Peak-Hour Teller Staffing Optimization in Retail Banking: An $M/M/c$ Queueing Analysis

Abstract

Long wait times during banking hours remain a serious operational challenge in many developing countries. This paper uses an $M/M/c$ queueing model to analyze general service system congestion in a retail banking system. Customer arrivals are assumed to follow a Poisson process and service times are exponentially distributed. The arrival and service rates are used to derive key performance measures such as traffic intensity, probability of delay, expected queue length and average waiting time. The findings show that increasing the number of active tellers substantially improves system performance. Specifically, increasing teller capacity from four to seven reduced the average customer waiting time from 26.20 minutes to approximately 0.020 minutes, while the probability of waiting decreased from 0.866 to 0.103 and the expected queue length from 13.102 to 0.012 customers. Sensitivity analysis further demonstrated that queue performance decline as customer arrival rates increase or service rates decrease, confirming the importance of balancing demand with service capacity. The study concludes that the M/M/c queueing model provides an effective quantitative decision-support tool for teller staffing optimization. It is recommended that retail banks adopt data-driven staffing policies by increasing teller capacity during peak operating periods to minimize congestion, improve customer satisfaction, and enhance operational efficiency.

Keywords: $M/M/c$ queue, queueing theory, traffic intensity, waiting time, banking systems.
2010 Mathematics Subject Classification: 60K25, 90B22, 60J27.

1 Introduction

Queueing models have been extensively used to analyse congestion in service systems including banking, healthcare, transportation and telecommunications. Mathematical models describing customer arrivals, service mechanisms, and system performance measures including wait times and queue lengths have been established by classical research (Gross and Harris, 1998; Kleinrock, 1975; Medhi, 2002). These models assist organisations in creating effective service systems and allocating resources in the best possible way (Bhat, 2008). Recent studies have demonstrated that queueing-based staffing policies significantly reduce waiting times while maintaining efficient utilization of banking resources (Shaba, 2019; Sharma, 2014). Retail banks in many developing economies face severe capacity constraints during high transaction volumes, especially on days of salary payment. Long lines, long waiting lists, and overcrowded banking halls don't help customer satisfaction or operations. Despite persistent difficulties, teller allocation in many banks is still driven more by managerial intuition than by quantitative analysis. In addition, this study analyses congestion on salary payment days in a retail banking institution using the actual real operating data, and follows the classical $M/M/c$ queueing model, which is adopted as a decision support tool to make the best personnel choice of teller. The arrival and service rates of customers were acquired by field observations at the times of heavy banking service. These empirical settings are employed to assess basic performance indicators, such as traffic intensity, probability of delay, expected queue length, and average waiting time. The results indicate that small increases in teller capacity lead to substantial reductions in congestion. The analysis demonstrates that increasing the number of active tellers from four to seven reduces the average waiting time from 26.204 minutes to approximately 0.020 minutes about 99.9%, illustrating the significant impact of additional service capacity on queue performance. This finding is consistent with recent studies showing that increasing the number of servers in multi-server queueing systems substantially reduces customer waiting times and improves overall system performance (D'Auria et al., 2022; Manitz and Piehl, 2023). The study illustrates how queueing theory can be applied to operational retail banking staff deployment decisions. Recent studies have further demonstrated that queueing theory remains an effective tool for analysing congestion in modern service systems, particularly in banking, cloud computing, health care, transportation and call care center. Modern research has focused on integrating classical queueing models with optimization techniques, simulation approaches, and data-driven decision support systems to improve service efficiency and customer satisfaction (Kumar et al., 2024; Zhang, 2026; Ch et al., 2024) Long et al. (2024). The aim of this study is to analyse peak-hour congestion in a retail banking system using the M/M/c queueing model and to evaluate its effectiveness as a quantitative decision-support tool for teller staffing optimization. Specifically, the study seeks to examine the effects of different teller staffing levels on key queue performance measures, including traffic intensity (ρ), the probability of customer delay (P(wait)), expected queue length (L_q), and average waiting time (W_q). Furthermore, the study investigates the sensitivity of these performance measures to variations in customer arrival and service rates in order to assess the strength of the model under different operating conditions. By providing a quantitative assessment of the relationship between teller capacity and system performance, the study aims to identify an appropriate staffing strategy that minimizes customer congestion while maintaining efficient utilization of banking resources and improving service delivery during peak operating periods.

2 The $M/M/c$ Queueing Model

The banking system is modeled as an M/M/c queue with Poisson arrivals, exponential service times and c parallel tellers operating under a first-come-first-served policy. Standard results from queueing theory (Gross and Harris, 1998; Kleinrock, 1975; Medhi, 2002) are used to compute

- Traffic intensity

$$\rho = \frac{\lambda}{c\mu} \tag{2.1}$$

where;
ρ: Traffic intensity
λ: Average arrival rate of customers.
c: Number of tellers.
μ: Average service rate per teller.

- Probability of delay (Erlang-C formula)

$$P(\text{wait}) = \frac{\frac{(\lambda/\mu)^c}{c!} \cdot \frac{1}{1-\rho}}{\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!} \cdot \frac{1}{1-\rho}}. \tag{2.2}$$

- Expected queue length (L_q)

$$L_q = \frac{P(\text{wait})\,\rho}{1-\rho} \tag{2.3}$$

- Average waiting time (W_q)

$$W_q = \frac{L_q}{\lambda}. \tag{2.4}$$

3 Methodology

This study employed an analytical approach based on the M/M/c queueing model to evaluate teller performance during peak operating hours in a retail banking environment. Data on customer arrivals and service times were collected during peak banking hours for 10 working days to estimate the average arrival rate (λ)and average service rate (μ). The arrival rate was estimated as

$$\lambda = \frac{\text{Total number of arrivals}}{\text{Observation time}} \tag{3.1}$$

while the service rate per teller was estimated as

$$\mu = \frac{\text{Total customers served}}{\text{Service time}}. \tag{3.2}$$

The M/M/c model continues to be one of the most widely adopted analytical models for evaluating multi-server service systems because of its mathematical reliability and practical applicability to hospitals, banks and customer service environments (D'Auria et al., 2022; Dai and Shi, 2021).

