SOME EFFECTS OF NONSTATIONARITY ON MULTISERVER MARKOVIAN QUEUEING SYSTEMS

LINDA GREEN, PETER KOLESAR and ANTHONY SVORONOS

Columbia University, New York, New York

(Received November 1987; revisions received June 1988, November 1989, January 1990; accepted February 1990)

We examine the effects of nonstationarity on the performance of multiserver queueing systems with exponential service times and sinusoidal Poisson input streams. Our primary objective is to determine when and how a stationary model may be used as an approximation for a nonstationary system. We focus on a particular question: How nonstationary can an arrival process be before a simple stationary approximation fails? Our analysis reveals that stationary models can seriously underestimate delays when the actual system is only modestly *nonstationary*. Other findings include confirmation and elaboration of S. M. Ross's conjecture that expected delays increase with nonstationarity, and the identification of easily computed and tight lower and upper bounds for expected delay and the probability of delay. These empirical results are based on a series of computer experiments in which the differential equations governing system behavior are solved numerically.

In most real service systems, the customer demand process is nonstationary. Examples can be found in Edie (1954), Koopman (1972), Segal (1974), Kolesar et al. (1975), Kleinrock (1976), Kolesar (1984) and Landauer and Becker (1989). Indeed, it is difficult to imagine actual service systems in which arrivals are truly stationary, yet virtually all theoretical queueing models make this assumption. This is not surprising because the mathematics of nonstationary stochastic processes is so complex that analytical results are very limited. Moreover, theoreticians since Erlang himself have presumed that stationary models can be used to assist managers and designers who must deal with a nonstationary world (see Brockmeyer, Halstrom and Jensen 1948).

Of course, if the arrival process is strongly nonstationary, a sensible analysis will have to deal with it explicitly. In practice, several approaches have been employed. First, one can isolate the period of peak arrivals and carry out a separate (stationary) analysis using as the arrival rate some average of the arrival rate during the peak period. Second, the entire time period can be segmented with average arrival rates estimated for each segment and used as inputs to a series of independent (and stationary) analyses. Third, the nonstationarity of the actual process may be explicitly captured in a simulation model. The choice among these approaches, decisions about how to execute them (how long a peak period, how many segments, etc.) and the utility of these approaches depend on an understanding of how nonstationarity affects the accuracy of stationary approximations. This paper proposes to contribute to that understanding by focusing on what appears to be a fundamental question: How nonstationary can an arrival process be before a simple stationary approximation fails? Our ongoing research builds on the work reported here and deals explicitly with issues, such as the accuracy of peak period analyses and segmented analyses. Results on these will be reported in subsequent papers.

There appears to be a broad agreement among theoreticians and practitioners that if the arrival rate fluctuations are *mild*, they can be ignored and a standard stationary analysis used. Yet, there has been no exploration of what level of nonstationarity is *sufficiently* mild for this to be true. Furthermore, though it is widely believed that delays increase as nonstationarity of the arrivals increases, there has been no proof of this even for the simplest Markovian single-server model. In fact, the literature does not contain a general definition of the degree of nonstationarity. (Even the existence of a limiting distribution for the waiting-time process for the single-server queue with periodic Poisson input was not proven until Harrison and Lemoine 1977.)

Thus, the question of when a stationary model may be reasonably used to estimate performance for a nonstationary system is clearly of practical as well as theoretical importance. There are situations in which use of a stationary model can lead to bad decision making, e.g., by suggesting two few servers to meet the desired performance. This may even be true when the stationary model is used in a more sophisticated manner, e.g., with time segmentation (see Green and Kolesar 1989). Yet, approaches that explicitly model nonstationarity, such as Monte Carlo simulations or numerical solution of the differential equations of the system, can be quite onerous in modeling and computational effort, especially for systems with a large number of servers.

Subject classifications: Queues, multichannel Markovian: effects of nonstationarity. Queues, nonstationary: behavior as a function of nonstationarity.

0030-364X/91/3903-0000 $01.25

The major goal of this paper is to gain a better understanding of how nonstationarity affects delays in queueing systems and, hence, to provide insights and guidelines on when nonstationarity can safely be ignored, and, perhaps more importantly, when it cannot. Although there is a substantial literature dealing with nonstationary queues, there has been no systematic study of this issue. Most of the literature primarily concerns numerical solutions of either exact (see, e.g., Koopman 1972, Luchak 1957) or approximate (see, e.g., Rothkopf and Oren 1979, Clark 1981) differential equations of nonstationary systems. Prime among the few papers that address the behavior of nonstationary systems relative to their stationary counterparts is that of Ross (1978) which puts forth two conjectures: First, that in a single-server infinite capacity queueing system, the "more nonstationary" the arrival process, the greater the average delay. Second, that in a finite capacity model, the proportion of lost customers is greater when arrivals are nonstationary than when they are stationary. Rolski (1981) proves that the average delay is at least as large in an infinite capacity, single-server system with nonstationary Poisson arrivals as it is in the comparable M/G/1 queue. Heyman (1982) shows that the average delay does not necessarily increase as the arrival process becomes "more nonstationary." He also shows by a counterexample that the second conjecture is not always valid. However, Rolski (1984) proves that the second Ross conjecture is true for pure loss systems with exponential service times and one or two servers, while Svoronos and Green (1988) show that for single-server loss systems with exponential service times and periodic Poisson input, the proportion of losses is convex increasing in the amplitude. Newell (1968, 1971) uses a diffusion approximation to examine the dynamic behavior of a queueing system in which the arrival rate increases at a nearly constant rate to a maximum that exceeds the (constant) service rate, and then decreases. He found that there exists a characteristic time T prior to the peak of the arrival rate such that for $t \ll T$, the queue distribution stays close to the equilibrium distribution that would prevail if the system were stationary with the arrival rate at time t. Newell (1971) also presents approximations for use when the distribution stays close to the "quasiequilibrium."

