A Note on Approximating Peak Congestion in $M_t/G/\infty$ Queues with Sinusoidal Arrivals

Linda V. Green • Peter J. Kolesar

Columbia Business School, Columbia University, New York, New York 10027-6902

We study the $M_t/G/\infty$ queue where customers arrive according to a sinusoidal function $\lambda_t = \lambda + A \sin(2\pi t / T)$ and the service rate is $\mu$. We show that the expected number of customers in the system during peak congestion can be closely approximated by $(\lambda + A) / \mu$ for service distributions with coefficient of variation between 0 and 1. Motivated by a result derived by Eick, Massey, and Whitt that the time lag of the peak congestion from the peak of the customer arrivals is $1/2\mu$ for models with deterministic service times, we show that the time lag for exponential service times is closely approximated by $1/\mu$. Based on a cycle length of 24 hours and regardless of the values of other system parameters, these approximations are of the order of 1% accuracy for $\mu = 1$, and the accuracy increases rapidly with increasing $\mu$.

(Queues; Nonstationarity; Approximations)

1. Introduction

Customer arrivals at many service facilities follow periodic patterns. Examples include: aircraft landings (Koopman 1972), post-office mail receipts (Oliver and Samuel 1962), emergency calls for police assistance (Larson 1972; Green and Kolesar 1984, 1989), fire alarms (Walker et al. 1979), telephone traffic (Segal 1974), transactions at automatic teller machines (Kolesar 1984), telemarketing (Andrews and Parsons 1989), and automobiles at bridge toll plazas (Edie 1954). Unfortunately, most queuing models that include time-varying arrivals are too complex to solve analytically. This has motivated our interest in simple approximate solutions for queues with cyclic arrivals (Green and Kolesar 1991, 1995; Green et al. 1991).

Eick, Massey, and Whitt (1993a) (EMW), building on earlier work of Palm and Khintchine, have derived closed-form expressions for the time-varying expected number of customers in an $M_t/G/\infty$ queue with sinusoidal arrivals. Using these exact results, we show that the magnitude and the time lag of peak congestion in an $M_t/G/\infty$ queue may, in many cases, be approximated very accurately by simple "back of the envelope" expressions. This is important for several reasons. First, the approximations provide obvious computational convenience for practitioners. This is particularly useful for cases other than exponential or deterministic service time distributions where EMW's exact results may be quite difficult for practitioners to evaluate. (See section 2 for the case of Erlang 2 service times.) Moreover, as our numerical results indicate, the approximations we propose are so accurate that in many practical situations there would be no advantage in having the exact values. Second, these approximations may also be important for use in other models where a simple structure is needed, as, for example, in resource allocation or optimization (Harel 1988). Indeed, a paper by a team that includes Whitt and Massey (Jennings et al. 1996) use their own more complex approximations for the exact $M_t/G/\infty$ results in this way. Third, and perhaps most importantly, our own experience has shown that such simple expressions are very useful for providing insight to practitioners in evaluating alternative operating designs (Kolesar and Green 1998).

From the practical point of view, the $M_t/G/\infty$ queue is, itself, almost always an approximation—real systems never have an infinite number of servers. An infinite server model may, however, give a reasonable
description of congestion in actual operations if the number of servers available is large relative to the demand. Consequently, using an approximate solution for the $M_t/G/\infty$ queue can be quite natural in an applied context. (For further justification and discussion of the use of infinite server models in finite server environments see Whitt (1991) and Jennings et al. (1994).)

2. The Model and Main Result
We study a queuing system in which customers arrive according to a sinusoidal Poisson process with arrival rate given by

$$\lambda_t = \lambda + A \sin(2\pi t/T).$$  \hspace{1cm} (1)

Here, $\lambda$ is the overall mean arrival rate, $A$ is the amplitude of the arrival function and $T$ is its period. (Motivated by the many practical cases in which a daily cycle is evident, we take $T = 24$ hours as standard.) We denote the relative amplitude by $RA = A/\lambda$. An infinite number of servers are available who commence serving each customer immediately upon its arrival. Service times are independent identically distributed random variables with a general service time distribution with cumulative distribution $G(t)$ and mean $1/\mu$. We restrict attention to service time distributions with coefficients of variation between 0 and 1. This range—from constant to exponentially distributed service times—covers a very broad class of applications and is also convenient to explore computationally.

