Optimization of a Call Centre Performance Using the Stochastic Queueing Models

Alenka Brezavšček  
Faculty of Organizational Sciences, University of Maribor, Slovenia  
Alenka Baggia  
Faculty of Organizational Sciences, University of Maribor, Slovenia

Abstract

Background: A call centre usually represents the first contact of a customer with a given company. Therefore, the quality of its service is of key importance. An essential factor of the call centre optimization is the determination of the proper number of operators considering the selected performance measure. Results of previous research show that this can be done using the queueing theory approach.  
Objectives: The paper presents the practical application of the stochastic queueing models aimed at optimizing a Slovenian telecommunication provider’s call centre.  
Methods/Approach: The arrival and the service patterns were analysed, and it was concluded that the call centre under consideration can be described using the M/M/r {infinity/infinity/FIFO} queueing model.  
Results: An appropriate number of operators were determined for different peak periods of the working day, taking into consideration the following four performance measures: the expected waiting time, the expected number of waiting customers, the probability that a calling customer will have to wait, and the call centre service level.  
Conclusions: The obtained results prove the usefulness and applicability of the queueing models as a tool for a call centre performance optimization. In practice, all the data needed for such a mathematical analysis are usually provided. This paper is aimed at illustrating how such data can be efficiently exploited.

Keywords: call centre; service quality; performance measure; optimization; stochastic queueing models

JEL main category: C  
JEL classification: C02, C44, C61  
Paper type: Research article

Received: 20, January, 2014  
Accepted: 18, June, 2014

DOI: 10.2478/bsrj-2014-0016

Introduction

The queueing theory is a special field of stochastic process theory. Many books discussing the fundamentals of the queueing theory were published in recent years, e.g. van Dijk and Bouchiere (2011); Gross et al. (2008); Tanner (1995); Tijms (2003).
Queues often represent an area where customers wait for service providers to be served, and can be used as inventories in manufacturing or any human related services. Applications of the queueing theory can be found in diverse areas such as: (1) telecommunications, e.g. Attahirusule (2010), Giambene (2005); (2) computer networks traffic studies, e.g. Robertazzi (2000); and (3) road traffic studies e.g. Ismail et al. (2011); Raheja (2010); and others.

Numerous contributions in the literature prove that the queueing theory can be applied successfully also to call centres performance optimization, e.g. Brown et al. (2005); Koole (2007). Koole and Mandelbaum (2001) consider that the world of call centres is a challenging fertile area for the applications of queueing models.

Call centres (or their contemporary successors contact centres) are frequently used by organizations for marketing purposes and technical support for end users. They are preferred and prevalent way for many companies to communicate with their customers (Koole and Mandelbaum, 2001). Call centres usually support banking, insurance, telecommunication companies, online catalogue shops, market research agencies, etc.

A call centre constitutes a set of resources (typically personnel, computers and telecommunication equipment), which enable the delivery of services via telephone. Therefore, every call centre can be described as a quite complex queueing system with its own particularity. Inbound calls arrive at random according to some complicated stochastic processes, call durations are also random, waiting calls may abandon after a random patience time, some agents may fail to show up to work for any reason, and so on (Avramidis et al., 2010). In some cases, emergency calls in call centre should also be considered (e.g. police, fireman, hospital call centre), where priorities play an important role. Additionally, in the research Aksin et al. (2007) operator’s ability to handle stress is taken into account.

An essential task of managers is to define the appropriate number of operators in the call centre to ensure optimal performance. As stated by Aksin et al. (2007) an appropriate number of operators for each period of the scheduling horizon in a call centre depends on both, how many customers and consequently how much work is arriving into the call centre at what times, and how quickly the operators seek to serve these customers. The main performance measures in call centres have to do with the quality of service and/or the operating cost. The widely used approach in a call centre optimization is operating cost minimization objective with the call centre service level constraint (Aksin et al., 2007). In general, the quality of call centre service is related to customer satisfaction with the service, and is frequently dependent on the waiting times of calls before they are answered.

