MANAGEMENT SCIENCE
Vol. 37, No. 1, January 1991
Printed in U.S.A.

THE POINTWISE STATIONARY APPROXIMATION FOR QUEUES WITH NONSTATIONARY ARRIVALS*

LINDA GREEN AND PETER KOLESAR
Graduate School of Business, Columbia University, New York, New York 10027

We empirically explore the accuracy of an easily computed approximation for long run, average performance measures such as expected delay and probability of delay in multiserver queueing systems with exponential service times and periodic (sinusoidal) Poisson arrival processes. The pointwise stationary approximation is computed by integrating over time (that is taking the expectation of) the formula for the stationary performance measure with the arrival rate that applies at each point in time. This approximation, which has been empirically confirmed as a tight upper bound of the true value, is shown to be very accurate for a range of parameter values corresponding to a reasonably broad spectrum of real systems.
(QUEUES; NONSTATIONARITY; APPROXIMATION)

Introduction

Most actual queueing systems do not have constant arrival rates; indeed, it is quite common to find systems in which there are predictable and often cyclic patterns by which demand varies with time. Computer systems, telecommunications networks, banks, airports, toll booths, and emergency systems are just a few examples of facilities with such time-varying demand processes. Yet, the bulk of literature in queueing theory deals with stationary arrival processes. Although practitioners and engineers have fashioned ad hoc methods to deal with nonstationarity, there is only a modest amount of theory dealing with time-varying arrival processes, and there are no known general formulas for computing performance measures such as expected delay, for even the simplest of single server nonstationary Markovian systems.

Of course, any nonstationary system with known parameter values can be studied using simulation or, if Markovian, by numerical integration of the differential equations which describe its dynamics (see e.g., Koopman 1972, Clarke 1956, and Luchak 1956). However, these methods are usually very costly in computational effort. Several methods have been proposed to reduce the computational burden by approximating the solution to the infinite set of equations of a Markovian system using closure techniques that require a relatively small number of equations. For example, see Rider (1976), Rothkopf and Oren (1979) and Clark (1981). More recently, similar approximation methods have been used to reduce the number of equations necessary for large finite capacity systems (see Taafe and Ong 1987, 1988, 1989). Other solution techniques include diffusion approximations, first presented in the pioneering work of Newell (1968 and 1971), and more recently used by Duda (1986) for studying computer-communications systems. Gaver (1966) and Kotiah (1978) suggested the use of transform approximation methods. Techniques based on approximating the actual time-varying arrival rate by a surrogate arrival rate generated by a stationary Markov process have been suggested by Neuts (1978), Rolski (1987), and Gelenbe and Rosenberg (1990), all for single server systems. Massey (1985) analyzed a single server Markovian queue by uniformly accelerating the time dependent arrival and service rates for a fixed time interval and examining the asymptotic behavior of the queue length process.

Most of the above work has focused on obtaining estimates for the time-varying behavior of the system. However, in many applications, such as those involving design decisions

* Accepted by Donald G. Morrison; received August 1989. This paper has been with the authors 3 months for 2 revisions.

0025-1909/91/3701/0084$01.25

about how much fixed capacity is needed, composite performance measures such as long run average expected waiting time in the system are appropriate and sufficient, at least for an initial analysis. (At later stages more detailed time dependent behavior may be required.) Until the present, there appeared to be agreement among some theoreticians and practitioners that for such broad uses, the nonstationarity could be ignored and a standard stationary analysis could be used if the time-varying fluctuations were "mild" (see e.g., Rothkopf and Oren 1979). Yet, in a recent paper, Green, Kolesar and Svoronos (1991) showed that a stationary model can seriously underestimate delays even when the arrival rate is only modestly nonstationary (for example, when the amplitude of the arrival process is only 10% of its average). This finding raises serious concerns about the use of a simple stationary model for almost any real system. Of course, variants on the simplest stationary approximation such as peak hour analysis or segmenting the time period and using a series of stationary models based on the average arrival rates for each segment can also be employed (see, for example, Kolesar et al. 1975). Yet, there has not been a general exploration of the accuracy of these methods. (Our current research builds on the results reported here and will focus on these issues.)

In this paper, we discuss an easy-to-compute approximation for determining long run average performance measures for multiserver Markovian queues with periodic arrival rates. This approximation is obtained by computing the expectation of the performance measure over the period using the stationary formula with the arrival rate that corresponds to each point in time. For this reason, we call it the pointwise stationary approximation (PSA). The PSA is, in effect, a limiting version of the segmentation approach discussed in the previous paragraph. It is a rather intuitive concept that most likely has been employed earlier. Rothkopf and Oren (1979) appear to suggest something of the kind. Yet in searching the literature after our rediscovery, we found no references before the work of Grassman (1983), who first conjectured that the PSA is an upper bound for the expected number of customers in queue as a consequence of his result that the mean queue size is convex with respect to the arrival rate. This was confirmed by Rolski (1986) for the case of single server systems.

In Green et al. (1991), we showed by numerical comparisons that when the PSA exists, it is an upper bound of the actual value. For probability of delay or probability of all servers busy, it always exists. For expected delay or expected number in system, etc., it results in a finite value only for systems in which the maximum traffic intensity is strictly less than one. Green et al. also empirically demonstrated that for Markovian multiserver systems with sinusoidal arrival rate, the simple stationary approximation, i.e., computed using the time average arrival rate as if it were constant over the entire period, results in a lower bound for the usual overall measures of performance—expected delay, expected number in system, probability of delay, etc. Therefore, since the PSA produces an upper bound, we have a pair of simple computations that bound actual results and the errors produced by any approximation technique. Moreover, we demonstrate that in a broad range of systems, the PSA produces reasonably accurate estimates of the actual performance measures. There are, however, parameter values for which the PSA is not a good approximation.