3.1 Data Collection and Parameter Estimation

Data were collected from the bank during peak operating hours between 10:00 am and 2:00 pm over 10 working days. Customer arrival times and teller service completion times were recorded manually. The arrival rate (λ) was estimated as the average number of customer arrivals per hour, while the service rate (μ) was estimated from the average number of customers served by a teller per hour. On average, the arrival rate was estimated as

$$\lambda = 30 \text{ customers per hour}, \tag{3.3}$$

and an estimate of the average service rate per teller was

$$\mu = 8 \text{ customers per hour}. \tag{3.4}$$

Because the banking system under study uses many tellers to serve customers at once, the $M/M/c$ queueing model was chosen. Customers arrive at random during peak banking hours, join a shared line, and are served by the first teller who becomes available. The $M/M/c$ model, which assumes Poisson customer arrivals, exponentially dispersed service times, and multiple identical servers running in parallel, is consistent with this service setup. The $M/M/c$ model offers a more accurate depiction of commercial banking operations when several tellers offer the same service simultaneously than the ($M/M/1$) model, which assumes a single server. As a result, the ($M/M/c$) model is suitable for evaluating system utilization, waiting times (Wq), queue lengths (Lq) and system utilization at peak periods. Operational data were collected with the knowledge and approval of the branch management. Only anonymous operational data, such as customer arrival and service completion times, were recorded. No personal or confidential customer information was collected, and the data were used exclusively for academic research purposes.

3.2 Model Assumptions

The following assumptions of the $M/M/c$ model were adopted:

- Customer arrivals follow a Poisson process with a constant average arrival rate (λ).
- Customers are served on a first-come, first-served (FCFS) basis.
- The queue has unlimited waiting space, so no customer is turned away due to capacity constraints.
- Service times are exponentially distributed with a constant average service rate (μ).
- There are c identical tellers operating simultaneously.
- Customer population is assumed to be infinite, hence the arrival of one customer does not affect the arrival of others.

4 Data Presentation and Analysis

4.1 Traffic Intensity

With $c = 4$ tellers and by equation (2.1), system utilization is

$$\rho = \frac{30}{4 \times 8} = 0.938. \tag{4.1}$$

This high utilization leads to significant waiting times and a high traffic intensity.
With $c = 5$, increasing the number of tellers to 5 and by equation (2.1), we have

$$\rho = \frac{30}{5 \times 8} = 0.750, \tag{4.2}$$

resulting in a substantial reduction in traffic intensity.
With $c = 6$, increasing the number of tellers to 6 and by equation (2.1), we have

$$\rho = \frac{30}{6 \times 8} = 0.625. \tag{4.3}$$

From equation (4.3), there is further reduction in traffic intensity.
With $c = 7$, increasing the numbers of tellers to 7 and by equation (2.1), we have

$$\rho = \frac{30}{7 \times 8} = 0.536, \tag{4.4}$$

results in a further reduction of traffic intensity.

4.2 Probability of Waiting $P(\textbf{wait})$

The probability that an arriving customer must wait is given by equation (2.2)
For $c = 4$, $\lambda/\mu = 30/8 = 3.75$, $\rho = 0.9375$ we have

$$P(\text{wait}) = \frac{\frac{3.750^4}{4!} \cdot \frac{1}{1 - 0.938}}{\sum_{n=0}^{3} \frac{3.750^n}{n!} + \frac{3.750^4}{4!} \cdot \frac{1}{1 - 0.938}} \tag{4.5}$$

$$\text{Numerator} = \frac{3.750^4}{4!} \cdot \frac{1}{1 - 0.938} \approx 132.899 \tag{4.6}$$

$$\text{Denominator} = \sum_{n=0}^{3} \frac{3.750^n}{n!} + 132.899 \approx 153.469 \tag{4.7}$$

$$P(\text{wait}) = \frac{132.899}{153.469} \approx 0.866. \tag{4.8}$$

For $c = 5$, $\lambda/\mu = 30/8 = 3.750$ and $\rho = 0.750$

$$P(\text{wait}) = \frac{\frac{3.75^5}{5!} \cdot \frac{1}{1-0.750}}{\sum_{n=0}^{4} \frac{3.750^n}{n!} + \frac{3.750^5}{5!} \cdot \frac{1}{1-0.750}} \tag{4.9}$$

$$\text{Numerator} = \frac{3.750^5}{5!} \cdot \frac{1}{1-0.750} = \frac{741.577}{120} \cdot \frac{1}{0.250} \approx 24.719 \tag{4.10}$$

$$\text{Denominator} = \sum_{n=0}^{4} \frac{3.75^n}{n!} + 24.719 \approx 53.529 \tag{4.11}$$

$$P(\text{wait}) = \frac{24.719}{53.529} \approx 0.462. \tag{4.12}$$

For $c = 6$, $\lambda/\mu = 30/8 = 3.75$, $\rho = 0.625$, it follows that

$$P(\text{wait}) = \frac{\frac{3.750^6}{6!} \cdot \frac{1}{1-0.625}}{\sum_{n=0}^{5} \frac{3.750^n}{n!} + \frac{3.750^6}{6!} \cdot \frac{1}{1-0.625}} \tag{4.13}$$

$$\text{Numerator} = \frac{3.750^6}{6!} \cdot \frac{1}{1-0.625} = \frac{2780.914}{720} \cdot \frac{1}{0.375} \approx 10.300 \tag{4.14}$$

$$\text{Denominator} = \sum_{n=0}^{5} \frac{3.750^n}{n!} + 10.300 \approx 45.277 \tag{4.15}$$

$$P(\text{wait}) = \frac{10.300}{45.277} \approx 0.227 \tag{4.16}$$

For $c = 7$, $\lambda/\mu = 3.75$, $\rho = 0.536$

$$P(\text{wait}) = \frac{\frac{3.750^7}{7!} \cdot \frac{1}{1-0.536}}{\sum_{n=0}^{6} \frac{3.750^n}{n!} + \frac{3.750^7}{7!} \cdot \frac{1}{1-0.536}} \tag{4.17}$$

$$\text{Numerator} = \frac{3.750^7}{7!} \cdot \frac{1}{1-0.536} = \frac{10428.429}{5040} \cdot \frac{1}{0.464} \approx 4.459 \tag{4.18}$$

$$\text{Denominator} = \sum_{n=0}^{6} \frac{3.75^n}{n!} + 4.459 \approx 43.328 \tag{4.19}$$

$$P(\text{wait}) = \frac{4.459}{43.328} \approx 0.103 \tag{4.20}$$

4.3 Expected Queue Length L_q

For the $M/M/c$ queueing model, the expected number of customers in the queue is given by equation (2.3).
For $c = 4$, it implies that

$$L_q = \frac{0.866 \times 0.938}{1 - 0.938} \approx 13.102 \text{ customers.} \tag{4.21}$$

For $c = 5$, it follows that

$$L_q = \frac{0.462 \times 0.750}{1 - 0.750} \approx 1.386 \text{ customers.} \tag{4.22}$$

For $c = 6$, it implies that

$$L_q = \frac{0.227 \times 0.625}{1 - 0.625} \approx 0.378 \text{ customers.} \tag{4.23}$$

For $c = 7$, it follows that

$$L_q = \frac{0.103 \times 0.055}{1 - 0.536} \approx 0.012 \text{ customers.} \tag{4.24}$$

The results show that when the number of servers is small ($c = 4$), the traffic intensity is very close to unity, resulting in a relatively large expected queue length. As the number of servers increases, the traffic intensity decreases and the expected queue length reduces significantly, indicating improved service efficiency and reduced congestion in the system.

