6-2014

Modeling queues with simulation versus M/M/C models

Kum Khiong YANG
Singapore Management University, kkyang@smu.edu.sg

Mei Wan LOW
Singapore Management University, joycelow@smu.edu.sg

Cayirli TUGBA
Ozyegin University

DOI: https://doi.org/10.1007/s12927-014-0007-3

Published in Journal of Service Science Research, June 2014, Volume 6, Issue 1, pp 173-192.
http://doi.org/10.1007/s12927-014-0007-3

Modeling Queues with Simulation versus *M/M/C* Models

Kum-Khiong Yang, Joyce M. W. Low, Tugba Cayirli

ABSTRACT

This paper examines the performance of single-queue service systems using a combination of computer simulation and *M/M/C* queuing models. Our results show that the accuracy of *M/M/C* models is significantly affected by the assumptions supporting the models. Managers should therefore exercise caution in using the *M/M/C* models for designing queuing systems when the models' assumptions are violated. Our results show that cost-centric and service-centric firms should manage their queues differently. While cost-centric firms should target higher arrival load, single service session, and front-loaded arrival pattern for higher efficiency, service-centric firms should strive for lower arrival load, multiple short sessions and even arrival pattern for better service. In addition, both cost-centric and service-centric firms can consider pooling servers together and reducing the variability of inter-arrival and service times to improve both cost and service simultaneously.

KEYWORDS

M/M/C Models, Computer Simulation, Queuing System Design, Operations Strategy.

Kum-Khiong Yang (✉), corresponding author
Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899
e-mail: kkyang@smu.edu.sg
Joyce MW Low
Singapore Management University
Tugba Cayirli
Ozyegin University

1. INTRODUCTION

Despite the current advances in modeling queues, the standard texts and courses in quantitative modeling and operations management have focused primarily on the simplest queuing models that assume exponential arrival and service times (Hillier & Hillier 2002; Hillier & Lieberman 2004; Jacobs & Chase 2010; Krajewski et al. 2012). These models typically assume a system with stationary arrival rate such that the average number of customer arrivals does not change with time. These models also frequently consider only scenarios with one to several servers serving a single common queue. Customers in the queue are served as first-come, first-served; and the queue has ample space such that no customers are turned away due to limited space. Queuing systems with the above characteristics are often modeled as *M/M/C* models.

As formulae for the *M/M/C* models are relatively simple and easy to use, they appeal to many students, teachers and practitioners of operations management. In their paper, Donnelly & McMullan (1994) used the *M/M/C* models to predict the mean waiting time and probability of no waiting at a service enquiry counter. The customer arrival rate was noted to vary both within a day and across days; but the authors still used the *M/M/C* models even though the service counter did not operate continuously, but encountered opening and closing transience every day. In another study, Goldstein (2009) also used the *M/M/C* models to predict the mean waiting times of customers when they are served at separate counters versus a single counter. In both papers, the authors made no attempt to validate the accuracy of the *M/M/C* models-suggesting that many practitioners believe in the robustness of *M/M/C* models even when the assumptions are violated.

To model the effects of non-stationary arrival rate and opening and closing transience in a queuing system, relatively more complex procedures are available. Two basic approaches are proposed in the literature. The first approach is to explicitly model the system transience and state transition over time as suggested by Abate & Whitt (1987), Lee & Roth (1993), Van Den Berg & Groenendijk (1991), Wang (1999), and Garcia et al. (2002). The second approach is to divide time into segments, estimate the performance in each segment using stationary queuing models such as the *M/M/C* models, and finally average the performance across all segments. This approach is suggested and used by Green & Kolesar (1991, 1995, 1997),

Green et al. (2001), and Green et al. (2007). However, both approaches are relatively complex and require recursive procedures to calculate and predict the system performance. It is thus not surprising that most teachers and practitioners of operations still prefer the *M/M/C* models and ignore the presence of non-stationary arrival rate and operating transience.

In queuing systems where the assumptions supporting the *M/M/C* models are violated, computer simulation offers a viable alternative to model the performance of real systems, especially with the advent of simple and easy-to-use simulation software. The *M/M/C* models and computer simulation are the two most preferred techniques for analyzing queues (Martinich 2002; Sheu et al. 2003; Treville & Ackere 2006; Wang et al. 2006). The *M/M/C* models may be easier to use but are less accurate than computer simulation when the assumptions supporting the analytical models are violated.

This research has two objectives. The first objective is to test the robustness of the *M/M/C* models against computer simulation in predicting the performance of queuing systems under different environments. Our results show that *M/M/C* models report sizable estimation errors when the assumptions supporting the models are violated. We therefore caution the indiscriminate use of *M/M/C* models for designing real systems where one or more of the model's assumptions are violated. While many will agree that *M/M/C* models provide good insights on understanding the tradeoff between cost and service in queue design, their ability to predict the actual system performance accurately should be cautioned.

