Guidelines for the use of infinite source queueing models in the analysis of computer system performance

by J. P. BUZEN
Harvard University
Cambridge, Massachusetts
and
Honeywell Information Systems
Waltham, Massachusetts
and
P. S. GOLDBERG
Case Western Reserve University
Cleveland, Ohio

INTRODUCTION

Mathematical models based on queueing theory are widely used in the analysis of computer system performance. As in the case of other engineering disciplines, these models never correspond exactly to the real systems they are intended to represent. However, if the associated error terms are sufficiently small the models can still serve as valuable tools for estimating performance levels in specific cases and for studying the factors which influence overall system behavior. In this paper we examine some error terms which arise when the familiar M/G/1 queueing model is used to predict expected response times and queue lengths in systems which contain only a finite number of sources.

To illustrate the essence of the problem suppose it is necessary to determine the mean response time (i.e., waiting time + service time) for a particular component or subsystem that is being considered for use within a larger system. Queueing theory provides a set of formulas for estimating this quantity once the service time distribution and the arrival process are known. Figure 1 illustrates the M/G/1 queueing model which is commonly used in such cases. To apply this model it is necessary to determine the mean rate at which requests will arrive at the subsystem and also the distribution of the subsystem's service time. The M/G/1 model then assumes the following two conditions are true:

A. Service times are given by independent identically distributed random variables having the empirically observed distribution.

B. Requests arrive according to a Poisson process which has the empirically observed arrival rate.

Given A and B it is possible to demonstrate that the expected response time of the subsystem is

$$R = \frac{1 + \rho(1 + c)}{\mu} \cdot \frac{1}{2\rho(1 - \rho)}$$

where

$\mu$ = the reciprocal of the mean service time

$\rho$ = the mean service time to mean inter-arrival time ratio (the load on the subsystem)

$c$ = the variance to mean squared ratio for the service time distribution

FINITE AND INFINITE SOURCE MODELS

Although equation 1 is relatively easy to apply, it is only of interest if it yields values of $R$ which are reasonably close to the response times one would actually observe in real systems. This depends in part on the extent to which assumptions A and B are satisfied in real systems. Assumption A should be reasonably accurate in many cases since individual service times are typically generated by independent processes and thus tend to be serially uncorrelated. However, assumption B is often suspect because of the finite source nature of the arrival process.

This point is illustrated in Figure 2. Assume that this figure presents an exact description of some existing system. The system contains $N$ customers (programs), each circulating counterclockwise through the loop and undergoing successive periods of independent activity, queueing delay and subsystem service. In the case where service times are given by arbitrarily distributed random variables and independent activity times by exponentially distributed random variables, the system depicted in Figure 2 is known as a finite source M/G/1/N queueing system.

One of the significant aspects of the M/G/1/N model is that the rate of arrival of new customers to the subsystem steadily decreases as queue length increases, and in fact
drops to zero when all \( N \) customers are waiting for or receiving service. However, the arrival rate will not vary greatly if \( N \) is fairly large and queue length remains relatively short most of the time. In such cases it is reasonable to assume, as an approximation, that the mean arrival rate remains constant regardless of the actual length of the queue.

This is precisely the approximation that is made in the M/G/1 model. It is further assumed in this model that customer service times are given by arbitrarily distributed random variables and that inter-arrival times are exponentially distributed. Thus the M/G/1 model can be regarded as the limiting case of the M/G/1/N model in which \( N \) approaches infinity. Of course it is necessary to assume that mean independent activity time also approaches infinity as part of the limiting process in order to prevent the subsystem from becoming overloaded as \( N \) grows large.

**INFINITE SOURCE ADVANTAGES**

Although it is possible to explicitly solve the "finite source" M/G/1/N model and derive formulas for quantities such as expected response time (see equation 2 in Appendix), there are several reasons why the "infinite source" M/G/1 model is often preferred. The primary reason is mathematical simplicity. For example, in the basic case of first come first served scheduling, the formula for expected response time is significantly easier to evaluate in the M/G/1 case (equation 1) than it is in the M/G/1/N case (equation 2). This greater complexity is present in almost all finite source formulas.

Another consequence of the relative mathematical simplicity of the M/G/1 model is the fact that infinite source solutions exist for certain problems which have not yet been solved in the finite source case. Schrage's analysis of the infinite level foreground/background scheduling discipline is a case in point. 4

A somewhat different reason for preferring the infinite source model is that it is sometimes easier to measure the mean arrival rate at a subsystem (a local measurement) than it is to measure the independent activity times at remote sites (e.g., user terminals in a time-sharing system).