In this paper, we numerically investigate the behavior of multiserver, exponential queueing systems with sinusoidal Poisson input. This class of models was chosen because the exact differential equations for the steady-state probabilities can readily be solved numerically and because this easily parameterized arrival process captures the essence of many actual periodic arrival processes. (In our own work on real-world queueing problems in police patrol, firefighting, banks, and telecommunications we have encountered arrival rate processes with strong unipeaked daily cycles.)

The only results we found that are in the same spirit as ours appear in Rothkopf and Oren (1979). As an application of their closure approximation method for systems with nonstationary Markovian queues, they note some effects of sinusoidal input on the behavior of the time average mean number in the system as the amplitude and arrival rate increase.

Our findings include: 1) a numerical confirmation and extension of Ross's conjecture to the effect that expected delay is convex increasing in the amplitude of the arrival process, 2) the establishment of easy-to-compute tight upper bounds for the probability of delay and expected delay that provide a simple method for bounding the errors produced by any estimate, including the stationary one, and 3) overwhelming evidence that if a queueing system has a periodic arrival process with a relative amplitude (amplitude normalized by the average arrival rate) of only 10% (which is not obviously *very* nonstationary), a stationary model is likely to produce poor estimates of performance. In an attempt to understand better what factors contribute to making performance difficult to predict from a stationary model, we also explore the effects of traffic intensity, system size and event frequency (i.e., the expected number of arrivals and departures per cycle) on relative system performance.

Following a description of our research methodology in Section 1, we present our results in Section 2. In Section 3, we use these findings to explore the accuracy of a *completely stationary* approximation under various conditions. We conclude in Section 4 with a summary of our findings and the resulting conjectures.

1. METHODOLOGY

1.1. Evaluation of Performance Measures

Our numerical results were obtained by solving the following set of differential equations for the $M(t)/M/s$ system:

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

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

$$p_n'(t) = \lambda(t)\,p_{n-1}(t) + s\mu p_{n+1}(t) - (\lambda(t) + s\mu)\,p_n(t), \quad n \geqslant s$$

where $\lambda(t)$ is the arrival rate at time t, which we assume varies according to a sinusoid, μ is the service rate, and $p_n(t)$ is the probability of n customers in the system at

time t. We will assume that $\bar{\lambda} = \int_0^T \lambda(t)\,dt/T < s\mu$, where T is the period of the sinusoid or the cycle length. Thus, the system will develop a periodic steady-state behavior (see Heyman and Whitt 1984).

The numerical integration was performed using the International Math-Science Library subroutine DVERK which recursively uses fifth- and sixth-order Runge-Kutta methods. The length of the recursion interval is determined internally so that a user specified global error is not exceeded. Our integrations were usually initialized using the steady-state M/M/s solution obtained with $\lambda(0)$ as the stationary arrival rate.

In order to obtain an accurate solution for the infinite capacity system, we truncated the number of equations employed by following a method suggested in Odoni and Roth (1983). A maximum dimension of N is used by DVERK as the number of equations to be solved on the first call. After each subsequent call of DVERK (having solved $N' \leqslant N$ equations), the probability $p_{N'}(t)$ of a saturated system is compared to a specified small number, ϵ (e.g., $\epsilon = 10^{-8}$). If $p_{N'}(t)$ is larger than ϵ, the number of equations solved at the next call is increased by m (e.g., $m = 5$). Conversely, if both $p_{N'}(t)$ and $p_{N'-m}(t)$ are less than ϵ, N' is reduced by m for the next call.

For conceptual convenience and ease of discussion, we assume that a cycle is 24 hours. DVERK is called every five minutes on the simulated clock and the solution vector of state probabilities is used to calculate the probability of all servers busy and the expected number in the queue. Thus, each cycle is divided into 288 five-minute segments. (Test cases reveal that the daily average measures of performance obtained by using a five-minute grid yield identical results to five decimal places as compared to those obtained by using a one-minute grid.)

Figure 1 shows a typical graph of the time-varying probability that all servers are busy (probability busy) produced by this method. We observe first that the probability busy peak occurs after the peak in the arrival rate process, and second, that the probability busy curve is clearly not a sinusoid. In particular, it is not symmetric, displaying a shorter time from maximum to minimum then vice versa. (In the case shown in Figure 1, the time from peak of probability busy to the minimum value is about 11.5 hours.) These patterns were evident in the other cases we examined as well. Further work on the behavior of time-varying measures and related implications for using stationary models will be reported in a subsequent paper.

