SYSTEMS & MANAGEMENT
Electronic magazine

MANAGEMENT OF QUEUES ASSISTED BY TWO SERVERS WITH DIFFERENT ASSISTANCE RATES

Fabio Favaretto
fabio.favaretto@unifei.edu.br
Federal University of Itajubá – UNIFEI, Itajubá, Minas Gerais, Brazil

ABSTRACT

The Queue Management literature deals with queues with more than one server where everyone has the same efficiency and the indicators used are difficult to obtain in practice. In a large part of real situations this situation is not observed, since it is natural that the people who play the role of servers act with different incomes. The objective of this work is to present a way of managing a system with two servers, which have different efficiencies and the indicators used are obtained in a practical way. A visual management method was proposed for a service system in which the manager needs to decide the moment of opening a second server to assist a queue and the resource that will be used, considering its efficiency. The results allow a practical management of this situation, with indicators that are easy to obtain and greater control over the need to open new servers.

Keywords: Queue Management; Service Management; Simulation.

PROPPI / LATEC
DOI: 10.20985/1980-5160.2018.v13n1.825

1. INTRODUCTION

The management of service provision has a process that is important and directly associated with customer satisfaction, which is queuing management (Fitzsimmons et Fitzsimmons, 2006). A queue is formed in a service delivery system whenever a service point (called server) is assisting a customer and there are one or more clients to be served. If it is desired that there is no waiting or that this is the smallest possible, the obvious solution is to open another server. On the other hand, a server that assists few clients or that is idle for some time burdens the service providing system. In such case is necessary to find a balance between fast service and low operational cost. Thus, it is necessary to find a balance between the rapid assistance and the system operating cost.

The Queuing Theory is a field of quantitative research that proposes the relation between several indicators. The indicators usually associated with the level of customer satisfaction are the average waiting time, the average queue size and the probability of there being n clients in the queue at any given time (Morabito *et* Lima, 2000; Chwif *et* Medina, 2006; Bouzada, 2009; Camelo *et al.*, 2010). Obtaining these indicators presupposes a constant data collection in the productive system, and this process can be complex and require dedicated resources for this purpose. In practical situations, the use of these indicators can be prevented due to these limitations, leaving to the manager the decision based mainly on his feeling or on rules devoid of conceptual and/or quantitative basis.

The problem to be addressed in this research is the definition of the moment when a service delivery system, serviced by a single server, must have a second server in order to perform the attendance of a queue. Moreover, among the human resources available for the operation of this server, which one should be chosen? These definitions are expected to increase the quality of the service provided by decreasing customer waiting time and queue size.

The main objective of this work is to present a way of managing a system with two servers that have different efficiencies and the indicators used are obtained in a practical way. For this, some indicators will be proposed that will allow the manager to decide whether or not to open a second server in different service and demand conditions. The different rates represent different efficiencies between available human resources.

The expected result of this work is a set of practical managerial implications for managers of queued systems. This paper presents a contribution to make the process of data collection for the management of queues and the use of mathematical tooling practical. The strict application of the quantitative methods requires data such as the average time spent in the queue, and in order to obtain it, it is necessary to monitor the clients individually, record the times of entry and exit in the queue of each customer and calculate the resulting average. Thus, obtaining this data in real time, especially in non-computerized service delivery systems, may make it impossible to make fast and efficient decisions.

This article is structured in the following way: after this introduction, a brief conceptual reference on the management of queues is addressed. In the following section an analysis of the indicators used in this type of management is made. Following is a description of the actual system that served as the basis for this study, as well as the simulation performed. Finally, we present the results obtained in the simulation, the proposed management implications and the conclusion of this work.

2. QUEUE MANAGEMENT

Queue Management is a Service Operations Management process. One of the activities of this process is to manage th indicator of customer waiting time (Fitzsimmons *et* Fitzsimmons, 2006). This indicator influences the overall customer satisfaction of a service system. Hwang *et* Lambert (2009) argue that in a perfect system there would be no waiting; however, in practice, it is common for customers to wait in a queue until the resource is available. Usually, administrators' actions are focused on decreasing the waiting time when the system is overloaded with large queues, and a common action in this regard is the use of additional servers, according to Stolletz *et* Manitz (2013).