A final advantage is derived from the fact that the parameter \( N \) does not explicitly enter into the infinite source model. Thus the exact value of \( N \) need not be considered when applying the model to a specific system. This is particularly useful in cases where the actual value of \( N \) fluctuates as a result of external factors.

**INFINITE SOURCE ERROR TERMS**

Given the advantages of infinite source models, it is natural to consider the question of how accurate these models are in predicting performance levels of real systems which, in fact, almost always contain only a finite number of sources. To study this question, suppose that an actual system which conforms exactly to the M/G/1/N description is being analyzed with the aid of an infinite source M/G/1 model. Since both models can be solved explicitly it is possible to compute both the "exact" mean response time \( R_N \) (equation 2) and the "approximate" mean response time \( R \) (equation 1). Assuming the M/G/1 model has been properly calibrated to the M/G/1/N system so that the service time distributions and mean arrival rates are identical, it is then possible to obtain the error term \( (R - R_N)/R_N \).

This error is directly attributable to the infinite source assumption. As an additional point, it can be shown that the relative error in expected response time is identical to

**TABLE I—Relative Errors for Constant Service Times**

<table>
<thead>
<tr>
<th>( N )</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.059</td>
<td>.127</td>
<td>.215</td>
<td>.332</td>
<td>.508</td>
<td>.742</td>
<td>1.172</td>
<td>1.723</td>
<td>2.309</td>
</tr>
<tr>
<td>2</td>
<td>.030</td>
<td>.068</td>
<td>.117</td>
<td>.188</td>
<td>.296</td>
<td>.456</td>
<td>.725</td>
<td>1.290</td>
<td>2.752</td>
</tr>
<tr>
<td>3</td>
<td>.014</td>
<td>.039</td>
<td>.066</td>
<td>.105</td>
<td>.171</td>
<td>.276</td>
<td>.458</td>
<td>.819</td>
<td>1.865</td>
</tr>
<tr>
<td>4</td>
<td>.011</td>
<td>.028</td>
<td>.053</td>
<td>.090</td>
<td>.142</td>
<td>.238</td>
<td>.389</td>
<td>.785</td>
<td>1.590</td>
</tr>
<tr>
<td>5</td>
<td>.010</td>
<td>.024</td>
<td>.045</td>
<td>.076</td>
<td>.124</td>
<td>.199</td>
<td>.329</td>
<td>.597</td>
<td>1.450</td>
</tr>
<tr>
<td>6</td>
<td>.009</td>
<td>.021</td>
<td>.038</td>
<td>.065</td>
<td>.108</td>
<td>.173</td>
<td>.297</td>
<td>.569</td>
<td>1.293</td>
</tr>
<tr>
<td>7</td>
<td>.008</td>
<td>.018</td>
<td>.034</td>
<td>.059</td>
<td>.095</td>
<td>.162</td>
<td>.267</td>
<td>.502</td>
<td>1.226</td>
</tr>
<tr>
<td>8</td>
<td>.007</td>
<td>.016</td>
<td>.030</td>
<td>.051</td>
<td>.088</td>
<td>.141</td>
<td>.248</td>
<td>.466</td>
<td>1.132</td>
</tr>
<tr>
<td>9</td>
<td>.006</td>
<td>.016</td>
<td>.030</td>
<td>.051</td>
<td>.088</td>
<td>.141</td>
<td>.248</td>
<td>.466</td>
<td>1.132</td>
</tr>
<tr>
<td>10</td>
<td>.006</td>
<td>.015</td>
<td>.028</td>
<td>.047</td>
<td>.078</td>
<td>.133</td>
<td>.230</td>
<td>.430</td>
<td>1.021</td>
</tr>
<tr>
<td>20</td>
<td>.003</td>
<td>.008</td>
<td>.014</td>
<td>.025</td>
<td>.042</td>
<td>.074</td>
<td>.134</td>
<td>.202</td>
<td>.716</td>
</tr>
<tr>
<td>30</td>
<td>.002</td>
<td>.005</td>
<td>.009</td>
<td>.017</td>
<td>.028</td>
<td>.052</td>
<td>.094</td>
<td>.202</td>
<td>.587</td>
</tr>
<tr>
<td>40</td>
<td>.001</td>
<td>.004</td>
<td>.007</td>
<td>.013</td>
<td>.022</td>
<td>.039</td>
<td>.073</td>
<td>.156</td>
<td>.436</td>
</tr>
<tr>
<td>50</td>
<td>.001</td>
<td>.003</td>
<td>.006</td>
<td>.010</td>
<td>.018</td>
<td>.032</td>
<td>.062</td>
<td>.139</td>
<td>.392</td>
</tr>
<tr>
<td>60</td>
<td>.001</td>
<td>.003</td>
<td>.005</td>
<td>.008</td>
<td>.015</td>
<td>.026</td>
<td>.054</td>
<td>.116</td>
<td>.349</td>
</tr>
<tr>
<td>70</td>
<td>.001</td>
<td>.002</td>
<td>.004</td>
<td>.007</td>
<td>.013</td>
<td>.025</td>
<td>.045</td>
<td>.098</td>
<td>.339</td>
</tr>
<tr>
<td>80</td>
<td>.001</td>
<td>.002</td>
<td>.003</td>
<td>.006</td>
<td>.011</td>
<td>.020</td>
<td>.040</td>
<td>.099</td>
<td>.297</td>
</tr>
<tr>
<td>90</td>
<td>.001</td>
<td>.002</td>
<td>.003</td>
<td>.006</td>
<td>.010</td>
<td>.018</td>
<td>.036</td>
<td>.083</td>
<td>.258</td>
</tr>
<tr>
<td>100</td>
<td>.001</td>
<td>.001</td>
<td>.003</td>
<td>.005</td>
<td>.009</td>
<td>.016</td>
<td>.033</td>
<td>.077</td>
<td>.250</td>
</tr>
<tr>
<td>120</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.004</td>
<td>.008</td>
<td>.014</td>
<td>.028</td>
<td>.068</td>
<td>.265</td>
</tr>
<tr>
<td>140</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.003</td>
<td>.006</td>
<td>.012</td>
<td>.023</td>
<td>.055</td>
<td>.196</td>
</tr>
<tr>
<td>160</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.003</td>
<td>.005</td>
<td>.010</td>
<td>.020</td>
<td>.036</td>
<td>.185</td>
</tr>
<tr>
<td>180</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.009</td>
<td>.018</td>
<td>.045</td>
<td>.153</td>
</tr>
<tr>
<td>200</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.004</td>
<td>.008</td>
<td>.016</td>
<td>.042</td>
<td>.145</td>
</tr>
</tbody>
</table>
the relative error in expected queue length. Thus the value of \((R - R_0)/R_N\) can be interpreted in either manner.