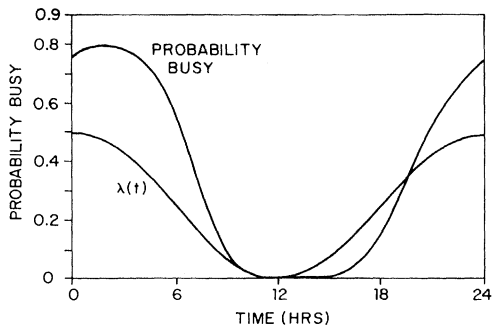

Figure 1. Probability of all servers busy versus time.

In this paper, we focus on three key performance measures: the daily (customer average) expected delay, the daily average queue length, and the daily (customer average) probability of delay. We use these composite measures, rather than time dependent ones, so that we can compare the performance of the nonstationary systems to their stationary counterparts. In addition to customer averages, we also examined time average measures (e.g., the probability of all servers busy) for which we reach the same conclusions. Since customer averages are more commonly used for managing systems, e.g., for capacity planning, we display our results for these in this paper.

For nonstationary systems, we calculate these measures by averaging the instantaneous measures provided at the 288 time segments of the five-minute grid. Assuming that segment 1 begins at midnight and that segments are numbered consecutively, let λ_i be the average arrival rate at the start of segment i, so that $\bar{\lambda} = \sum_{i=1}^{288} \lambda_i/288$. Let p_{ni} be the probability that n customers are in the system at the start of segment i. Then the daily average probability of delay is defined as

$$p_d = \sum_{i=1}^{288} \lambda_i \left(1 - \sum_{n=0}^{s-1} p_{ni}\right) \Big/ 288\bar{\lambda}$$

$$= \sum_{i=1}^{288} \lambda_i p_{di}/288\bar{\lambda} \qquad (1)$$

where p_{di} is the probability that all servers are busy at the start of segment i.

Similarly, the daily average expected delay is defined as

$$Wq = \sum_{i=1}^{288} \sum_{n=s}^{N} (n-s)\,p_{ni}/288\bar{\lambda} \qquad (2)$$

or

$$Wq = \sum_{i=1}^{288} \frac{\lambda_i\left(\sum_{n=s}^{N}(n-s+1)\,p_{ni}\right)/s\mu}{288\bar{\lambda}}$$

$$= \sum_{i=1}^{288} \frac{\lambda_i Wq_i}{288\bar{\lambda}} \qquad (3)$$

where N is the maximum allowable system size as specified in the DVERK computer program and Wq_i is the expected delay for a customer arriving at the start of segment i. From Little's formula, the expected queue length is $Lq = \bar{\lambda} Wq$. For the stationary models, we obtain the usual long-run measures from the equations for the M/M/s system.

1.2. Experimental Strategy

The generic model used in our study has four parameters. These are s, the number of identical exponential servers, μ, the service rate of each server, $\bar{\lambda}$, the daily average arrival rate, and A (> 0) the amplitude of the sinusoidal arrival process given by

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

Without loss of generality, the period is assumed to be 24 hours.

The interpretation of the parameters and their various combinations is key to understanding our experimental strategy and results. Consistent with tradition, the average traffic intensity, $\bar{\rho} = \bar{\lambda}/s\mu$ measures the average load on the system, and s, the number of servers, is a fundamental measure of system size. We will also use the measure of maximum traffic intensity, $\rho_{\max} = (\bar{\lambda} + A)/s\mu$. The amplitude of the arrival process, A, is intuitively a measure of the nonstationarity, though we often found it more useful to discuss our results in terms of the relative amplitude, $RA = A/\bar{\lambda}$, which normalizes the amplitude with respect to the average arrival rate. Relative amplitude provides a uniform scale for the degree of nonstationarity which varies between 0 and 1, and thus allows for easier comparison of results across different systems. Another possible measure of nonstationarity is the frequency of events per cycle as measured by both $\bar{\lambda}$ and μ, the average arrival and departure rates. This idea will be discussed further in Section 2.

In the specific model instances investigated, we select parameter values with $\bar{\lambda} < s\mu$ to assure the existence of a limiting distribution (see Heyman and Whitt 1984), and $\bar{\lambda} - A \geqslant 0$ so that $\lambda(t) \geqslant 0$ 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.

We began our experiments on a set of trial models with whose results we did a preliminary exploration of the effects of varying the amplitude and frequency of the arrival process, as well as the roles of traffic intensity and system size on behavior. The results of these initial runs confirmed some hypotheses, modified others, and gave one negative result.

We then specified a broader range of experimental runs that would confirm our modified set of conjectures over a wide range of conditions. Though numerical results can never truly prove a result, our intent was to cover a broad enough spectrum of models to constitute a very convincing case. This spectrum of models was based on four service rates reflecting time ranges that we had experienced in real-world applications of queueing theory.

1. $\mu = 0.2$, or equivalently a mean service time of 5 hours, represents situations such as field maintenance or service of mechanical systems or processing times for some chemical batch production systems.
2. $\mu = 2$, or equivalently mean service times of half an hour, corresponds to some emergency systems, such as police, fire or ambulance services.
3. $\mu = 20$, or equivalently a mean service time of 3 minutes, approximates certain factory operations or transaction times for automatic teller machines.
4. $\mu = 200$, which gives mean service times of a third of a minute, reflects operations in some computer and telecommunications systems.

