Hence, from Theorem 1 we obtain $E[X_n(t)^j] \to E[Y_t(\infty)^j]$ as $n \to \infty$ for almost all t. The rest of the proof is as for Corollary 1. □

5. Discussion

We conclude with a few additional remarks.

5.1. *Finite State Spaces*

We considered TDBD's with infinite state spaces. Corresponding results hold for finite state spaces. Suppose that the state space is the finite set of integers $\{k: k_1 \leq k \leq k_2\}$. Without loss of generality (by changing the labels), we can assume that $k_1 = 0$. Now we do not need Lemma 2 to establish a stochastic upper bound; we can use k_2 instead, e.g., let $X_{\epsilon n}^{+}(t - \epsilon) = k_2$ before (8).

5.2. *Instantaneous Instability*

It is important to note that, because of (1) and (2), Theorem 1 does not apply to an $M_t/M/s$ queue if the instantaneous traffic intensity $\rho_t \equiv \lambda(t)/s\mu$ ever can exceed one. Of course, PSA is well defined whenever the performance measure is bounded, but we think that the use of PSA is much more questionable in this case. For example, consider an $M_t/M/1$ queue with service rate $\mu = 1$ and a periodic Poisson arrival process with cycle of length 1 and arrival rate

$$\lambda(t) = \begin{cases} 99, & 0.00 < t \leq 0.01, \\ 0, & 0.01 < t \leq 1. \end{cases} \tag{14}$$

Then the average arrival rate per cycle is $\bar{\lambda} = \int_0^1 \lambda(t)dt = 0.99$ and the long-run probability of finding the server busy is $p_b = 0.99$ by Little's law. On the other hand, a direct PSA approximation is $p_b^{\infty} = 0.01$. Green and Kolesar (in press) have cleverly circumvented this difficulty for p_b by using $\lambda(t)/\mu$ for the instantaneous probability that the server is busy instead of max $\{1, \lambda(t)/\mu\}$, and similarly for the multi-server case. For the $M_t/M/1$ queue, this gives the exact result and for $M_t/M/s$ queues it evidently gives an upper bound. Perhaps other bounded performance measures can also be successfully modified to produce reasonable PSA-type approximations when there is instantaneous instability, but that needs to be shown. For example, consider the long-run average value of $e^{-\alpha N(t)}$ where α is a positive constant and $N(t)$ is the number of customers in the system at time t. Our example illustrates that there can be serious difficulties with a direct PSA approximation when instantaneous instability is allowed. When periodicity is dropped, even more bizarre behavior is possible, as shown by §2 of Heyman and Whitt (1984).

5.3. *Another Approximation*

GKS observe that PSA performs much better than a *simple stationary approximation* (SSA, the stationary model with the long-run average arrival rate) for periodic $M_t/M/s$ queues when the service rate is relatively high, but the reverse can be true when the service rate is relatively low; see Tables 1 and 2 of Green and Kolesar (in press). We now propose an approximation for $M_t/M/s$ queues (and more general $M_t/G/s$ models) which may be more robust; i.e., that may perform reasonably well for $M_t/M/s$ as the service rate increases or decreases with the instantaneous traffic intensity held fixed. The new approximation is a generalization of both PSA and SSA, which we call the *average stationary approximation* (ASA). To determine the arrival rate to use in the stationary model, ASA stipulates that we average the arrival rates over an interval. As the interval gets long, ASA approaches SSA; as the interval concentrates about the point of interest, ASA approaches PSA. To determine the effective arrival rate at time t for ASA, we suggest averaging the arrival rate $\lambda(s)$ over the interval $[t - m, t]$, where m is chosen to be proportional to (e.g., equal to) the mean service time. (Other finite intervals could be used, but this seems to be a good candidate.) Assuming that the instantaneous traffic intensity is bounded above sufficiently below 1, the arrival process up to some time should be nearly independent of the system state several mean service times later.

To be more specific, suppose that we want to approximate $\mathrm{E}f[X(t_0)]$ where $\{X(t): t \geq 0\}$ represents the number of customers in a $M_t/G/s$ queue with arrival rate $\lambda(t)$ and mean service time μ^{-1}. We suggest letting

$$\bar{\lambda}_{t_0} = \frac{1}{\alpha\mu^{-1}} \int_{t_0-\alpha\mu^{-1}}^{t_0} \lambda(s)\,ds \tag{15}$$

for some positive constant α (e.g., $\alpha = 1$) and approximating $\mathrm{E}f[X(t_0)]$ by $\mathrm{E}f[Z(\infty; \bar{\lambda}_{t_0})]$ where $Z(\infty; \bar{\lambda}_{t_0})$ is the steady-state number in system of an $M/G/s$ queue with constant arrival rate $\bar{\lambda}_{t_0}$ and the same service-time distribution with mean μ^{-1}. If we want the average performance measure over some interval $[a, b]$, then we suggest computing

$$\frac{1}{(b-a)} \int_a^b \mathrm{E}f[Z(\infty; \bar{\lambda}_t)]\,dt. \tag{16}$$

To compute (16), we need to be able to observe $\lambda(t)$ for $a - \alpha\mu^{-1} \leq t \leq b$. If $a = 0$ and the system is initially empty, then let $\lambda(t) = 0$ for $t \leq a$ in (15).