For a call centre performance optimization many theoretical mathematical models are available in the literature, e.g. Chassioti (2005); Dombacher (2010). Some of them are rather simple and provide closed form expressions for most of its performance measures. Application of such model to practice is therefore quite easy. However, in order to obtain relevant and useful results, it should be assured that underlying assumptions of the model do not contradict the properties of the call centre and its operation. The basis for proper selection of a theoretical model to describe a particular call centre in practice represents the knowledge of the probability density functions of inter-arrival times (i.e. times between two successive incoming calls) and service times (i.e. calls length). These functions can be obtained if accurate and complete data about the call centre operation are available. Since contemporary technology enables automatic logging of all the events in the call centre, data needed for the mathematical analysis are usually provided. However, lack of expert knowledge in practice prevents the companies from efficient usage
of them. In our opinion the number of operators in a shift is often based on the rule-of-thumb decision, and is frequently not an optimal solution.

The paper represents the practical application of the queueing theory for optimization of a call centre performance. It is a continuation of our prior researches (Brezavšček et al., 2012; Brezavšček and Baggia, 2013), conducted on the case of Slovenian telecommunication provider's call centre. The field data on the call centre operation will be used to analyse the arrival and service patterns (Žugaj et al., 2006). On the basis of this analysis a suitable theoretical queueing model will be selected to describe the call centre under consideration. An appropriate number of operators will be determined for several peak periods of the working day regarding different call centre performance measures.

**Methodology**

A typical queueing system comprises of one or more service units (i.e. servers), arrivals of customers demanding the service, and the service process. Whenever all the customers cannot be served at once, queues are formed. This leads to costs (losses) due to waiting which increase with the number of customers in the queue. To decrease the waiting costs and raise the service level ensuring better system performance different improvements can be implemented. However, any improvement often comes with a certain investment leading to higher costs of the queueing system operation. Figure 1 shows that it is always possible to determine the optimal service level which ensures the total costs of the queueing system performance are minimal.

**Figure 1**

Costs of queueing system performance

![](image.png)

*Source: Author’s illustration*

To determine the optimal service level of a queueing system, different quantitative characteristics (i.e. performance measures) can be used. The values of these measures can be calculated using a suitable mathematical model. Proper selection of the mathematical model is based on the following elements of the queueing system:
o Arrival process: Population of customers can be considered either limited (closed systems) or unlimited (open systems). Most mathematical models presume individual arrivals of customers and independent identically distributed inter-arrival times.

o Service mechanism is determined with the system capability, availability and probability density function of service times. Most of the mathematical models assume that service times are independent identically distributed random variables.

o Queueing discipline represents the way the queue is organised (e.g. First-In-First-Out (FIFO), Last-In-First-Out (LIFO), random selection of customers or selection based on customer priorities).

When there is only one sever, or there are a number of equivalent and parallel servers, the queueing system is called simple. Simple queueing models use the standard notation for describing the probability density function of inter-arrivals and service times:

- **M** – a Poisson process of the number of events (i.e. customer arrivals or end of services); an exponential density function of times between two successive events.
- **G** – a general distribution of times between two successive events (with a known mean and variance; e.g. a normal density function).
- **D** – a deterministic situation; times between two successive events are constant.

Notation **M/M/r {infinity/infinity/FIFO}** therefore describes the queueing system with **r** parallel servers, unlimited population, unlimited queue, FIFO queueing discipline, while both, inter-arrival and service times are distributed according to the exponential density function, e.g. Chassioti (2005). For many types of simple queueing model there exists closed form expressions for most system performance measures.

Assuming the **M/M/r {infinity/infinity/FIFO}** queueing model the closed form of all four performance measures are available, e.g. Tijms, (2003).