For each of these values of μ, we designed a target set of models that spans a range from low (0.25) to high (0.75) average traffic intensities, a range of relative amplitudes from 0 to 1, and a range of system sizes from 1 to 12 servers.

Each experimental model can be viewed as a point in the four dimensional space $\bar{\lambda}$, μ, s, A. Our original goal was to experiment at *extreme* and *interpolation* points of this space. Unfortunately, due to computational limitations, we were unable to carry out all runs in this target range of experiments. For systems with a high value of μ (20 or 200), high traffic intensity, high relative amplitude and many servers, the CPU time needed to accurately estimate the limiting distributions at peak congestion grew beyond practical limits.

A few words about the nature of these limitations: As with stationary systems, in these models the number of customers in the system expands without bound as the traffic intensity approaches 1, yet our computer program was limited to a maximum of 600 customers in the system. Worse, in highly nonstationary systems, a large excess of customers can arrive during a peak period and drive up the number of customers in the system even when the average traffic intensity is moderate. Moreover, in systems that are otherwise well behaved, a large number of servers would often take us beyond the maximum number of customers in the system even when the traffic intensity and relative amplitude were moderate. On the other hand, for some of the systems with slow service rates (e.g., $\mu = 0.2$), service times are so long

that the number of periods needed before the state probabilities approach the limiting periodic steady-state behavior was intolerably large.

As an alternative, we adopted (or, if you will, retreated to) an experimental strategy in which for each hypothesis to be tested we selected a *central test case* and several *surrounding test cases* that would confirm the result at a point inside our experimental domain and demonstrate that it also holds for perturbations from this point in each parameter. These runs are described in the following sections. A data set listing the results for all experimental runs in this exercise is contained in Green and Kolesar (1990) and may be obtained from the authors.

2. EFFECTS OF NONSTATIONARITY

Do delays increase as the arrival process becomes "more nonstationary?" In addressing this question, it is important to note that "degree of nonstationarity" is not a well defined concept. For a sinusoidal-Poisson arrival process, it is intuitive that for a fixed average rate $\bar{\lambda}$, the larger the amplitude, the greater the nonstationarity. In his 1978 paper, Ross suggested that the frequency of events per cycle is also a measure of nonstationarity. In particular, he discussed a queueing system with a Poisson arrival process that alternates between two arrival rates (states) according to a continuous-time Markov process, and he argued that "intuitively," the more frequent the transitions between the two states, the "more stationary" the process. Thus, we will consider both amplitude and frequency as measures of nonstationarity. We first investigate the effects of amplitude on delays. For ease of comparison among systems, most of the following discussion will be in terms of the relative amplitude, $RA = A/\bar{\lambda}$.

2.1. Amplitude

Figure 2 illustrates our major experimental result on the effects of amplitude—expected delay is convex increasing with amplitude. Moreover, the greater the traffic intensity, the steeper the rate of increase. The figure illustrates these findings for our central test case of $\bar{\lambda} = 6$, $\mu = 2$ and $s = 6$, and in addition shows expected delay curves for several other values of $\bar{\lambda}$ corresponding to average traffic intensities ranging from 0.25 to 0.75. Such graphs generated for other models including cases with $\mu = 0.2, 2, 20$; $\bar{\lambda} = 3, 6, 7.8, 9$; and $s = 3, 6, 9$ confirm this result for perturbations of parameter values from our central case of $\mu = 2$, $\bar{\lambda} = 6$, $s = 6$.

The convexity and the steepness of the curves have strong implications for the appropriateness of using a stationary model based on the daily average arrival rate

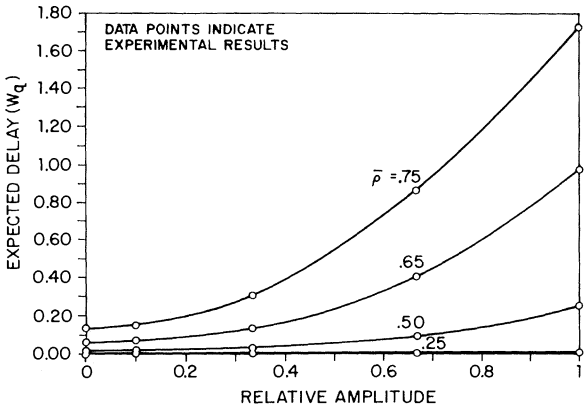

Figure 2. Expected delay versus relative amplitude when $\mu = 2$ and $s = 6$.

to estimate expected delays. Looking at the top curve which corresponds to an average traffic intensity $\bar{\rho}$ of 0.75, we see that with a relative amplitude of only $^1/_3$, the actual expected delay (0.308) is more than twice that of the stationary expected delay (0.141). At a relative amplitude of 100%, the actual expected delay is more than ten times the stationary estimate.

Figure 3 illustrates the effect of relative amplitude on the probability of delay for the same parameter values just discussed. Though we again observe a monotone deterioration in performance as relative amplitude increases, the curves level off since the probability of delay is bounded above by 1.

