UNDER PEER REVIEW

Peak-Hour Teller Staffing Optimization in Retail Banking: An MMc Queuing Analysis

Abstract

Long wait times during banking hours remain a serious operational challenge in many developing countries. This paper uses an $M/M/c$ queuing model to analyze general service system congestion in a retail banking system. Customer arrivals are believed to follow a Poisson process and service times are widely distributed. The arrival and service rates are used to determine key performance measures such as traffic intensity, probability of delay, expected queue length and average waiting time. The findings demonstrate that whereas seven tellers greatly lessen congestion, four tellers lead to high congestion and lengthy wait times. This research paper demonstrates the effectiveness of queuing theory as a quantitative decision support tool to optimize teller capacity in banking systems.

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

1 Introduction

Queueing model has been extensively used to analyse congestion in service systems including banking, healthcare, transportation and telecommunications. Mathematical models that explain 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). Queueing models have been applied in banking settings to study staffing techniques, wait times and teller utilization in an effort to increase customer satisfaction (Shaba, 2019; Sharma, 2014). Retail banks in many developing economies face extreme capacity issues at high volumes of transactions, especially on days of salary payment. Long lines, long waiting lists, and overcrowded banking halls don't help customer satisfaction or operations. And yet for all the years we have been struggling with this issue, the allocation of tellers in many banks remains driven more by managerial intuition than quantitative evaluation. In addition, this study analyses the salary payment day congestion in a retail banking institution using the actual real operating data, and follows the classical $M/M/c$ queuing 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 changes in teller capacity lead to the most important decreases in congestion. Increasing the number of active tellers from four to seven, in particular, reduces the average waiting time by more than 85%. The study illustrates how queuing theory can be applied to operational retail banking staff deployment decisions.

2 The $M/M/c$ Queuing 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 Data Collection and Parameter Estimation

The data was gathered from a retail banking hall at the peak hours of salary payments between the hours of 9:00 am to 1:00 pm. Customer arrivals were collected on several observation days and the inter-arrival times were computed. Time stamps from transaction start to completion were also recorded for service times at each teller. On average, the arrival rate was estimated as

$$\lambda = 30 \text{ customers per hour,} \tag{3.1}$$

and an estimate of the average service rate per teller was

$$\mu = 8 \text{ customers per hour.} \tag{3.2}$$

4 Data Presentation and Analysis

4.1 Traffic Intensity

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

$$\rho = \frac{30}{4 \times 8} = 0.9375. \tag{4.1}$$

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

$$\rho = \frac{30}{5 \times 8} = 0.75, \tag{4.2}$$

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 to further reduction of traffic intensity.

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

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

$$P(\text{wait}) = \frac{\frac{3.75^4}{4!} \cdot \frac{1}{1 - 0.9375}}{\sum_{n=0}^{3} \frac{3.75^n}{n!} + \frac{3.75^4}{4!} \cdot \frac{1}{1 - 0.9375}}. \tag{4.5}$$

$$\text{Numerator} = \frac{3.75^4}{4!} \cdot \frac{1}{1 - 0.9375} \approx 131.7 \tag{4.6}$$

$$\text{Denominator} = \sum_{n=0}^{3} \frac{3.75^n}{n!} + 131.7 \approx 152.27 \tag{4.7}$$

$$P(\text{wait}) = \frac{131.7}{152.27} \approx 0.865. \tag{4.8}$$

For $c = 5$, $\lambda/\mu = 30/8 = 3.75$ and $\rho = 0.75$

$$P(\text{wait}) = \frac{\frac{3.75^5}{5!} \cdot \frac{1}{1-0.75}}{\sum_{n=0}^{4} \frac{3.75^n}{n!} + \frac{3.75^5}{5!} \cdot \frac{1}{1-0.75}} \tag{4.9}$$

$$\text{Numerator} = \frac{3.75^5}{5!} \cdot \frac{1}{1 - 0.75} = \frac{759.375}{120} \cdot \frac{1}{0.25} = 6.3281 \cdot 4 \approx 25.31 \tag{4.10}$$

$$\text{Denominator} = \sum_{n=0}^{4} \frac{3.75^n}{n!} + 25.31 = 1 + 3.75 + 7.03125 + 8.7891 + 8.23 + 25.31 \approx 54.11 \tag{4.11}$$

$$P(\text{wait}) = \frac{25.31}{54.11} \approx 0.468 \tag{4.12}$$

For $c = 6$, $\lambda/\mu = 30/8 = 3.75$, $\rho = 0.625$, we haves

$$P(\text{wait}) = \frac{\frac{3.75^6}{6!} \cdot \frac{1}{1-0.625}}{\sum_{n=0}^{5} \frac{3.75^n}{n!} + \frac{3.75^6}{6!} \cdot \frac{1}{1-0.625}} \tag{4.13}$$