**The expected waiting time** can be calculated according to the following equation:

\[
E(W_q) = \frac{1}{S} \frac{(r\rho)^r}{r!(1-\rho)^2 r\sigma} \quad (1)
\]

**The expected number of waiting customers** is given by the expression:

\[
E(N_q) = \frac{1}{S} \frac{(r\rho)^r}{r!(1-\rho)^2} \quad (2)
\]

**The probability that the calling customer will have to wait** because all operators are busy can be calculated as follows:

\[
P_{wait} = \frac{1}{S} \frac{(r\rho)^r}{r!} \frac{1}{1-\rho} \quad (3)
\]
In the literature, the equation (3) is also known as the Erlang C formula (e.g. Tanner, 1995; Garnett et al., 2001) and plays an important role in the performance of telephone systems. The service level is the most common measure of quality of the call centre service. It is defined by a given percentile of the waiting time distribution, and can be calculated according to the following expression:

\[
SL(t_0) = P[W_q \leq t_0] = 1 - \frac{(r \rho)^r}{S \cdot r!} \exp\left(\frac{-r \cdot (1-\rho) \cdot r \cdot \sigma t_0}{S}\right)
\]  

The equation (4) gives the long-term fraction of customers whose waiting time \( W_q \) in the queue is no larger than a given threshold \( t_0 \) (Avramidis, 2010).

The symbols used in equations (1), (2), (3) and (4) denote:
\[r\] – the number of servers
\[\alpha\] – the arrival rate; \(1/\alpha\) is the expected time between two successive arrivals
\[\sigma\] – the service rate; \(1/\sigma\) is the expected service time
\[\rho\] – the traffic intensity calculated as \(\rho = \frac{\alpha}{r \cdot \sigma}\)
\[S\] – the sum which can be calculated as follows:

\[
S = 1 + r \rho + \frac{(r \rho)^2}{2!} + \cdots + \frac{(r \rho)^{r-1}}{(r-1)!} + \frac{1}{r!} \frac{1}{1-\rho}
\]  

Equations (1), (2), (3) and (4) make sense when \( S < \infty \). This condition is met when \( \rho < 1 \). The condition \( \rho < 1 \) ensures that the steady state distribution exists. In such a case the infinite queues are not formed and the queueing system still operates after a long run. The minimum number of servers \( r_{\text{min}} \) needed to satisfy the steady state condition is the lowest integer that fulfils the equation

\[
r > \frac{\alpha}{\sigma}
\]  

The case study: Call centre as a queueing system

The presented research was conducted on the case of a Slovenian telecommunication provider’s call centre. The call centre is opened from 8:00AM till 12:00PM. It employs 8 full time operators while additional contractors are hired when needed. The schedule of operators is defined based on prior experiences. No analysis of the schedule has been performed up to now. Customers are calling a single phone number. If at least one of the operators is available at the time of the call, he answers the call and serves the customer. If all of the operators are busy the calling customer is not rejected but can wait for a free operator regardless the number of customers in the queue. The principal scheme of the call centre is presented in Figure 2.
The call centre under consideration can be treated as a simple queueing system, where the number of servers is determined with the number of active operators and the queueing discipline is FIFO.

The key element of the call centre optimization is the determination of the adequate number of active operators for different periods of the day. For this purpose the number of calls in different time periods of the day on a typical working week was analysed. The working day of the call centre was divided into the following four periods: from 8:00 AM to 10:00 AM, from 10:00 AM to 1:00 PM, from 1:00 PM to 6:00 PM and from 6:00 PM to 12:00 PM. The results of the analysis are presented in Figure 3.

The analysis showed that the number of incoming calls is significantly lower during the weekend than during the working days. Therefore, the weekend days were excluded from our further analysis. The number of calls in the particular period is similar for all working days. The exception is the last period on Tuesday. It was found out that this deviation was caused by the unexpected downtime of one of the main
provider’s services. The third period (1:00PM to 6:00PM) is the most frequent period, while the morning and evening periods are less burdened.

**Queueing model selection**

To select a suitable theoretical mathematical model to describe the call centre under consideration, the distribution of inter-arrival times and the distribution of service times have to be analysed. Figure 4 shows the frequency distribution of inter-arrival times while Figure 5 shows the frequency distribution of service times in different periods of a typical working day.