Our findings are based on extensive numerical investigation of multi-server exponential queueing systems with sinusoidal Poisson input. This class of models was chosen because the exact differential equations for the steady-state probabilities can be readily solved numerically and because this easily parameterized arrival process captures the essence of many actual periodic arrival processes. However, as discussed in Green et al., there is no reason to believe that the PSA would not result in an upper bound for other periodic Poisson arrival processes or more general service distributions. Similarly, as will be discussed, there is good reason to believe that the findings reported in this paper are generally applicable to other nonstationary systems.

In §1, we briefly describe our research methodology. In §2, we discuss some characteristics of the PSA and present examples of its usefulness for decision-making as compared with the strictly stationary approximation. The accuracy of the PSA is systematically explored in §3 and we present a summary of our conclusions in §4.

1. Definitions and Methodology

We consider an $M(t)/M/s$ system where $\lambda(t)$ is the arrival rate at time t (which we assume varies according to a sinusoid), μ is the service rate and s is the number of servers. We will assume that

$$\bar{\lambda} = \frac{1}{T}\int_0^T \lambda(t)dt < s\mu, \tag{1}$$

where T is the period of the sinusoid, also called its cycle length. Thus, the system will develop a periodic steady-state behavior (see Heyman and Whitt 1984).

Let $p_n(t)$ be the periodic steady-state probability of n customers in the system at time t. Our numerical results were obtained by solving the following standard set of differential equations:

$$p_0'(t) = -\lambda(t)p_0(t) + \mu p_1(t),$$

$$p_n'(t) = \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - (\lambda(t) + n\mu)p_n(t), \qquad 1 \leqq n < s,$$

$$p_n'(t) = \lambda(t)p_{n-1}(t) + s\mu p_{n+1}(t) - (\lambda(t) + s\mu)p_n(t), \qquad n \geqq s. \tag{2}$$

The details of the numerical integration methodology are described in Green et al. (1991).

For convenience, assume that the period of the arrival rate process, T, is 24 hours. Let L_q, W_q, p_d and p_b be the daily average queue length, the daily expected delay, the daily probability of delay and the daily probability of all servers being busy, respectively. Specifically,

$$L_q = \frac{1}{T}\int_0^T \left(\sum_{n=s}^{\infty}(n-s)p_n(t)\right)dt,$$

$$W_q = L_q/\bar{\lambda},$$

$$p_d = \frac{1}{\bar{\lambda}T}\int_0^T \lambda(t)\left(1 - \sum_{n=0}^{s-1} p_n(t)\right)dt, \qquad \text{and}$$

$$p_b = \frac{1}{T}\int_0^T \left(1 - \sum_{n=0}^{s-1} p_n(t)\right)dt.$$

We denote L_q^{∞}, W_q^{∞}, p_d^{∞} and p_b^{∞} as the *pointwise stationary approximations* (PSAs) for L_q, W_q, p_d and p_b and define them as follows:

$$L_q^{\infty} = \frac{1}{T}\int_0^T L_q(\lambda(t))dt, \tag{3}$$

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

$$p_d^{\infty} = \frac{1}{\bar{\lambda}T}\int_0^T \lambda(t)p_d(\lambda(t))dt, \qquad \text{and} \tag{5}$$

$$p_b^{\infty} = \frac{1}{T}\int_0^T p_b(\lambda(t))dt, \tag{6}$$

where $W_q(\lambda(t))$, $L_q(\lambda(t))$, $p_d(\lambda(t))$ and $p_b(\lambda(t))$ are given by the formulae for expected delay expected queue length, probability of delay and probability of all servers busy in a stationary $M/M/s$ with an arrival rate of $\lambda(t)$ and μ and s as given. It is important to note that even for $\lambda(t) \geqq s\mu$, $p_d(\lambda(t)) = p_b(\lambda(t))$ is computed from the stationary $M/M/s$ delay formula (see Kleinrock 1975, equation 3.40), although this formula will result in a value greater than one in this case. In Green et al. (1991), we determined empirically that when the maximum traffic intensity is less than one, that is, when

$$\rho_{\max} \equiv \sup_t \frac{\lambda(t)}{s\mu} < 1, \tag{7}$$

then

$$W_q \leqq W_q^{\infty} \qquad \text{and} \qquad L_q \leqq L_q^{\infty}.$$

This result was substantiated for over 250 models we examined with widely varying parameter values (See Green and Kolesar (1990) for the data that support our findings.) These empirical findings extended an analytical result by Rolski (1986) which established that (4) holds for single server systems with doubly stochastic Poisson arrivals. (Rolski's result also holds for general service time distributions in which case $W_q(\lambda(t))$ is the expected delay for the corresponding $M/G/1$ system.) We also found that p_d^{∞} and p_b^{∞} result in a finite upper bound even when the maximum traffic intensity exceeds one. (Occasionally the computed value exceeds one which, of course, should be interpreted as producing an upper bound of one.)

The goal of this paper is to establish under what conditions the PSA gives reasonable estimates of the actual performance measure. In particular, the objective of our work was to determine how the accuracy of the PSA varies as a function of the parameters of our generic model which assumes that there are s identical exponential servers, each serving a rate μ, and a time dependent Poisson arrival process with rate given by

$$\lambda(t) = \bar{\lambda} + A \cos(2\pi t/24), \tag{8}$$

where $\bar{\lambda}$ is the daily average arrival rule, and A (>0) is the amplitude. Without loss of generality, the period T, is assumed to be 24 hours. In the specific model instances we investigated, we selected parameter values such that

$$\bar{\rho} = \bar{\lambda}/s\mu < 1 \tag{9}$$

to assure the existence of a limiting distribution, and with the relative amplitude

$$\text{Relative Amplitude (RA)} = A/\bar{\lambda} < 1, \tag{10}$$

so that $\lambda(t) \geqq 0$ for all t.