$$\text{Numerator} = \frac{3.75^6}{6!} \cdot \frac{1}{1 - 0.625} = \frac{2847.97}{720} \cdot \frac{1}{0.375} \approx 3.955 \cdot 2.667 \approx 10.55 \tag{4.14}$$

$$\text{Denominator} = \sum_{n=0}^{5} \frac{3.75^n}{n!} = 1 + 3.75 + 7.03125 + 8.7891 + 8.23 + 6.25 + 10.55 \approx 44.6 \tag{4.15}$$

$$P(\text{wait}) = \frac{10.55}{44.6} \approx 0.236 \tag{4.16}$$

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

$$P(\text{wait}) = \frac{\frac{3.75^7}{7!} \cdot \frac{1}{1-0.536}}{\sum_{n=0}^{6} \frac{3.75^n}{n!} + \frac{3.75^7}{7!} \cdot \frac{1}{1-0.536}} \tag{4.17}$$

$$\text{Numerator} = \frac{3.75^7}{7!} \cdot \frac{1}{1 - 0.536} = \frac{10647.38}{5040} \cdot \frac{1}{0.464} \approx 2.114 \cdot 2.155 \approx 4.56 \tag{4.18}$$

$$\text{Denominator} = \sum_{n=0}^{6} \frac{3.75^n}{n!} = 1 + 3.75 + 7.03125 + 8.7891 + 8.23 + 6.25 + 3.57 + 4.56 \approx 43.18 \quad (4.19)$$

$$P(\text{wait}) = \frac{4.56}{43.18} \approx 0.105 \quad (4.20)$$

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 follows that

$$L_q = \frac{0.865 \times 0.9375}{1 - 0.9375} = \frac{0.81094}{0.0625} \approx 12.98 \text{ customers.} \quad (4.21)$$

For $c = 5$, it follows that

$$L_q = \frac{0.468 \times 0.75}{1 - 0.75} = \frac{0.351}{0.25} \approx 1.40 \text{ customers.} \quad (4.22)$$

For $c = 6$, it follows that

$$L_q = \frac{0.236 \times 0.625}{1 - 0.625} = \frac{0.1475}{0.375} \approx 0.39 \text{ customers.} \quad (4.23)$$

For $c = 7$, it follows that

$$L_q = \frac{0.105 \times 0.536}{1 - 0.536} = \frac{0.05628}{0.464} \approx 0.12 \text{ customers.} \quad (4.24)$$

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 $c = 4$,

$$W_q = \frac{12.98}{30} = 0.433 \text{ hours } \approx 25.96 \text{ minutes.} \quad (4.25)$$

For $c = 5$,

$$W_q = \frac{1.40}{30} = 0.0467 \text{ hours } \approx 2.80 \text{ minutes.} \quad (4.26)$$

For $c = 6$,

$$W_q = \frac{0.39}{30} = 0.013 \text{ hours } \approx 0.78 \text{ minutes.} \tag{4.27}$$

For $c = 7$,

$$W_q = \frac{0.12}{30} = 0.004 \text{ hours } \approx 0.24 \text{ minutes.} \tag{4.28}$$

Table 1: Queue Performance Measures

Number of Tellers (c)	ρ	P(wait)	L_q(customers)	W_q (minutes)
4	0.9375	0.865	13	25.96
5	0.7500	0.468	1	2.80
6	0.6250	0.236	0	0.78
7	0.5360	0.105	0	0.24

From Table 1, the results show that when the number of tellers is small ($c = 4$), the traffic intensity is very close to unity ($\rho = 0.9375$), resulting in longer waiting times in the queue. As the number of tellers increases, congestion in the system decreases significantly. When $c = 5$, the average waiting time drops to about three minutes, and for $c = 6$ and $c = 7$, the waiting time becomes less than one minute. This demonstrates that increasing teller capacity in an $M/M/c$ queueing system significantly improves service efficiency and reduces customer delay.