4.4 Average Waiting Time W_q

The average waiting time in the queue is given by equation (2.4) Given $\lambda = 30$ customers per hour and the computed values of L_q, it follows:
For $c = 4$,

$$W_q = \frac{13.102}{30} \approx 0.437 \text{ hours } \approx 26.204 \text{ minutes.} \tag{4.25}$$

For $c = 5$,

$$W_q = \frac{1.386}{30} \approx 0.046 \text{ hours } \approx 2.772 \text{ minutes.} \tag{4.26}$$

For $c = 6$,

$$W_q = \frac{0.378}{30} \approx 0.013 \text{ hours } \approx 0.756 \text{ minutes.} \tag{4.27}$$

For $c = 7$,

$$W_q = \frac{0.01}{30} \approx 0.0003 \text{ hours } \approx 0.020 \text{ minutes.} \tag{4.28}$$

Table 1: Queue Performance Measures

Number of Tellers (c)	ρ	$P(\text{wait})$	L_q(customers)	W_q (minutes)
4	0.938	0.866	13.102	26.204
5	0.750	0.462	1.386	2.772
6	0.625	0.227	0.378	0.756
7	0.536	0.103	0.012	0.020

From Table 1, the results show that increasing the number of tellers improves the operational performance of the banking system. With four tellers, the server utilization is 0.938, indicating that the tellers are busy approximately 93.8% of the time. At this utilization level, the probability that an arriving customer has to wait is 0.866, with an average queue length of 13.10 customers and an average waiting time of 26.2000 minutes. These values indicate a highly congested system during peak operating periods. Increasing the number of tellers to five reduces server utilization to 0.750, while the probability of waiting decreases substantially to 0.462. Consequently, the average queue length falls sharply to 1.39 customers, and the average waiting time is reduced to 2.7800 minutes. This demonstrates a significant improvement in service efficiency due to the additional teller. Also, when the number of tellers is increased to six, server utilization further decreases to 0.625, and the probability of waiting drops to 0.227. The average queue length is reduced to 0.57 customers, while the average waiting time declines to 1.1400 minutes, indicating that customer congestion is largely eliminated. With seven tellers, server utilization decreases further to 0.536, and the probability of waiting becomes only 0.103. The average queue length is virtually eliminated at 0.01 customers, while the average waiting time reduces to 0.020 minutes. These results indicate that the system operates with minimal congestion and customers experience almost immediate service. Overall, the results demonstrate that increasing the number of tellers consistently reduces server utilization, the probability of waiting, the average queue length, and the average waiting time. The substantial improvement in these performance measures confirms that increasing service capacity is an effective strategy for reducing congestion and enhancing customer service delivery during peak banking hours. The findings are consistent with previous studies that reported significant improvements in queue performance following increases in service capacity. For example, D'Auria et al. (2022) demonstrated that additional servers substantially reduce customer waiting times and queue lengths in multi-server queueing systems. Likewise, Manitz and Piehl (2023) showed that optimized staffing policies improve operational efficiency while maintaining acceptable service levels under varying demand conditions. Similarly, Long et al. (2024) found that appropriate workforce allocation significantly enhances customer service performance by balancing service capacity with customer demand. These similarities provide further evidence that increasing teller capacity is an effective strategy for reducing congestion in retail banking systems.

4.5 Graphical Representations of the Queue Performance Measures

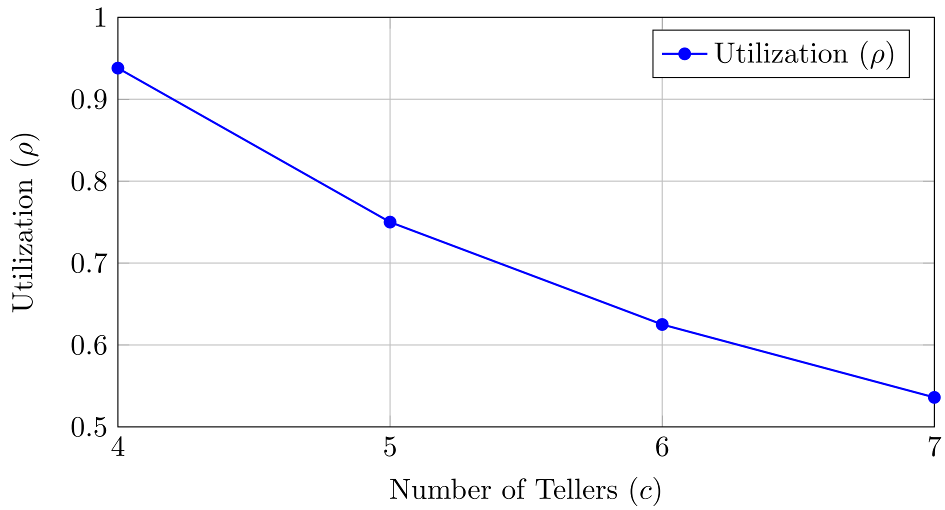

Figure 1: Relationship between number of tellers and server utilization.