Aside from these constraints, our choices of experimental models were governed by three major considerations—correspondence to actual service systems, a desire to be as general as possible, and computational feasibility.

Based on our previous work (Green et al. 1991), we formulated some conjectures on how the accuracy of the PSA would vary with parameters such as the amplitude and frequency of the input. Out experimental strategy was to confirm each conjecture first at a "central case," and if confirmed there, then to determine its validity in a region surrounding the central case by perturbing each of the key parameters.

The resulting models span a fairly broad spectrum of parameter values: the number of servers ranges from 1 to 12, the service rate varies from .2 to 20, average traffic intensities range between .25 and .75 and relative amplitudes between .1 and 1.0. Yet, in some specific tests of a hypothesis, we were further constrained in the range of one or more of the parameters due to theoretical constraints such as equations (9) and (10), or to

computational limitations arising from either the time necessary to achieve steady-state behavior (when the frequency of events is very low, this can be extremely long) or from the computational effort required to accurately estimate the solution to the differential equations at peak congestion epochs for systems with high traffic intensities, high service rates and many servers. However, in all we have solved over 300 models in this corroboration effort. The data are available in Green and Kolesar (1990).

2. The Pointwise Stationary Approximation

In Green et al. (1991), we observed that not only is L^{∞} an upper bound whenever the maximum traffic intensity was less than one, but that it is asymptotically approached as the frequency of events per period increased, i.e., as $\lambda(t)$ and μ increased simultaneously so that both $\bar{\rho} = \bar{\lambda}/s\mu$ and the relative amplitude (RA = $A/\bar{\lambda}$) remained fixed. This is illustrated in Figure 1 where we use a semi-log plot of expected queue length against frequency as measured by $\bar{\lambda}$.

The intuition that led us to this observation is that as $\bar{\lambda}$ and μ grow larger together, the number of both arrivals and departures during any given small interval Δt becomes so large that the system approaches steady-state behavior during Δt and for any t' as $\Delta t \to 0$, $\lambda(t)$ for $t \in (t', t' + \Delta t)$ will be almost constant. Thus, we reason that as event frequency goes to infinity, the overall expected queue length approaches the expectation over time of the expected queue length in a system which at every time t behaves like a stationary $M/M/s$ with arrival rate $\lambda(t)$.

Another important observation, that is illustrated in Figure 1 (which has a logarithmic scale), is that the rate of increase of the expected queue length is quite sharp so that even at moderate arrival rates, actual behavior is close to that predicted by the upper bound (PSA). This led us to believe that the PSA might be a good approximation for many real systems.

Another finding in Green et al. (1991) was that the simple stationary approximation obtained by using the overall mean arrival rate in the stationary $M/M/s$ model, results

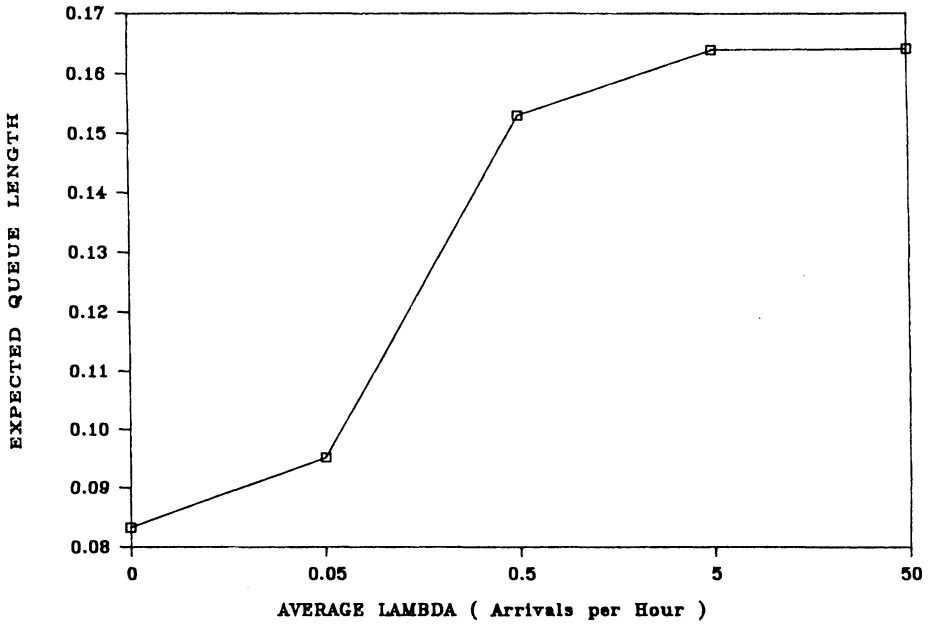

FIGURE 1. L_q v Frequency of Arrivals.
($S = 1$, RA = 1, $\bar{\rho} = 0.25$)

in a lower bound for any performance measure. This confirmed a general conjecture first made by Ross (1978) and proved by Rolski (1981) specifically for the $M(t)/G/1$ system. More importantly, we found that the completely stationary approximation *substantially* underestimates the actual delay in most cases. Therefore, the "obvious" simple approximation for estimating overall performance measures for systems with nonstationary arrivals fails, *even when the degree of nonstationarity is quite modest.*

Table 1 gives two examples for which the completely stationary approximation seriously underestimates the delays, but the PSA (where it exists) is quite good. These examples were constructed to imitate arrival and service time parameters of police patrol operations in New York City. Calls for police service in New York's 911 emergency telephone system arrive in a periodic daily pattern and behave very much like a time-varying Poisson process. This arrival process is strongly unipeaked and it is not uncommon for the maximum arrival rate to be twice the daily average (see Green and Kolesar 1989). Service times are very close to exponential and, in most precincts, average about thirty minutes. Table 1A illustrates predicted probabilities of delay and expected delays for all "reasonable" staffing levels in a precinct with a "light" demand level averaging one arrival per hour, while Table 1B corresponds to a precinct with a "heavy" demand of six calls per hour on average. The stationary delays were obtained from the stationary $M/M/s$ model, the "actual" delays resulted from numerical integration of the differential equations given by (2) with $\lambda(t)$ given by (8), and the PSA predictions were calculated using equations (4) and (5).

