Simulation Modelling of the Addition of Servers in Queueing Systems

Mihir Dash 1*

1Alliance University, Bangalore, India.

Author’s contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Received: 21 February 2020
Accepted: 27 April 2020
Published: 29 April 2020

Abstract

The objective of the study was to analyse the improvement in operating characteristics of an M/M/1 queueing system with the addition of a server, as a function of the utilisation rate λ/µ. The study has applied a simulation model for M/M/1 and M/M/2 systems using the same generated set of random inputs to examine the impact of the addition of servers in queueing systems. The improvement in system length ΔL was analysed using four proposed models: ln(ΔL) as linear and quadratic functions of λ/µ, and as linear and quadratic functions of ln(λ/µ). The Chow test was used to examine structural breaks at λ/µ = 1 and λ/µ = 2.

Keywords: M/M/1 and M/M/2 queueing systems; simulation; utilization rate λ/µ; Chow test; structural breaks.

1 Introduction

The operating characteristics of queueing systems have been studied extensively, particularly for stationary queueing systems (Gross and Harris, [1]). The most important determinant of the operating characteristics of queueing systems is the utilisation rate λ/µ, that is, the ratio of the arrival rate λ and the service rate µ. This parameter is used in practice to take decisions about several service arrangements, particularly the number of channels/servers.

The improvement in operating characteristics of queueing systems when additional servers are introduced has been studied by several authors. Chao and Scott [2] showed that the mean waiting time in a queueing system decreases with each addition of a server, reaching a minimum only in the limiting case of an infinite number of servers. Scheller-Wolf [3] showed that the tail of the stationary delay distribution is significantly reduced by increasing the number of servers under highly variable service demand distributions. Several authors have proposed conditions under which a single server is optimal: Morse [4] for the M/M/k/FCFS

*Corresponding author: E-mail: mihirda@rediffmail.com;

The problem of optimising the number of channels.servers is of critical importance in service facility planning and management. Hillier [9] proposed some models for optimising service in different queueing systems, including the cost model $TC = c_1k + c_2L_s$ for the M/M/k/FCFS system, where $k$ represents the number of servers, $L_s$ represents the average number of customers in the system (at any point of time), and $c_1$ and $c_2$ represent the direct cost per server per unit time and the indirect/waiting cost per customer per unit time, respectively. Beckman [10] proposed a model for M/M/1 systems using dynamic programming, balancing the opportunity cost $c_1$ per server per unit of time and the waiting cost $c_2$ per customer per unit of time; viz. to use no server when the number of customers in the system is less than a critical value $m$, and to use one server if greater than $m$. The critical value $m$ is taken to be equal to 1 when the arrival rate is large compared to the service rate, but $m$ can exceed 1 under certain conditions.

2 Simulation Model and Analysis

2.1 Motivation

The operating characteristics for an M/M/1 system in steady-state (i.e. when $\lambda < \mu$) are given by:

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} L_q = \frac{\lambda^2}{\mu^2 - \lambda^2} W_s = \frac{1}{\mu - \lambda} L_s = \frac{\lambda}{\mu - \lambda}$$

(1)

Similarly, the operating characteristics of an M/M/2 system in steady-state (i.e. when $\lambda < 2\mu$) are given by:

$$W_q = \frac{\lambda^2}{\mu^2(\mu - \lambda)} L_q = \frac{\lambda^3}{\mu^3 - \lambda^3} W_s = \frac{4\mu}{4\mu^2 - \lambda^2} L_s = \frac{4\mu}{4\mu^2 - \lambda^2}$$

(2)

Clearly, when $\lambda < \mu$, both systems are eventually in steady-state, so that the improvement in operating characteristics when an additional server is introduced can be expressed as

$$\Delta W_q = \Delta W_s = \frac{\lambda(4\mu - \lambda)}{(\mu - \lambda)(4\mu^2 - \lambda^2)} \Delta L_q = \Delta L_s = \frac{\lambda^2(4\mu - \lambda)}{(\mu - \lambda)(4\mu^2 - \lambda^2)}$$

(3)

Thus, following Hillier [9], if the cost model for the system is $TC = c_1k + c_2L_s$, the condition for reduction of total cost with the introduction of an additional server in an M/M/1 system becomes

$$\frac{\lambda^2(4\mu - \lambda)}{(\mu - \lambda)(4\mu^2 - \lambda^2)} > \frac{c_1}{c_2}$$

(4)

where $c_1$ and $c_2$ represent the direct/opportunity cost per server per unit time and the indirect/waiting cost per customer per unit time, respectively.