**Figure 4**

_Frequency distribution of inter-arrival times in different periods of a typical working day_

![Frequency distribution of inter-arrival times](image)

Source: Author’s illustration

We can conclude from Figure 4 that inter-arrivals times in all four periods fit the exponential density function. Besides, it can be seen from Figure 5 that probability density function of service times follows an asymmetric function (e.g. lognormal density function). When calls shorter than one minute are omitted, we can assume that also the distribution of service times can be described by the exponential density function in all four periods of a working day. Since these short calls do not cause queues and therefore do not threaten the efficiency of the call centre, our assumption is justified.

From the description of the call centre organization and from the analysis of arrival and service patterns we can conclude that the call centre under consideration can be described by the M/M/r {infinity/infinity/FIFO} queueing model.
Definition of performance measures
To analyse the call centre service quality different performance measures can be used. We restrict ourselves to performance measures that depend on customers’ waiting before their calls are answered. The following performance measures will be used:
- the expected waiting time (1)
- the expected number of waiting customers (2)
- the probability that the calling customer will have to wait (3)
- the call centre service level (4)

Optimization Results
The first step of the optimization is definition of the steady state conditions. From the field data the parameters $\alpha$ and $\sigma$ were estimated for a particular period of a working day. The value $r_{\min}$ needed to satisfy the steady state condition was determined according to (6). Results are given in Table 1.

<table>
<thead>
<tr>
<th>Period</th>
<th>$\alpha$ [min$^{-1}$]</th>
<th>$\sigma$ [min$^{-1}$]</th>
<th>$\alpha/\sigma$</th>
<th>$r_{\min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM - 10:00AM</td>
<td>0.847</td>
<td>0.336</td>
<td>2.53</td>
<td>3</td>
</tr>
<tr>
<td>10:00AM - 1:00PM</td>
<td>1.053</td>
<td>0.342</td>
<td>3.10</td>
<td>4</td>
</tr>
<tr>
<td>1:00PM - 6:00PM</td>
<td>0.877</td>
<td>0.356</td>
<td>2.46</td>
<td>3</td>
</tr>
<tr>
<td>6:00PM - 12:00PM</td>
<td>0.532</td>
<td>0.356</td>
<td>1.49</td>
<td>2</td>
</tr>
</tbody>
</table>

The next step is setting up the optimization requirement regarding the chosen performance measures. Considering four service measures we have selected, we define four optimization requirements.
The optimization requirement 1: The expected waiting time in a particular time period should not be longer than 20 seconds. The mathematical formulation of this requirement is:

\[ E(W_q) \leq 20 \text{sec} = 0.33 \text{min} \]

Optimization results assuming the requirement 1 are given in Table 2. In the first column of Table 2 the expected waiting time is calculated according to (1) and (5) considering the steady state condition \( r_{\text{min}} \) from Table 1. Then the minimal number of servers \( r \) needed to fulfill the stated requirement, and corresponding expected waiting time are determined by iteration. Results are given in the second and in the third column of Table 2.

Table 2
Results of the call centre optimization assuming the requirement \( E(W_q) \leq 20 \text{sec} = 0.33 \text{min} \)

<table>
<thead>
<tr>
<th>Period</th>
<th>( E(W_q) ) [min] - steady state</th>
<th>Minimal ( r ) to fulfil the requirement</th>
<th>( E(W_q) ) [min]</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM - 10:00AM</td>
<td>4,34</td>
<td>5</td>
<td>0,16</td>
</tr>
<tr>
<td>10:00AM - 1:00PM</td>
<td>1,72</td>
<td>6</td>
<td>0,11</td>
</tr>
<tr>
<td>1:00PM - 6:00PM</td>
<td>3,57</td>
<td>5</td>
<td>0,14</td>
</tr>
<tr>
<td>6:00PM - 12:00PM</td>
<td>3,55</td>
<td>4</td>
<td>0,08</td>
</tr>
</tbody>
</table>

Source: Authors’ work

The optimization requirement 2: The expected number of waiting customers in a particular time period should not exceed 1,5. The mathematical formulation of this requirement is:

\[ E(N_q) \leq 1,5 \]

Optimization results assuming the requirement 2 are given in Table 3. In the first column of Table 3 the expected number of waiting customers is calculated according to (2) and (5) considering the steady state condition \( r_{\text{min}} \) from Table 1. Then the minimal number of servers \( r \) needed to fulfill the optimization requirement 2, and corresponding expected number of waiting customers are determined by iteration. Results are given in the second and in the third column of Table 3.