2.2. Frequency of Events

We next consider the effects on delays of increases in the frequency of events. To do this, we assume that the cycle length T is fixed (at 24 hours) and that the number of servers s is fixed, and we examine cases in which we

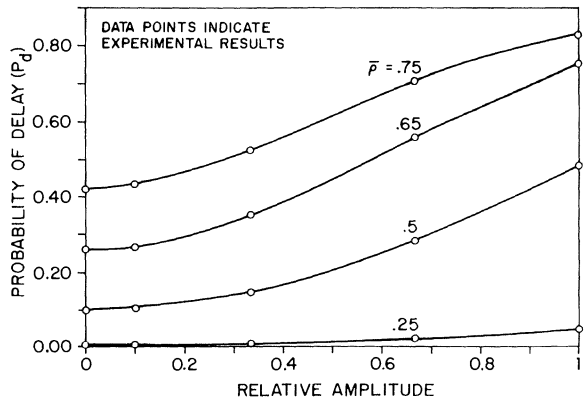

Figure 3. Probability delay versus relative amplitude when $\mu = 2$ and $s = 6$.

This result was first conjectured by Grassmann (1983) for the M/M/s queue as a consequence of the convexity of the mean queue size with respect to the arrival rate. The intuition here is obtained by noting that when $\bar{\lambda}$ and μ are both large, the number of events during any given small interval Δt becomes so large that the system approaches steady-state behavior during Δt. As $\Delta t \to 0$, $\lambda(t)$ for $t \in (t', t' + \Delta t)$ for any given t' will be almost constant. Thus, as frequency goes to infinity, the overall expected queue length approaches the expectation over time of the expected queue length in a system that at every time t behaves like a stationary M/M/s with arrival rate $\lambda(t)$.

Another interesting observation seen in Figure 4 is that the rate of increase of expected queue length is quite sharp so that the upper bound behavior starts appearing even at moderate average arrival rates—at about 5 arrivals per hour in this case.

In Figure 5, the maximum traffic intensity is greater than one and we see that expected queue length is convex increasing in the frequency of arrivals. Graphs generated for other systems in which $\rho_{\max}$ is greater than one, including models with 3 and 6 servers and relative amplitudes of 0.5 and 1.0, confirmed this finding.

Now we consider how the probability of delay behaves as a function of arrival frequency. Not surprisingly, we found that there exists an upper bound analogous to (7) for the probability of delay. That is, as $\bar{\lambda} \to \infty$ (or $T \to \infty$), the probability of delay approaches

$$p_d^{\infty} = \frac{1}{\bar{\lambda} T} \int_0^T \lambda(t)\, p_d(\lambda(t))\, dt \tag{8}$$

where $p_d(\lambda(t))$ is the Erlang delay formula for a stationary system; see, e.g., Gross and Harris (1985). However, unlike our result for expected delay, (8) results in a finite upper bound even when $\rho_{\max}$ exceeds one. This was confirmed empirically for all of the more than 300 models that we examined. Figure 6 shows the probability of delay versus the frequency curve for a system with $\rho_{\max}$ less than one, and Figure 7 for a system with $\rho_{\max}$ greater than one.

3. ACCURACY OF THE SIMPLE STATIONARY APPROXIMATION

We next examine the implications of using a stationary model based on the daily average arrival rate to estimate delays in a nonstationary system. That is, how good an approximation is obtained by ignoring the nonstationarity? More specifically, we are interested in determining under what conditions, if any, this stationary approximation will yield reasonably accurate results for the expected delay and the probability of delay. Thus, in addition to exploring accuracy as a function of the degree of nonstationarity (i.e., amplitude and frequency of the arrival process), we discuss how, for a given degree of nonstationarity, the accuracy of a stationary approximation may be affected by the basic system characteristics of traffic intensity and system size.

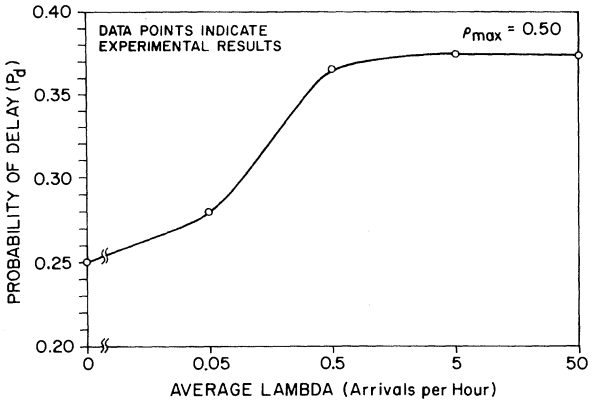

Figure 6. Probability of delay versus frequency of arrivals when $s = 1$, $RA = 1$ and $\bar{\rho} = 0.25$.

For this purpose, we define a relative error measure for expected delay as

$$RE = \frac{\text{Actual } E(\text{delay}) - \text{Stationary } E(\text{delay})}{\text{Actual } E(\text{delay})}$$

where the stationary E(delay) is obtained from the stationary M/M/s model using $\lambda = \bar{\lambda}$ (see Gross and Harris, Equation 2.50). The relative error for expected queue size and probability of delay are analogously defined.