Clearly, as the arrival rate $\lambda$ approaches the service rate $\mu$, this inequality would hold, and empirical studies have shown that the addition of a server when the system starts getting congested would lead to a drastic reduction of system congestion. Further, the improvement in operating characteristics is expected to be even more dramatic when $\mu \leq \lambda < 2\mu$, as the M/M/1 system can no longer reach steady-state; even in the case when $\lambda \geq 2\mu$, although both the M/M/1 and M/M/2 systems would be non-stationary, an improvement in operating characteristics would be expected with the addition of a server.

2.2 Methodology

The objective of the study was to analyse the improvement in operating characteristics of an M/M/1 queueing system with the addition of a server, as a function of the utilisation rate $\lambda/\mu$. The study has applied
a simulation model for M/M/1 and M/M/2 systems using the same generated set of random inputs, implemented on MS Excel. The simulation was replicated for five hundred runs using arrival rates \( \lambda \) and service rates \( \mu \) randomly generated from the negative exponential distribution using the Monte Carlo method. Each simulation run was carried out for one thousand customers, and the operating characteristics for both systems were estimated by taking average waiting times for the second half of the sample, to reduce biases due to initial system transience. The improvement in waiting time (\( \Delta W \)) was computed for each simulation run as the difference in the average waiting time in the queue between the two systems, and the improvement in system length (\( \Delta L \)) was computed by multiplying \( \Delta W \) by \( \lambda \), as per Little’s formula [11].

The simulation results were analysed by dividing the data into three groups: the first group was characterised by \( \lambda/\mu < 1 \), i.e. for which both the M/M/1 and M/M/2 systems were eventually stationary; the second group was characterised by \( 1 \leq \lambda/\mu < 2 \), i.e. for which the M/M/1 system was non-stationary, but the M/M/2 system was eventually stationary; and the third group was characterised by \( \lambda/\mu \geq 2 \), i.e. for which both the M/M/1 and M/M/2 systems were non-stationary. In each of the groups, the improvement in operating characteristics was modelled as function of the utilisation rate \( \lambda/\mu \).

Following Dash [12], the improvement in system length \( \Delta L \) was analysed using four proposed models: \( \ln(\Delta L) \) as linear and quadratic functions of \( \lambda/\mu \):

\[
\ln(\Delta L) = a + b \, \lambda/\mu + c(\lambda/\mu)^2
\]

and \( \ln(\Delta L) \) as linear and quadratic functions of \( \ln(\lambda/\mu) \):

\[
\ln(\Delta L) = a + b \ln(\lambda/\mu) + c(\ln(\lambda/\mu))^2
\]

The Chow test was used to examine structural breaks at \( \lambda/\mu = 1 \) and \( \lambda/\mu = 2 \), using the test statistic

\[
F_{\text{cal}} = \frac{(SSE_1 - SSE_2 - SSE_3) / k}{(SSE_1 - SSE_3) / (n_1 + n_2 - 2k)}
\]

The null hypothesis of no structural break was rejected if \( F_{\text{cal}} > F^* \), with numerator degree of freedom \( k \) and denominator degree of freedom \( n_1 + n_2 - 2k \).

**2.3 Analysis and findings**

**2.3.1 Group I**

The following graphs explore the relationship between \( \ln(\Delta L) \) and \( \lambda/\mu \) & \( \ln(\lambda/\mu) \) for group I.

![Fig. 1. \( \ln(\Delta L) \) as a function of \( \lambda/\mu \) in group I](image.png)
The results of the regression models are presented in the table below.

Table 1. Regression results for Group I

<table>
<thead>
<tr>
<th>Model</th>
<th>(Constant)</th>
<th>$\lambda/\mu$</th>
<th>$(\lambda/\mu)^2$</th>
<th>$\ln(\lambda/\mu)$</th>
<th>$(\ln(\lambda/\mu))^2$</th>
<th>F Stat</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I</td>
<td>-4.528***</td>
<td>7.382**</td>
<td>-2.594***</td>
<td>2.796**</td>
<td>0.616**</td>
<td>3761.636**</td>
<td>94.4%</td>
</tr>
<tr>
<td>Model II</td>
<td>-5.200**</td>
<td>10.149**</td>
<td>-2.594***</td>
<td>4.269**</td>
<td>0.616**</td>
<td>2179.826**</td>
<td>95.2%</td>
</tr>
<tr>
<td>Model III</td>
<td>1.500***</td>
<td>-2.594***</td>
<td>2.796**</td>
<td>2817.839**</td>
<td>0.616**</td>
<td>2746.434**</td>
<td>92.8%</td>
</tr>
<tr>
<td>Model IV</td>
<td>2.024**</td>
<td>-2.594***</td>
<td>2.796**</td>
<td>2817.839**</td>
<td>0.616**</td>
<td>2746.434**</td>
<td>96.2%</td>
</tr>
</tbody>
</table>