The second objective in this paper is to examine the impact of different operating factors on the performance of queuing systems. The operating factors are represented by six factors, namely the number of servers, arrival load, session length, arrival pattern, arrival time variability, and service time variability. Our results show that these factors should be managed differently depending on the cost and service orientation of a firm. A cost-centric firm should target higher arrival load, single rather than multiple shorter sessions, and front-loaded arrival pattern for greater efficiency. In contrast, a service-centric firm should strive for lower arrival load, multiple short sessions, and even arrival pattern to keep customer waiting times in check. While it is a common belief that a firm can choose either cost or service, but not both, pooling servers together for a common queue improves both the cost and service performance of a firm. Reducing the variability of the inter-arrival times and service times is another

option to improve both cost and service performance simultaneously.

The rest of the paper is organized as follows: Section 2 describes the simulation model and the experimental design. Section 3 presents the performance measures used in the study, followed by Section 4 which discusses the results on the robustness of *M/M/C* models and the impact of each experimental factor on the performance of queuing systems. Section 5 discusses the managerial implications, and Section 6 ends with the conclusions.

2. SIMULATION MODEL AND EXPERIMENTAL DESIGN

A simulation model of a service system with a single queue is built using the simulation software ARENA (Kelton et al. 2010). In total, six factors are examined for their impact on system performance. These include: (1) number of servers, (2) arrival load, (3) session length, (4) arrival pattern, (5) arrival time variability, and (6) service time variability.

2.1 Number of Servers (NS)

In order to assess the impact of number of servers on the estimation accuracy of *M/M/C* models, this factor is examined at two levels, with one and four servers. As a result, the potential benefit of pooling servers together for a common queue can be investigated.

2.2 Arrival Load (AL)

The arrival load is examined at three levels by adjusting the mean customer arrival rate to achieve a mean load of 65, 80 and 95% of the total servers' capacity. In systems where congestion is costly, the arrival load may be kept low deliberately by limiting the customer arrivals or by expanding the service capacity.

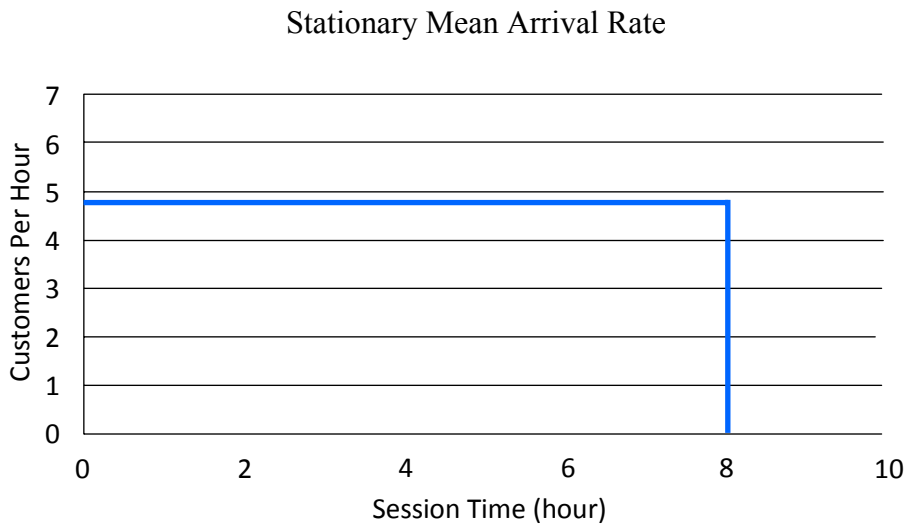
2.3 Session Length (SL)

While most queuing systems start and end a session with no customers in the system, some may operate continuously. For example, an emergency department of a hospital operates 24 hours a day, whereas an outpatient clinic normally operates about 8 hours per day. Within a day, the same clinic may also choose to close for lunch and change the 8-hour session into two 4-hour sessions. Three different session lengths are investigated, representing 4, 8 and 24

hours of operations. In systems that operate 4 or 8 hours, late arrivals after the scheduled session end are denied entry into the system. The system, however, continues to operate beyond the scheduled session end and closes only after the last customer in the system is served.

