On the Timing of the Peak Mean and Variance for the Number of Customers in an *M(t)/M(t)*1 Queueing System

K. Malone and A. Ingolfsson

OR 301-94 October 1994

On the Timing of the Peak Mean and Variance for the Number of Customers in an $M(t)/M(t)/1$ Queueing System

Kerry M. Malone*
Massachusetts Institute of Technology

Ármann Ingólfsson†
Massachusetts Institute of Technology

July 1994

Revised October 1994

Abstract

This paper examines the time lag between the peak in the arrival rate and the peaks in the mean and variance for the number of customers in an $M(t)/M(t)/1$ system. We establish a necessary condition for the time at which the peak in the mean is achieved. In cases in which system utilization exceeds one during some period, we show that the peak in the mean occurs after the end of this period.

Keywords: Queues, Nonstationary: Timing of the Peak Mean and Variance; Queues, Markovian: Effects of Nonstationarity on Mean and Variance.

Motivation

The validity of most analytical results in Queueing Theory is contingent on a series of strict assumptions, one of which is that arrival and service rates do not vary with time. Most real-world

*address: Operations Research Center, Massachusetts Institute of Technology, E40-139, Cambridge, MA 02139, e-mail: kmm@mit.edu

†address: Operations Research Center, Massachusetts Institute of Technology, E40-135, Cambridge, MA 02139, e-mail: armann@mit.edu

queueing systems, however, lack this critical characteristic. Airports, air terminals, manufacturing processes, roads and highways, automated teller machines and telecommunication networks are examples of systems for which the conditions underlying classical, analytical results from Queueing Theory are not tenable because the demand for service and the available capacity vary strongly with time. Hence it is often difficult to compute performance measures pertaining to these systems.

When managers design or analyze facilities with strongly time-varying demand and capacity, they commonly focus on the performance of the facility during *peak utilization:* that period during which the facility is busiest. For example, airport authorities strive to have their facilities designed so that aircraft and passenger delays in the peak periods are within tolerable limits. Therefore it is important to develop an understanding of when the peak in congestion (= expected number of customers) will occur in relation to the arrival peak. The difference between the time at which the arrival rate is highest and the time at which a system performance measure (for example number in system) reaches its highest value is called the *time lag.* The time lag can be matter of minutes or hours, depending on the type of queueing system, the average utilization rate, and how much the utilization peak rises over the average utilization level. In addition, insight into the relationship between the time-dependent mean, $m(t)$, and variance, $v(t)$, for the number of customers in the system will allow more effective management of congestion at facilities where demand and capacity vary strongly with time.

Researchers have already begun addressing this time lag issue in nonstationary queueing systems. In the case of oversaturated queues, Newell [8] conjectured that the peak mean number of customers in the system should occur at about the end of the period of oversaturation, without making assumptions about the particular form of the arrival process or service-time distribution. Using the diffusion approximation to nonstationary queues, he observed that the maximum variance for the number of customers in the system occurs later than the maximum mean number in system, based on numerical calculations. Green and Kolesar [4] and Green, Kolesar and Svoronos [5] addressed the behavior of several performance measures of $M(t)/M(t)/s$ queueing systems. They noted that not only did the arrival rate peak *not* coincide with peaks in other measures such as expected queue length and probability of delay, but the measures also behave differently from one another. They also noted that as the *event frequency* (the number of arrivals or service completions per cycle) increases, the lag between the peak in the arrival rate and the number in queue *decreases*. Eick, Massey and Whitt [2, 3] examined $M(t)/G/\infty$ queueing systems. Using the exact result for $m(t)$ derived by Palm [10] and Khintchine [6], they found exact expressions for $m(t)$, the extreme values of $m(t)$, and for the time lag between the peak in the arrival rate and the number of customers in the system, in the case of a sinusoidal arrival rate function. This result can also be used for approximating finite-server systems where the customer time in queue is small relative to the length of the service time. However, in many real-world systems with arrival and service rates which vary strongly as a function of time, the queueing time is not insignificant relative to the service time. Therefore, because there is no queueing in an $M(t)/G/\infty$ system (customers' system times are independent of one another), the time lag approximation with infinite-servers will underestimate the actual time lag in a finite-server system. Eick, Massey and Whitt also prove that the mean number of customers in an $M(t)/G/\infty$ system with a sinusoidal arrival rate is symmetric about its extremes, i.e., if an extreme occurs at time t_m^*, then $m(t_m^* - t) = m(t_m^* + t)$ for all t. In contrast, Malone and Odoni, and Green, Kolesar and Svoronos have empirical results showing that none of their system performance measures for finite server systems ($1 - 12$ servers) are symmetric about their extremes.

The purpose of this paper is to provide theoretical insight and computational results for the time lag between the peak in the arrival rate and the peaks in the mean and variance of the number in the system for $M(t)/M(t)/1$ and other single-server nonstationary queueing systems. We have found little in the literature regarding the behavior of the variance and standard deviation for the number of customers in nonstationary queueing systems.

The paper has three sections. Section 1 presents conditions for the extremes (local maxima and minima) of $m(t)$ to be achieved in an $M(t)/M(t)/1$ system. The relationship between $m(t)$ and $v(t)$, and the timing of the peak, m^*, of $m(t)$ are explored. Section 2 sets forth a hypothesis regarding the relationship between m^* and v^*, the peak of $v(t)$, along with supporting numerical results for the $M(t)/M(t)/1$ queueing system under a variety of conditions. Section 3 summarizes the results.