According to Jones et Peppiat (1996), the first option for service system managers is to design an operation in such a way that the waiting time is as small as possible, to the point where the cost increase is greater than the value added by small waits. Another type of action used is to divide the clients into different classes, with specific queues for service (Alotaibi *et* Liu, 2013).

Houston et al. (1998) present a series of considerations about the perception of waiting time in a queue and the evaluation of the service performed by the consumer. One of these considerations is that a long waiting time may influence more on a negative evaluation than poor service. Another consideration made by the authors, for service managers is that, reducing the time of service and the perception of waiting time (through parallel activities or items for distraction) causes less impact

Electronic Journal of Management & System
Volume 12, Number 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

than the de facto decrease of the waiting time. When a long queue is required, an explanation of why or some form of apology diminishes the negative rating. Still, for these authors, the perception of *unnecessary* waiting, motivated by the vision of employees or unemployed resources or in activities that do not contribute to the service, strongly impacts on the negative evaluation of the service provided.

3. INDICATORS USED FOR QUEUE MANAGEMENT IN THIS WORK

In order to meet the objective of this paper, it is necessary to use indicators to manage the queues. It is intended to provide a set of applicable observations to service systems managers where queuing occurs and, for this, these indicators need to be meaningful and practical. Significant so that, with few indicators, managers have an overview of the situation and the possibilities of the queues. It is practical in order to allow them to obtain easily and without the need for great effort in terms of collecting and processing data.

The quantitative study of the queues is based on the relations between some parameters, mainly the arrival rate to the system (λ) and the assistance rate (μ). The ratio of the arrival rate divided by the attendance rate shows the occupation of the system, represented by ρ (Chwif *et* Medina, 2006). An arrival rate that is greater than the assistance rate clearly shows that the system will not be able to meet the arrivals, generating queues that never decrease while maintaining the rates. It is said that a system is stable or in a permanent state when the system occupancy is between 0 and 0.8 (Chwif *et* Medina, 2006). A stable system may have queues; however, these queues will remain the same size at worst. In the case of systems with more than one server, their quantity must be considered. Equation 1 (Chwif et Medina, 2006; Fitzsimmons et Fitzsimmons, 2006) calculates the occupation (ρ) of the system, where λ is the *average arrival rate* (number of arrivals per time interval), μ is the *assistance average rate* (number of calls per time interval) and *c* is the number of servers in the system. All the distinguished servers are considered with the same efficiency or average service rate.

$$\rho = \frac{\lambda}{c\mu} \tag{1}$$

The probability of *n* customers in the system (P_n) is another indicator used by Chwif et Medina (2006) and Fitzsimmons et Fitzsimmons (2006). This indicator considers all the clients in the system, both in-service and in-line. As the focus of this work is queue management, an indicator will be used that shows the probability of not queuing at the time of arrival of a new client, which is equivalent to the *arrivals served without waiting*. For this to happen on a system with two servers, at least one of them must be vacant. The use of this indicator is justified because it is of particular interest to the new customer, who always wants to be served without waiting.

Another indicator used in this work is the number of customers in queue upon arrival. Chwif *et* Medina (2006) and Fitzsimmons *et* Fitzsimmons (2006) present the average row size indicator (L_q), considered as an average number of people in the queue throughout the period considered. Obtaining the L_q indicator from observations of the real system is impractical, since it depends on constant counts. The proposed indicator - *average queue size at arrival* (L_{qa}) can be obtained in a relatively more practical manner by counting queue size at the time each new customer arrives at the system. In addition, the queue upon arrival is of particular interest to the arriving customer as it defines their perception of waiting.

The simulation used seeks to reproduce a service delivery system where there is usually queuing to pay for the service used in a cashier. There is always an open cashier (called the first server), and it is possible to open a second cashier. The first server is operated by a standard clerk, trained for the function. The second station can be operated by another trained attendant, but the availability of such a resource is not always available. If there is no attendant trained specifically for this function and the system manager wishes the second station to be opened, the available operator will be used, even if it is less efficient than the standard. The variation in terms of efficiency equals a second server who is roughly able to perform the task. A second server with 0.5 (or 50%) efficiency equates to a resource with less training or aptitude, and would spend double the time compared to the reference server performing the same task. Thus, the queuing discipline implies that the first server is always preferred for service and the second server is only used when the first server is busy. If both servers are free, the priority is the first. If both servers are busy there will necessarily be a wait and consequently the formation of a queue.