Table 3
Results of the call centre optimization assuming the requirement \( E(N_q) \leq 1,5 \)

<table>
<thead>
<tr>
<th>Period</th>
<th>( E(N_q) ) - steady state</th>
<th>Minimal ( r ) to fulfil the requirement</th>
<th>( E(N_q) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM - 10:00AM</td>
<td>3,76</td>
<td>4</td>
<td>0,56</td>
</tr>
<tr>
<td>10:00AM - 1:00PM</td>
<td>1,82</td>
<td>5</td>
<td>0,41</td>
</tr>
<tr>
<td>1:00PM - 6:00PM</td>
<td>3,13</td>
<td>4</td>
<td>0,49</td>
</tr>
<tr>
<td>6:00PM - 12:00PM</td>
<td>1,89</td>
<td>3</td>
<td>0,23</td>
</tr>
</tbody>
</table>

Source: Authors’ work
The optimization requirement 3: It should be assured that at least 80% calling customers are served immediately, and will not have to wait to the operator. This implies that the probability of waiting should not exceed 20%:

\[ P_{\text{wait}} \leq 0.2 \]

Optimization results assuming the requirement 3 are given in Table 4. In the first column of Table 4 the probability of waiting is calculated according to (3) and (5) considering the steady state condition \( r_{\text{min}} \) from Table 1. Then the minimal number of servers \( r \) needed to fulfil the optimization requirement 3, and corresponding waiting probability are determined by iteration. Results are given in the second and the third column of Table 4.

<table>
<thead>
<tr>
<th>Period</th>
<th>( P_{\text{wait}} ) - steady state</th>
<th>Minimal ( r ) to fulfil the requirement</th>
<th>( P_{\text{wait}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM - 10:00AM</td>
<td>0.714</td>
<td>5</td>
<td>0.134</td>
</tr>
<tr>
<td>10:00AM - 1:00PM</td>
<td>0.543</td>
<td>6</td>
<td>0.110</td>
</tr>
<tr>
<td>1:00PM - 6:00PM</td>
<td>0.682</td>
<td>5</td>
<td>0.124</td>
</tr>
<tr>
<td>6:00PM - 12:00PM</td>
<td>0.639</td>
<td>4</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Source: Authors’ work

The optimization requirement 4: An industry standard for telephone service seems to be 80/20 rule, under which at least 80% of the customers must wait no more than 20 seconds (Koole and Mandelbaum, 2001). The mathematical formulation of this requirement is:

\[ SL(20\text{sec}) = SL(0,33\text{min}) \geq 0.8 \]

Optimization results assuming the requirement 4 are given in Table 5. In the first column of Table 5 the service level \( SL(20\text{sec}) \) considering the steady state condition \( r_{\text{min}} \) from Table 1 is calculated according to (4) and (5). Then the minimal number of servers \( r \) needed to fulfil the optimization requirement 4, and corresponding service level is determined by iteration. Results are given in the second and in the third column of Table 5.

<table>
<thead>
<tr>
<th>Period</th>
<th>( SL(20\text{sec}) ) - steady state</th>
<th>Minimal ( r ) to fulfil the requirement</th>
<th>( SL(20\text{sec}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00AM - 10:00AM</td>
<td>0.323</td>
<td>5</td>
<td>0.898</td>
</tr>
<tr>
<td>10:00AM - 1:00PM</td>
<td>0.511</td>
<td>6</td>
<td>0.921</td>
</tr>
<tr>
<td>1:00PM - 6:00PM</td>
<td>0.360</td>
<td>5</td>
<td>0.908</td>
</tr>
<tr>
<td>6:00PM - 12:00PM</td>
<td>0.398</td>
<td>3</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Source: Authors’ work
Discussion
The arrival and the service patterns of Slovenian telecommunication provider’s call centre were analysed. It was established that the call centre under consideration can be described by the \( M/M/r \{\infty/\infty/FIFO\} \) queueing model. Four performance measures, dependent on customers’ waiting, were used to analyse the call centre service quality. Considering the selected performance measures, four optimization requirements were stated as follows:
- the expected waiting time should not exceed 20 seconds,
- the expected number of waiting customers should not exceed 1.5,
- at most 20% calling customers will have to wait to the operator, at least 80% of them should be served immediately,
- at least 80% of the customers should wait no more than 20 seconds.