The linear and quadratic regressions of $\ln(\Delta L)$ on $\lambda/\mu$ in models I and II were statistically significant, explaining 94.4% and 95.2% of the variation in $\ln(\Delta L)$, respectively. The quadratic coefficient of $\lambda/\mu$ was significant and negative, indicating a convex parabolic trend with $\lambda/\mu$. Similar results were obtained in model III with linear regression of $\ln(\Delta L)$ on $\ln(\lambda/\mu)$, while the results of model IV with quadratic regression of $\ln(\Delta L)$ on $\ln(\lambda/\mu)$ has a significant positive quadratic coefficient of $\ln(\lambda/\mu)$, i.e. a concave parabolic trend with $\ln(\lambda/\mu)$.

2.3.2 Group II

The following graphs explore the relationship between $\ln(\Delta L)$ and $\lambda/\mu$ & $\ln(\lambda/\mu)$ for group II.
The results of the regression models are presented in the table below.

**Table 2. Regression results for Group II**

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>0.827**</td>
<td>-8.133**</td>
<td>3.874**</td>
<td>3.381**</td>
</tr>
<tr>
<td>$\lambda/\mu$</td>
<td>3.235**</td>
<td>16.129**</td>
<td>-4.460**</td>
<td></td>
</tr>
<tr>
<td>$(\lambda/\mu)^2$</td>
<td></td>
<td></td>
<td>4.682**</td>
<td>9.494**</td>
</tr>
<tr>
<td>$\ln(\lambda/\mu)$</td>
<td></td>
<td></td>
<td></td>
<td>-7.410**</td>
</tr>
<tr>
<td>$(\ln(\lambda/\mu))^2$</td>
<td>492.456**</td>
<td>530.590**</td>
<td>671.761**</td>
<td>601.835**</td>
</tr>
<tr>
<td>$R^2$</td>
<td>78.9%</td>
<td>89.5%</td>
<td>83.6%</td>
<td>90.2%</td>
</tr>
</tbody>
</table>

The linear and quadratic regressions of $\ln(\Delta L)$ on $\lambda/\mu$ in models I and II were statistically significant, explaining 78.9% and 89.5% of the variation in $\ln(\Delta L)$, respectively. The quadratic coefficient of $\lambda/\mu$ was significant and negative, indicating a convex parabolic trend with $\lambda/\mu$. The stationary point for the estimated quadratic regression for group II was similar to that for the estimated quadratic regression in group I. Similar results were obtained in models III and IV with linear and quadratic regressions of $\ln(\Delta L)$ on $\ln(\lambda/\mu)$.

The results of the Chow tests for structural break at $\lambda/\mu = 1$ are presented in the table below.

**Table 3. Chow test results Group I - Group II**

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>“jump”</td>
<td>1.208</td>
<td>1.181</td>
<td>2.374</td>
<td>1.357</td>
</tr>
<tr>
<td>F Stat</td>
<td>323.056</td>
<td>107.922</td>
<td>653.686</td>
<td>201.731</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

All of the models were found to have a structural break at $\lambda/\mu = 1$, indicating a shift in the trend. However, the shift is somewhat ambiguous, as model I indicates a decrease in the slope, while model III indicates an increase in the sensitivity/elasticity; model II indicates an increase in the curvature, whereas model IV indicates a transition from a concave to a convex parabolic trend. The “jump” discontinuity at $\lambda/\mu = 1$ was estimated by the models to be between 1.18 and 2.37.

### 2.3.3 Group III

The following graphs explore the relationship between $\ln(\Delta L)$ and $\lambda/\mu$ & $\ln(\lambda/\mu)$ for group II.
The results of the regression models are presented in the table below.