2.4 Arrival Pattern (AP)

Three arrival patterns, namely, stationary, front-loaded and back-loaded, are explored. Front-loaded pattern is a common sight in post offices and banks where customers rush in during the early opening hours. Back-loaded arrival pattern can also be observed where shoppers rush in to buy groceries after their workday. Arrival pattern is not totally uncontrollable by the management. Some organizations, for example, may intentionally publish the expected waiting times for different periods of their operations to elicit a more even and stationary arrival pattern. Others may offer various incentives or differential pricing to achieve their desired arrival patterns. While there are many possible arrival patterns, the purpose of this study is to examine the effect of ignoring a varying, i.e. non-stationary, arrival pattern when it is present. The exact form of the non-stationary arrival pattern is thus of less importance. The mean arrival rate pattern is modeled with a stationary rate (μ) and with peaks occurring at either the front (F) or back (B) of the session. Figure 1 illustrates the three arrival patterns for a single server working on a session length of 8 hours and an arrival load of 80%.



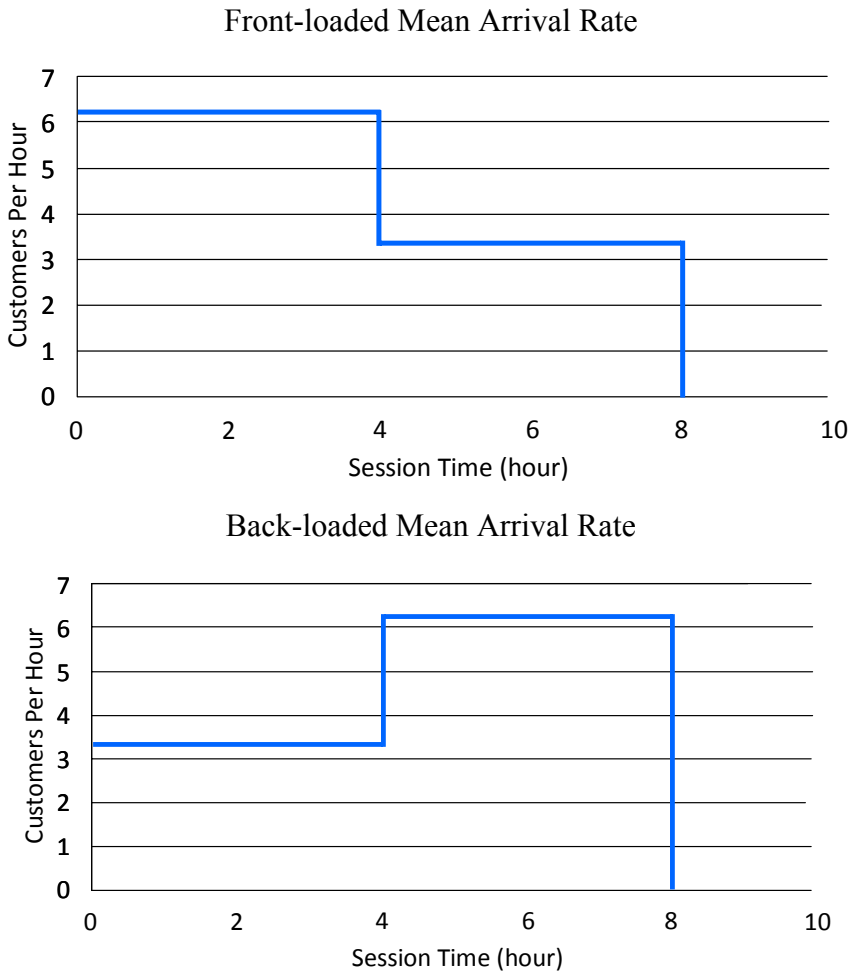


Figure 1. Arrival Patterns for a Single-Server, 8-Hour Session Length and 80% Arrival Load

2.5 Arrival Time Variability (AT)

Variability of the arrival times between customers is expected to affect the performance of a queuing system. A system that faces highly variable inter-arrival times is more likely to experience sporadic congestion and idleness. To examine the effect of this factor, a uniform distribution with a coefficient of variation of 0.4 and an exponential distribution with a coefficient of variation of 1.0 are used to generate the arrival times between customers. It should be noted that the actual probability density functions used is not important as the variability of arrival times can be characterized fairly accurately by the coefficient of variation. Ho & Lau (1992), for instance, found that system performance is affected primarily by the mean and coefficient of variation but not by the skewness, kurtosis and other shape parameters

of the probability density function.

2.6 Service Time Variability (ST)

The mean service rate of each server is fixed at 6 customers per hour by generating the service time for each customer from a probability density function with a mean of 10 minutes. The variability of service times is modeled at two levels. A lognormal distribution with a coefficient of variation of 0.4 is used to represent a less variable service time that is common in practice, while an exponential distribution is used to represent a highly variable service time with a coefficient of variation of 1.0 assumed by the *M/M/C* models. It is again noted the variability of service times can be represented fairly accurately by the coefficient of variation since the other shape parameters have minimal impact on the system performance (Ho & Lau 1992).