3.1. The Effect of Amplitude

How is the relative error of expected delay affected by amplitude? Since we already know that expected delay

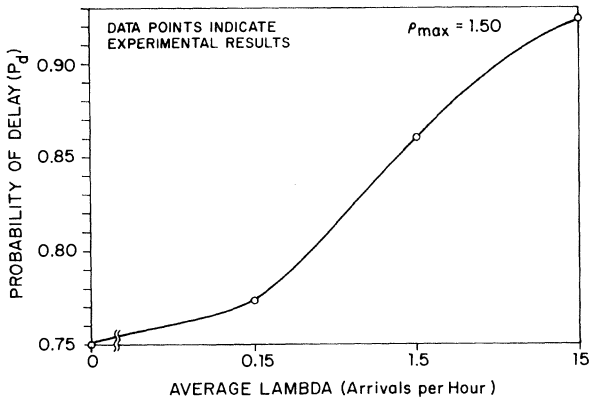

Figure 7. Probability of delay versus frequency of arrivals when $s = 1$, $RA = 1$ and $\bar{\rho} = 0.75$.

increases convexly with amplitude, it follows that the relative error will increase as well. Figure 8 illustrates this behavior for a typical case. The most interesting observation here is that for all cases in this figure, when relative amplitudes are greater than 10%, the relative error is greater than 10%. In *all* of the models we studied with a relative amplitude of 25% or more, the relative error was always greater than 10%, and usually significantly so. Since, from a practical perspective, 25% is a rather small degree of nonstationarity, this finding implies that *using the average arrival rate to estimate expected delays in a system with a time-varying arrival process is likely to be quite misleading.*

Figure 9 shows the corresponding relative error plot for the probability of delay. Though, not unexpectedly, the errors are smaller (due to the boundedness of this measure), they too become quite significant for relative amplitudes of greater than 10%. Thus, again it appears that using the stationary model with the average arrival rate is likely to produce unacceptably inaccurate estimates.

3.2. The Effect of Frequency of Events

A somewhat positive interpretation of our results on relative error versus relative amplitude is that using a model based on the average arrival rate to estimate expected delay is generally fairly accurate (i.e., results in a relative error of $\leqslant 10\%$) if the relative amplitude is $\leqslant 10\%$. (We did find some exceptions to this.) Is there a similarly *safe* event frequency? Our empirical evidence says no. Figure 10 shows a typical graph for relative error versus frequency of events which indicates the contrary. The relative error curve rises so steeply that it is over 12% for $\bar{\lambda} = 0.05$ (i.e., an average of 1 arrival every 20 hours—an extreme case indeed, considering a cycle length of 24 hours). At $\bar{\lambda} = 0.5$, the relative error is already close to its upper bound of about 48.5%.

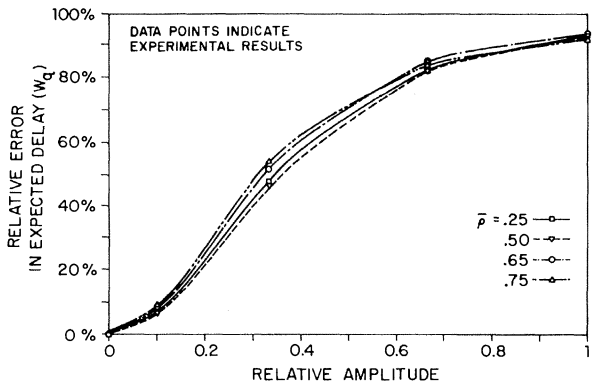

Figure 8. Error in expected delay versus relative amplitude when $\mu = 2$ and $s = 6$.

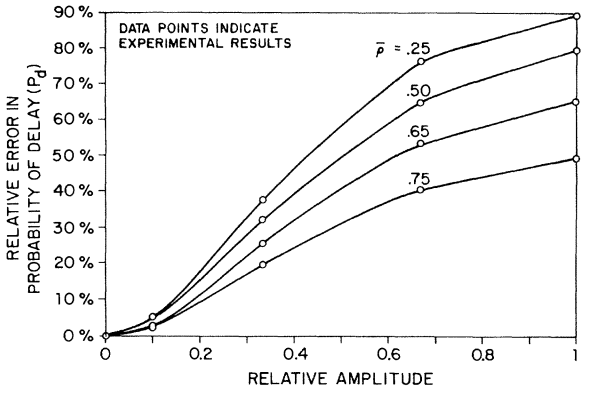

Figure 9. Error in probability of delay versus relative amplitude when $\mu = 2$ and $s = 6$.

(Since $\rho_{\max}$ here is less than 1, the expected queue length is bounded above from (7) and thus the relative error will be correspondingly bounded.)

3.3. The Effect of Traffic Intensity

For a fixed level of nonstationarity, i.e., relative amplitude and frequency, how is the relative error affected by traffic intensity? We had first conjectured that as congestion increases, the accuracy of the simple stationary estimate would get worse. Figure 11 shows that this is not the case. Indeed, there seems to be no easily predictable effect due to traffic intensity.