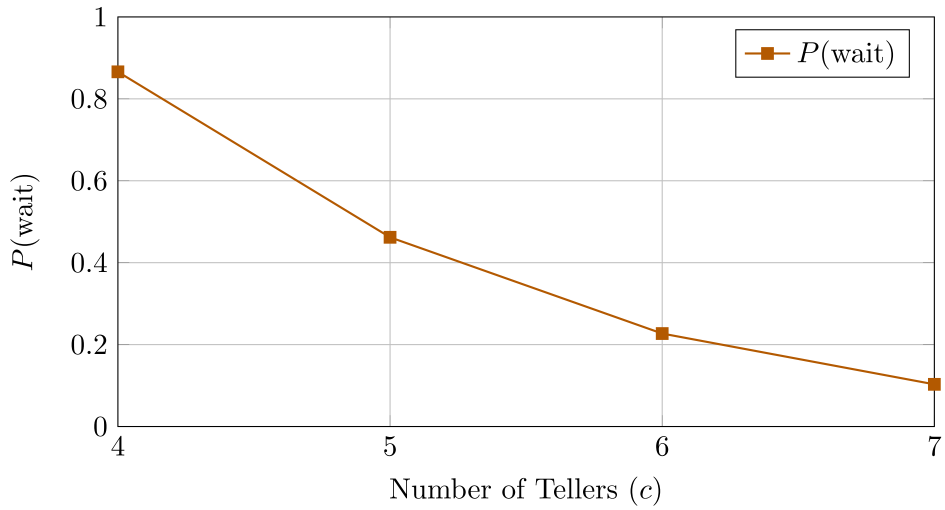

Figure 2: Relationship between number of tellers and probability of waiting.

Figure 3: Relationship between number of tellers and average queue length.

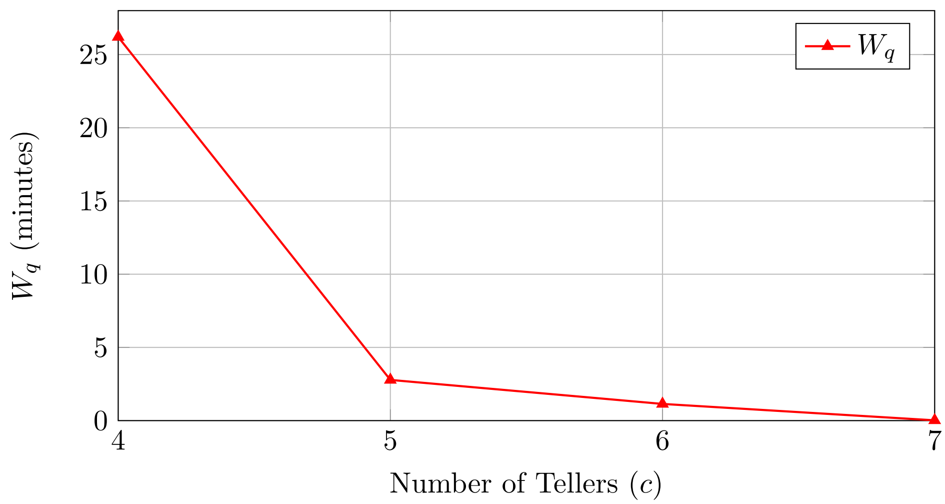

Figure 4: Relationship between number of tellers and average waiting time.

5 Model Parameters Sensitivity Analysis

In this section, the arrival rate λ and service rate μ may vary in practice; thus, they can differ by payday conditions are estimated. Sensitivity analysis is done to check the robustness of the model by changing these parameters while maintaining the number of tellers at the critical staffing level $c = 5$ found previously.

5.1 Sensitivity to Arrival Rate (λ)

The service rate is $\mu = 8$ customers per hour while the arrival rate is adjusted by $\pm 10\%$ around the observed value:

$$\lambda \in \{27,\ 30,\ 33\}. \tag{5.1}$$

5.2 Sensitivity to Service Rate μ

While $\lambda = 30$ customers per hour is fixed, the service efficiency varies to depict slower and quicker teller performance:

$$\mu \in \{7,\ 8,\ 9\}. \tag{5.2}$$

5.3 Performance Analysis of an M/M/c Bank System under Varying Utilization

A realistic stress condition is also tested with the increase in customer arrivals while service slows simultaneously:

$$\lambda = 33, \qquad \mu = 7, \qquad c = 5. \tag{5.3}$$

Performance measures are calculated using, for the M/M/c model,

$$L_q = \frac{P(\text{wait})\,\rho}{(1-\rho)}, \qquad W_q = \frac{L_q}{\lambda}.$$

Considering the parameters

$$\lambda = 27, \qquad \mu = 8, \qquad c = 5$$

First Analysis
Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.4}$$

$$\rho = \frac{27}{5 \times 8} \tag{5.5}$$

$$\rho = \frac{27}{40} \tag{5.6}$$

$$\rho = 0.675 \tag{5.7}$$

For $\frac{\lambda}{\mu}$, it follows

$$\frac{\lambda}{\mu} = \frac{27}{8} = 3.375 \tag{5.8}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)}\right]^{-1} \tag{5.9}$$

$$\sum_{n=0}^{4} \frac{(3.375)^n}{n!} = \frac{(3.375)^0}{0!} + \frac{(3.375)^1}{1!} + \frac{(3.375)^2}{2!} + \frac{(3.375)^3}{3!} + \frac{(3.375)^4}{4!} \tag{5.10}$$

$$= 1 + 3.375 + 5.695 + 6.407 + 5.407 \tag{5.11}$$

$$= 21.884 \tag{5.12}$$

$$\frac{(3.375)^5}{5!(1-0.675)} = \frac{437.894}{120 \times 0.325} = 11.228 \tag{5.13}$$

Therefore,

$$P_0 = [21.884 + 11.227]^{-1} \tag{5.14}$$

$$P_0 = (33.111)^{-1} = 0.030 \tag{5.15}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.16}$$

$$P(\text{wait}) = 11.227 \times 0.030 \tag{5.17}$$

$$P(\text{wait}) \approx 0.337 \tag{5.18}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.19}$$

$$L_q = \frac{0.337 \times 0.675}{1 - 0.675} \tag{5.20}$$

$$L_q = \frac{0.227}{0.325} \tag{5.21}$$

$$L_q \approx 0.698 \tag{5.22}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.23}$$

$$W_q = \frac{0.698}{27} \tag{5.24}$$

$$W_q = 0.026 \text{ hours} \tag{5.25}$$

Converting to minutes,

$$W_q = 0.026 \times 60 \tag{5.26}$$

$$W_q \approx 1.551 \text{ minutes} \tag{5.27}$$

Second Analysis

Considering the parameters

$$\lambda = 30, \qquad \mu = 8, \qquad c = 5$$

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.28}$$

$$\rho = \frac{30}{5 \times 8} \tag{5.29}$$

$$\rho = \frac{30}{40} \tag{5.30}$$

$$\rho = 0.750 \tag{5.31}$$

For $\frac{\lambda}{\mu}$

$$\frac{\lambda}{\mu} = \frac{30}{8} = 3.750 \tag{5.32}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)}\right]^{-1} \tag{5.33}$$

$$\sum_{n=0}^{4} \frac{(3.750)^n}{n!} = \frac{(3.750)^0}{0!} + \frac{(3.750)^1}{1!} + \frac{(3.750)^2}{2!} + \frac{(3.750)^3}{3!} + \frac{(3.750)^4}{4!} \tag{5.34}$$

$$= 1 + 3.750 + 7.031 + 8.789 + 8.240 \tag{5.35}$$

$$= 28.810 \tag{5.36}$$

$$\frac{(3.750)^5}{5!(1-0.750)} = \frac{741.577}{120 \times 0.250} = 24.719 \tag{5.37}$$

