Application of Queuing Theory in Public Service Counters

Jiayi Ji [1, *]

[1] Xi'an Jiaotong-liverpool University, Xi'an, 215000, China
*Jiayi.Ji21@student.xjtlu.edu.cn

Abstract. Queuing theory is a mathematical field concerned with the dynamics of waiting lines, or queues. Its applications are pivotal in analyzing and optimizing systems where the timing of customer arrivals and service is of utmost importance, such as in banks, hospitals, airports, and call centers. By employing queuing theory, system designers can craft service setups that are more efficient and effective, consequently reducing waiting times, boosting customer satisfaction, and containing operational costs. Public service counters, which offer a variety of services including passport applications, business transactions, and consultancy services, stand as an exemplary setting for the application of queuing theory. These counters often grapple with high demand, limited resources, and complex procedures. Without a competent management system in place, they can easily succumb to issues like service interruptions, unsatisfactory customer experiences, and low efficiency. The essential goal when tackling queuing challenges is to balance the costs of waiting against the expenditures associated with increasing resources. In such a scenario, the implementation of queuing theory proves invaluable. It enhances both the performance and service quality at public service counters, affirming the theory's significant role in public service management.

Keywords: Queuing Theory, Public Service Counters, Service Systems

1 Introduction

The primary objective of this paper is to delve into the application of queuing theory in public service counters, with a particular focus on two common models: the single-server model and the multiple-server model. The basic functionality of these models will be elucidated and their applicability across various public service counter scenarios will be assessed. The investigation will also scrutinize the influence of the second-service rate, which represents the likelihood of a customer requiring an additional service after the initial one. To evaluate the advantages and shortcomings of applying queuing theory in public service counters, a selection of numerical examples and case studies will be utilized [1].

The paper unfolds in a structured manner, first introducing fundamental aspects of queuing models, such as the arrival pattern, service pattern, queue discipline, system capacity, and service channel. It then proceeds to discuss the implementation of the

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_52

single-server model in public service counters, supplemented by a relevant case study. This is followed by an exploration of the multiple-server model's application in similar settings, once again illuminated by a case study. The penultimate section analyzes the impact of the second-service rate on the performance of public service counters, providing an illustrative example for clarity [2]. The paper concludes by summarizing the findings and proposing avenues for future research in the domain.

2 Basic Features of Queuing Models

Queuing models are mathematical tools that capture and analyze the dynamics of waiting lines or queues. These models provide a deep understanding of how various factors can affect the performance of a service system [3]. Key variables typically include the average arrival rate (λ), the average service rate (μ), and the average response time ($E[T]$). While there are numerous types of queuing models, each tailored to the assumptions and characteristics of different service systems, common features tend to emerge across the spectrum. These shared attributes often fall into categories such as the arrival pattern, the service pattern, and the queue discipline. By studying these categories, one can compare and contrast the different queuing models to better understand their application to particular service environments [4].

2.1 Arrival Pattern

The arrival pattern represents the way customers approach the service system. This typically involves an examination of the distribution of interarrival times, which are the periods between consecutive customer arrivals. These interarrival times can follow various probability distributions, including but not limited to Poisson, exponential, and normal distributions. The most frequent scenario is that the interarrival times are independently and identically distributed. This suggests that arrivals are random and adhere to a Poisson process [5].

Alongside this, it's important to consider the total number of customers and the structure of arrival batches. The customer pool can be either finite or infinite. Customers might arrive individually or in groups. For instance, in the case where each customer arrives independently, the batch size is simply one. However, if customers come in groups, the batch size can differ accordingly, introducing further variability into the system [6].

2.2 Service Pattern

The service pattern refers to how customers are served by the service system. It can be described by two aspects: the average service times and the quantity of servers.

The average time it takes to serve each customer is according to an exponential distribution, which means that the service times are random and memoryless 7.

The quantity of servers is the number of service facilities that can serve customers simultaneously. It can be constant or variable. Therefore, single-server model means

there is only one server in the system and multiple-server model means there are more than one server in the system. In addition, multiple services can be parallel, string, mixed arrangement 8.

2.3 Queue Rule

The common queue rules are first-come first-served (FCFS), last-come first-served (LCFS), Random services (RSS), shortest processing time (SPT), priority service (PS), etc. The most common assumption is that the queue rule is FCFS, which means that customers are served depends on the order of arrival.

2.4 Queuing System

Different queuing systems can be formed from the above basic features [3]. As shown in Figures 1-4.