One unmistakable conclusion to be drawn from this table is that the stationary approximation fails miserably in both precincts and at all staffing levels. Decisions on the capacity required to meet a given performance standard would, if based on such estimates, be quite erroneous. In contrast, the PSA estimates are, in almost all instances, very close to the actual delays and would result in reliable capacity decisions. The obvious exceptions are cases where the PSA for the expected delay does not exist because the maximum traffic intensity exceeds one. Interestingly, the PSA for the probability of delay in these

TABLE 1

$\mu = 2$, RA = 1.0

A. $\bar{\lambda} = 1$

	Probability of Delay			Expected Delay		
s	Stationary	Actual	PSA	Stationary	Actual	PSA
1	.5000	.6748	.7500	.5000	1.131	—
2	.1000	.2137	.2180	.0333	.0936	.0977
3	.0152	.0519	.0527	.0030	.0123	.0125
4	.0018	.0155	.0107	.0003	.0017	.0017

B. $\bar{\lambda} = 6$

	Probability of Delay			Expected Delay		
s	Stationary	Actual	PSA	Stationary	Actual	PSA
6	.0991	.4815	.5446	.0165	.2539	—
7	.0376	.2951	.3139	.0047	.0894	.1166
8	.0129	.1650	.1717	.0013	.0329	.0363
9	.0040	.0860	.0888	.0003	.0125	.0132
10	.0012	.0420	.0434	.0001	.0048	.0050
11	.0003	.0193	.0200	.0000	.0018	.0019
12	.0001	.0084	.0087	.0000	.0007	.0007

instances exists and is still reasonably accurate. Another useful observation is that the accuracy of the PSA appears to improve as the number of servers increases and thus the PSA seems to be particularly useful for identifying staffing levels that would keep delays "reasonable." We discuss this phenomenon and its application further in the next section.

Are there situations in which the stationary approximation is better than the PSA? The examples given in Table 2 show that while there are situations where the stationary approximation is better, in such cases it is not necessarily good enough to be useful. In the example in Table 2A, the stationary approximation is quite good, at least at low staffing levels, and is much better than the PSA. But the frequency of events is very low—average service times are five hours and arrivals average fewer than one every six hours. The nonstationarity appears to be significant since the relative amplitude is one, yet with such infrequent arrivals, it would be difficult to diagnose. (An observer would have to collect data over a very long time to verify the nonstationary when events are this infrequent.) As we showed in Green et al. (1991) as the event frequency decreases the stationary approximation improves, and in the limit, the system converges to the corresponding stationary one. Since the PSA is the limiting behavior in the opposite direction, it is not surprising to see that the PSA is quite bad for the example in Table 2A. However, it is important to note that the stationary approximation also becomes very inaccurate as the number of servers increases.

Table 2B shows a system for which *neither* the stationary approximation nor the PSA are good. Though service times are again very long, and the relative amplitude is not very high, the average arrival rate of one per hour, though still low, is not low enough for the system to behave in a manner that is closely approximated by its stationary counterpart.

Examples such as those in Table 1 led us to believe that the PSA would provide a potentially valuable and practical basis for decision making for many actual systems with time-varying arrivals. In the next section we systematically explore how the accuracy of the PSA varies as a function of the system parameters. We focus on the performance measures of daily expected delay, expected queue length and the probability of delay since these are the most common measures for managing systems. The results for

TABLE 2

$\mu = .2$

A. $\bar{\lambda} = .15$ RA = 1.0

	Proability of Delay			Expected Delay		
s	Stationary	Actual	PSA	Stationary	Actual	PSA
1	.7500	.7731	1.125	15.000	15.697	—
2	.2045	.2578	.4267	.818	1.099	3.294
3	.0441	.0743	.1405	.098	.1790	.4262

B. $\bar{\lambda} = 1$ RA = 1.0

	Probability of Delay			Expected Delay		
s	Stationary	Actual	PSA	Stationary	Actual	PSA
8	.1673	.2188	.3871	.2788	.4109	2.2927
9	.0805	.1214	.2298	.1006	.1740	.5962
10	.0361	.0639	.1305	.0361	.0738	.2226
11	.0151	.0318	.0708	.0126	.0306	.0903
12	.0059	.0149	.0365	.0042	.0122	.0373
13	.0021	.0066	.0180	.0013	.0047	.0153

probability that all servers are busy are generally similar to those described for probability of delay though the PSA appears to be more accurate for probability busy. In fact, for the special case of single server systems, the PSA given by (6) is exact. However, so is the stationary approximation. This is easily proven by noting that from Little's formula

$$P(\text{server busy}) = \bar{\lambda} \times E(\text{service time}) = \bar{\lambda}/\mu$$

for both the stationary and nonstationary system. Equation (7) results in

$$p_b^{\infty} = \frac{1}{T}\int_0^T \frac{\lambda(t)}{\mu}\,dt = \frac{\bar{\lambda}}{\mu}.$$

3. Accuracy of the Pointwise Stationary Approximation

In this section, we systematically explore how the accuracy of the PSA changes with the parameters of the system. Our goal is to determine general conditions under which the PSA produces reasonable estimates of expected delay, expected queue length and probability of delay.

For this purpose we define a relative error measure as follows:

$$\text{Relative error} = \frac{\text{Actual value} - \text{PSA}}{\text{Actual value}}.$$

Note that by Little's formula, the relative error of expected delay will be identical to the relative error of expected queue length, and thus we will use these measures interchangeably.