Based on the above considerations, the *received arrivals indicator* will be used *without waiting*, equivalent to the probability of a customer arriving at the system and not finding queue - with the notation P_0. This indicator is aimed at the manager to define a level of service or level of assistance, through immediate customer service. This indicator is formed by the proportion (in percentage) of the clients served without waiting for the total number

Electronic Journal of Management & System
Volume 12, Number 1, 2018, pp. 2-9
DOI: 10.20985/1980-5160.2018.v13n1.825

of clients, as seen in Equation 2. It is proposed to use this indicator because it is not an average, facilitating its obtaining.

$$P_0 = \frac{Number\_of\_clients\_assisted\_without\_wait}{Total\_number\_of\_clients\_assisted} \quad (2)$$

4. SIMULATED ENVIRONMENT

The simulated situation portrays the service system of a university restaurant. In this system, data were collected on the time of service and the intervals between arrivals, which allowed the calculation of the average arrival and attendance rates. These data were collected through observations, timekeeping, photographs and filming during the years of 2013 and 2014. The individual values allowed to analyze the probability distribution with better adherence and, for both cases, the suggested distribution (Morabito *et* Lima, 2000) was the exponential of probability. Observations were stratified to cover several days of the week, periods of the month and different months. This allowed obtaining representative averages of the whole period.

The average service time measured was 30 seconds, which is equivalent to an average service rate (µ) of 0.033 clients per second. This time did not show significant variations in different situations. The interval between the arrivals presents great variations during the day, related to the intervals between classes, periods of meals and beginning and end of work of university employees. By analyzing the data, we can distinguish periods with different level of movement, resulting in three situations: small, medium and large movement of the restaurant. The intervals between the arrivals and the average arrival rates are presented in Table 1, which also shows the occupation of the system (ρ) for each situation, considering only one active server and the average attendance rate presented above. These data were obtained in face-to-face observations and with the use of counts and timings. The period analyzed comprises the months between April and October 2014. In addition to the face-to-face observations, the author obtained data observing the daily videos of the security circuit, made available by the administration.

Table 1 describes the motion situations used in this work. The situation of great movement corresponds to the peak times, corresponding to the lunch schedule and the class intervals. At these times the largest queues occur, with the average arrival time of 30 seconds. In addition, the system occupancy, if only one server is used for the service, is 1 (one), which means an unstable system.

At less busy times queues rarely occur, since the 120-second interval between arrivals practically guarantees immediate service to the system. The occupation of the system at 0.25 depicts this situation.

5. SIMULATION

Three simulation scenarios were created, one for each of the motion situations presented in Table 1. For each scenario, simulations were performed considering one or two servers, and in situations with two servers efficiency will be varied. The efficiency variation of the second server was from zero (equivalent to a single server) to 1 (equivalent to a second server with the same efficiency of the reference server), with increment of 0.1.

The *occupation* (ρ) of the system was calculated through Equation 3, where λ is the *mean rate of arrival* (number of arrivals per time interval) and µ is the *average attendance rate* (number of arrivals per time interval). The denominator is an adaptation of Equation 1 above, in which the efficiency of the first server is equal to 1, because it is the reference, and the efficiency of the second server (in percentage) is represented by E_2.

$$\rho = \frac{\lambda}{(1+E_2)\mu} \quad (3)$$

Each scenario was simulated with 5000 replications, using Crystal Ball software, and the results are presented below. Each replication shows a period followed by one hour in each of the movement situations and con-

Table 1. System data analyzed

Motion Situation	Interval between arrivals (seconds)	Average arrival rate (arrivals / seconds)	System occupation with a server (ρ)
Big	30	0,033	1
Medium	50	0,02	0,6
Small	120	0,00833	0,25

Source: The author

siders that the average arrival rate follows the Exponential Probability Distribution. The independent variable used was the efficiency of the second server. In each scenario a variation of the efficiency of the second server was made, with values between zero (when only the first server is available) and 1 (when the second server is equal to the standard), with a variation of 0.1.