3.4. The Effect of System Size

What is the effect, if any, of increasing system size, as measured by the number of servers s, on the accuracy of the simple stationary approximation? First, our numerical results indicate that, as in stationary M/M/s systems, for a fixed traffic intensity, delays decrease as s

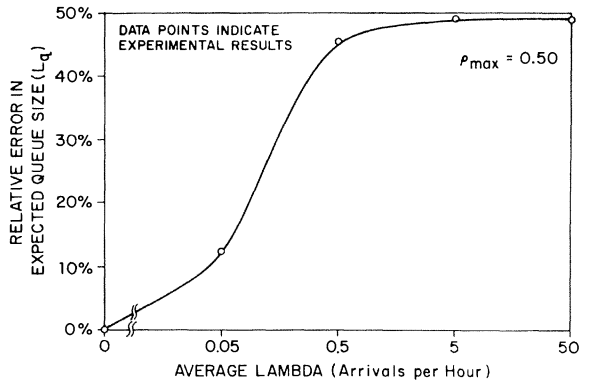

Figure 10. Error in expected queue size versus event frequency when $s = 1$, $RA = 1$ and $\bar{\rho} = 0.25$.

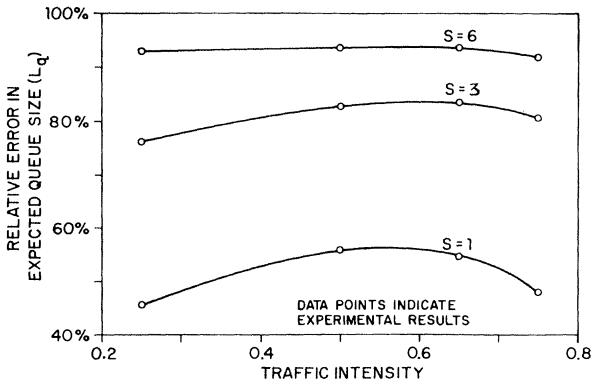

Figure 11. Error in expected queue size versus traffic intensity when $RA = 1$, $\mu = 2$ and $\bar{\rho} = 0.25$.

increases. However, it was not intuitively clear to us how the performance in a nonstationary system relative to its stationary counterpart might be affected by its size. To explore this, we compared the relative error in expected delays for queueing systems in which the service rate and relative amplitude were held fixed, and both the number of servers and the arrival rate were increased proportionally so that the traffic intensity remained constant. Figure 12 illustrates our general finding that relative error increases as system size increases by showing two curves generated for our central case of $\mu = 2$ and an average traffic intensity of 0.65. This finding was confirmed by similar graphs for models with $\mu = 0.2, 2, 20$; $\bar{\rho} = 0.25, 0.5, 0.65, 0.75$; and relative amplitudes of 0.1, $^1/_3$, $^2/_3$ and 1, all of which are perturbations from our central case of $\mu = 2$, $\bar{\rho} = 0.65$, and relative amplitude $= {}^1/_3$.

Does it make sense that as the system size increases the stationary approximation becomes worse? We offer an explanation similar to the one concerning the effect of frequency of events. As the number of servers and hence the arrival rate increases more customers arrive at the peak of the cycle producing congestion, which then propagates throughout the cycle. Conversely, as system size decreases, the corresponding decrease in arrivals, particularly during the peak, attenuates the effect of the nonstationarity and results in delays that are closer to those arising in the stationary system. Of course, it is important to note that at the relative amplitudes shown in Figure 12, the relative errors are generally unacceptably high even with systems with only three servers.

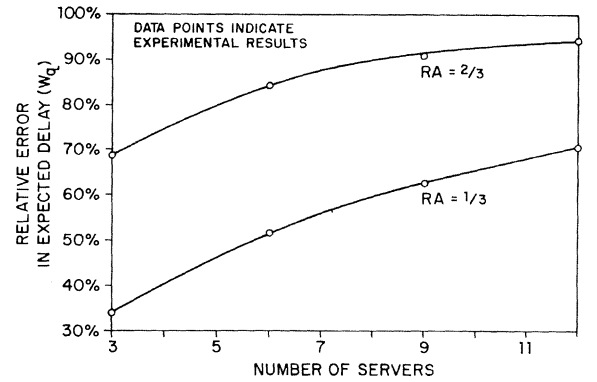

Figure 12. Error in expected delay versus system size when $\bar{\rho} = 0.65$ and $\mu = 2$.

4. FINDINGS AND CONJECTURES

The (mostly) empirical results described in this paper provide very strong evidence of the existence of some practically and theoretically important structural characteristics of Markovian queueing systems with periodic input which are summarized here.

1. Expected delay and probability of delay increase as amplitude increases. For expected delay, the increase is convex and the rate of increase is steeper for larger traffic intensities.
2. Expected queue length and probability of delay also increase as the event frequency, or equivalently, the cycle length increases. For both measures, we have proven that as the event frequency tends toward zero, the measures converge to the values obtained from the corresponding stationary model based on the average arrival rate.
3. For systems in which the maximum traffic intensity is strictly less than one, the expected queue length approaches an upper bound given by (7) as the event frequency, or equivalently, the cycle length approaches infinity, and otherwise, diverges. The probability of delay approaches a limit given by (8) as the event frequency goes to infinity, even for systems where the instantaneous traffic intensity exceeds one.