Let $L_t$ denote the periodic stationary expected number of customers in the system at time $t$. Motivated by the many applications in which system capacity is fixed over the period and therefore must be sufficient to provide good performance when demand is highest, we focus on the "peak-epoch" and "peak-hour" values of $L_t$. We denote the peak-epoch, or instantaneous maximum value of $L_t$ by $L_{peak}$ which is

$$L_{peak} = \max_{0 \leq t \leq T} L_t.$$  \hspace{1cm} (2)

The average value of $L_t$ during the peak hour of the day, which we denote by $L_{peak-hour}$ is

$$L_{peak-hour} = \max_{0 \leq t \leq T} \int_{t}^{t+1} L_v \, dv.$$  \hspace{1cm} (3)

Transforming the results of EMW which are stated in terms of systems with $\mu = 1$ to our frame of reference which is systems with $T = 24$, we get for the case of exponential service times (EMW Equation 18) that $L_{peak}$ is

$$L_{peak} = \frac{\lambda}{\mu} + A \left( \frac{1}{\sqrt{1 + (\gamma/\mu)^2}} \right),$$  \hspace{1cm} (4)

where the arrival frequency $\gamma = 2\pi / 24$. With sinusoidal arrivals on a 24 hour period, customer arrivals peak at $t = 6$ hours. However, the peak congestion occurs at a time lag of $t_{lag}$ hours after the peak of the customer arrival rate where (EMW eq. 16)

$$t_{lag} = \frac{\cot^{-1}(\mu/\gamma)}{\gamma}.$$  \hspace{1cm} (5)

For the model with deterministic service times, we get (EMW eq. 27)

$$L_{peak} = \frac{\lambda}{\mu} + A \left( \frac{2 \sin(\gamma/2\mu)}{\gamma} \right),$$  \hspace{1cm} (6)

while the peak congestion occurs at a lag of (EMW eq. 26)

$$t_{lag} = 1/2\mu.$$  \hspace{1cm} (7)

We also analyze an Erlang 2 service time model which has a coefficient of variation of $1/\sqrt{2}$ as an "intermediary" case between exponential service times with coefficient of variation of 1 and deterministic service times with coefficient of variation of 0. The Erlang 2 model with mean $1/\mu$ has density function

$$f(x) = (2\mu)^2 x e^{-2\mu x}, \quad x \geq 0.$$  \hspace{1cm} (8)

Using eq. 12 in EMW we derive

$$L_{peak} = \frac{\lambda}{\mu} + A \sqrt{D},$$  \hspace{1cm} (9)

where $D$ is

$$D = \left( \frac{4\mu^2}{2(4\mu^2 + \gamma^2)} + \frac{4\mu^2(4\mu^2 - \gamma^2)}{2(4\mu^2 + \gamma^2)^2} \right)^2$$

$$+ \left( \frac{2\mu\gamma}{2(4\mu^2 + \gamma^2)} + \frac{8\mu^3\gamma}{(4\mu^2 + \gamma^2)^2} \right)^2.$$  \hspace{1cm} (10)

We define the Simple Peak Epoch Approximation (SPEA) for the expected number of customers in the system at the epoch of peak congestion as
\[ L_{\text{SPEA}} = \frac{\lambda + A}{\mu}. \quad (11) \]

We will use equation (11) to approximate equations (4), (6), and (9).

Noting that the arrival rate of a sinusoid on a 24 hour period peaks at \( t = 6 \), we define
\[ \lambda_{\text{peak-hour}} = \int_{5.5}^{6.5} \lambda_t \, dt, \quad (12) \]

and, using \( \lambda_{\text{peak-hour}} \), the Simple Peak Hour Approximation (SPHA) for expected number of customers in the system during the hour of peak congestion is defined as
\[ L_{\text{SPHA}} = \frac{\lambda_{\text{peak-hour}}}{\mu}. \quad (13) \]

Note that \( L_{\text{SPEA}} \) is the long run mean number in the system for a stationary \( M/G/\infty \) queue operating steadily with the peak epoch arrival rate. Similarly, \( L_{\text{SPHA}} \) is the mean number in the system for a stationary \( M/G/\infty \) queue operating steadily with an arrival rate that is the average over the peak hour. In Green and Kolesar (1995), we showed that such approximations were accurate in finite server systems for large service rate values, i.e., where \( \mu > 20 \). We show here that for infinite server systems, the approximations are good for much smaller service rates as well, i.e., when \( \mu \approx 1 \).