Tables I through IV present values of \((R - R_0)/R_N\) for a wide range of model parameters. The service time distributions used in the four tables are constant, Erlang-2, exponential and hyperexponential respectively. Thus the tables present distributions whose variance to mean squared ratios increase progressively through the range: 0, \(\frac{1}{2}, 1, \frac{3}{2}\).

The other parameters in the tables are the number of sources in the actual finite source system \((N)\) and the empirically observed load \((\rho)\) placed on the subsystem. Conceptually, \(\rho\) is first measured for the M/G/1/N system and then automatically duplicated in the M/G/1 model when the input rate and service time distribution are assigned. Note that \(\rho\) is also the subsystem utilization factor (i.e., the proportion of time the subsystem is busy). For a more detailed description of the parameters involved and the manner in which the tables were constructed, see the Appendix.

### DISCUSSION

The description of finite and infinite source models presented earlier suggests that the M/G/1 model should be a relatively good approximation to an M/G/1/N system when \(N\) is large. Although the tables support this general observation, they also point out the sensitivity of the error factors to the value of \(\rho\). That is, the value of \(N\) necessary to attain a given error level is much smaller for low and moderate values of \(\rho\) than it is for values of \(\rho\) which are close to one. Thus the infinite source assumption is of questionable value when \(\rho\) is small.
TABLE V—Minimum Number of Sources Necessary to Attain a 5 Percent Error Level

<table>
<thead>
<tr>
<th>Distribution</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>17</td>
<td>32</td>
<td>61</td>
<td>100</td>
</tr>
<tr>
<td>Erlang-2</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>16</td>
<td>27</td>
<td>49</td>
<td>101</td>
<td>200+</td>
</tr>
<tr>
<td>Exponential</td>
<td>3</td>
<td>6</td>
<td>12</td>
<td>21</td>
<td>38</td>
<td>70</td>
<td>142</td>
<td>200+</td>
</tr>
<tr>
<td>Hyperexponential</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>27</td>
<td>48</td>
<td>93</td>
<td>188</td>
<td>200+</td>
</tr>
</tbody>
</table>

The greatest differences are associated with the disciplines to parameters such as quantum size, one usually finds that the largest values of \( \rho \) are associated with the larger values of \( \rho \). However, these are precisely the values of \( \rho \) for which infinite source models are least accurate and therefore most misleading. Thus, even though infinite source models may be acceptable in a variety of engineering applications that involve small- or mid-range values of \( \rho \), the large error factors that arise as \( \rho \) approaches one may seriously undermine the value of these models in many theoretical contexts.