The analyzed dependent variables are presented in the sequence. The first variable analyzed is the average wait in seconds, which shows the average wait times of all users. This variable is used only as a reference, especially for the system manager. The second variable used is the probability that the user will be immediately assisted after arriving at the system or finding a queue of zero size (P_0), according to Equation 2. The last dependent variable analyzed is the average queue size at the time of arrival, which considers an average of the existing queue size at the time of each user's arrival.

6. RESULTS

The results obtained with the simulations are presented below. In all of them, the attendance rate (μ) of the first constant server, with a value of 0.033 clients per second, was considered.

In the small-motion situation, the average arrival rate (λ) is a random variable that follows the Exponential Distribution with rate 0.0083 arrivals per second. The results obtained for this scenario are presented in Table 2.

Table 2. Simulation results for the small-motion scenario.

Number of servers	Efficiency of the second server	Avg wait (seconds)	P_0	Average queue size upon arrival
1	0	9,8	0,75	0,08
2	0,1	3,56	0,9	0,03
2	0,2	2,17	0,93	0,02
2	0,3	1,6	0,94	0,01
2	0,4	1,26	0,95	0,01
2	0,5	1,04	0,96	0,01
2	0,6	0,85	0,96	0,01
2	0,7	0,72	0,97	0,01
2	0,8	0,63	0,97	0,01
2	0,9	0,55	0,97	0,01
2	1	0,48	0,97	0,01

Source: The author

With only one server assistance, the average wait was 9.8 seconds and the probability of no queuing at the time of arrival of a new user (P_0) is 0.75 or 75%. With the addition of a second server, even with only 0.1 or 10% of the efficiency of the first server, there is a significant reduction in the average wait time to 3.56 seconds. The P0 is increased to 0.9 or 90%. The average wait as a function of the efficiency of the second server can be seen in Figure 1. The average wait is appreciably reduced by increasing the efficiency of the second server to values close to 0.5 or 50%. From this point, increasing the efficiency of the second server allows a small reduction in the average waiting time.

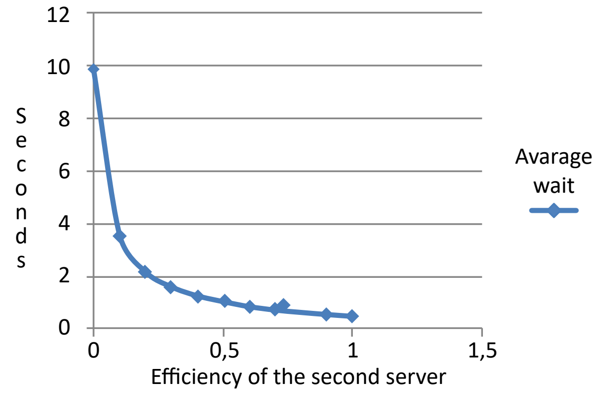

Figure 1. Relation between the average waiting time and the efficiency of the second server in case of small movement.

Source: The author

In a small-motion situation, the wait for service with only one server is not significantly high, and is on the order of 10 seconds. Even so, opening a second server with half the efficiency of the first server can lead to average wait times of approximately 1 second and a 96% chance of no queue on arrival of a new user.

In the average moving situation, the average arrival rate (λ) is a random variable that follows the Exponential Distribution with a rate of 0.02 arrivals per second. The results obtained in the simulation can be seen in Table 3.

With only one server attending, the average wait was 40.44 seconds, and the probability of no queuing on arrival of a new user (P0) is 0.42 or 42%. With the addition of a second server with 0.1 or 10% of the efficiency of the first server, there is a reduction in the average wait time to 25.44 seconds. The P0 is increased to 0.57 or 57%. The average wait due to the efficiency of the second server can be seen in Figure 2.

Table 3. Results of the mean motion simulation.