1 Extremal Conditions for the Mean and Variance of the Number of Customers in the System

A time lag between the peak in the arrival rate and the peak in system congestion has been observed for $M(t)/M(t)/s, M(t)/E_k(t)/1, M(t)/D(t)/1$, and other types of nonstationary queueing systems. Figure 1 shows an example of this lag for an $M(t)/M/1$ queueing system where $\lambda(t) = 75 + 50\sin(2\pi/24)$ and $\mu(t) = 100$. Figure 2 shows the variance for the number of customers in the system for an $M(t)/M/1$ queueing system with the same parameters as in figure 1. The peak in the arrival rate occurs at $t = 102$, the peak in the mean at about $t = 106$, and the peak in the variance at about $t = 109$. Our computational results indicate that the variance (and, of course, standard deviation) for the number in the system peaks later than the mean number in system does for all the nonstationary single-server queueing systems we have examined. In this section, we focus on the $M(t)/M(t)/1$ system and establish conditions for the peak in the mean and variance for the number of customers in the system, as well as the relationship between the two.

Notation: Let $\lambda(t)$ be the arrival rate of customers to the queueing system at time t, let $\mu(t)$ be the service rate at time t, and let $\rho(t)$ be the instantaneous system utilization. The probability that there are i customers in the system at time t will be denoted $P_i(t)$. The mean, variance, and second moment of the number of customers in the system will be denoted $m(t)$, $v(t)$, and $m_2(t)$, respectively. Primes will be used to denote derivatives, e.g., $m_2'(t) = dm_2(t)/dt$. Peak values (local maxima) will be denoted with asterisks; for example m^* will denote a peak value of $m(t)$. The time at which m^* is achieved will be denoted t_m^*.

After proving the following preliminary lemma, we will derive conditions for when the expected number of customers in system peaks.

Lemma 1 *In an $M(t)/M(t)/1$ queueing system, if $P_0(0) > 0$ and $\mu(t) \geq 0$ for $t \geq 0$, then*

$$P_0(t) > 0 \text{ for all } t \geq 0$$

Proof: The familiar Chapman-Kolmogorov forward equation for state 0 in an $M(t)/M(t)/1$ system is

$$P_0'(t) = -\lambda(t)P_0(t) + \mu(t)P_1(t).$$

Let $\hat{P}_0(0) = P_0(0)$, where $\hat{P}(t)$ satisfies

$$\hat{P}_0'(t) = -\lambda(t)\hat{P}_0(t). \tag{1}$$

The quantity $\hat{P}_0(t)$ will no greater than $P_0(t)$ for all $t \geq 0$, as we now show. Since $\mu(t)P_1(t) \geq 0$ it follows that $P_0'(t) \geq \hat{P}_0'(t)$ for all $t \geq 0$. Integrating on both sides of $P_0'(t) \geq \hat{P}_0'(t)$ we obtain

$$P_0(t) \geq \hat{P}_0(t) \text{ for all } t \geq 0. \tag{2}$$

The solution to equation (1) is (see, e.g., Luenberger [7]):

$$\hat{P}_0(t) = \exp\left[-\int_{\tau=0}^{t} \lambda(\tau)d\tau\right] P_0(0).$$

Since $P_0(0) > 0$ and $\exp\left[-\int_{\tau=0}^{t} \lambda(\tau)d\tau\right] > 0$ for all $t \geq 0$, we have that $\hat{P}_0(t) > 0$ for all $t \geq 0$. Finally, inequality (2) implies that $P_0(t) > 0 \; \forall t \geq 0$. ■

Theorem 1 *In an $M(t)/M(t)/1$ Queueing System, a necessary condition for the times at which the expected number of customers in the system $m(t)$ takes on its extreme values is:*

$$\frac{\lambda(t)}{\mu(t)} = 1 - P_0(t) \tag{3}$$

Proof: By differentiating both sides of $m(t) = \sum_{i=0}^{\infty} iP_i(t)$, which defines the expected value $m(t)$, we obtain

$$m'(t) = \sum_{i=0}^{\infty} iP_i'(t). \tag{4}$$

The Chapman-Kolmogorov forward equations for an $M(t)/M(t)/1$ system are:

$$\begin{aligned} P_0'(t) &= -\lambda(t)P_0(t) + \mu(t)P_1(t) \\ P_i'(t) &= \lambda(t)P_{i-1}(t) - (\lambda(t)+\mu(t))P_i(t) + \mu(t)P_{i+1}(t) \text{ for } i = 1, 2, \ldots \end{aligned}$$

Substituting in equation (4) we obtain

$$\begin{aligned} m'(t) &= \sum_{i=0}^{\infty} i \cdot P_i'(t) \\ &= \lambda(t)\sum_{i=1}^{\infty} iP_{i-1}(t) - \lambda(t)\sum_{i=0}^{\infty} iP_i(t) - \mu(t)\sum_{i=1}^{\infty} iP_i(t) + \mu(t)\sum_{i=0}^{\infty} iP_{i+1}(t) \\ &= \lambda(t)\sum_{i=0}^{\infty}(i+1)P_i(t) - \lambda(t)\sum_{i=0}^{\infty} iP_i(t) - \mu(t)\sum_{i=1}^{\infty} iP_i(t) + \mu(t)\sum_{i=1}^{\infty}(i-1)P_i(t) \\ &= \lambda(t) - \mu(t)\left[1 - P_0(t)\right] \end{aligned}$$

This was shown by Clarke [1] in 1956 and used by Rothkopf and Oren [11] in the derivation of their closure approximation for the nonstationary $M/M/s$ queue.

To find when $m(t)$ achieves its extreme values, we simply set $m'(t) = 0$ and find the following condition:

$$\begin{aligned} m'(t) = 0 \;&\Leftrightarrow\; \lambda(t) - \mu(t)[1 - P_0(t)] = 0, \\ &\Leftrightarrow\; \lambda(t) = \mu(t)[1 - P_0(t)], \\ &\Rightarrow\; \frac{\lambda(t)}{\mu(t)} = 1 - P_0(t) \text{ (if } \mu(t) > 0) \end{aligned} \tag{5}$$

Equation (5) must hold for $m(t)$ to achieve its local maximum, m^*, or local minimum, m_*. ■

Theorem 1 has the following important corollary.