Table 1 summarizes the factors and factor-levels examined in this paper. A total of 216 factor combinations are examined in the simulation experiments (i.e., 2 NS×3 AL×3 SL×3 AP×2 AT×2 ST). The "base case" is represented by factor combinations with session length of 24 hours, stationary arrival pattern, and coefficient of variation of arrival and service times of 1.0, which correspond to the assumptions of *M/M/C* models. For each factor combination, the simulation model is run to produce 20 observations of 2000 sessions each. Five performance measures are collected as described in Section 3.

Table 1. Experimental Design

Factors	Levels
Number of Servers (NS)	1 & 4
Arrival Load (AL)	65%, 80% & 95%
Session Length (SL)	4, 8 & 24 hours
Arrival Pattern (AP)	Stationary, Front-loaded & Back-loaded
Arrival Time (AT)	0.4 & 1.0
Service Time (ST)	0.4 & 1.0

3. PERFORMANCE MEAURES

Five performance measures are collected to examine the effects of the experimental factors. The measures include the mean number in queue (NIQ), probability of no waiting on

arrival (PNW), mean session overtime (SOT), mean overtime per customer served (COT), and mean server utilization (SUT).

The mean number in queue (NIQ) measures the mean queue length that a customer will encounter on arrival. Waiting time is one of the more important factors affecting customer satisfaction (Karaca et al. 2011). A long queue length will increase not only the actual but also the perceived waiting time of customers. The probability of no waiting on arrival (PNW) measures the probability of a customer being able to enter service immediately on arrival, i.e. when there is at least one idle server in the system. A high probability of no waiting is an indication of fast service but low server utilization.

The mean session overtime (SOT) is the extra time needed beyond the official session length to serve all customers in system. All customers who arrive during the official session length are allowed entry into the system, and the session ends only after the last customer is served. SOT is a measure of the extra time, i.e. overtime, to keep the system open to serve all customers admitted into the system. The mean overtime per customer served (COT) measures the overtime cost incurred per customer, and it is computed by dividing the total servers' overtime by the number of customers served per session. When the session length is fixed at 24 hours, SOT and COT are always zero as system operates continuously.

The mean server utilization (SUT) measures the percentage of the time that servers are busy from the beginning to the end of each session. SUT is a measure of cost efficiency, i.e. the proportion of the servers' capacity used productively to serve customers.

4. RESULTS

The results are presented in two parts. First, the accuracy and robustness of the *M/M/C* models in predicting the performance of queuing systems are examined when the assumptions supporting the models are violated. Second, the performance of the queuing systems is examined under changing factor levels to understand the impact of each factor on system performance.

4.1 Robustness of *M/M/C* Models

The *M/M/C* models provide easy-to-use formulae to compute the mean number in queue and probability of no waiting in queuing systems. These formulae can be found in standard

texts (Hillier & Hillier 2002; Hillier & Lieberman 2004); and their accuracy is assessed by computing (1) the percentage estimation error of the mean number in queue (ENIQ) and (2) the percentage estimation error of the probability of no waiting (EPNW) for various scenarios. Table 2 and Table 3 tabulate the results for ENIQ and EPNW, respectively. In both tables, the third column shows the percentage estimation errors for different number of servers and arrival loads when all assumptions of the *M/M/C* models are valid in the simulation (i.e. 24-hr session length, stationary arrival pattern and exponentially-distributed inter-arrival and service times with CV = 1). The percentage estimation errors for these "base cases" are expectedly close to zero, which validate the accuracy of the *M/M/C* and simulation models when all assumptions are valid.

Table 2. Percentage Estimation Error of *M/M/C* models for Mean Number in Queue

NS (*C*)	**AL**	***M/M/C***	**SL**		**AP**		**AT**	**ST**
			4-hr	**8-hr**	**FL**	**BL**	**CV = 0.4**	**CV = 0.4**
1	**65**	-0.23	59.25	27.35	-33.48	-32.99	116.28	71.77
1	**80**	-0.01	140.96	73.66	-35.86	-35.97	92.63	73.44
1	**95**	-0.99	747.52	448.16	-16.45	-16.86	76.98	72.67
4	**65**	-0.01	26.38	12.35	-59.85	-59.93	184.63	61.49
4	**80**	0.53	54.74	25.51	-72.67	-72.79	115.95	67.45
4	**95**	0.46	358.13	202.84	-56.00	-55.91	81.64	70.51
	Mean:	**-0.04**	**231.16**	**131.65**	**-45.55**	**-45.91**	**111.35**	**69.56**

Table 3. Percentage Estimation Error of *M/M/C* models for Mean Probability of No Waiting