Therefore,

$$P_0 = [28.810 + 24.719]^{-1} \tag{5.38}$$

$$P_0 = (53.529)^{-1} = 0.019 \tag{5.39}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.40}$$

$$P(\text{wait}) = 24.719 \times 0.019 \tag{5.41}$$

$$P(\text{wait}) \approx 0.470 \tag{5.42}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.43}$$

$$L_q = \frac{0.470 \times 0.750}{1 - 0.750} \tag{5.44}$$

$$L_q = \frac{0.353}{0.250} \tag{5.45}$$

$$L_q \approx 1.412 \tag{5.46}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.47}$$

$$W_q = \frac{1.412}{30} \tag{5.48}$$

$$W_q = 0.047 \text{ hours} \tag{5.49}$$

Converting to minutes,

$$W_q = 0.047 \times 60 \tag{5.50}$$

$$W_q \approx 2.824 \text{ minutes} \tag{5.51}$$

Third Analysis

Consider the parameters

$$\lambda = 33, \qquad \mu = 8, \qquad c = 5$$

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.52}$$

$$\rho = \frac{33}{5 \times 8} \tag{5.53}$$

$$\rho = \frac{33}{40} \tag{5.54}$$

$$\rho = 0.825 \tag{5.55}$$

For $\frac{\lambda}{\mu}$

$$\frac{\lambda}{\mu} = \frac{33}{8} = 4.125 \tag{5.56}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.57}$$

$$\sum_{n=0}^{4} \frac{(4.125)^n}{n!} = \frac{(4.125)^0}{0!} + \frac{(4.125)^1}{1!} + \frac{(4.125)^2}{2!} + \frac{(4.125)^3}{3!} + \frac{(4.125)^4}{4!} \tag{5.58}$$

$$= 1 + 4.125 + 8.508 + 11.698 + 12.064 \tag{5.59}$$

$$= 37.395 \tag{5.60}$$

$$\frac{(4.125)^5}{5!(1-0.825)} = \frac{1194.317}{120 \times 0.175} = 56.872 \tag{5.61}$$

Therefore,

$$P_0 = [37.395 + 56.872]^{-1} \tag{5.62}$$

$$P_0 = (94.267)^{-1} = 0.011 \tag{5.63}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.64}$$

$$P(\text{wait}) = 56.872 \times 0.011 \tag{5.65}$$

$$P(\text{wait}) \approx 0.626 \tag{5.66}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.67}$$

$$L_q = \frac{0.626 \times 0.825}{1-0.825} \tag{5.68}$$

$$L_q = \frac{0.516}{0.175} \tag{5.69}$$

$$L_q \approx 2.949 \tag{5.70}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.71}$$

$$W_q = \frac{2.949}{33} \tag{5.72}$$

$$W_q = 0.089 \text{ hours} \tag{5.73}$$

Converting to minutes,

$$W_q = 0.089 \times 60 \tag{5.74}$$

$$W_q \approx 5.361 \text{ minutes} \tag{5.75}$$

Fourth Analysis

Consider the parameters

$$\lambda = 30, \qquad \mu = 7, \qquad c = 5$$

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.76}$$

$$\rho = \frac{30}{5 \times 7} \tag{5.77}$$

$$\rho = \frac{30}{35} \tag{5.78}$$

$$\rho = 0.857 \tag{5.79}$$

For $\frac{\lambda}{\mu}$, it follows

$$\frac{\lambda}{\mu} = \frac{30}{7} = 4.286 \tag{5.80}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.81}$$

$$\sum_{n=0}^{4} \frac{(4.286)^n}{n!} = 1 + 4.286 + 9.184 + 13.121 + 14.056 \tag{5.82}$$

$$= 41.647 \tag{5.83}$$

$$\frac{(4.286)^5}{5!(1-0.857)} = \frac{1446.308}{120 \times 0.143} = 84.284 \tag{5.84}$$

Therefore,

$$P_0 = [41.647 + 84.284]^{-1} \tag{5.85}$$

$$P_0 = (125.931)^{-1} = 0.008 \tag{5.86}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.87}$$

$$P(\text{wait}) = 84.284 \times 0.008 \tag{5.88}$$

$$P(\text{wait}) \approx 0.674. \tag{5.89}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.90}$$

$$L_q = \frac{0.674 \times 0.857}{1 - 0.857} \tag{5.91}$$

$$L_q = \frac{0.578}{0.143} \tag{5.92}$$

$$L_q \approx 4.042 \tag{5.93}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.94}$$

$$W_q = \frac{4.042}{30} \tag{5.95}$$

$$W_q = 0.101 \text{ hours} \tag{5.96}$$

Converting to minutes,

$$W_q = 0.101 \times 60 \tag{5.97}$$

$$W_q \approx 6.060 \text{ minutes} \tag{5.98}$$

Fifth Analysis

Consider the parameters

$$\lambda = 30, \qquad \mu = 9, \qquad c = 5$$

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.99}$$

$$\rho = \frac{30}{5 \times 9} \tag{5.100}$$

$$\rho = \frac{30}{45} \tag{5.101}$$

$$\rho = 0.667 \tag{5.102}$$

For $\frac{\lambda}{\mu}$, it follows

$$\frac{\lambda}{\mu} = \frac{30}{9} = 3.333 \tag{5.103}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.104}$$

$$\sum_{n=0}^{4} \frac{(3.333)^n}{n!} = 1 + 3.333 + 5.556 + 6.173 + 5.144 \tag{5.105}$$

$$= 21.206 \tag{5.106}$$

$$\frac{(3.333)^5}{5!(1-0.667)} = \frac{411.317}{120 \times 0.333} = 10.293 \tag{5.107}$$