Corollary 1 *Suppose that* $P_0(0) > 0$ *and* $\rho(t) > 1$ *for* $t \in (t_1, t_2)$ *for an* $M(t)/M(t)/1$ *system. Then the first congestion peak* m^* *after* t_1 *will occur after* t_2*, i.e.,* $t_m^* > t_2$*.*

Proof: From the relation $m'(t) = \lambda(t) - \mu(t)[1 - P_0(t)]$ we see that $m'(t) > 0$ (i.e., $m(t)$ will increase) whenever

$$\rho(t) = \frac{\lambda(t)}{\mu(t)} > [1 - P_0(t)]$$

Since $P_0(t)$ is a probability, it is bounded above by 1 and below by 0. Therefore, $1 \geq 1 - P_0(t) \geq 0$. When $\rho(t) > 1$,

$$\rho(t) > 1 \geq [1 - P_0(t)] \Rightarrow m'(t) > 0 \tag{6}$$

Therefore, $m(t)$ does not peak while $\rho(t) > 1$. The peak value m^* must then occur at or after the end of the period for which $\rho(t) > 1$, i.e., $t_m^* \geq t_2$. By Lemma 1, we know that $P_0(t) > 0$ for all $t \geq 0$ for systems in which $P_0(0) > 0$. Thus, $P_0(t_2) > 0$, which implies

$$1 - P_0(t_2) < 1 = \rho(t_2).$$

Therefore, the condition of equation (5) is not met, so m^* does not occur at t_2. We conclude that $t_m^* > t_2$. ∎

Figures 3A–C illustrate Corollary 1. They correspond to an $M(t)/M/1$ system with $\lambda(t) = 90 + 30\sin(2\pi/24)$ and $\mu(t) = 100$. Figures 3A, 3B, and 3C depict $\rho(t), m'(t)$, and $m(t)$, respectively, over one period of $\lambda(t)$. The times t_1 and t_2 mark the beginning and the end of the period during which $\rho(t) > 1$, t_3 is the time t_m^* at which $m(t)$ peaks, and t_4 is the time at which the minimum value m_* of $m(t)$ is achieved. Note that $t_3 - t_2$ is very small in this particular case, but positive nevertheless.

The expression for $m'(t)$ derived in the proof of theorem 1 provides insight into the transient behavior of $m(t)$ for a stationary $M/M/1$ system. Consider a system that starts out empty and has utilization less than one. Then $P_0(0) = 1$, so $\lambda/\mu > 1 - P_0(0) = 0$, implying that $m'(0) > 0$, i.e., the expected number in the system starts to grow. As $m(t)$ grows, $P_0(t)$ must decrease, causing $m'(t) = \lambda - \mu(1 - P_0(t))$ to decrease. Eventually $P_0(t)$ reaches its limiting value P_0 when $\lambda/\mu = 1 - P_0$, or $P_0 = 1 - \frac{\lambda}{\mu}$, which is the familiar steady-state probability of an empty system. This scenario and others with different initial conditions are depicted in Odoni and Roth [9].

In our work with nonstationary queueing systems, we have also observed that the behavior of the variance and standard deviation ($\sigma(t)$) for the number of customers in the system can be quite different from that of $m(t)$. One of the most salient differences is that $v(t)$ and $\sigma(t)$ peak later – sometimes much later – than $m(t)$. This behavior was not expected. Theorem 2 establishes a condition under which the variance peak v^* occurs later than the peak in the mean m^*. Let t_v^* be the time at which v^* is achieved.

Theorem 2 *In an $M(t)/M(t)/1$ system, $t_v^* \geq t_m^*$ iff*

$$\frac{1}{m^*+1} \geq P_0(t_m^*). \tag{7}$$

We remark that in all our numerical computations to date, $t_v^* > t_m^*$.

<u>Proof:</u> The proof consists of showing that $v'(t_m^*) \geq 0$ if condition (7) holds. First, we derive an expression for $v'(t)$ (previously derived by Clarke [1] and Rothkopf and Oren [11]) using the Chapman-Kolmogorov forward equations:

$$\begin{aligned}
m_2(t)' &= \sum_{i=0}^{\infty} i^2 P_i'(t) \\
&= \lambda(t)\sum_{i=1}^{\infty} i^2 P_{i-1}(t) - \lambda(t)\sum_{i=0}^{\infty} i^2 P_i(t) - \mu(t)\sum_{i=1}^{\infty} i^2 P_i(t) + \mu(t)\sum_{i=0}^{\infty} i^2 P_{i+1}(t) \\
&= \lambda(t)\sum_{i=0}^{\infty}(i^2+2i+1)P_i(t) - \lambda(t)\sum_{i=0}^{\infty} i^2 P_i(t) - \mu(t)\sum_{i=1}^{\infty} i^2 P_i(t) + \mu(t)\sum_{i=1}^{\infty}(i^2-2i+1)P_i(t) \\
&= \lambda(t) + \mu(t)\,[1-P_0(t)] + 2(\lambda(t)-\mu(t))\underbrace{\sum_{i=1}^{\infty} i\cdot P_i(t)}_{m(t)} \\
&= \lambda(t) + \mu(t)[1-P_0(t)] + 2m(t)(\lambda(t)-\mu(t))
\end{aligned}$$

By differentiating the relation $v(t) = m_2(t) - (m(t))^2$, we obtain

$$\begin{aligned}
v'(t) &= m_2(t)' - 2m(t)m'(t) \\
&= \lambda(t) + \mu(t)[1-P_0(t)] + 2m(t)(\lambda(t)-\mu(t)) - 2m(t)(\lambda(t)-\mu(t)[1-P_0(t)])
\end{aligned}$$

$$= \mu(t)\left[\frac{\lambda(t)}{\mu(t)} + 1 - P_0(t)[1 + 2m(t)]\right] \tag{8}$$