NS (*C*)	**AL**	***M/M/C***	**SL**		**AP**		**AT**	**ST**
			4-hr	**8-hr**	**FL**	**BL**	**CV = 0.4**	**CV = 0.4**
1	**65**	-0.16	-16.89	-9.67	-0.04	0.14	-0.08	0.07
1	**80**	0.03	-36.71	-25.98	0.24	0.26	0.01	-0.13
1	**95**	0.12	-78.51	-71.75	0.39	0.37	0.15	1.39
4	**65**	0.02	-5.48	-2.97	7.93	7.87	-11.41	-1.65
4	**80**	-0.10	-17.27	-10.03	12.43	12.73	-13.50	-2.88
4	**95**	0.35	-64.31	-54.10	12.86	12.96	-15.16	-3.54
	Mean:	**0.04**	**-36.53**	**-29.08**	**5.64**	**5.72**	**-6.67**	**-1.12**

The remaining columns in the tables show the percentage estimation errors when one of the assumptions of the *M/M/C* models is violated at a time. Columns 4 and 5, for example,

show the percentage estimation errors when the session length is 4 and 8 hours, respectively, instead of 24 hours. Positive (or negative) errors indicate overestimation (or underestimation) of the performance measures.

Some interesting patterns in the results are observed when one assumption is violated at a time. The main findings are summarized as follows:

i. Overall, the percentage estimation errors are substantially larger for the mean number in queue (ENIQ), ranging from -45.91 to 231.16% on average (Table 2), compared to the probability of no waiting (EPNW) ranging from -36.53 to 5.72% (Table 3). For ENIQ, the SL, AT, ST and AP have the largest impact on the percentage estimation errors in the order as listed, whereas for EPNW, the order changes to SL, AT, AP and ST.

ii. *Session Length* (*SL*): The shorter the SL, the higher the percentage estimation errors for both performance measures. Relatively larger positive errors (i.e. overestimation) of ENIQ are observed in Table 2, compared to smaller but still sizable negative errors (i.e. underestimation) of EPNW in Table 3. A queuing system starts empty when its operation is not continuous. Consequently, it is not surprising that *M/M/C* models overestimate the number in queue and underestimate the probability of no waiting. The impact of shorter SL on higher estimation errors is further exacerbated by smaller number of servers and/or higher arrival loads. As a result, the highest percentage estimation errors occur for the extreme case with SL = 4, NS = 1 and AL = 95% (e.g. 747.52% and -78.51% for ENIQ and EPNW, respectively).

iii. *Arrival Pattern* (*AP*): The percentage estimation errors of front-loaded and back-loaded arrival pattern on the mean number in queue, ENIQ (and probability of no waiting, EPNW) are practically equal, since the effect of a peak arrival occurring at the beginning or end of session on the mean performance measure is the same when session length is fixed as 24 hours, i.e. continuous. An underestimation of ENIQ occurs when the arrival pattern is non-stationary and the magnitudes are smaller at both extremes of arrival loads with 65 and 95% (See Table 2). This is intuitive, as ignoring the presence of non-stationary arrival pattern is relatively less important when the system is relatively idle (or very busy) which occurs at low (or very high) arrival load. On the other hand, overestimation is observed for

the effect of non-stationary arrival pattern on EPNW. The estimation error is negligible with one server, but increases as both the number of servers and arrival load increase (See Table 3). This result is not surprising given that in a single-server system, the probability of no waiting depends largely on the probability of the single server being free, i.e. the mean server's utilization. In contrast, in a multiple-server system, a new customer has to wait only if all servers are busy; and the probability of one to all servers being busy is a more complex function of the mean server utilization, arrival pattern, arrival time variability, and service time variability. Therefore, ignoring the arrival time variability, service time variability and non-stationary arrival pattern when there are multiple servers introduces larger percentage estimation errors of EPNW. For both ENIQ and ENPN, the percentage estimation errors increase as the number of servers (NS) increase from one to four servers.

iv. *Arrival Time* (*AT*): When the arrival times between customers are less variable with a coefficient of variation of 0.4, the *M/M/C* models overestimate the mean queue length. The percentage estimation errors in ENIQ decrease as AL increases and/or NS decreases (See Table 2). This suggests that the arrival time variability has less impact on the mean number in queue in single-server systems when the arrival load is high. With regards to EPNW, the errors are negligible when NS = 1, but increase significantly for the multiple-server system simulated with NS = 4, especially when the arrival loads are also higher (See Table 3).

v. *Service Time* (*ST*): When service time variability is reduced to CV = 0.4, there is substantial overestimation for ENIQ, whereas the impact is almost negligible for EPNW. Unlike the other factors, the impact of ST is rather robust to changes in NS and/or AL with similar percentage estimation errors (around 60-70% for ENIQ; 0-4% for EPNW in Table 2 & Table 3). Overall, it is safer to use *M/M/C* models to estimate the probability of no waiting for single-server systems when only one of the assumptions related to arrivals (i.e. AL or AT) or service times (ST) is violated.