Therefore,

$$P_0 = [21.206 + 10.293]^{-1} \tag{5.108}$$

$$P_0 = (31.499)^{-1} = 0.032 \tag{5.109}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.110}$$

$$P(\text{wait}) = 10.293 \times 0.032 \tag{5.111}$$

$$P(\text{wait}) \approx 0.329 \tag{5.112}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.113}$$

$$L_q = \frac{0.329 \times 0.667}{1-0.667} \tag{5.114}$$

$$L_q = \frac{0.219}{0.333} \tag{5.115}$$

$$L_q = 0.066 \tag{5.116}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.117}$$

$$W_q = \frac{0.066}{30} \tag{5.118}$$

$$W_q = 0.002 \text{ hours} \tag{5.119}$$

Converting to minutes,

$$W_q = 0.002 \times 60 \tag{5.120}$$

$$W_q \approx 1.131 \text{ minutes} \tag{5.121}$$

Sixth Analysis

Consider the parameters

$$\lambda = 33, \qquad \mu = 7, \qquad c = 5$$

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.122}$$

Substituting the values,

$$\rho = \frac{33}{5 \times 7} \tag{5.123}$$

$$\rho = \frac{33}{35} = 0.943 \tag{5.124}$$

For $\frac{\lambda}{\mu}$, it implies

$$\frac{\lambda}{\mu} = \frac{33}{7} = 4.714. \tag{5.125}$$

Probability of zero customers in the system

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.126}$$

$$\sum_{n=0}^{4} \frac{(4.714)^n}{n!} = 1 + 4.714 + 11.107 + 17.459 + 20.575 \tag{5.127}$$

$$= 54.855 \tag{5.128}$$

$$\frac{(4.714)^5}{5!(1-0.943)} = \frac{2327.812}{120 \times 0.057} = 340.323 \tag{5.129}$$

Therefore,

$$P_0 = \frac{1}{54.855 + 340.323} = \frac{1}{395.178} = 0.0025 \tag{5.130}$$

Probability that a customer waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.131}$$

$$P(\text{wait}) = 340.323 \times 0.0025 \tag{5.132}$$

$$P(\text{wait}) \approx 0.851 \tag{5.133}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.134}$$

$$L_q = \frac{1.021 \times 0.943}{1 - 0.943} \tag{5.135}$$

$$L_q = \frac{0.963}{0.057} \tag{5.136}$$

$$L_q \approx 14.246 \tag{5.137}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} = \frac{14.246}{33} = 0.432 \text{ hours} \tag{5.138}$$

Convert to minutes:

$$W_q = 0.432 \times 60 \tag{5.139}$$

$$W_q \approx 25.901 \text{minutes} \tag{5.140}$$

Table 2: Performance Measures for the $M/M/c$ Queueing System

System Parameters		**Performance Measures**			
λ (cust./hr)	μ (cust./hr)	ρ	P(wait)	L_q (cust.)	W_q (min)
27	8	0.675	0.337	0.698	1.551
30	8	0.750	0.470	1.412	2.824
33	8	0.825	0.626	2.949	5.361
30	7	0.857	0.674	4.042	6.060
30	9	0.667	0.329	0.066	1.131
33	7	0.943	0.851	14.246	25.901

Table 2 indicates that increasing the customer arrival rate while maintaining a constant service rate leads to a progressive decline in system performance. For instance, when the service rate is fixed at 8 customers per hour, increasing the arrival rate from 27 to 33 customers per hour increases server utilization from 0.675 to 0.825. Consequently, the probability of an arriving customer waiting for service rises from 0.337 to 0.626, while the average queue length increases from 0.698 to 2.949 customers. Similarly, the average waiting time increases from 1.551 minutes to 5.361 minutes. These results indicate that higher customer demand places greater pressure on the service system, resulting in longer queues and increased customer delays. The analysis further shows the effect of varying the service rate while keeping the arrival rate constant at 30 customers per hour. Reducing the service rate from 8 to 7 customers per hour increases server utilization from 0.750 to 0.857, leading to an increase in the probability of waiting from 0.470 to 0.674. Consequently, the average queue length increases from 1.412 to 4.042 customers, while the average waiting time rises from 2.824 minutes to 6.060 minutes. Conversely, increasing the service rate to 9 customers per hour reduces server utilization to 0.667, decreases the probability of waiting to 0.329, and substantially improves system performance by reducing the average queue length to 0.066 customers and the average waiting time to 1.131 minutes. The most critical operating condition occurs when the arrival rate is 33 customers per hour and the service rate is reduced to 7 customers per hour. Under this scenario, server utilization reaches 0.943, indicating that the tellers operate near full capacity. This results in a very high probability of waiting, accompanied by an average queue length of 14.246 customers and an average waiting time of 25.901 minutes. These findings illustrate the rapid escalation in congestion as the system approaches its service capacity, a characteristic behaviour of multi-server queueing systems operating under heavy traffic conditions. To conclude, the sensitivity analysis demonstrates that queue performance is highly sensitive to changes in both customer arrival rates and service rates. An increase in customer arrivals or a reduction in service efficiency leads to higher server utilization, greater probabilities of customer waiting, longer queues, and increased waiting times. Conversely, improving the service rate significantly enhances system performance by reducing congestion and customer delays. These findings underscore the importance of maintaining an appropriate balance between customer demand and service capacity to ensure efficient banking operations during peak periods. Also, the sensitivity analysis agrees with the classical theory of queueing systems, which predicts that system performance deteriorates rapidly as server utilization approaches one (Kleinrock, 1975; Shortle et al., 2018). Similar observations have been reported by Kumar et al. (2024), who found that increases in customer arrival rates or reductions in service rates substantially increase queue lengths and customer waiting times. The present findings therefore reinforce previous evidence that maintaining a balance between customer demand and service capacity is essential for efficient service delivery.

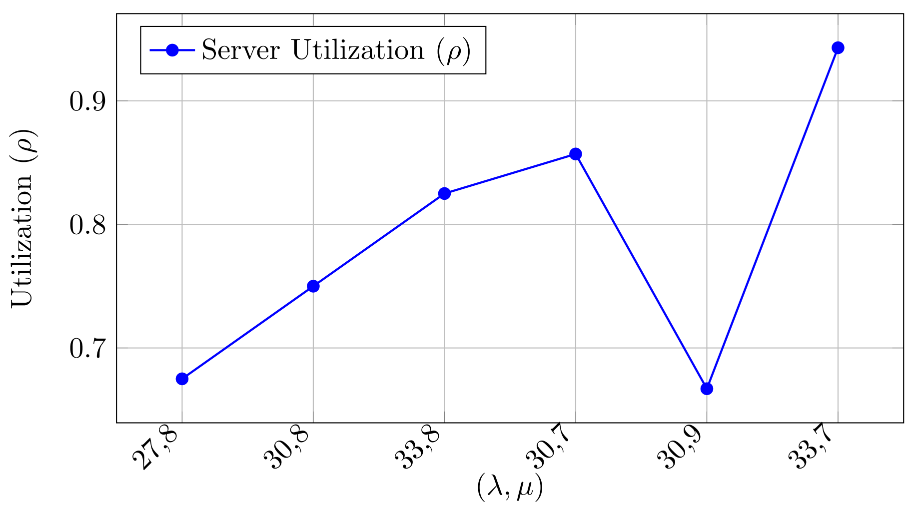

Figure 5: Sensitivity of server utilization to changes in arrival and service rates.