The hypotheses we tested resulted from numerical results that were generated during our previous work on the effects of nonstationarity, and from the nature of the PSA itself. Two premises guided our work. The first is based on an interpretation of the PSA as the expected performance measure for a system that at every instant t behaves like a stationary $M/M/s$ with arrival rate $\lambda(t)$ in steady-state. Thus, we would expect that the PSA improves as system conditions better approximate this situation. Secondly, for expected delay, we know that the integral expression for the PSA diverges whenever the traffic intensity is greater than or equal to one. Therefore we expected that the PSA for expected delay, at least, would deteriorate as the maximum traffic intensity increases. The specific tests for these conjectures are detailed below.

A. *The Effect of Event Frequency*

Since as shown in Figure 1, the expected queue length asymptotically approaches the PSA given by (5), as the event frequency increases, it follows that the relative error will decrease. To formally test this, we examined cases in which we fix the cycle length (at 24 hours) and the number of servers while $\lambda(t)$ and μ are simultaneously increased so that $\bar{\lambda}/\mu$ (and hence the traffic intensity) and the relative amplitude (RA in the figures) are both constant. The results for our central case of average traffic intensity (rho) equal to .43, relative amplitude of $\frac{1}{3}$ and 7 servers, is shown in Figure 2. The general result was confirmed for all of our surrounding cases which included systems with rho ranging from .25 to .75, relative amplitudes from .1 to 1.0 and from 1 to 12 servers. Note that the horizontal scale is measured in units of $\bar{\lambda}$ and that the $\bar{\lambda} = 0$ point corresponds to the stationary system (see Green et al. 1991).

Also consistent with our earlier findings, our tests for probability of delay confirmed that the relative error decreases as event frequency increases.

Note that letting $\lambda(t)$ and μ increase proportionally is mathematically equivalent to increasing the cycle length T. For example, the system defined by $s = 6$, $\bar{\lambda} = 100$,

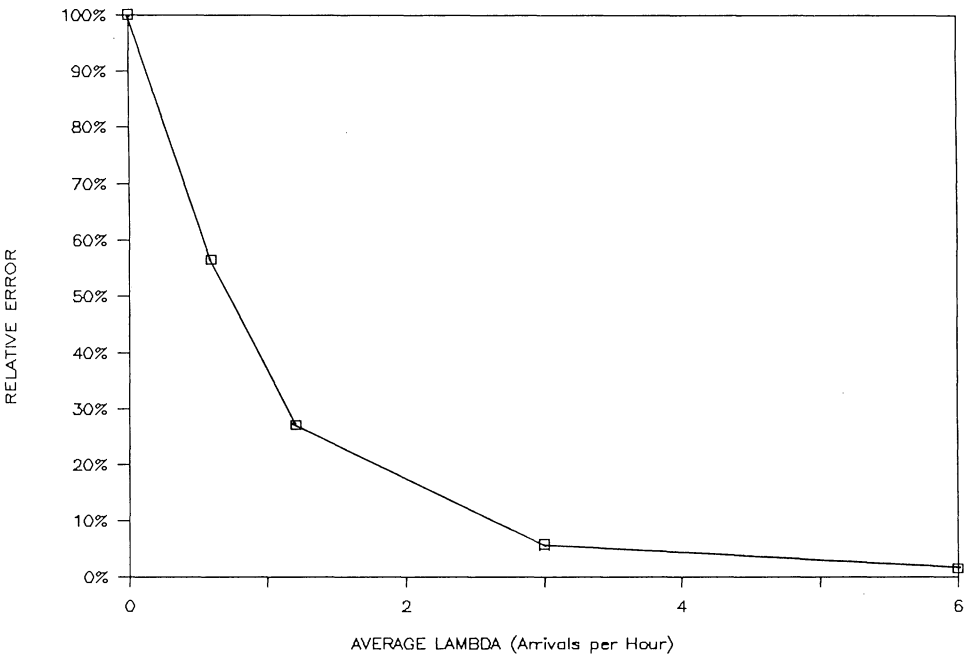

FIGURE 2. Relative Error in L_q v Event Frequency.
($\bar{\rho}$ = 0.43, RA = 0.333, S = 7)

$50 \leqq \lambda(t) \leqq 150$, $\mu = 20$ and $T = 24$ is identical in performance to one defined by $s = 6$, $\bar{\lambda} = 10$, $5 \leqq \lambda(t) \leqq 15$, $\mu = 2$ and $T = 240$. Therefore, the results discussed in this section can also be stated as establishing that the accuracy of the PSA improves as the cycle length increases.

B. *The Effect of Service Rate*

Since our tests for event frequency involved increasing $\lambda(t)$ and μ simultaneously, it is logical to determine whether increasing one of these while holding the other constant will have the same effect. We hypothesized that increasing the service rate μ, would result in the PSA being more accurate since faster clearing of customers from the system should cause consecutive time intervals to be more independent of each other (which is consistent with approaching the limiting condition of the PSA). Our experiments confirmed that the relative error of expected delay decreases as μ increases. This is illustrated by Figure 3. In these tests, $\lambda(t)$ remained constant and the number of servers was decreased as μ was increased so that $s\mu$ remained constant and hence the traffic intensity remained fixed. Not surprisingly, the same phenomenon was found to be true for probability of delay, as well.

What is the effect of holding the service rate constant and increasing the arrival rate, while proportionally increasing the number of servers, so that all other parameters remain constant? We initially surmised that the PSA would improve under these circumstances since the increased arrival frequency would move the system closer to being in steady-state at every moment (implicit for the PSA). However, our tests produced no consistent effect. While increasing the arrival rate served to improve the PSA's accuracy when the service rate was large enough (e.g. $\mu = 2$) to result in small relative errors (i.e. less than 10%), the opposite appeared to be true for every small service rates (e.g. $\mu = .2$) which resulted in large relative errors (e.g. over 50%). However, we did not pursue this line of testing enough to establish a definitive pattern.