While it is interesting to outline the causes and reasons for the estimation errors, most of the percentage estimation errors of the *M/M/C* models in Table 2 and Table 3 are sizable,

even though only one assumption is violated at a time. Computer simulation is thus a more reliable and accurate tool to estimate the performance of queuing systems when one or more assumptions supporting the *M/M/C* models are violated.

4.2 Impact of Operating Factors on Queue Performance

Each of the six factors, namely the number of servers, arrival load, session length, arrival pattern, arrival time variability, and service time variability is examined for its impact on the five performance measures (See Section 3). In order to identify potential interactions among these factors, analysis of variances (ANOVA) is conducted on each performance measure.

Table 4. Main Effects of the Six Operating Factors

Number of Servers, NS	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
1	2.846	25.11	17.81	0.6304	75.88
4	5.246	45.38	15.71	0.5738	76.37

Arrival Load, AL	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
65	0.649	52.89	8.916	0.4232	63.32
80	2.363	35.09	14.81	0.5595	76.66
95	9.127	17.75	26.58	0.8238	88.40

Session Length, SL	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
4	1.542	39.49	22.11	1.1095	73.03
8	2.498	36.41	28.19	0.6970	75.39
24	8.098	29.82	0.000	0.0000	79.96

Arrival Pattern, AP	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
Stationary	2.244	35.91	14.27	0.5241	76.60
Front-loaded	5.282	31.78	9.534	0.3554	77.61
Back-loaded	4.613	38.04	26.50	0.9269	74.18

Arrival Time, AT	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
CV = 0.4	3.270	36.04	14.63	0.5243	76.72
CV = 1.0	4.822	34.45	18.91	0.6800	75.54

Service Time, ST	**NIQ**	**PNW (%)**	**SOT (min)**	**COT (min)**	**SUT (%)**
CV = 0.4	3.448	34.85	13.21	0.4687	76.94
CV = 1.0	4.644	35.64	20.32	0.7356	75.31

The results show that all main effects and many higher-order interactions are statistically significant at 1%. The interactions are examined and found to affect only the relative performance but not the rankings of the factor-levels. In other words, for each performance measure, the same factor-level performs the best across all interactions. It is therefore sufficient to present only the main effects in Table 4 since the interactions do not affect the choice of the best factor-level.

As shown in Table 4, a 4-server system always performs better than a single-server system with higher probability of no waiting, higher server utilization, lower session overtime and lower overtime per customer served. While a 4-server system has 1.84 times (i.e. 5.246/ 2.846) the mean number in queue of a 1-server system, the mean waiting time in queue of the 4-server system is only 0.46 times that of the 1-server system using Little's Law. Pooling servers together for a common queue is thus preferred for both cost- and service-centric firms with no trade-offs.

As the arrival load increases, the number in queue, session overtime, and overtime per customer served increase, whereas the probability of no waiting decreases significantly. In other words, both customer service and overtime cost will deteriorate when the arrival load increases. However, on a more positive note, the mean server utilization increases as the arrival load increases. Consequently, a service-centric firm should favor a lower arrival load while a cost-centric firm should favor a higher arrival load, i.e. higher server utilization as long as the overtime premium is not excessive.

As the session length increases from 4 to 8 hours, the number in queue increases and the probability of no waiting decreases, while the overtime per customer served decreases and the server utilization increases. Overall, the results show that customer service deteriorates while cost efficiency improves with session length (assuming that the fixed cost per unit time to keep system open during the session overtime is not excessive). Service-centric firms should therefore favor multiple short sessions while cost-centric firms should favor a single long session.

The front-loaded arrival pattern produces the lowest session overtime, lowest overtime per customer served, and highest server utilization. Therefore, it is a good choice for cost-centric firms, even though it produces the largest number in queue and lowest probability of no

waiting. Table 4 also shows that the stationary arrival pattern produces the smallest number in queue, but a slightly lower probability of no waiting compared to the back-loaded arrival pattern. Although it does not dominate the back-loaded pattern on both measures of customer service (i.e. with the smallest number in queue but a slightly smaller probability of no waiting), the stationary arrival pattern is preferred due to its significantly lower session overtime, lower overtime per customer served, and higher server utilization. Promoting a stationary and less variable arrival pattern is a good strategy for service-centric firms.

The last two factors-variability of the arrival times and service times have smaller impact on the performance measures relative to the other factors. Table 4 shows that the performance measures, especially PNW and SUT, change only marginally at different levels of these two factors. Specifically, reducing the variability of the arrival times and service times exhibits a small, but positive, impact on the number in queue, session overtime, and overtime per customer served. Both cost- and service-centric firms are thus encouraged to reduce variability of the arrival and service times.