The aim was to determine the minimal number of servers in a particular period of a working day to fulfil the stated requirement. Results obtained prove that all performance measures can be applied to practice rather easily. Especially useful seem to be the probability that the calling customer will have to wait (so called Erlang C function) and the service level which is the most common measure of the call centre service quality.

The optimization results requirements showed that the number of servers (operators) needed to satisfy the steady state condition is relatively low (from two to four operators in the particular time period). This ensures that the queueing system still operates after a long run without forming infinite queues. However, if the call centre under consideration wants to improve the quality of its service, the number of the operators has to be raised. For each time period, two additional operators have to be employed.

In the first period, three out of four optimization requirements identified the minimum number of operators to be five. Further, the second period, being the most occupied, six operators are needed to satisfy three out of four requirements. In the third period, again three out of four requirements showed that the minimum number of operators needed is five. The fourth, night period has the lowest frequency of calls. Therefore, in two out of four requirements the minimum number of operators proved to be four.

Considering the presented results, the call centre can operate at a satisfying service quality level during the weekdays with 10 operators. A typical working day should employ three shifts: a) five operators for the shift from 8:00 AM to 4:00 PM; b) four operators for the shift from 4:00 PM to 24:00 operators; and c) one operator for the shift from 10:00 AM to 6:00 PM.

Conclusion
The paper represents a case study of a call centre performance optimization using the queueing theory approach. The research was conducted on the case of a Slovenian telecommunication provider’s call centre. Using the field data on call centre operation it was found out that the call centre under consideration can be described by the \( M/M/r \{\infty/\infty/FIFO\} \) queueing model. Four performance measures, dependent on customers’ waiting, were used to analyse the call centre service quality. The minimal number of servers needed to fulfil the stated requirements in a particular period of a working day was defined.
Results obtained prove that stochastic queueing models represent a useful tool for a call centre performance optimization. Since all the data needed for mathematical analysis are usually available, implementation of such models is rather simple while the information they provided is of great importance. Namely, determination of the adequate number of active operators regarding a specific performance measure is a preliminary condition to ensure the optimal service level and therefore the minimal cost of the queueing system performance.

Recent studies demonstrate that customer waiting time is not the only measure for the service level quality. As stated by Aksin et al. (2007), customers tend to place a high value on other dimensions of their experience (first call resolution, perceived agent competency, politeness or friendliness). Hence, if a need to model service quality consistent with concurrent customer values occurs, the function for optimal service level should be reconsidered. Further research on scheduling and schedule adjustments, as for example presented by (Mehrotra et al., 2010) has to be conducted to efficiently define the personnel schedules for the entire time of call centre operation.

Discrete event simulation is also a viable option for accurate performance modelling and subsequent decision support. Some authors, e.g. Akhtar and Latif (2010) argue that the analytical approach is not accurate enough, as it does not mimic randomness. In future research we will simulate the presented case with a discrete event simulation tool, where for describing the probability density function of service times an asymmetric function will be used. In addition we will try to involve other performance measures considering abandonment and retrials.

References


About the authors
Alenka Brezavšček is an assistant professor at the Faculty of Organizational Sciences, University of Maribor. Her research fields are stochastic processes, system reliability and availability, and information system security. The author can be contacted at: alenka.brezavscek@fov.uni-mb.si

Alenka Baggia is employed as a teaching assistant at the Faculty of Organizational Sciences, University of Maribor. Her main research fields are resource scheduling and simulation. The author can be contacted at: alenka.baggia@fov.uni-mb.si