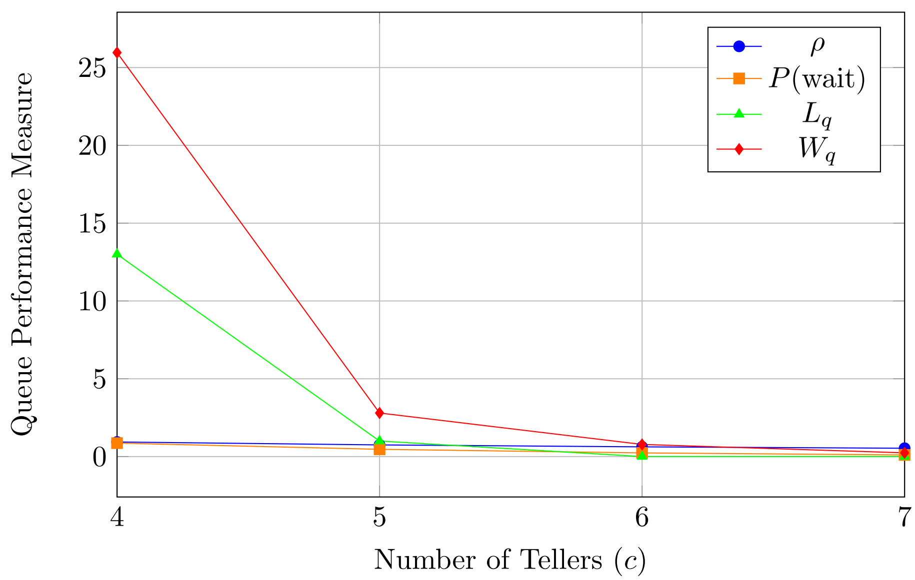

Figure 1: Queue Performance Measures for Different Numbers of Tellers

Figure 1 shows the impact on the 4 performance measures of server utilization (ρ), probability of waiting P(wait), expected queue length L_q and average wait W_q when we introduce more tellers (c). When we increase the number of tellers from 4 to 7, all congestion-related measures decrease significantly. On $c = 4$, it is very high utilization ($\rho = 0.9375$), meaning that, the tellers are busy a lot of the time. However, at that level of utilization, the expected queue length ($L_q = 13$ customers) and average waiting time ($W_q = 25.96$ minutes) become very high. This indicates that the congestion of the $M/M/c$ system increases quickly with ρ getting near unity. When number of tellers increase to 5, congestion reduces markedly. Estimated queue length falls sharply to 1 customer, and the average waiting time falls to 2.80 minutes. Congestion is also lessened by increasing that number to 6 and 7 tellers. The system exhibits low utilization ($\rho = 0.5360$), minimal queue length (0.24), low waiting probability (0.105), and average wait time (0.24 minutes) by $c = 7$. These findings imply that when usage is close to unity, the system is extremely sensitive. Performance will be significantly improved by even a small change in the number of tellers near this important location. However, the marginal benefits of having more tellers decrease as one moves away from this point. This implies that a better realistic balance between operational costs and service efficiency may be achieved with 5 or 6 tellers.

5 Model Parameters Sensitivity Analysis

In this section, the arrival rate λ and service rate μ in the field and 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

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

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

$$\rho = \frac{27}{40} = 0.675 \tag{5.6}$$

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

$$\frac{\lambda}{\mu} = \frac{27}{8} = 3.375 \tag{5.7}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.8}$$

$$\sum_{n=0}^{4} \frac{(3.375)^n}{n!} = \frac{(3.375)^0}{0!} + \frac{(3.375)^1}{1!} + \frac{(3.375)^2}{2!} + \frac{(3.375)^3}{3!} + \frac{(3.375)^4}{4!} \tag{5.9}$$

$$= 1 + 3.375 + 5.695 + 6.407 + 5.407 \tag{5.10}$$

$$= 21.884 \tag{5.11}$$

$$\frac{(3.375)^5}{5!(1-0.675)} = \frac{437.89}{120 \times 0.325} = 11.227 \tag{5.12}$$

Therefore,

$$P_0 = [21.884 + 11.227]^{-1} \tag{5.13}$$

$$P_0 = (33.111)^{-1} = 0.0302 \tag{5.14}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.15}$$

$$P(\text{wait}) = 11.227 \times 0.0302 \tag{5.16}$$

$$P(\text{wait}) = 0.339 \tag{5.17}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.18}$$

$$L_q = \frac{0.339 \times 0.675}{1-0.675} \tag{5.19}$$

$$L_q = \frac{0.229}{0.325} \tag{5.20}$$

$$L_q = 0.70 \tag{5.21}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.22}$$