5. Managerial Implications

A service firm can choose to adopt a cost-, service-, or value-centric proposition. Generally, cost-centric firms will seek to achieve higher server utilization as well as lower session overtime and overtime per customer served; whereas service-centric firms will seek better customer service with lower queue length and higher probability of no waiting. Measures that these firms can adopt are as follows:

i. *Expanding the scale of operations:* With multiple servers serving a larger pool of customers, firms can achieve higher cost efficiency through increased server utilization and decreased session overtime and overtime per customer served. In particular, if demand is price-elastic, a virtuous cycle can be generated if the firms share the cost savings with their customers to stimulate higher demand resulting in the installation of multiple-server systems to reap the cost and service benefits of resource pooling.

ii. *Increasing session length:* A longer session length increases the session overtime, but decreases the mean overtime per customer served and increases the overall server utilization. Consequently, unless the overtime premium in keeping a system open is high, reducing

the mean overtime per customer served and increasing the mean server utilization are more beneficial in curbing the total operating cost. Hence, it is preferable for cost-centric firms to run a single long session than multiple short sessions of equivalent total length. While a long session length offers customers greater access to service, it also increases the mean queue length and reduces the probability of no waiting. To offer customers good access and better queuing experience, service-centric firms can thus consider offering multiple short sessions with a total length equivalent to a single long session.

iii. *Scheduling arrivals:* Cost-centric firms may foster a front-loaded arrival pattern by introducing incentives such as early bird discounts to reduce the chance and magnitude of overtime. Similarly, service-centric firms may also introduce tailored reward and penalty schemes to solicit a more even arrival pattern. Firms may also try to control the arrival patterns and variability of inter-arrival times between customers by scheduling appointments. The results attainable will, however, depend on whether the firms have complete or partial control over the arrivals. Implementing an appointment system that mandates customers to adhere strictly to the schedule is useful in establishing the desired arrival patterns and less variable inter-arrival times.

iv. *Standardizing service:* Standardization of service represents another means for firms to achieve less variable service times, which shortens the session overtime and overtime per customer served. Standardization of service can be achieved by establishing a set of standard protocols and procedures for serving customers. It can also be achieved by segregating customers into similar groups for standardized processing.

Table 5. Strategic Choices for Cost-Centric and Service-Centric Firms

Factor	Cost Strategy	Service Strategy
Number of servers	• Multiple servers	• Multiple servers
Arrival Load	• High	• Low
Session Length	• Single, long	• Multiple, short
Arrival Pattern	• Front-Loaded	• Stationary
Arrival Time	• Less variability	• Less variability
Service Time	• Less variability	• Less variability

In summary, Table 5 summarizes the strategic directions of cost-centric and service-centric

firms. Between the two extremes, a value-centric firm can seek to find a balance within the continuum of cost versus service. A trade-off has to be made between high and low arrival load. In addition, a value-centric firm has to weigh the cost and benefits of having a single versus multiple shorter sessions. A front-loaded arrival pattern may help to reduce cost, but hurt customer service relative to a more even, stationary arrival pattern. Unambiguously, less variable inter-arrival and service times are beneficial in reducing cost and enhancing service to customers. Pooling single-server systems into a single multiple-server system also helps to improve both cost and service performance.

6. CONCLUSIONS

This paper seeks to examine the robustness of *M/M/C* models and the influence of various factors on the performance of single queue systems with one or multiple servers. To accomplish these objectives, this study examines the mean number in queue, probability of no waiting on arrival, mean overtime per session, mean overtime per customer served and mean server utilization of various simulated queuing systems, and provides results on the percentage errors in estimating the mean queue length and mean probability of no waiting by the *M/M/C* models when the assumptions supporting the analytical models are violated. Our results show that session length, inter-arrival time variability, service time variability and arrival pattern have the largest impact on the estimation errors of the mean queue length in the order as listed. The sequence changes to session length, inter-arrival time variability, arrival pattern, and service time variability for the estimation errors of the probability of no waiting.

One of the main objectives of this paper is to test the robustness of the *M/M/C* models via simulation when the assumptions supporting the analytical models are violated. The comparison is conducted on the percentage errors in estimating the mean number in queue (NIQ) and probability of no waiting (PNW). The results reveal that the session length (SL) and inter-arrival time variability (AT) have the largest impacts on the estimation errors of both measures, as listed in order of significance. This means that any violation in the assumptions of these two factors indicates a serious caution on the accuracy of *M/M/C* models. The highest errors are observed for shorter session lengths combined with higher arrival load and

lower number of servers (i.e. single-server systems). The inter-arrival time variability is the second most critical factor in terms of accuracy and robustness of these analytical models, such that it results in very high estimation errors of NIQ, especially when the number of servers increases and/or the arrival load decreases.