Our approximations are based on a concept that we have named the “Pointwise Stationary Approximation,” or PSA for short (Green and Kolesar 1991). The PSA idea is to approximate the behavior of the actual cyclic queue at each point in time by a stationary version of the system model that uses the parameter values that apply at that instant. Therefore, the PSA for the expected number in the system at the peak epoch in the \( M_t/G/\infty \) queue is just \( L_{\text{SPEA}} \). This PSA approach assumes that the process is, in effect, dynamically in equilibrium at its time-dependent parameter values. Intuitively, the PSA will be a better approximation of systems where customers clear quickly and therefore act relatively independently at each time epoch, i.e., when the service rate is high. Indeed, Whitt (1991) proved that for finite server systems the PSA is asymptotically correct as the arrival and service rates increase. Similarly, infinite server models should be better approximated by the PSA approach than finite server models since each customer’s residence in the system is independent of that of other customers and therefore of the past.

3. Approximating Peak-Epoch Congestion

Table 1 and Table 2 show relative errors for \( L_{\text{SPEA}} \) for the exponential and deterministic models, respectively. Relative error is defined by
\[ \text{RelErr} = \frac{L_{\text{SPEA}} - L_{\text{peak}}}{L_{\text{peak}}} \times 100. \quad (14) \]

Consistent with our earlier findings for approximating peak-epoch behavior of finite server systems, two factors determine the accuracy of the approximation. The first and more important is the service rate. The tables show that relative errors decrease dramatically as the service rate increases. This is consistent with EMW’s result that \( L_t \) converges uniformly in \( t \) to the PSA as the service rate increases. The second factor is the relative amplitude. Not surprisingly, relative errors decrease with decreasing relative amplitudes, although somewhat more slowly than with the service rate. These conclusions are true for both service time models. Note that, except for an anomaly that occurs for deterministic service times with \( \mu = 1/24 \) when all customers stay in the system for exactly 24 hours, the deterministic service time model is better approximated than the exponential service time model.

Note that the parameter ranges in the tables are broad: Values of \( \mu \) range from a low of 1/24, where a service takes 24 hours to complete on average, to a high of 512, where a service takes about 6 seconds on average. Relative amplitudes range from 0.1 to 1.0 in steps of 0.1, and there is also a relative amplitude of 0.01. For the exponential model with \( \mu = 1.0 \), which is an average service time of an hour, the relative errors are of the order of a percent for the higher relative amplitudes (Table 1). For the deterministic model with \( \mu = 1.0 \) the relative errors are of the order of a tenth of a percent for the higher relative amplitudes (Table 2).

Our results for the Erlang 2 case (not tabulated), as expected, fall between the exponential and deterministic values. For example, when \( \mu = 1.0 \) and relative amplitude = 0.5 the Erlang 2 has a relative error of
0.494% while the relative errors for the exponential and deterministic cases are 1.099% and 0.107%, respectively. (The relative error for the Erlang 2 model is approximately 44% of the relative error for the exponential case whenever $\mu > 1.0$, regardless of the relative amplitude.)
Our concept of a "good" approximation can now be made explicit. By good, we mean relative errors on the order of a percent and one can see from the tables that all models with $\mu > 1.333$ satisfy this definition.

4. Peak-Hour Approximations
Having demonstrated that the PSA approach works well for peak epochs, we now examine its accuracy for peak hours. Since, as demonstrated in section 3, the PSA approach is worst for the exponential service time model it will be sufficient to demonstrate the utility of $L_{\text{SPHA}}$ by an analysis of the exponential service time case. In the following we make use of the exact solution for $L_t$ in the exponential case (EMW eq. 15.) Since $L_t$ is symmetric about its peak epoch, $L_{\text{peak-hour}}$, the peak-hour average number of customers in the system, is obtained by

$$
L_{\text{peak-hour}} = \int_{\alpha}^{\alpha + 1} L_t \, dt = \int_{\alpha}^{\alpha + 1} \lambda / \mu \\
+ A / \mu \left[ \frac{\sin(\gamma t) - \gamma / \mu \cos(\gamma t)}{1 + (\gamma / \mu)^2} \right] dt \\
= \frac{\lambda}{\mu} + \frac{A}{\mu} \frac{1}{1 + (\gamma / \mu)^2} \\
\times \left[ \frac{\sin(\gamma \alpha) - \sin(\gamma (\alpha + 1))}{\gamma} \\
+ \frac{\cos(\gamma \alpha) - \cos(\gamma (\alpha + 1))}{\gamma} \right],
$$

(15)

where

$$
\alpha = 6 + \frac{\cot^{-1}(\mu / \gamma)}{\gamma} - 1/2
$$

(16)
is the beginning of the peak hour.