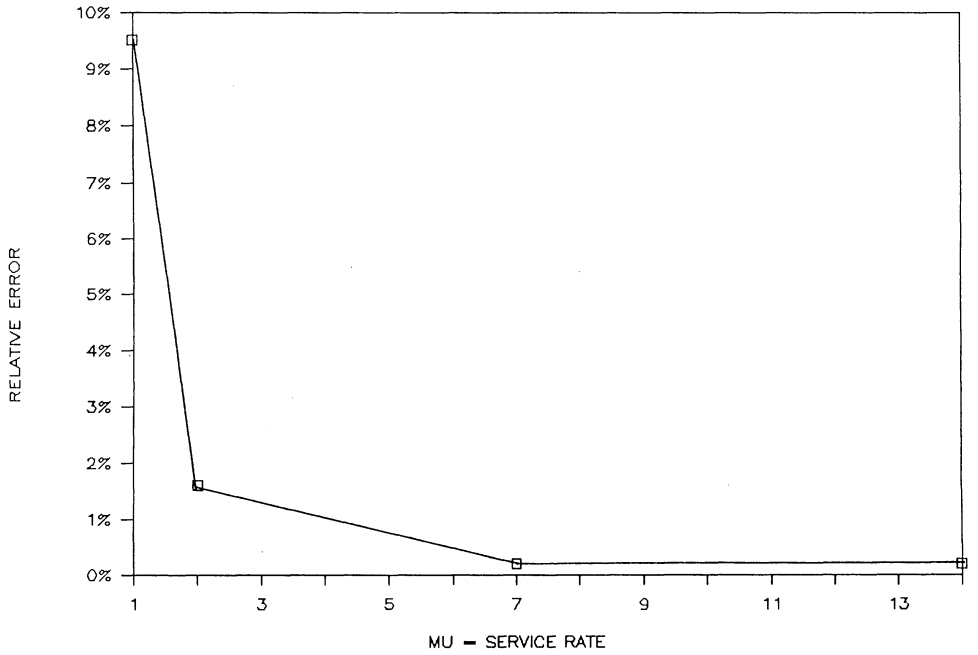

FIGURE 3. Relative Error in L_q v Service Rate.
($\bar{\rho}$ = 0.43, RA = 0.333, Lambda = 6)

C. *The Effect of the Maximum Traffic Intensity*

As mentioned previously, the PSA for expected delay does not produce a finite value when the maximum traffic intensity is greater than or equal to one. Furthermore, our previous work indicated that for maximum traffic intensities close to one, the PSA

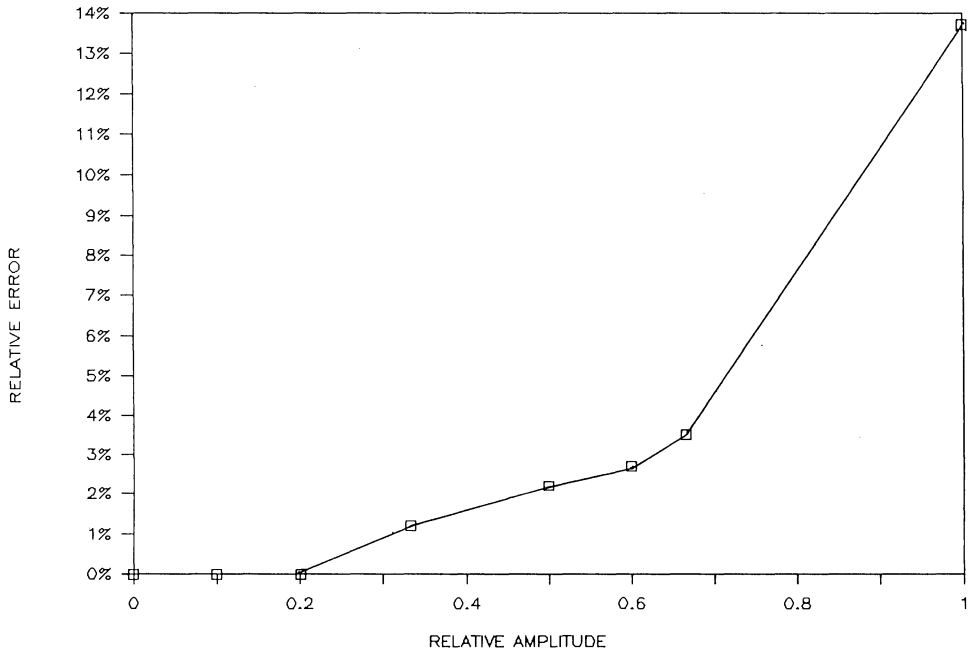

FIGURE 4. Relative Error in L_q v Maximum Rho.
(Lambda = 6.0, S = 6, Mu = 6.5)