Although violations of the other two factors, namely service time variability (ST) and arrival pattern (AP), also result in significant errors for NIQ, the *M/M/C* models are more robust in estimating the PNW. In fact, the *M/M/C* models may be used for estimating the PNW in single-server systems with very high accuracy of greater than 99%, when *only* one of the assumptions-related to the arrival time variability, arrival pattern or service time variability -is violated at a time.

This study further investigates the impact of various key operating factors on the performance of single queue systems based on their cost-efficiency (i.e. mean overtime per session, mean overtime per customer served, and mean server utilization) as well as customer-service measures (i.e. mean number in queue and probability of no waiting). The six operating factors include the number of servers, arrival load, session length, arrival pattern, arrival time variability, and service time variability. The results suggest that cost-centric firms should encourage a front-loaded arrival pattern and operate a single long session with heavy load; whereas service-centric firms should undertake measures to elicit a more stationary arrival pattern and operate multiple, short sessions with lighter load. Meanwhile, value-centric firms need to weigh the cost and service trade-offs arising from arrival load, session length and arrival pattern. Regardless of the strategy pursuits of a firm, pooling multiple servers into a single queue system and reducing the variability of inter-arrival and service times are always desirable, resulting in shorter mean queue length, higher server utilization and lower overtime.

REFERENCES

Abate J & Whitt W (1987) Transient behavior of the M/M/1 queue: starting at the origin. Queuing Systems 2(1):41-65.

Donnelly M & McMullan CH (1994) Setting service standards for local government reception services. Managing Service Quality 4(5):42-47.

Garcia J-M, Brun O, & Gauchard D (2002) Transient analytical solution of M/D/1/N queues.

Journal of Applied Probability 39(4):853-864.

Green L & Kolesar P (1991) The pointwise stationary approximation for queues with non-stationary arrivals. Management Sciences 37(1):84-97.

Green L & Kolesar P (1995) On the accuracy of the simple peak hour approximation for Mark-ovian queues. Management Sciences 41(8):1353-1370.

Green L, Kolesar P (1997) The lagged PSA for estimating perk congestion in multiserver Markovian queues with periodic arrival rates. Management Sciences 43(1):80-87.

Green L, Kolesar P, & Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49(4):549-564.

Green L, Kolesar P, & Whitt W (2007) Coping with time-varying demand when setting staffing requirement for a service system. Production and Operations Management 16(1):13-39.

Goldstein SD (2009) Improve customer satisfaction through dedicated service channels. Journal of Applied Business and Economics 9(1):11-21.

Hillier FS & Hillier MS (2002) Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets, 2nd Edition, McGraw-Hill/Irwin.

Hillier FS, Lieberman GJ (2004) Introduction to Operations Research, 8th Edition, McGraw-Hill.

Ho C-J & Lau H-S (1992) Minimizing total cost in scheduling outpatient appointments. Management Sciences 38(12):1750-1764.

Jacobs FR & Chase RB (2010) Operations and Supply Chain Management, 13th Edition, McGraw-Hill/Irwin.

Karaca MA, Erbil B, & Özmen MM (2011) Waiting in the Emergency Room: Patient and Attendant Satisfaction and Perception. European Journal Surgery Science 2(1):1-4.

Krajewski LJ, Ritzman LP, & Malhotra MK (2012) Operations Management-Processes and Supply Chains. 10th Edition, Pearson.

Kelton WD, Sadowski RP, & Swets NB (2010) Simulation with ARENA, 5th Edition, McGraw Hill.

Lee I-J & Roth E (1993) A heuristic for the transient expected queue length of Markovian queuing systems. Operations Research Letter 14(1):25-27.

Martinich JS (2002) The critical few minutes in scheduling time-varying queuing systems.

Decision Sciences 33(3):415-431.

Sheu C, McHaney R, & Babbar S (2003) Service process design flexibility and customers waiting time. International Journal of Operations and Production Management 23(7/8): 901-917.

Treville S & Ackere A van (2006) Equipping students to reduce lead times: The role of queuing-theory-based modeling. Interfaces 36(2):165-173.

Van Den Berg TJL & Groenendijk WP (1991) Transient Analysis of an M/M/1 Queue with Regularly Changing Arrival and Service Intensities. In: Jensen A and Iversen VB, Tele-traffic and Datatraffic in a Period of Change, ITC-13. Elsevier Science Publishers B.V. (North-Holland):677-681.

Wang C-H, Lee Y-D, Lin W-I, & Liu P-M (2006) Application of queuing model in healthcare administration with incorporation of human factors. Journal of American Academy of Business 8(1):304-310.

Wang C-L (1999) On the transient delays of M/G/1 queues. Journal of Applied Probability 36(3):882-893.