It is not difficult to see that ASA is exact for the $M_t/D/\infty$ model when $\alpha = 1$. (The $M_t/G/\infty$ model could be used as a basis for approximations for the $M_t/G/s$ model too as suggested before by Palm, Newell (1973) and Jagerman (1975).) Moreover, it is not difficult to see, by a minor modification of Theorem 1 and its corollaries, that it is also asymptotically correct for the $M_t/M/s$ model at t_0 as $\mu \rightarrow \infty$ with $\lambda(t)/\mu$ held fixed provided that t_0 is a left-continuity point of $\lambda(t)$. Similarly, the approximation appears to be asymptotically correct as $\mu \rightarrow 0$ with $\lambda(t)/\mu$ held fixed, which corresponds to very short cycles. Theorem 1 of Whitt (1984) implies that the arrival process then is asymptotically stationary Poisson with the long-run average rate, which is what is determined by (15) as $\mu \rightarrow 0$; also see Proposition A of Rolski (1989, p. 127).

5.4. *Engineering Practice*

GKS note that in most real service systems, the customer demand process is nonstationary, whereas the vast majority of queueing theory papers concern stationary models. Nevertheless, there is a long history of addressing this nonstationarity in a fairly sensible way; e.g., in telecommunications over the last 70 years. Indeed, something like what is proposed in §5.3 above seems to be common practice. Certainly, in telecommunications one would never expect to see the simple stationary approximation (SSA) of GKS with a daily average arrival rate when the mean service time (holding time) is a few minutes and the arrival rate at night is negligible. The traditional way to address this problem is to focus on appropriate busy hours, recognizing that the system tends to approach steady state in minutes when the rates stabilize; e.g., see §§2.3, 9, 12.2 of Bear (1980). Nevertheless, we still need to better understand time-dependent queues.[1]

[1] I thank Linda Green and Bhaskar Sengupta for helpful comments.

References

BEAR, D., *Principles of Telecommunication—Traffic Engineering*, Peter Peregrinus, Ltd., London, 1980.

BILLINGSLEY, P., *Convergence of Probability Measures*, Wiley, New York, 1968.

GREEN, L. AND P. KOLESAR, "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals," *Management Sci.*, 37, 1 (1991), 84–97.

———, ———, AND A. SVORONOS, "Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems," *Oper. Res.*, in press.

HEYMAN, D. P. AND W. WHITT, "The Asymptotic Behavior of Queues with Time-Varying Arrival Rates," *J. Appl. Probab.*, 21 (1984), 143–156.

JAGERMAN, D. L., "Nonstationary Blocking in Telephone Traffic." *Bell System Tech. J.*, 54 (1975), 625–661.

KARR, A. F., "Weak Convergence of a Sequence of Markov Chains," *Z. Wahrsch. und Verw. Gebiete*, 33 (1975), 41–48.

KEILSON, J., *Markov Chain Models—Rarity and Exponentiality*, Springer-Verlag, New York, 1979.

MASSEY, W. A., "Asymptotic Analysis of the Time Dependent $M/M/1$ Queue," *Math. Oper. Res.*, 10 (1985), 305–327.

McCLISH, D., "Queues and Stores with Non-homogeneous Input," Ph.D. dissertation, University of North Carolina, Chapel Hill, 1979.

NEWELL, G. F., "Queues with Time-Dependent Arrival Rates, I, II and III," *J. Appl. Probab.*, 5 (1968), 436–451, 579–606.

———, *Approximate Stochastic Behavior of n-Server Service Systems with Large n.* Lecture Notes in Econ. and Math. Systems 87, Springer-Verlag, Berlin, 1973.

———, *Applications of Queueing Theory*, 2d ed., Chapman and Hall, London, 1982.

PALM, C., *Intensity Variations in Telephone Traffic,* North-Holland, Amsterdam, 1988 (transl. of 1943 article in *Ericsson Technics* 44, 1–189).

ROLSKI, T., "Upper Bounds for Single Server Queues with Doubly Stochastic Poisson Arrivals," *Math. Oper. Res.*, 11 (1986), 442–450.

———, "Queues with Nonstationary Inputs," *Queueing Systems*, 5 (1989), 113–130.

WHITT, W., "Comparing Counting Processes and Queues," *Adv. Appl. Probab.*, 13 (1981), 207–220.

———, "Continuity of Generalized Semi-Markov Processes," *Math. Oper. Res.*, 5 (1980), 494–501.

———, "Departures from a Queue with Many Busy Servers," *Math. Oper. Res.*, 9 (1984), 534–544.