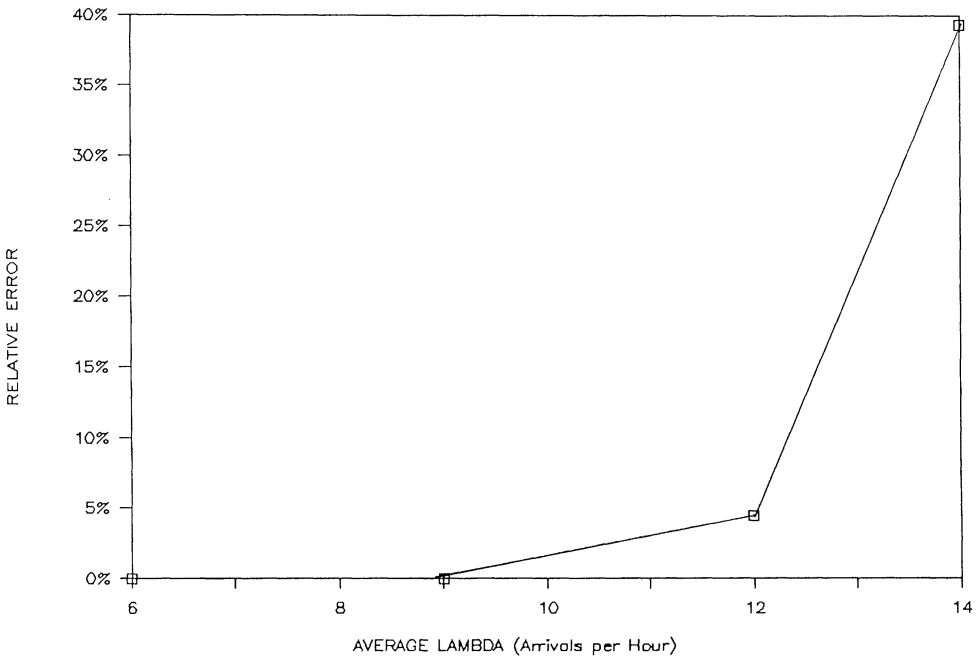

FIGURE 5. Relative Error in L_q v Maximum Rho.
(RA = 0.333, S = 5, Mu = 4.0)

significantly overestimated the actual expected delay. This is not surprising, since as can be seen from equation (4), the PSA for expected delay is most heavily weighted by the stationary expected delay for the highest values of $\lambda(t)$. Therefore, it is reasonable to conjecture that the PSA for expected delay becomes less accurate as the maximum traffic intensity increases.

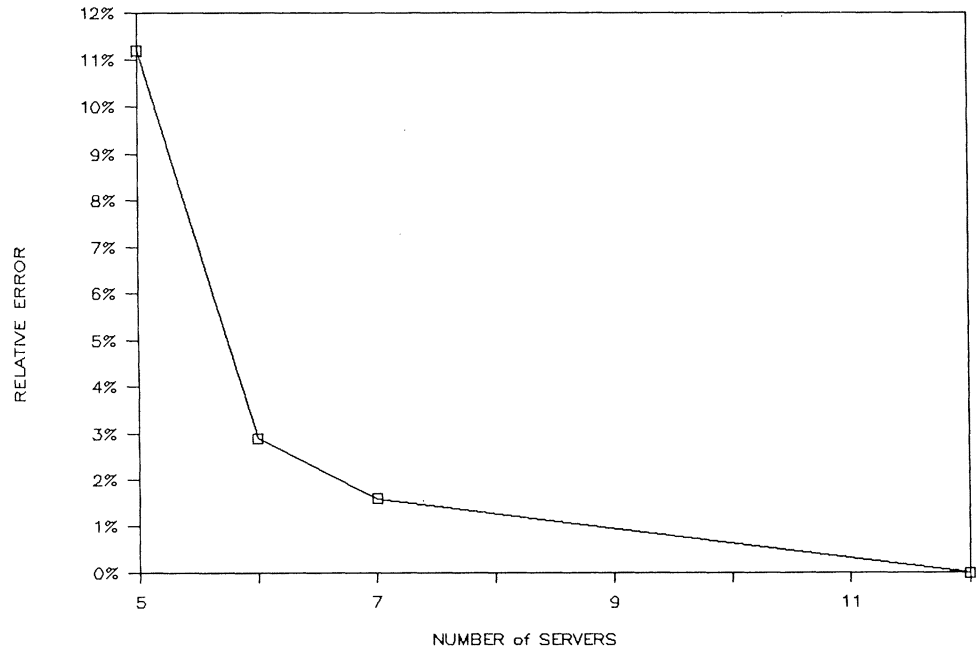

FIGURE 6. Relative Error in L_q v Number of Servers.
(RA = 0.333, Mu = 2.0, Lambda = 6.0)

We performed three sets of tests for this hypothesis. In the first, we raised the maximum traffic intensity by increasing the relative amplitude and keeping all other parameters fixed. Figure 4 illustrates our general finding that, as expected, the relative error increases as the relative amplitude increases. Figure 5 shows our central case for the second set of tests in which the relative amplitude was held constant, but the average arrival rate, $\bar{\lambda}$ was increased. Thus, in these cases, both the average and maximum traffic intensities increase. Again, the result is that the PSA becomes less accurate. Finally, we examined the situation in which the number of servers increases while all other parameters are held constant. Observations of results such as those in Table 1 led us to conjecture that the PSA was increasingly accurate at higher staffing levels. Since increasing the number of servers also decreases the maximum traffic intensity, this conjecture is consistent with our general hypothesis. Our tests confirmed this and our central case is shown in Figure 6.

All of these findings regarding the effect of the maximum traffic intensity pertain only to expected delay and expected queue length. Tests of these hypotheses for probability of delay did not yield consistent results.

4. General Conclusions

Green, Kolesar and Svoronos (1991) established that the simple stationary approximation is quite inaccurate for a large class of nonstationary Poisson queueing systems, even for many systems in which the nonstationarity is very modest, i.e. the relative amplitude is only 10%. The only instances in which ignoring the nonstationarity seems reasonably "safe" are those in which the frequency of events is very low (e.g. $\lambda \ll 1$ and $\mu \ll 1$) and the relative amplitude is not too high (e.g. RA $\leqq .25$). Since most real service systems have significant nonstationarity in their arrival processes, this finding indicates the criticality of some analysis beyond one based only upon a simple stationary approximation.

The PSA is easy to compute and provides tight upper bounds on key performance measures. Furthermore, the results reported here indicate that, for many nonstationary Poisson service systems, the PSA provides good estimates for several key performance measures. We summarize these findings here:

(1) For expected delay, expected queue length, probability of delay and probability of all servers busy, the PSA improves as the frequency of events (or equivalently, the cycle length) increases.