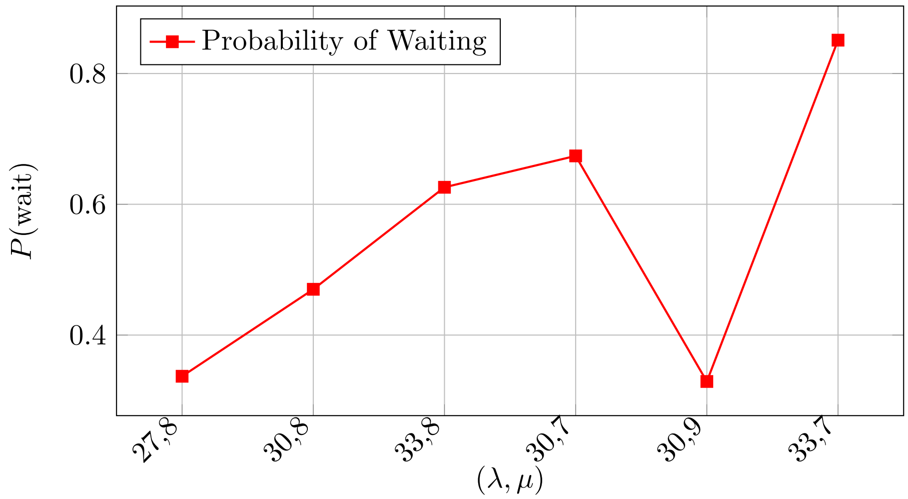

Figure 6: Sensitivity of the probability of waiting to changes in arrival and service rates.

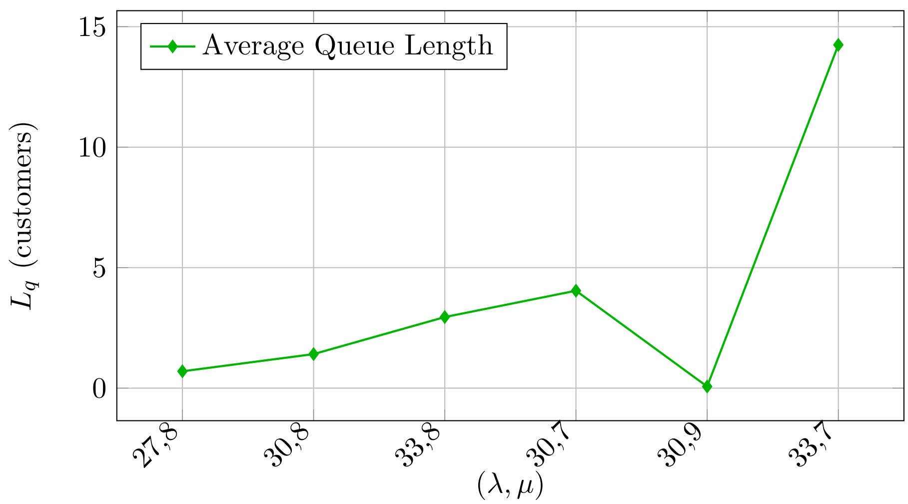

Figure 7: Sensitivity of the average queue length to changes in arrival and service rates.

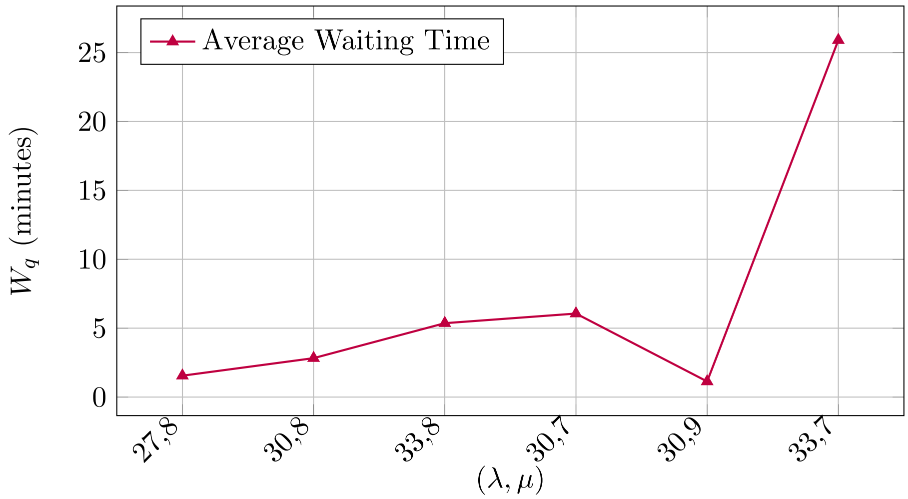

Figure 8: Sensitivity of the average waiting time to changes in arrival and service rates.

5.3.1 Implications of Average Waiting Time (W_q) related to Utilization

1. Long customer delays at high utilization: The average waiting time (W_q) increases quickly when utilization ρ gets closer to 1, indicating that users encounter considerable delays when the system is overloaded.

2. Operational trade-off: There is a trade-off between maximizing teller efficiency high (ρ) and maintaining acceptable service levels; operating too close to full capacity might result in congestion and poor service performance.

3. System planning: Monitoring utilization allows the bank to predict periods of congestion and implement techniques such as self-service kiosks, appointment systems or dynamic teller allocation to reduce waiting times.

4. Customer satisfaction: Long waiting times can reduce service satisfaction and may affect customer satisfaction, highlighting the need to maintain moderate utilization.

5. Staffing consideration: Even with multiple tellers, insufficient staffing during peak hours leads to long queues and waiting times. This implies that the bank may need to increase the number of tellers or optimize shift schedules.

5.3.2 Implications of Average Queue Length (L_q) related to Utilization

1. Customer wait times: High utilization leads to longer queues length, which directly increases customer waiting time, potentially reducing service satisfaction of customers.