At the time t_m^*, when $m(t)$ achieves its peak m^*, the relation $\frac{\lambda(t_m^*)}{\mu(t_m^*)} = 1 - P_0(t_m^*)$ will hold, by Theorem 1. Substituting this relation into equation (8), we get:

$$v'(t_m^*) = 2\mu(t_m^*)\left(1 - P_0(t_m^*)[m^* + 1]\right). \tag{9}$$

Assuming that $\mu(t_m^*) > 0$, the right-hand-side of equation (9) will nonnegative iff $1 - P_0(t_m^*)[m^* + 1] \geq 0$, that is, if $\frac{1}{m^*+1} \geq P_0(t_m^*)$ ■

We note that

$$m^* \geq 0 \Longrightarrow 1 \geq \frac{1}{m^* + 1} \geq 0$$

i.e., $\frac{1}{m^*+1}$ is never greater than one so condition (7) is not trivially true. We also note that for stationary $M/M/1$ systems, $m^* = \rho/(1-\rho)$. Therefore, in this case, $\frac{1}{m^*+1} = 1 - \rho = P_0$, i.e., the relationship (7) is an equality.

We have provided results for the $M(t)/M(t)/1$ queueing system. In the course of investigating $M(t)/M/1$, $M(t)/E_k/1$, and $M(t)/D/1$ queueing systems, and three approximation methods for $M(t)/E_k/1$ systems, we obtained results consistent with the Theorems 1 and 2. Therefore, we conjecture that the results hold for a more general class of nonstationary systems, including the ones we investigated.

2 Computational Results for $M(t)/M/1$ Systems

In this section we present some of the computational results for $M(t)/M/1$ queueing systems which led us to investigate the time lags for the mean and variance for the number of customers in the system and to derive the results of Section 1. We also make two conjectures which we have not yet been able to prove but for which we have consistent results in all cases we have tried. We first outline our approach to nonstationary queueing systems by defining our parameters and the 19 cases examined. We then provide computational results for these cases. All programs were run on a SUN

SPARCstation 10. We used the Visual Numerics C/Math/Library ordinary differential equation (ODE) solver to solve the Chapman-Kolmogorov forward equations for the $M(t)/M/1$ system. This subroutine solves the ODE's using the Runge-Kutta-Verner fifth-order and sixth-order method.

2.1 Parameter Definitions and Cases Examined

We used a sinusoidal Poisson arrival process with parameter (cf. Green, Kolesar and Svoronos [5])

$$\lambda(t) = \overline{\lambda} + A \cdot \sin\left(\frac{2\pi t}{24}\right)$$

The sinusoidal arrival process has a period of 24 hours. The long-run average arrival rate, $\overline{\lambda}$, is also the average arrival rate over the 24 hour period: $\overline{\lambda} = \frac{1}{24}\int_{t=0}^{24} \lambda(t)dt$. The amplitude of the sine wave is A; it is restricted to lie between zero and $\overline{\lambda}$ to ensure that $\lambda(t) \geq 0$. Note that $\lambda(t)$ is a smooth differentiable function with one peak over each period.

The amount by which the peak instantaneous arrival rate $\lambda(t)$ exceeds $\overline{\lambda}$ has an effect on the performance measures of the $M(t)/M/1$ system. Therefore, we define the parameter *Relative Amplitude* (RA) [5] of the arrival process to be

$$RA = \frac{A}{\overline{\lambda}}.$$

Note that $0 \leq RA \leq 1$.

We kept the service rate parameter $\mu(t) = \mu$ constant in the 19 cases examined. In the future, we intend to allow $\mu(t)$ to vary with time, as well.

Define the average utilization over the period to be $\overline{\rho} = \frac{1}{24}\int_{t=0}^{24} \rho(t)dt = \frac{\overline{\lambda}}{\mu}$. The maximum utilization over the period will be denoted by $\rho_{\max} = \max_{0 \leq t \leq 24}\{\rho(t)\}$ [5].

The 19 cases examined are combinations of the following:

- $\mu = 10, 100$, corresponding to low- and high-frequency event systems.
- $\overline{\rho} = 0.5, 0.7, 0.75, 0.9$, ranging from moderate to high average utilization. In 15 of the 19 cases, $\rho_{\max} > 1$.

- $RA = \frac{1}{3}, \frac{2}{3}, 1$, ranging from moderate to maximum nonstationarity.

Note that selecting $\mu, \overline{\rho}$, and RA for a system automatically determines $\overline{\lambda}$ and A. In each of the 19 cases, we recorded the peak in the mean and standard deviation for the number of customers in the system, and the times at which they occurred.

The system started out empty, that is, the probability of 0 customers in the system was 1.0, and was allowed to run until

$$\left|\frac{m(h) - m(h-24)}{m(h)}\right| < 0.02 \tag{10}$$

and

$$\left|\frac{\sigma(h) - \sigma(h-24)}{\sigma(h)}\right| < 0.02 \tag{11}$$

for $h = 24i, 24i+1, 24i+2, \ldots, 24i+23$, for some $i \in \{1, 2, \ldots\}$, i.e., until two consecutive daily profiles of $m(t)$ and $\sigma(t)$ are no more than 2% apart. Note that this implies that once we find i satisfying inequalities (10) and (11), these inequalities also hold for all $j > i, j$ integer, i.e., the system is essentially at equilibrium.