Another point which the tables illustrate is that the value of \( N \) necessary to attain a given error level increases as the variance to mean squared ratio increases. This effect is explicitly illustrated in Table V which presents the minimum value of \( N \) necessary to attain a 5 percent error level as a function of the variance to mean squared ratio and also the value of \( \rho \). Similar tables can be constructed for any other error levels of interest by using the data in Tables I through IV. In addition, error values for any service time distribution with a variance to mean squared ratio between 0 and \( 3\frac{1}{2} \) can be estimated from the tables by interpolation.

REFERENCES


APPENDIX

The expected response time for a finite source M/G/1/N queuing model is given by

\[
R_N = \frac{N}{\mu} \left( 1 - \frac{1}{\lambda} \sum_{r=0}^{N-1} \frac{(N-r)}{H_r} \right)^{-1} \tag{2}
\]

where

- \( \mu \) = the reciprocal of the mean service time
- \( \lambda \) = the reciprocal of the mean independent activity time

\( N \) = the number of sources

\[
H_r = \prod_{i=1}^{r} \frac{\varphi(\lambda)}{1 - \varphi(\lambda)} \quad \text{for } r = 1, 2, \ldots, N-1
\]

\[
\varphi(s) = \int_0^s e^{-t} dF(t)
\]

\( F(t) \) = the cumulative service time distribution

In order to compute the relative error values \( (R-R_S)/R_N \) which appear in Tables I-IV it is necessary to specify the following quantities: the service time distribution, the load on the server (\( \rho \)) and, for the M/G/1/N case, the number of sources (\( N \)). The service time distributions used in the four tables are as follows:

Table I. Constant service time, mean = 1, variance to mean squared ratio = 0.
Table II. Erlang-2 service time, distribution = \( 4e^{-2t} \) (sum of two exponentials each having a mean of \( \frac{1}{2} \)), mean = 1, variance to mean squared ratio = \( \frac{1}{2} \).
Table III. Exponential service time, distribution = \( e^{-t} \), mean = 1, variance to mean squared ratio = 1.
Table IV. Hyperexponential service time, distribution = \( e^{-t} + \frac{1}{2} e^{-\frac{2t}{3}} \) (equal mixture of exponentials with means of \( \frac{1}{2} \) and \( \frac{2}{3} \)), mean = 1, variance to mean squared ratio = \( \frac{1}{2} \).

Once \( \rho \) is specified, computation of \( R \) by use of equation 1 is entirely straightforward. The major difficulty arises in the computation of \( R_N \). This is because equation 2 does not depend explicitly on the value of \( \rho \) but rather on \( N, \lambda \) and the cumulative service time distribution \( F(t) \). Hence it is necessary to compute \( \lambda \) from \( \rho, N \) and \( F(t) \) before applying equation 2.

In principle, \( \lambda \) can be computed from \( \rho, N \) and \( F(t) \) by using equation 3.

\[
\rho = \lambda N \left[ \lambda N + \sum_{r=0}^{N-1} \frac{(N-r)}{H_r} \right]^{-1} \tag{3}
\]

However, the complex dependency of \( H_r \) on \( \lambda \) makes an exact solution for \( \lambda \) impossible. Thus the basic strategy used in constructing the tables is to first obtain an approximate solution to equation 3 by numerically determining a value \( \lambda' \) which satisfies equation 4.

\[
\left| \rho - \lambda' N \left[ \lambda' N + \sum_{r=0}^{N-1} \frac{(N-r)}{H_r} \right]^{-1} \right| < .005 \tag{4}
\]

This value of \( \lambda' \) is then used in equation 2 to compute \( R_N \), and the corresponding value of \( R \) is computed from equation 1. The value of \( \rho \) used in the computation of \( R \) is obtained by substituting \( \lambda' \) into equation 3. Thus the finite and infinite source models will have the identical value of \( \rho \), and this value will always be within .005 of the corresponding column headings which appear in the tables.