These results have strong implications for using the average arrival rate in a stationary model to estimate delays for time-varying systems. For small systems (e.g., one or two servers) with small relative amplitudes (e.g., less than 10%) and infrequent events (or equivalently short cycle lengths), such a stationary approximation may give reasonable estimates. However, relative error increases as each of these parameters increase and the sharpest increases occur as event frequency increases. In most systems we examined, and particularly those which are most likely to approximate real systems, the relative errors for both expected delay and probability of delay are unacceptably high.

The convergence of expected queue length and probability of delay to (7) and (8), respectively, suggest use of those equations as approximations to actual behavior. We explore this in Green and Kolesar (1991).

Our choice of a sinusoidal arrival rate function and exponential service times were largely determined by issues of computational and experimental convenience. There is nothing we know of to indicate that our basic findings would not hold in more general models. (We have proven that our result on the convergence of performance measures to the stationary case when cycle length goes to zero holds more generally. Also, Rolski's result (1986) for single-server systems holds for more general arrival and service processes.) In particular, we conjecture the following:

1. In multiserver queueing systems with periodic Poisson arrival processes and well behaved continuous service distributions (i.e., the first two moments exist), expected delay is convex increasing and probability of delay is monotone increasing in the amplitude.
2. For the same class of systems, as arrival frequency per cycle approaches infinity, expected queue length asymptotically approaches the upper bound given by (7) when the instantaneous traffic intensity is always strictly less than one, and otherwise, diverges. Similarly, the probability of delay approaches the upper bound given by (8), but for all cases.

REFERENCES

BROCKMEYER, E., H. L. HALSTROM AND A. JENSEN. 1948. *The Life and Works of A. K. Erlang*. Transactions of the Danish Academy of Technical Science **2**.

CLARK, G. M. 1981. Use of Polya Distributions in Approximate Solutions to Nonstationary M/M/s Queues. *Commun. ACM* **24,** 206–217.

EDIE, L. C. 1954. Traffic Delays at Toll Booths. *Opns. Res.* **2,** 107–138.

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

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

GREEN, L., AND P. KOLESAR. 1990. A Database for the Study of Nonstationary Sinusoidal Markovian Queues. Research Working Paper, Graduate School of Business, Columbia University, New York.

GREEN, L., AND P. KOLESAR. 1991. The Pointwise Stationary Approximation for Queues With Nonstationary Arrivals. *Mgmt. Sci*, **37,** 84–97.

GROSS, D., AND C. M. HARRIS. 1985. *Fundamentals of Queueing Theory*, 2nd ed. John Wiley & Sons, New York.

HARRISON, J. M., AND A. J. LEMOINE. 1977. Limit Theorems for Periodic Queues. *J. Appl. Prob.* **14,** 566–576.

HEYMAN, D. P. 1982. On Ross's Conjectures About Queues With Nonstationary Poisson Arrivals. *J. Appl. Prob.* **19,** 245–249.

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

KLEINROCK, L. 1976. *Queueing Systems, Vol. II: Computer Applications*. John Wiley, New York.

KOLESAR, P. J. 1984. Stalking the Endangered CAT: A Queueing Analysis of Congestion at Automated Teller Machines. *Interfaces* **14,** 16–26.

KOLESAR, P. J., K. L. RIDER, T. B. CRABILL AND W. E. WALKER. 1975. A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars. *Opns. Res.* **23,** 1045–1062.

KOOPMAN, B. O. 1972. Air-Terminal Queues Under Time-Dependent Conditions. *Opns. Res.* **20,** 1089–1114.

LANDAUER, E. G., AND L. C. BECKER. 1989. Reducing Waiting Time at Security Checkpoints. *Interfaces* **19,** 57–65.

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

NEWELL, G. F. 1968. Queues With Time-Dependent Arrival Rates I–III. *J. Appl. Prob.* **5,** 436–606.

NEWELL, G. F. 1971. *Applications of Queueing Theory*. Chapman & Hall, London.

ODONI, A. R., AND E. ROTH. 1983. An Empirical Investigation of the Transient Behavior of Stationary Queueing Systems. *Opns. Res.* **31,** 432–455.

ROLSKI, T. 1981. Queues With Non-Stationary Input Stream: Ross's Conjecture. *Adv. Appl. Prob.* **13,** 603–618.

ROLSKI, T. 1984. Comparison Theorems for Queues With Dependent Interarrival Times. In *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (eds.). *Lecture Notes in Control and Information Sciences* **60,** 42–67.

ROLSKI, T. 1986. Upper Bounds for Single Server Queues With Doubly Stochastic Poisson Arrivals. *Math. Opns. Res.* **11,** 442–450.

ROSS, S. M. 1978. Average Delay in Queues With Non-Stationary Poisson Arrivals. *J. Appl. Prob.* **15,** 602–609.

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

SEGAL, M. 1974. The Operator Scheduling Problem: A Network Flow Approach. *Opns. Res.* **22,** 808–823.

SVORONOS, A., AND L. GREEN. 1988. A Convexity Result for Single Server Exponential Loss Systems With Nonstationary Arrivals. *J. Appl. Prob.* **25,** 224–227.

WHITT, W. 1974. The Continuity of Queues. *Adv. Appl. Prob.* **6,** 175–183.