From equations (12) and (13) we have

$$
L_{\text{SPHA}} = \frac{\lambda}{\mu} + A / \mu \gamma \left[ \cos \left( \frac{11\pi}{24} \right) - \cos \left( \frac{13\pi}{24} \right) \right] \\
= \frac{\lambda}{\mu} + 0.997 A.
$$

(17)

We note that $L_{\text{SPEA}}$ and $L_{\text{SPHA}}$ are nearly identical because of the flatness of the sine function in the hour surrounding the peak. Table 3 contains our comparisons of $L_{\text{SPHA}}$ to the exact peak hour average expected number in the system for the case of exponential service times with $RA = 1$ (a worst case) and the same range of service rates as in Table 1. Table 3 shows that $L_{\text{SPHA}}$ is a very good approximation for the average peak hour expected number in the system whenever $\mu$ is greater than one. Furthermore, because of the flatness of the exact solution in the hour surrounding its peak, there is little difference between the peak-epoch and peak-hour values of either the approximate or exact solutions. Thus, for all practical purposes, the simpler $L_{\text{SPEA}} = (\lambda + A) / \mu$ can be used to approximate either the peak-epoch or peak-hour congestion.

5. Time-Lags of the Peak-Epoch Congestion
In queuing systems where customers arrive in a periodic pattern, congestion builds to the peak epoch which occurs some time after the peak in the arrival rate function, itself. In the $M_i/G/\infty$ system, with sinusoidal arrivals, the time lag from the peak of the arrival rate to the epoch of peak congestion depends only on the service rate—decreasing rapidly in $\mu$ for both the exponential and deterministic service time models. This is shown by equations (5) and (7), and also in Table 4. The time lags are generally less for the deterministic service time model. As is demonstrated in Table 4, for $\mu > 1$ one can approximate the time lag for exponential service quite closely by twice the time lag for deterministic service, that is, simply by $1 / \mu$. The error is about 2% at $\mu = 1$ and decreases rapidly as $\mu$ increases.

6. Closing Remarks
We feel confident that our conclusions that the $L_{\text{SPEA}}$ and $L_{\text{SPHA}}$ are excellent approximations apply to a very broad range of practical situations. Our own experience has been with applications of queuing models to telemarketing, police patrol, fire fighting, hospitals, copy machine repairs, and automatic teller machine operations. In all these situations service times were close to exponential, relative amplitudes were close to one, cycle lengths were 24 hours, and service rates were in the range of say 0.5 (copy machine repairs) to 100 (automatic teller machine transactions). The arrival rate
### Table 3  Comparing SPEA and SPHA to Exact Peak Epoch and Peak Hour Congestion

<table>
<thead>
<tr>
<th>Mu</th>
<th>Exact Peak Epoch Congestion</th>
<th>Peak Hour Congestion</th>
<th>Actual Peak Point-Hour Congestion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPEA</td>
<td>RelErr</td>
<td>Exact</td>
</tr>
<tr>
<td>0.042</td>
<td>111.089</td>
<td>192.000</td>
<td>72.834</td>
</tr>
<tr>
<td>0.083</td>
<td>62.559</td>
<td>96.000</td>
<td>53.455</td>
</tr>
<tr>
<td>0.167</td>
<td>36.889</td>
<td>48.000</td>
<td>30.121</td>
</tr>
<tr>
<td>0.333</td>
<td>21.437</td>
<td>24.000</td>
<td>11.955</td>
</tr>
<tr>
<td>0.500</td>
<td>15.087</td>
<td>16.000</td>
<td>6.050</td>
</tr>
<tr>
<td>0.667</td>
<td>11.585</td>
<td>12.000</td>
<td>3.584</td>
</tr>
<tr>
<td>1</td>
<td>7.870</td>
<td>8.000</td>
<td>1.657</td>
</tr>
<tr>
<td>1.333</td>
<td>5.944</td>
<td>6.000</td>
<td>0.946</td>
</tr>
<tr>
<td>2</td>
<td>3.983</td>
<td>4.000</td>
<td>0.425</td>
</tr>
<tr>
<td>4</td>
<td>1.998</td>
<td>2.000</td>
<td>0.107</td>
</tr>
<tr>
<td>8</td>
<td>1.000</td>
<td>1.000</td>
<td>0.027</td>
</tr>
<tr>
<td>16</td>
<td>0.500</td>
<td>0.500</td>
<td>0.007</td>
</tr>
<tr>
<td>32</td>
<td>0.250</td>
<td>0.250</td>
<td>0.002</td>
</tr>
<tr>
<td>64</td>
<td>0.125</td>
<td>0.125</td>
<td>0.000</td>
</tr>
<tr>
<td>128</td>
<td>0.062</td>
<td>0.063</td>
<td>0.000</td>
</tr>
<tr>
<td>256</td>
<td>0.031</td>
<td>0.031</td>
<td>0.000</td>
</tr>
<tr>
<td>512</td>
<td>0.016</td>
<td>0.016</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### Table 4  Time Lags for Peak Congestion (Hours)