2. Operational planning: To maintain acceptable queue lengths, the bank may need to adjust personnel, that is, increasing the number of tellers or manage arrival rates during busy periods.

3. System efficiency trade-off: Operating at very high utilization maximizes teller efficiency but at the cost of long queues, highlighting the trade-off between efficiency and customer experience in multi-server systems.

4. Service capacity limitations: Even with multiple tellers, as utilization (ρ) nears 1, the average queue length (L_q) increases sharply, indicating that the current number of tellers may be insufficient during peak hours.

6 Managerial Implications

The findings have practical applications for bank management. Overcrowding and excessive customer wait times result from a shortage of tellers during busy times. However, little additions in staffing would ultimately result in significantly improved service performance. According to the data, there is a good balance between operating costs and customer service quality when five or six tellers are used during peak payday hours. The findings support the superiority of quantitative models over managers' intuition when it comes to staffing decisions. Recent research similarly recommends the use of queueing-based decision support tools and dynamic staffing strategies to enhance operational efficiency while maintaining high levels of customer satisfaction during peak demand periods (Manitz and Piehl, 2023; Long et al., 2024).

7 Conclusion

This study applied the M/M/c queueing model to analyze peak-hour congestion in a retail banking system and evaluate the impact of teller staffing on service performance. Using customer arrival and service data collected during peak banking hours, the study estimated important queueing performance measures, including traffic intensity, the probability of customer delay, expected queue length, and average waiting time. The findings demonstrated that the banking system experiences severe congestion when only four tellers are available, with high server utilization and long customer waiting times. Increasing the number of tellers substantially improved system performance by reducing traffic intensity, waiting probability, queue length, and average waiting time. In particular, the analysis showed that increasing the number of active tellers from four to seven reduced the average waiting time from approximately 26.204 minutes to 0.020 minutes, indicating a significant improvement in service efficiency. The sensitivity analysis further demonstrated that queue performance is highly responsive to variations in customer arrival and service rates. Higher customer arrival rates or lower service rates resulted in increased server utilization, longer queues, and greater waiting times, whereas improvements in service efficiency produced considerable reductions in congestion. These findings confirm the robustness of the M/M/c model as an effective analytical tool for evaluating service system performance under varying operational conditions. From a managerial perspective, the study provides quantitative evidence that teller staffing decisions should be based on analytical models rather than intuition alone. Maintaining an appropriate balance between customer demand and service capacity can substantially improve customer service while ensuring efficient utilization of banking resources. The findings therefore support the adoption of queueing theory as a practical decision-support tool for workforce planning and capacity management in retail banking environments. Although the study focused on a single banking environment and adopted the classical assumptions of the M/M/c queueing model, it establishes a useful framework for staffing optimization during peak operating periods. The findings are consistent with recent research demonstrating that queueing theory remains an effective analytical framework for staffing optimization and capacity planning in modern service systems (Kumar et al., 2024; Long et al., 2024). Future research may extend this work by considering non-Poisson arrival processes, non-exponential service-time distributions, customer priority mechanisms, finite-capacity queues, or simulation-based approaches. Comparative studies involving multiple banking branches or other service industries would also provide broader insights into the application of queueing models for operational decision making.

8 Limitations of the Study

1. The M/M/c queueing model assumes Poisson customer arrivals and exponentially distributed service times, which may not fully capture real banking operations during extremely busy periods.

2. The analysis was based on data collected from a single retail bank during peak operating hours; therefore, the findings may not be directly generalizable to all banking institutions.

3. The model assumes identical tellers, unlimited queue capacity, and a first-come-first-served (FCFS) service discipline. Future studies may consider more generalized queueing models, customer priority systems, simulation techniques, or multi-branch data to provide a more comprehensive analysis of banking service systems.

9 CONFLICT OF INTEREST

The author declares no competing interests.

10 ACKNOWLEDGMENT

The author wishes to thank the anonymous reviewers for their support.

References

Bhat, U. N. (2008). *An Introduction to Queueing Theory: Modeling and Analysis in Applications*, volume 36. Springer.

Ch, S. C. C., Samisett, S., et al. (2024). Service Delivery In Select Public Sector Banks Using Queuing Model-A Comparative Study. *Chandra Ch, S., Kumar, AP, &Samisetty, S.(2025). Service delivery in select public sector banks using queuing model: A comparative study. Academy of Marketing Studies Journal*, 29(1):1–17.

Dai, J. G. and Shi, P. (2021). Recent Modeling and Analytical Advances in Hospital Inpatient Flow Management. *Production and Operations Management*, 30(6):1838–1862.

D'Auria, B., Adan, I. J., Bekker, R., and Kulkarni, V. (2022). An M/M/c Queue with Queueing-time Dependent Service Sates. *European Journal of Operational Research*, 299(2):566–579.

Gross, D. and Harris, C. M. (1998). *Fundamentals of Queueing Theory*. John Wiley & Sons, New York, 3 edition.

Kleinrock, L. (1975). *Queueing Systems, Volume I: Theory*. John Wiley & Sons, New York.

Kumar, A., Prikshit, P., and Gonder, S. S. C. (2024). Queueing System Optimization and Computational Analysis: A Comprehensive Investigation into Research Evaluation. In *2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE)*, pages 391–396. IEEE.

Long, Z., Tezcan, T., and Zhang, J. (2024). Routing and Staffing in Customer Service Chat Systems with Generally Distributed Service and Patience Times. *Manufacturing & Service Operations Management*, 26(5):1674–1691.

Manitz, M. and Piehl, M.-P. (2023). A Fast Staffing Algorithm for Multistage Call Centers with Impatient Customers and Time-Dependent Overflow. *Central European Journal of Operations Research*, 32(3):763–791.

Medhi, J. (2002). *Stochastic models in queueing theory*, pages 1–46. Elsevier.

Shaba, Y. (2019). Queuing Theory and ATM Service Optimization: Empirical Evidence from First Bank Plc, Kaura Namoda Branch, Zamfara State. *American Journal of Operations Management and Information Systems*, 4(3):80–86.

Sharma, R. (2014). Mathematical Analysis of Queue with Phase Service: An overview. *Advances in operations research*, 2014(1):1–19.

Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M. (2018). *Fundamentals of queueing theory*. John Wiley & Sons.

Zhang, Z. G. (2026). Toward next-generation queueing theory: research directions for intelligent service and manufacturing systems. *Quality Technology & Quantitative Management*, pages 1–18.