Number of servers	Efficiency of the second server	Avg wait (seconds)	P_0	Average queue size upon arrival
1	0	40,44	0,42	0,7
2	0,1	25,24	0,57	0,49
2	0,2	16,9	0,66	0,33
2	0,3	12,14	0,71	0,24
2	0,4	9,18	0,75	0,18
2	0,5	7,19	0,78	0,14
2	0,6	5,7	0,81	0,11
2	0,7	4,76	0,82	0,09
2	0,8	4,05	0,84	0,08
2	0,9	3,42	0,85	0,07
2	1	3	0,86	0,06

Source: The author

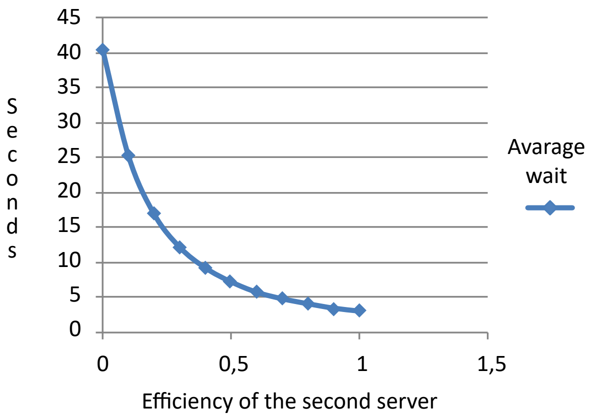

Figure 2. Relation between the average wait time and the efficiency of the second server for the medium-motion case.

Source: The author

In the situation of great movement, the average arrival rate (λ) is a random variable that follows the Exponential Distribution with a rate of 0.033 arrivals per second, corresponding to the average arrival of a customer every 30 seconds. As the average service time is of the same order, we have that the occupation of the system (ρ) is equal to 1, which indicates an unstable system with the queue sizes tending to infinity. This is the situation that requires more attention. In some situations observed in the real system, there were queues of more than 40 people with waiting times exceeding 15 minutes. The results obtained in the simulation of this scenario are presented in Table 4.

Table 4. Results of the large motion simulation.

Number of servers	Efficiency of the second server	Avg wait (seconds)	P_0	Average queue size upon arrival
1	0	198,7	0,11	5,35
2	0,1	128,6	0,19	3,88
2	0,2	81,35	0,28	2,54
2	0,3	54,34	0,36	1,73
2	0,4	39,18	0,43	1,26
2	0,5	28,68	0,49	0,93
2	0,6	22	0,54	0,71
2	0,7	17,55	0,58	0,57
2	0,8	14,21	0,62	0,47
2	0,9	11,39	0,65	0,37
2	1	9,49	0,67	0,31

Source: The author

With only one server serviced, the average wait was 198.7 seconds and the probability of no queuing on arrival of a new user (P0) is 0.11 or 11%. With the addition of a second server with 0.1 or 10% of the efficiency of the first server, there is a reduction in the average wait time to 128.6 seconds. The P0 is increased to 0.19 or 19%. The average wait due to the efficiency of the second server can be seen in Figure 3.

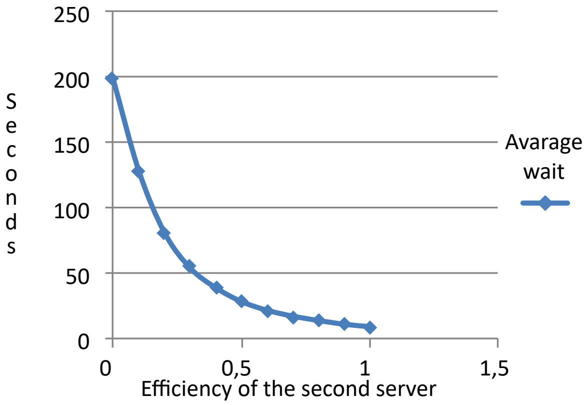

Figure 3. Relation between the average wait time and the efficiency of the second server for the case of great movement.

7. MANAGEMENT IMPLICATIONS

As observed in Table 4, the most critical situation is in the case of large movement, in which only one server is in service. In this case, an average queue of 5.35 customers is expected, which for practical purposes is approximated to 6 clients. Thus, when the average queue reaches this mark and the system is in the period of great movement, the occupation of the system (ρ) is equal to 1 and this means that the system is unstable. Thus, the existence of 6 clients, on average, in the queue marks the need to open a new server. For this, a delimitation of the queue can be made using poles and chains that force the formation of a straight line where necessarily a customer is behind the client that arrived just before him. With an evaluation of the distance occupied by customers in this queue, a mark can be made on the floor that should be located between the sixth and seventh customer in the queue. While the queue is before the end of the mark, there are six or fewer customers. When the queue goes beyond the mark it is because the seventh customer is in the queue. Because it is a decision based on the average queue size, it can eventually exceed this limit in situations with $\rho<1$. For this reason, the manager must observe the queue for some time or a few times. If a queue with more than 7 clients is confirmed, it is necessary to open a new server.