(2) For these same performance measures, the PSA becomes more accurate as the service rate increases.

(3) For expected delay and expected queue length, the PSA worsens as the maximum traffic intensity increases. When the maximum traffic intensity is greater than or equal to one, the PSA does not produce finite values.

So for probability of delay (or probability that all servers are busy), the accuracy of the PSA is primarily affected by the service rate. From our many observations, it is generally quite good for systems with service rates exceeding 2 per hour (assuming a 24 hour cycle).

The accuracy of the PSA for expected delay or expected queue length is primarily a function of both the service rate and the maximum traffic intensity. Our observations indicate that the PSA will produce reasonable estimates for systems in which the service rate is 2 or higher and the maximum traffic intensity is less than .83. For systems in which the service rate is considerably higher, e.g. $\mu = 20$, the estimates will be good at even higher maximum intensities. Thus, for service systems in which there are many

transactions per hour such as telecommunications systems, computer systems, banks, toll booths, and supermarkets, the PSA will generally produce very reliable estimates, particularly for identifying system capacity levels to keep long run average delays reasonable.

It is important to note that although our findings are based on experimental models with sinusoidal arrival rates and exponential service times, there is nothing to indicate that they are not more generally applicable. Rolski's (1986) proof that (4) is an upper bound for the single server system with general service distribution and doubly stochastic Poisson arrival processes leads us to believe that the PSA for each of the performance measures is an upper bound in the multiple server case under the same broad assumptions. In a paper based on this one, Whitt (1991) has verified our conjecture that the PSA is asymptotically correct as the rates increase and has also extended it to include general time-dependent birth- and death-processes. Furthermore, since many of our findings on the accuracy of the PSA follow from an interpretation of it as the expectation over time of a pointwise stationary system, it is reasonable to believe they hold more generally.

References

CLARK, G. M., "Use of Polya Distributions in Approximate Solutions to Nonstationary $M/M/s$ Queues," *Comm. of the ACM*, 24 (1981), 206–217.

CLARKE, A. B., "A Waiting Line Process of Markov Type," *Ann. Math. Statist.*, 27 (1956), 452–459.

DUDA, A., "Diffusion Approximations for Time-Dependent Queueing Systems," *IEEE J. on Selected Areas in Comm.*, 4 (1986), 905–918.

GAVER, D. P., "Observing Stochastic Processes and Approximate Transform Inversion," *Oper. Res.*, 14 (1966), 444–459.

GELENBE, E. AND C. ROSENBERG, "Queues with Slowly Varying Arrival and Service Processes," *Management Sci.* (1990), 928–937.

GRASSMANN, W., "The Convexity of the Mean Queue Size of the $M/M/c$ Queue with Respect to the Traffic Intensity," *J. Appl. Prob.*, 20 (1983), 916–919.

GREEN, L. AND P. KOLESAR, "Testing the Validity of a Queueing Model of Police Patrol," *Management Sci.*, 35 (1989), 127–148.

——— AND ———, "A Database for the Study of Nonstationary Sinusoidal Markovian Queues," Research Working Paper, Columbia University, 1990.

———, ——— AND A. SVORONOS, "Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems," *Oper. Res.* 39 (1991), forthcoming.

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

KLEINROCK, L., *Queueing Systems Volume* 1: *Theory*, John Wiley & Sons, New York, 1975.

KOLESAR, P., K. L. RIDER, T. B. CRAYBILL AND W. W. WALKER, "A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars," *Oper. Res.*, 23 (1975), 1045–1062.

KOOPMAN, B. O., "Air-Terminal Queues under Time-Dependent Conditions," *Oper. Res.*, 20 (1972), 1089–1114.

KOTIAH, T. C., "Approximate Transient Analysis of Some Queueing Systems," *Oper. Res.*, 26 (1978), 333–346.

LUCHAK, G., "The Solution of the Single Channel Queueing Equations Characterized by a Time-Dependent Arrival Rate and a General Class of Holding Times," *Oper. Res.*, 4 (1956), 711–732.

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

NEUTS, M. F., "The $M/M/1$ Queue with Randomly Varying Arrival and Service Rates," *Opsearch*, 14 (1978), 139–157.

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

———, *Applications of Queueing Theory*, Chapman and Hall, London, 1971.

RIDER, K. L., "A Simple Approximation to the Average Queue Size in the Time-Dependent $M/M/1$ Queue," *J. Assoc. Comput. Math.*, 23 (1976), 361–367.

ROLSKI, T., "Queues with Nonstationary Input Stream: Ross's Conjecture," *Adv. Appl. Prob.*, 13 (1981), 603–618.

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

———, "Approximation of Periodic Queues," *Adv. Appl. Prob.*, 19 (1987), 691–707.

ROSS, S. M., "Average Delay in Queues with Nonstationary Poisson Arrivals," *J. Appl. Prob.*, 15 (1978), 602–609.

ROTHKOPF, M. H. AND S. S. OREN, "A Closure Approximation for the Nonstationary $M/M/s$ Queue," *Management Sci.*, 25 (1979), 522–534.

TAAFE, M. R. AND K. L. ONG, "Approximating Nonstationary $Ph(t)/M(t)/s/c$ Queueing Systems," *Annals of Oper. Res.*, 8 (1987), 103–116.

——— AND ———, "Approximating $Ph(t)/Ph(t)/l/c$ Nonstationary Queueing Systems," *Math. and Computers and Simulation*, 30 (1988), 441–452.

——— AND ———, "Nonstationary Queues with Interrupted Poisson Arrivals and Unreliable/Repairable Servers," *QUESTA*, 4 (1989), 27–46.

WHITT, W., "The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues is Asymptotically Correct as the Rates Increase," *Management Sci.*, 37 (1991), forthcoming.