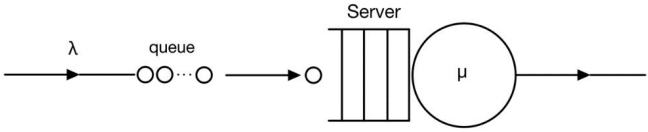

Fig. 1. Single-server model

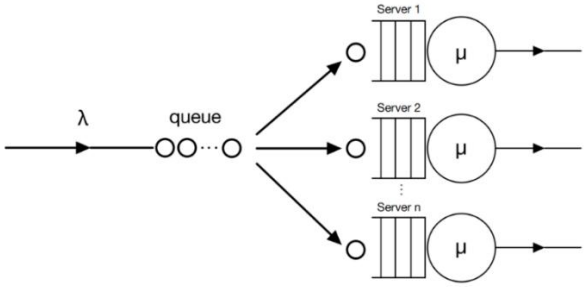

Fig. 2. Multiple-server model 1

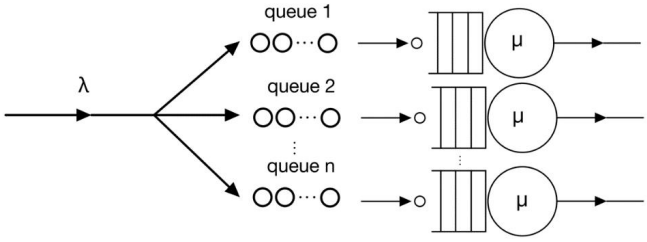

Fig. 3. Multiple-server model 2

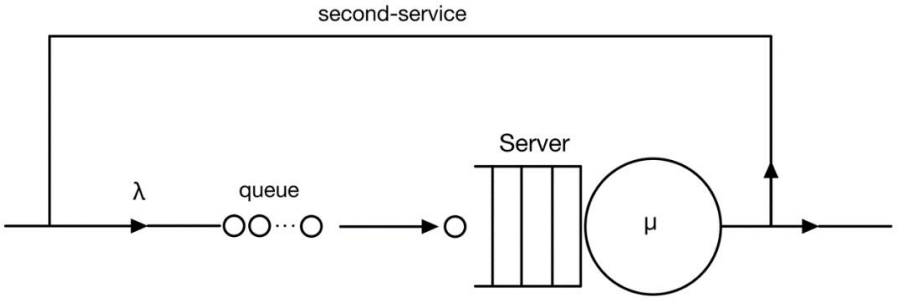

Fig. 4. Second-server model 1

3 Application of Single-server Model (M|M|1 Model) in Public Service Counters

3.1 Example 1

There are four chairs for people to line up for a haircut. When all four chairs are filled, subsequent customers leave without entering the store. Customers arrive at an average rate of 4 people per hour, and haircuts take an average of 10 minutes per person. Let the arrival process be Poisson distribution and the service time obey negative exponential distribution. It is easy to calculate the data in the Table 1.

Table 1. Summary of Key Performance Metrics for the Queuing System

Performance Measure	Result
Overall system utilization	63.46%
Average number of customers in the system $\boldsymbol{L}$	1.42
Average number of customers in the queue $\boldsymbol{L_q}$	0.79
Average number of customers in the queue for a busy system $\boldsymbol{L_b}$	1.24
Average time customer spends in the system $\boldsymbol{W}$	0.37 hours
Average time customer spends in the queue $\boldsymbol{W_q}$	0.21 hours
Average time customer spends in the queue for a busy system $\boldsymbol{W_b}$	0.33 hours
The probability that all servers are idle $\boldsymbol{P_0}$	36.54%
The probability an arriving customer waits $\boldsymbol{P_W}$ or system is busy $\boldsymbol{P_b}$	63.46%

Obviously, just knowing these data gives you an idea of the utilization of the system and how busy it is. The following example will focus on how to balance server costs and customer wait times for optimal utilization.

3.2 Example 2

Numerous businesses adopt a combined strategy for their refueling and car washing operations. Free car washes are provided for vehicles that get a full tank of gas, while

a charge of $0.50 is applied for vehicles that require only a car wash without fueling up. Based on surveys, the number of customers who opt for both fuel and car wash is roughly the same as those who seek just a car wash [9].

The average cost of refueling is $0.70, and a car wash costs $0.10. The robot designed to perform these services operates for 14 hours each day. It has three power and drive levels: Level A can wash a car every five minutes and costs $12 per day to operate; Level B can wash a car every four minutes and costs $16 per day; and Level C can wash a car every three minutes and costs $22 per day. Given that each customer prefers not to wait more than five minutes for a car wash, an extended waiting time may cause the company to lose customers. Let's consider a scenario where 10 customers arrive at the car wash every hour. The question is which robot offers the best service. If waiting time is the sole deciding factor, then Robot B should be the choice. However, a company must compare the two robots' profitability before making a final decision. For Robot A, the waiting time extends to 12.5 minutes, which might deter some customers from availing the service. A decrease in revenue can be anticipated if Robot A is selected. The arrival rate can be determined by increasing t_l = 5 minutes (average customer waiting time), which will yield the highest customer arrival rate efficiency.