2.2 Results

In Table 1, we provide data which confirm Theorems 1 and 2 and Corollary 1. Note that because the ODE solver takes discrete time steps, $m'(t_m^*)$ is not exactly equal to 0, but is very close. In Table 1, column 7, we show how small the difference $t_m^* - t_2$ is for the cases in which $\rho_{\max} > 1$. For the cases in which $\rho_{\max} \leq 1$, we left the entry in column 7 blank. Note that in the cases in which $\mu = 100$ and $\rho_{\max} > 1$, which correspond to heavily-stressed systems, there was no discernible difference (to two decimal places) between t_m^* and t_2. Although $P_0(t_m^*)$ is positive, it is extremely small in these cases, as can be seen in column 9 of Table 1. Based on the condition in equation (5) and Corollary 1, we expect $t_m^* - t_2$ to be very small in these cases.

In Table 1, we also provide support for the following hypothesis we have not yet proven: In nonstationary single-server systems under equilibrium conditions, $v(t)$ peaks after $m(t)$. The evidence supporting this is the value of $v'(t_m^*)$, listed in the sixth column of Table 1. In all of the 19 $M(t)/M/1$ cases examined, $v'(t_m^*) > 0$, indicating that $v(t)$ is still increasing at the moment

μ	$\bar{\rho}$	RA	t_m^*	$m'(t_m^*)$	$v'(t_m^*)$	$t_m^* - t_2$	$\frac{1}{m^*+1}$	$P_0(t_m^*)$
100	0.5	$\frac{1}{3}$	30.2	0.0038	0.03	N/A	0.33352	0.33346
		$\frac{2}{3}$	30.5	-0.0006	0.11	N/A	0.1695	0.1694
		1	103.5	0.0196	43.21	N/A	0.0494	0.0388
	0.75	$\frac{1}{3}$	104.0	-0.0022	42.49	N/A	0.0429	0.0338
		$\frac{2}{3}$	106.0	0.0014	199.61	0.00	0.0068	0.0000
		1	154.7	0.0355	200.04	0.00	0.0031	0.0000
	0.9	$\frac{1}{3}$	154.7	0.0196	198.50	0.00	0.0071	0.0001
		$\frac{2}{3}$	155.4	0.0061	200.01	0.00	0.0028	0.0000
		1	227.6	-0.0068	199.99	0.00	0.0171	0.0000
10	0.5	$\frac{1}{3}$	55.1	-0.0032	0.14	N/A	0.3424	0.3399
		$\frac{2}{3}$	103.7	-0.0008	1.09	N/A	0.2105	0.1991
		1	152.4	0.0013	4.15	N/A	0.1207	0.0956
	0.7	1	346.3	0.0050	17.59	0.02	0.0340	0.0041
	0.75	$\frac{1}{3}$	153.2	0.0006	4.17	N/A	0.1046	0.0828
		$\frac{2}{3}$	226.1	0.0008	13.66	0.13	0.0472	0.0150
		1	274.7	-0.0023	19.06	0.01	0.0269	0.0013
	0.9	$\frac{1}{3}$	226.9	0.0014	12.72	0.21	0.0431	0.0157
		$\frac{2}{3}$	347.4	-0.0073	19.33	0.01	0.0231	0.0008
		1	347.6	0.0111	20.00	0.00	0.0154	0.0000

Table 1: Numerical results: derivatives of $m(t)$ and $v(t)$ at the time t_m^* when $m(t)$ peaks. Note: N/A = Not Applicable for those cases in which $\rho_{\max} \leq 1$.

that $m(t)$ peaks. Figure 4 shows a plot of $\frac{1}{m^*+1}$. The shaded area corresponds to $\frac{1}{m^*+1} \geq P_0(t_m^*)$. Intuitively, as m^* gets larger, we expect $P_0(t_m^*)$ to get smaller and fall into the shaded region, thus guaranteeing that $v'(t_m^*) > 0$ (see Theorem 2). This did occur in the 19 cases we examined.

Table 2 lists the time lag between the peak in the arrival rate and the mean in column 4, and between the peak in the arrival rate and standard deviation in column 5. In all 19 cases, $\sigma(t)$ peaked later than $m(t)$. This is also true for all the other nonstationary single-server systems we mentioned at the end of Section 1, leading us to believe that this behavior may be typical of general nonstationary queueing systems.

Figures 5A–C are graphical representations of Table 2 for the 19 cases examined. Figures 5A and B plot $\rho_{\max}$ vs. the time lag between the peak in the arrival rate and the times at which m^* and σ^* occur, respectively, where σ^* is the peak in the standard deviation for the number in the system. In Figure 5C, t_{sd}^* is the time at which σ^* occurs. Figure 5C plots $\rho_{\max}$ vs. $t_{sd}^* - t_m^*$ and shows that t_{sd}^* always exceeds t_m^* and increases significantly faster than t_m^* for cases in which $\rho_{\max} > 1$. This

μ	$\bar{\rho}$	RA	lag in m^*	lag in σ^*
100	0.5	$\frac{1}{3}$	0.15	0.22
		$\frac{2}{3}$	0.49	0.72
		1	1.51	2.21
	0.75	$\frac{1}{3}$	2.01	2.94
		$\frac{2}{3}$	4.00	6.88
		1	4.70	9.10
	0.9	$\frac{1}{3}$	4.70	8.12
		$\frac{2}{3}$	5.36	10.89
		1	5.57	12.12
10	0.5	$\frac{1}{3}$	1.10	1.50
		$\frac{2}{3}$	1.70	2.50
		1	2.40	3.40
	0.7	1	4.33	6.80
	0.75	$\frac{1}{3}$	3.20	4.70
		$\frac{2}{3}$	4.13	6.00
		1	4.71	7.70
	0.9	$\frac{1}{3}$	4.91	7.69
		$\frac{2}{3}$	5.37	9.10
		1	5.60	10.20

Table 2: Time lags in hours between the peak in arrival rate and peaks in the mean and standard deviation of number of customers in system.

observation again suggests that $m(t)$ behaves differently from $v(t)$ and $\sigma(t)$, in significant ways.