### Table 4. Regression results for Group III

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>6.595**</td>
<td>6.453**</td>
<td>6.032**</td>
<td>5.754**</td>
</tr>
<tr>
<td>(\lambda/\mu)</td>
<td>0.112**</td>
<td>0.267**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\lambda/\mu)^2)</td>
<td></td>
<td>-0.005**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln((\lambda/\mu))</td>
<td></td>
<td></td>
<td>1.088**</td>
<td>1.456**</td>
</tr>
<tr>
<td>ln(((\lambda/\mu)^2))</td>
<td></td>
<td></td>
<td></td>
<td>-0.100**</td>
</tr>
<tr>
<td>F Stat</td>
<td>479.814**</td>
<td>999.794**</td>
<td>10309.142**</td>
<td>9021.445**</td>
</tr>
<tr>
<td>R²</td>
<td>78.2%</td>
<td>93.8%</td>
<td>98.7%</td>
<td>99.3%</td>
</tr>
</tbody>
</table>

*dependent variable: ln(\(\Delta L\))*

The linear and quadratic regressions of ln(\(\Delta L\)) on \(\lambda/\mu\) in models I and II were statistically significant, explaining 78.2% and 93.8% of the variation in ln(\(\Delta L\)), respectively. The quadratic coefficient of \(\lambda/\mu\) was significant and negative, indicating a convex parabolic trend with \(\lambda/\mu\). However, the stationary point for the
estimated quadratic regression for group III was different from that for the estimated quadratic regression in groups I and II. Similar results were obtained in models III and IV with linear and quadratic regressions of \( \ln(\Delta L) \) on \( \ln(\lambda/\mu) \).

The Chow tests for structural break at \( \lambda/\mu = 2 \) are presented in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;jump&quot;</td>
<td>-0.478</td>
<td>0.682</td>
<td>-0.333</td>
<td>0.314</td>
</tr>
<tr>
<td>F Stat</td>
<td>1103.453</td>
<td>1058.960</td>
<td>592.723</td>
<td>279.924</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

All of the models were found to have a structural break at \( \lambda/\mu = 2 \), indicating a shift in the trend. Model I indicates a further decrease in the slope, and model III indicates a decrease in the sensitivity/elasticity; on the other hand, models II and IV both indicate a decrease in the curvature. Thus, for large \( \lambda/\mu \), the trend is asymptotically linear, with unit elasticity; i.e. a 1\% increase/decrease in \( \lambda/\mu \) results in a 1\% increase/decrease in \( \Delta L \), respectively. The "jump" discontinuity at \( \lambda/\mu = 2 \) was estimated to be negative by the linear models, between -0.48 and -0.33, while it was estimated to be positive by the quadratic models, between 0.31 and 0.68.

3 Conclusion

The study proposes a model of the form \( \ln(\Delta L) = a + b \ln(\lambda/\mu) + c(\ln(\lambda/\mu))^2 \), that is, \( \ln(\Delta L) \) expressed as a quadratic function of \( \ln(\lambda/\mu) \), for modelling the decrease in system length and queue length with the addition of a server in an M/M/1 system. The model was found to explain at least 90\% of the variation in \( \ln(\Delta L) \).

The results of the study also indicate structural breaks at \( \lambda/\mu = 1 \) and \( \lambda/\mu = 2 \) in the improvement in operating characteristics of the M/M/1 queueing system with the introduction of an additional server. This is due to the failure of stationarity in the M/M/1 and M/M/2 systems, respectively, at these points.

Further, for large utilisation rates \( \lambda/\mu \), the coefficient of the quadratic term in the model was found to be significant and negative, so that, asymptotically, \( \Delta L \) tends to zero as \( \lambda/\mu \) increases, i.e. the improvement in operating characteristics with the addition of a server is inconsequential. On the other hand, for small utilisation rates \( \lambda/\mu \), the coefficient of the quadratic term in the model is significant and positive, so that \( \ln(\Delta L) \) increases until \( \lambda/\mu \) reaches the range 1.71–1.95 and subsequently decreases.

The implication of the above observations is that, if \( c_1/c_2 \) is too large, where \( c_1 \) and \( c_2 \) represent the direct/opportunity cost per server per unit time and the indirect/waiting cost per customer per unit time, respectively, then a single server will be optimal; that is, the opportunity cost per server may outweigh the decrease in waiting cost. This is consistent with several studies in the literature that assert that a single server is optimal under certain conditions in different contexts [4,5,6,7,8].

There are several limitations inherent in the present study. The study has employed simulation methodology with a sample of five hundred runs, each with one thousand customers. Though this is a sizeable sample, the results are nevertheless sample results, and need to be replicated with different sample sizes to test for consistency. Further, the study considered specific functional forms for the analysis, viz. \( \ln(\Delta L) \) as linear and quadratic functions of \( \lambda/\mu \), and linear and quadratic functions of \( \ln(\lambda/\mu) \). Other functional forms may be appropriate, and should be investigated. Also, several other interesting issues in queueing theory may be addressed using the simulation methodology.
Competing Interests

Author has declared that no competing interests exist.

References