Consequently, considering the initial estimated arrival rate (λ) of 10 people per hour, two customers will be lost per hour. The daily loss can be calculated as follows: 2 customers/hour x 14 hours x 1/2 (0.7+0.4) = $15.40. The daily cost increment for choosing Robot B is just $4, which is significantly lower than the $15.40 loss incurred with Robot A. Thus, Robot B, meeting the initial 5-minute wait limit, is the better choice, while Robot C can be disregarded unless a substantial increase in the arrival rate is expected.

4 Application of Multiple-server Model (M|M|S Model) in Public Service Counters

4.1 Example 1 （M|M|3|∞）

The ticket office has three Windows, and the arrival of customers is Poisson flow, with an average arrival rate of 0.9 people per minute; Service times conform to a negative exponential distribution, with an average service rate of 0.4 people per minute 10. When customers arrive, they form a queue and purchase tickets in turn at an open window. p_n = P{N = n} (n = 0, 1, 2, …) is the probability distribution of length N after the system reaches the equilibrium state. λ_n = λ, n = 0, 1, 2, … and. Assume $\boldsymbol{\rho_s} = \frac{\boldsymbol{\rho}}{\boldsymbol{s}} = \frac{\boldsymbol{\lambda}}{\boldsymbol{s}\mu}$, when $\boldsymbol{\rho_s} < 1$, get that

$$C_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!}, & n = 1,2,\ldots,s \\ \frac{(\frac{\lambda}{\mu})^s}{s!} = (\frac{\lambda}{s\mu})^n = \frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}, & n \geq s \end{cases} \quad (1)$$

and

$$p_n = \begin{cases} \frac{(\rho)^n}{n!} p_0, & n=1,2,\ldots,s \\ \frac{(\rho)^n}{s!s^{n-s}} p_0, & n \geq s \end{cases} \quad (2)$$

$$p_0 = [\sum_{n=0}^{s-1} \frac{\rho^n}{n!} + \frac{\rho^s}{s!(1-\rho_s)}]^{-1} \quad (3)$$

(1) and (2) give the probability that the customer is n in the system under the equilibrium condition, when n ≥ s, that is, the number of customers in the system is greater than or equal to the number of service stations, and then the customers who come again must wait.

$$c(s,\rho) = \sum_{n=s}^{\infty} p_n = \frac{\rho^s}{s!(1-\rho_s)} p_0 \quad (4)$$

Formula (4) is called Erlang waiting formula, which gives the probability of waiting when the customer arrives at the system. For the multi-service waiting queue system, the average queue length $\boldsymbol{L_q}$ can be obtained from the obtained stationary distribution, it is given by:

$$L_q = \frac{p_0 \rho^s \rho_s}{s!(1-\rho_s)^2} (5) \text{ or } L_q = \frac{c(s,\rho)\rho_s}{1-\rho_s} \quad (5)$$

In this example, if the queuing mode of customers is changed to that they can line up at any window after arriving at the ticket office and do not change lines after joining the queue, three queues can be formed. At this time, the original M|M|3|∞ system has become a queuing system composed of three M|M|1|∞ subsystems. Table 2 shows a comparison of the two data.

Table 2. Comparison of Performance Measures Between an M|M|3|∞ System and Three M|M|1|∞ Subsystems

Performance Measure	M|M|3|∞	Three M|M|1|∞ subsystems
The probability that all servers are idle P_0	0.0748	0.25(each subsystem)
The probability that customers should wait	0.57	0.75
Average number of customers in the system	3.95	9 (the whole system)
Average number of customers in the queue	1.70	2.25(each subsystem)
Mean duration of stay	4.39 min	10 min
Average time customer spends in the queue	1.89 min	7.5 min

Clearly, in this case, multi-server systems are more efficient than the combination of multiple single-server systems at the same system cost.

4.2 Influence of Second-Service rate

In some public service counters, customers may need to receive more than one service in a sequence. For example, after registering for medical services, customers may need to see a doctor, take some tests, get some prescriptions, etc. In this case, it is necessary to consider the probability that a customer needs to receive another service after the first one.

The second-service rate can affect the performance of public service counters in different ways. If the second-service rate is high, then customers may spend more time in the system and occupy more resources. This may increase the waiting time and queue length for other customers, which is a common problem in real life need to solve.

4.3 Evaluation Indicators

To measure the influence of second-service rate on the performance of public service counters, some performance evaluation indicators are:

L_s: the average number of customers in service. It is given by $L_s = s\rho + \frac{\rho^2}{1-\rho}$.

W_s: the average waiting time in service. It is given by $W_s = \frac{L_s}{\lambda}$.

L_r: the average number of customers who need to receive another service after the first one. It is given by $L_r = \frac{\rho^2}{1-\rho}$.