3 Summary

This paper begins to explore the time lag between the peak in the arrival rate and the times at which the peaks m^*, v^* and σ^* for the mean, variance, and standard deviation of the number of customers in the system occur. Our overall conjecture is that for nonstationary queueing systems in equilibrium, with smooth, periodic arrival and service rates, the mean peak m^* will occur later than the arrival rate peak and the variance peak v^* will occur later than the mean peak, under fairly general conditions.

We demonstrated analytically that if $\rho_{\max} > 1$ and $P_0(0) > 0$, then the mean peak m^* will occur later than the arrival rate peak; in fact it will occur strictly later than the time t_2 at which $\rho(t)$ passes one on its way down. Computational results for $M(t)/M(t)/1$ and other nonstationary single-server queueing systems confirm our analytical results and support our overall conjecture.

Acknowledgment

The authors thank Professor Amedeo Odoni of the Massachusetts Institute of Technology for his valuable comments.

References

[1] A.B. Clarke. A Waiting Line Process of the Markov Type. *Annals of Mathematical Statistics*, 27:452–459, 1956.

[2] S.G. Eick, W.A. Massey, and W. Whitt. $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Management Science*, 39(2):241–252, 1993.

[3] S.G. Eick, W.A. Massey, and W. Whitt. Physics of the $M_t/G/\infty$ Queue. *Operations Research*, 41(4):731–742, 1993.

[4] L.V. Green and P.J. Kolesar. On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues. Preliminary Version of a Paper prepared by the authors at the Graduate School of Business, Columbia University, New York, NY, 1993.

[5] L.V. Green, P.J. Kolesar, and A. Svoronos. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research*, 39(3):502–511, 1991.

[6] A.Y. Khintchine. *Mathematical Models in the Theory of Queueing.* Charles Griffin and Co., London, 1960. (Translation of 1955 Russian book).

[7] D.G.. Luenberger. *Introduction to Dynamic Systems. Theory, Models, and Applications.* John Wiley & Sons, 1979.

[8] G.F. Newell. Queues With Time-Dependent Arrival Rates I – III. *Journal of Applied Probability*, 5:436–606, 1968.

[9] Odoni, A.R. and Roth, E. An Empirical Investigation of the Transient Behavior of Stationary Queueing Systems. *Operations Research*, 31(3):432–455, 1983.

[10] C. Palm. Intensity Variations in Telephone Traffic. *North-Holland*, 1988. (Translation of 1943 article in *Ericsson Technics*, 44, 1–189).

[11] M.H. Rothkopf and S.S. Oren. A Closure Approximation for the Nonstationary $M/M/s$ Queue. *Management Science*, 25(6):522–534, 1979.