The opening of a second server can be done using the most efficient human resource available. The lowest efficiency observed among the human resources available in the analyzed system is in the order of 20% (0.2) of the standard server efficiency. In this case, observing Table 4, it is verified that the average size of the expected queue is 2.54 clients, and the value 3 clients is used. There should be another mark in the queue between the third and fourth customer. When the system operates with two servers and this mark is exceeded, it means that a server with more efficiency is needed. In this case, the manager should replace the second human resource with one more efficient.

8. CONCLUSIONS

This work presented a calculation of the occupancy rate (ρ) of a service delivery system with two servers where their efficiency is not necessarily the same. This is a contribution to the area, since the problem with two servers is widely handled, considering the servers with the same efficiency. When studying and understanding a real system, a computational simulation model was created to reproduce cases of small, medium and large movements. The results obtained allowed understanding the situation in which it is necessary to open a second server in situations of great movement. A practical rule has been presented so that the system manager can easily identify this situation. A practical rule has also been presented to identify whether it is necessary to replace the second server with a more efficient one. It is concluded that it is possible to develop effective practical managerial actions when using the application of Queue Theory knowledge, guaranteeing the agility in the decisions necessary to the system manager.

9. ACKNOWLEDGEMENTS

The author is grateful for the support received from CNPq and FAPEMIG.

REFERENCES

Alotaibi, Y., Liu, F. (2013), “Average waiting time of customers in a new queue system with different classes”, Business Process Management Journal, Vol. 19, No. 1, pp. 146-68.

Bouzada, M. A. C. (2009) “Dimensionamento de um call center: simulação ou Teoria das Filas?”, Anais... SIMPOI 2009: Simpósio de Administração da Produção, São Paulo, SP.

Camelo, G. R., Coelho, A. S. et al. (2010) “Teoria das filas e da simulação aplicada ao embarque de minério de ferro e manganês no terminal marítimo de ponta da madeira”, Cadernos do IME, Vol. 29, disponível em: http://www.e-publicacoes.uerj.br/index.php/cadest/article/view/15733/11904 (acesso em 18 jan. 2018).

Chwif, L., Medina, A. C. (2006) “Uma análise crítica da Lei Municipal 13.948 ou ‘Lei das Filas’ sob a ótica da Pesquisa Operacional: conclusões derivadas de modelos de simulação de eventos discretos”, Anais... XXVI ENEGEP – Encontro Nacional de Engenharia de Produção, Fortaleza, CE, 2006.

Fitzsimmons, J. A., Fitzsimmons, M. J. (2006), Service management: Operations, strategy, and information technology, 5th ed., McGraw Hill, New York.

Hwang, J., Lambert, C. U. (2009), “The use of acceptable customer waiting times for capacity management in a multistage restaurant”, Journal of Hospitality & Tourism Research, Vol. 33, No. 4, pp. 547-61.

Houston, M. B., Bettencourt, L. A. et al. (1998), “The relationship between waiting time in a service queue and evaluations of service quality: a field theory perspective”, Psychology & Marketing, Vol. 15, No. 8, pp. 735-753.

Jones, P., Peppiat, E. (1996), "Managing perceptions of waiting times in service queues", International Journal of Service Industry Management, Vol. 7, No. 5, pp. 47-61.

Morabito, R., Lima, F. C. R. (2000) "Um modelo para analisar o problema de filas em caixas de supermercados: um estudo de caso", Pesquisa Operacional, Vol. 20, No. 1, pp. 59-71.

Stolletz, R., Manitz, M. (2013), "The impact of a waiting-time threshold in overflow systems with impatient customers", Omega, Vol. 41, No. 2, pp. 280-86.

Received: Sept 10, 2014

Approved: Jan 15, 2018

DOI: 10.20985/1980-5160.2018.v13n1.825

How to cite: Favaretto, F. (2018), "Management of queues assisted by two servers with different assistance rates", Sistemas & Gestão, Vol. 13, No. 1, pp. 2-9, available from: http://www.revistasg.uff.br/index.php/sg/article/view/825 (access day month. year).