W_r: the average waiting time for customers who need to receive another service after the first one. It is given by $W_r = \frac{L_r}{\lambda}$.

L_f: the average number of customers who finish all their services and leave the system. It is given by $L_f = \rho - \frac{\rho^2}{1-\rho}$.

W_f: the average waiting time for customers who finish all their services and leave the system. It is given by $W_f = \frac{L_f}{\lambda}$.

4.4 Example Analysis

Consider an example of a hospital that employs four staff members to register customers for medical services. Each staff member can serve an average of 12 customers per hour. The customers' arrival rate follows a Poisson distribution with an average of 40 customers per hour. To examine the impact of the second-service rate on the performance of public service counters, two scenarios with different second-service rate values are compared: $p = 0$ and $p = 0.5$. The performance evaluation indicators will be used to measure and contrast the system's performance under these conditions.

In Scenario 1, where $p = 0$, no customer requires a second service after the first. This situation is equivalent to a multiple-server model with $\lambda = 40$, $\mu = 12$, and $s = 4$. Calculating the performance measures, it's observed that the system has a high utilization factor of 0.833, indicating each staff member is busy 83.3% of the time.

The average number of customers in service is 3.333, and the average waiting time in service is 5 minutes. The average number of customers requiring another service after the first is zero, as is their average waiting time. The average number of customers who complete all their services and exit the system is 0.667, and their average waiting time is 1 minute.

In Scenario 2, where p = 0.5, it signifies that 50% of customers require a second service after the first. This scenario equates to a multiple-server model with $\lambda = 40$, $\mu = 12$, and $s = 4$, but with a feedback loop that routes half of the customers back to the queue after receiving the first service. The calculated performance measures indicate that the system maintains the same utilization factor, average number of customers in service, and average waiting time in service as when p = 0. However, the average number of customers requiring another service after the first escalates to 4.167, and their average waiting time increases to 6 minutes. The average number of customers who complete all their services and leave the system remains unchanged from when p = 0, as does their average waiting time.

These two scenarios highlight how the second-service rate influences the performance of public service counters with varying values of p. They also demonstrate how performance evaluation indicators can effectively measure and contrast the impact of the second-service rate on the system.

5 Conclusion

In conclusion, queuing theory provides a valuable tool for designing efficient and effective public service counter systems. The single-server and multiple-server queuing models can assist in calculating crucial performance measures of the system, such as utilization factor, average waiting time, and probability of customers being served or departing without service. However, queuing theory relies on some simplifying assumptions that may not be reflective of real-world conditions, prompting the need for more sophisticated and dynamic queuing models. Moreover, future research should focus on providing prescriptive and normative guidance on improving and optimizing the performance of public service counters, including resource allocation, policy design, and customer and server management. Despite its limitations, queuing theory remains a critical analytical tool for decision-makers and managers in the public service sector to enhance customer satisfaction, minimize waiting times and reduce operational costs..

References

1. Chen, H., Wang, R., & Hu, B. (2021). Optimization of public service counters based on queuing theory and queuing model: An empirical study. International Journal of Industrial Engineering Computations, 12(3), 513-530.

2. Yadav, S., & Katiyar, V. (2021). Application of queuing theory to optimize public service counter: A case study. International Journal of Advanced Science and Technology, 30(1), 5775-5782.
3. Xia, Y., Guo, Y., & Hu, G. (2020). Research on optimization of public service counter based on queuing theory. Journal of Intelligent & Fuzzy Systems, 38(4), 4739-4748.
4. Al-Rawahi, A., & Al-Obeidani, B. (2020). Queuing theory application for improving public service in Oman. Journal of King Saud University-Computer and Information Sciences, 32(6), 659-664.
5. Zhang, F., & Wang, Y. (2019). Application of queuing theory in optimizing dispatch of public service counter. Journal of Ambient Intelligence and Humanized Computing, 10(4), 1469-1477.
6. Zhang, L., Li, Y., & Pan, K. (2019). Modeling and optimization of public service counter based on queuing theory. Journal of Physics: Conference Series, 1392(3), 032092.
7. Zhang, L., & Xie, C. (2018). Optimization of public service efficiency based on queuing theory: A case study in China. Future Cities and Environment, 4(1), 1-16.
8. Zhang, L., & Heng, Y. (2018). Design of public service queues based on queuing theory. IOP Conference Series: Materials Science and Engineering, 332(1), 012100.
9. Xu, J., & Ma, J. (2017). Using queuing theory to improve public service performance. Journal of Service Science Research, 9(2), 109-124.
10. Sabah, A. H., Sabir, A., & Aslam, M. (2017). Optimization of public service counter using queuing theory: A case study of Pakistan Railways. ARPN Journal of Engineering and Applied Sciences, 12(4), 1156-1163.