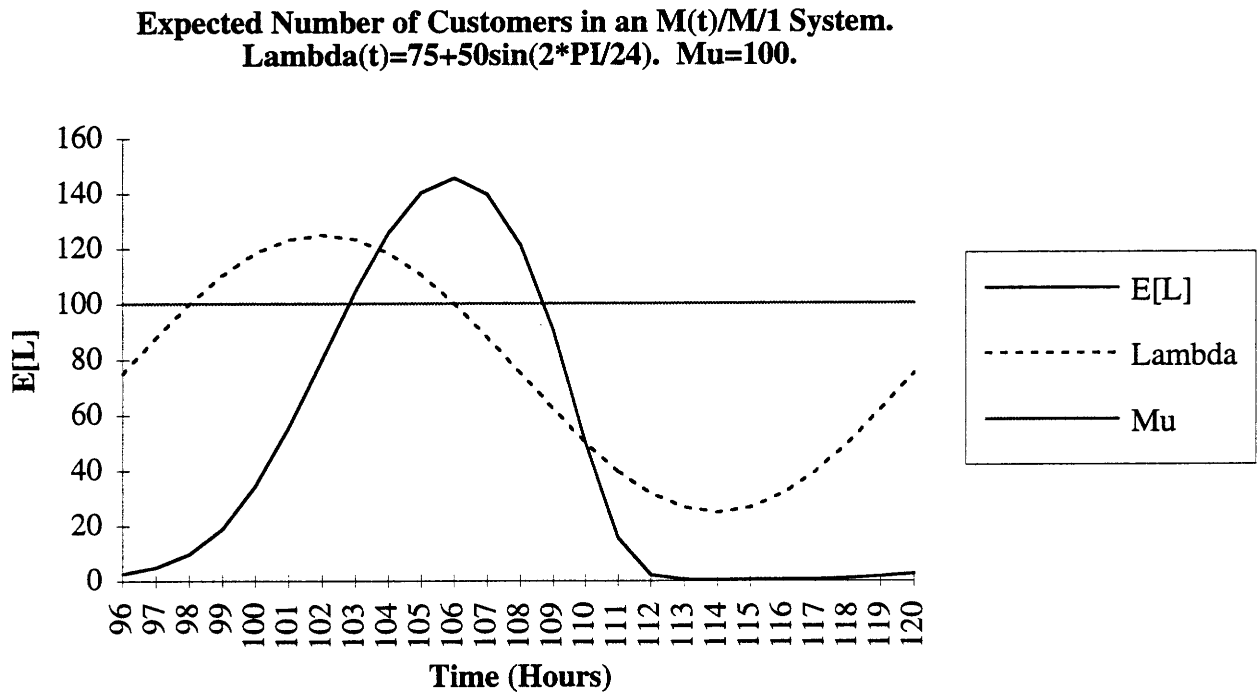

Figure 1

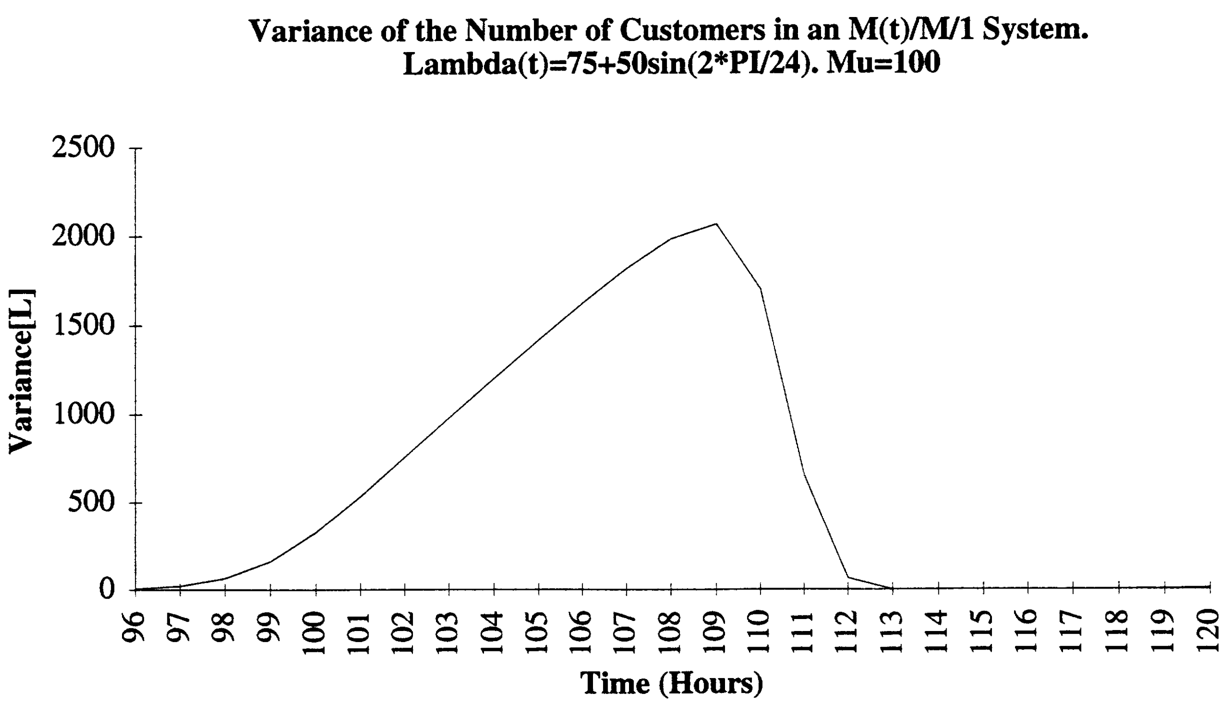

Figure 2

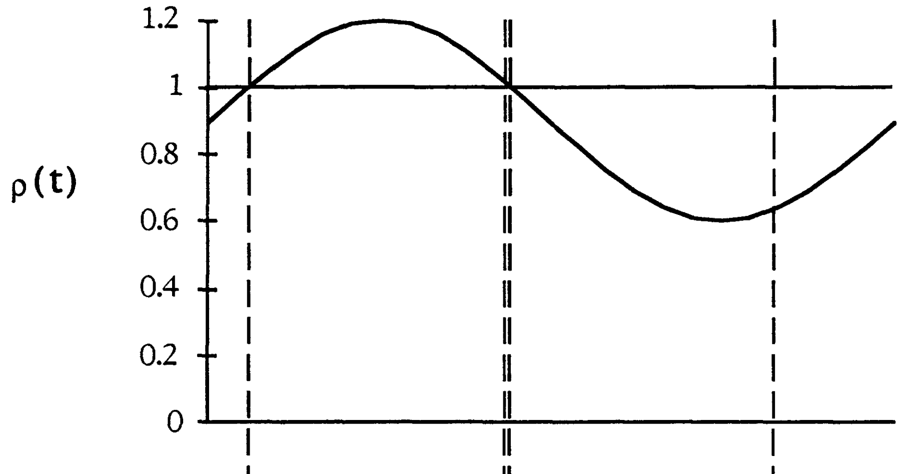

Figure 3A

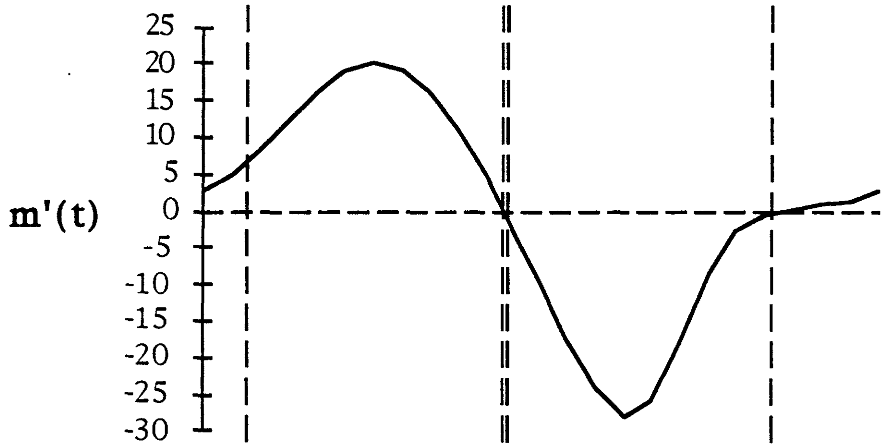

Figure 3B

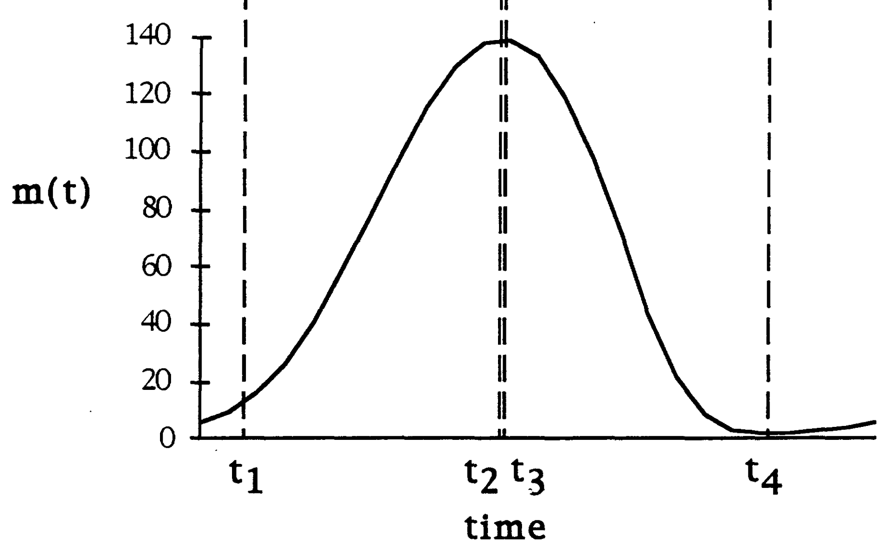

Figure 3C

Figure 3: $\rho(t)$, $m'(t)$, $m(t)$ for $\lambda(t)=90+30\sin(2\pi t/24)$, $\mu=100$

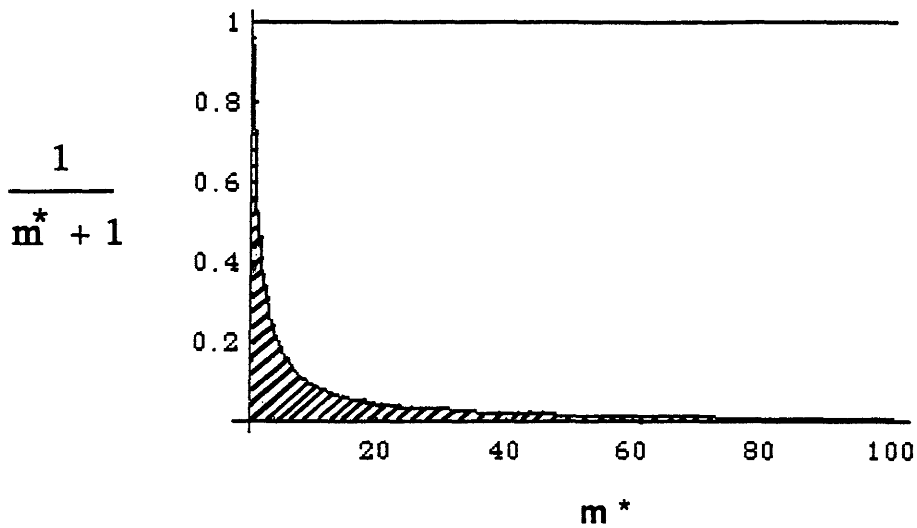

Figure 4

Time Lag Between Peak of Mean Number in an M(t)/M(t)/1 System and Peak in Arrival Rate