<table>
<thead>
<tr>
<th>Mu</th>
<th>Deterministic Time Lags</th>
<th>Exponential Time Lags</th>
<th>Approximation for Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1/μ)</td>
<td>Percent Error</td>
<td></td>
</tr>
<tr>
<td>0.042</td>
<td>12.000</td>
<td>5.3971</td>
<td>24.0000</td>
</tr>
<tr>
<td>0.083</td>
<td>6.0000</td>
<td>4.8329</td>
<td>12.0000</td>
</tr>
<tr>
<td>0.167</td>
<td>3.0000</td>
<td>3.8346</td>
<td>6.0000</td>
</tr>
<tr>
<td>0.333</td>
<td>1.5000</td>
<td>3.8346</td>
<td>3.0000</td>
</tr>
<tr>
<td>0.500</td>
<td>1.0000</td>
<td>1.8424</td>
<td>2.0000</td>
</tr>
<tr>
<td>0.667</td>
<td>0.7500</td>
<td>1.4293</td>
<td>1.5000</td>
</tr>
<tr>
<td>1</td>
<td>0.5000</td>
<td>0.9780</td>
<td>1.0000</td>
</tr>
<tr>
<td>1.333</td>
<td>0.3750</td>
<td>0.7406</td>
<td>0.7500</td>
</tr>
<tr>
<td>2</td>
<td>0.2500</td>
<td>0.4972</td>
<td>0.5000</td>
</tr>
<tr>
<td>4</td>
<td>0.1250</td>
<td>0.2496</td>
<td>0.5000</td>
</tr>
<tr>
<td>8</td>
<td>0.0625</td>
<td>0.1250</td>
<td>0.1250</td>
</tr>
<tr>
<td>16</td>
<td>0.0313</td>
<td>0.0625</td>
<td>0.0625</td>
</tr>
<tr>
<td>32</td>
<td>0.0156</td>
<td>0.0312</td>
<td>0.0313</td>
</tr>
<tr>
<td>64</td>
<td>0.0078</td>
<td>0.0156</td>
<td>0.0156</td>
</tr>
<tr>
<td>128</td>
<td>0.0039</td>
<td>0.0078</td>
<td>0.0078</td>
</tr>
<tr>
<td>256</td>
<td>0.0020</td>
<td>0.0039</td>
<td>0.0039</td>
</tr>
<tr>
<td>512</td>
<td>0.0010</td>
<td>0.0020</td>
<td>0.0020</td>
</tr>
</tbody>
</table>
process in each instance was smooth enough and flat enough at the peak that our conclusions about the accuracy of the approximation should apply. Furthermore, in many of these applications there was an operating policy to keep customer delays close to zero—a situation consistent with the use of an infinite server model.

Our numerical results for the three service time distributions—deterministic, gamma (2) and exponential—illustrate that the difference between the approximate and exact solution increases as the coefficient of variation increases. Thus, the reader is advised that, for coefficients of variation greater than 1, the $L_{\text{SPEA}}$ and $L_{\text{SPHA}}$ will be worse than Tables 1, 2 and 3 indicate.

References


Accepted by G. Bitran; received November 22, 1994. This paper has been with the authors for 7 months for 2 revisions.