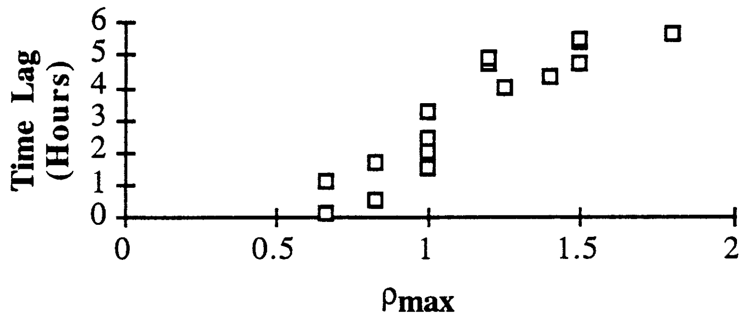

Figure 5A

Time Lag Between Peak of Standard Deviation for the Number in an M(t)/M(t)/1 System and Peak in Arrival Rate

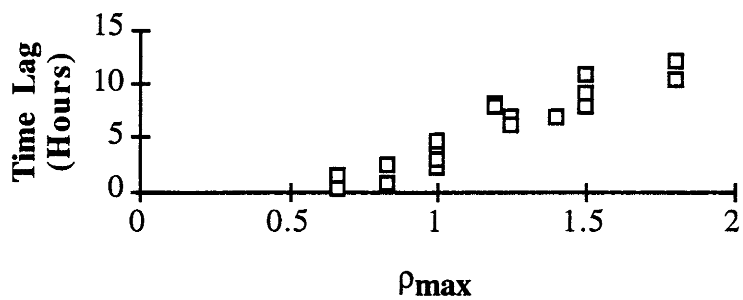

Figure 5B

Difference Between Standard Deviation and Mean Time Lags for Corresponding M(t)/M(t)/1 Systems

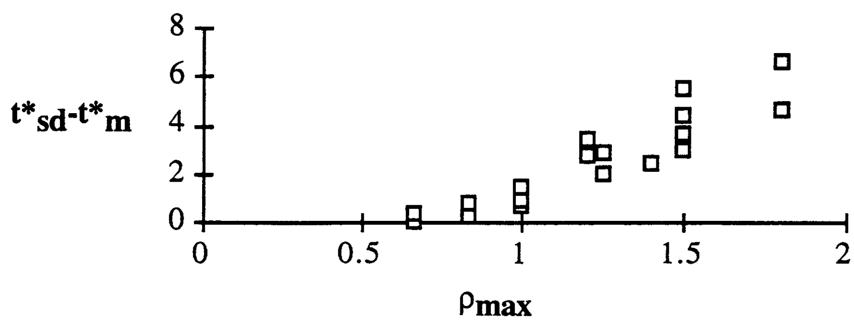

Figure 5C

Figure 5: Time Lags Between Peaks in Mean and Standard Deviation for the Number in an M(t)/M(t)/1 System and the Peak in the Arrival Rate