$$W_q = \frac{0.70}{27} \tag{5.23}$$

$$W_q = 0.0259 \text{ hours} \tag{5.24}$$

Converting to minutes,

$$W_q = 0.0259 \times 60 \tag{5.25}$$

$$W_q = 1.57 \text{ minutes} \tag{5.26}$$

Second Analysis

Considering the parameters

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

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.27}$$

$$\rho = \frac{30}{5 \times 8} \tag{5.28}$$

$$\rho = \frac{30}{40} = 0.750 \tag{5.29}$$

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

$$\frac{\lambda}{\mu} = \frac{30}{8} = 3.75 \tag{5.30}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)}\right]^{-1} \tag{5.31}$$

$$\sum_{n=0}^{4} \frac{(3.75)^n}{n!} = \frac{(3.75)^0}{0!} + \frac{(3.75)^1}{1!} + \frac{(3.75)^2}{2!} + \frac{(3.75)^3}{3!} + \frac{(3.75)^4}{4!} \tag{5.32}$$

$$= 1 + 3.75 + 7.031 + 8.789 + 8.239 \tag{5.33}$$

$$= 28.809 \tag{5.34}$$

$$\frac{(3.75)^5}{5!(1-0.75)} = \frac{741.58}{120 \times 0.25} = 24.719 \tag{5.35}$$

Therefore,

$$P_0 = [28.809 + 24.719]^{-1} \tag{5.36}$$

$$P_0 = (53.528)^{-1} = 0.0187 \tag{5.37}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.38}$$

$$P(\text{wait}) = 24.719 \times 0.0187 \tag{5.39}$$

$$P(\text{wait}) = 0.462 \tag{5.40}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.41}$$

$$L_q = \frac{0.462 \times 0.75}{1-0.75} \tag{5.42}$$

$$L_q = \frac{0.3465}{0.25} \tag{5.43}$$

$$L_q = 1.39 \tag{5.44}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.45}$$

$$W_q = \frac{1.39}{30} \tag{5.46}$$

$$W_q = 0.0463 \text{ hours} \tag{5.47}$$

Converting to minutes,

$$W_q = 0.0463 \times 60 \tag{5.48}$$

$$W_q = 2.77 \text{ minutes} \tag{5.49}$$

Third Analysis

Consider the parameters

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

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.50}$$

$$\rho = \frac{33}{5 \times 8} \tag{5.51}$$

$$\rho = \frac{33}{40} = 0.825 \tag{5.52}$$

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

$$\frac{\lambda}{\mu} = \frac{33}{8} = 4.125 \tag{5.53}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.54}$$

$$\sum_{n=0}^{4} \frac{(4.125)^n}{n!} = \frac{(4.125)^0}{0!} + \frac{(4.125)^1}{1!} + \frac{(4.125)^2}{2!} + \frac{(4.125)^3}{3!} + \frac{(4.125)^4}{4!} \tag{5.55}$$

$$= 1 + 4.125 + 8.508 + 11.700 + 12.066 \tag{5.56}$$

$$= 37.399 \tag{5.57}$$

$$\frac{(4.125)^5}{5!(1-0.825)} = \frac{1199.66}{120 \times 0.175} = 57.13 \tag{5.58}$$

Therefore,

$$P_0 = [37.399 + 57.13]^{-1} \tag{5.59}$$

$$P_0 = (94.529)^{-1} = 0.0106 \tag{5.60}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.61}$$

$$P(\text{wait}) = 57.13 \times 0.0106 \tag{5.62}$$

$$P(\text{wait}) = 0.603 \tag{5.63}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.64}$$

$$L_q = \frac{0.603 \times 0.825}{1 - 0.825} \tag{5.65}$$

$$L_q = \frac{0.497}{0.175} \tag{5.66}$$

$$L_q = 2.84 \tag{5.67}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.68}$$

$$W_q = \frac{2.84}{33} \tag{5.69}$$

$$W_q = 0.0861 \text{ hours} \tag{5.70}$$

Converting to minutes,

$$W_q = 0.0861 \times 60 \tag{5.71}$$

$$W_q = 5.17 \text{ minutes} \tag{5.72}$$

Fourth Analysis

Consider the parameters

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

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.73}$$

$$\rho = \frac{30}{5 \times 7} \tag{5.74}$$

$$\rho = \frac{30}{35} = 0.857 \tag{5.75}$$

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

$$\frac{\lambda}{\mu} = \frac{30}{7} = 4.286 \tag{5.76}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)}\right]^{-1} \tag{5.77}$$

$$\sum_{n=0}^{4} \frac{(4.286)^n}{n!} = 1 + 4.286 + 9.184 + 13.121 + 14.056 \tag{5.78}$$

$$= 41.647 \tag{5.79}$$

$$\frac{(4.286)^5}{5!(1-0.857)} = \frac{1445.1}{120 \times 0.143} = 84.21 \tag{5.80}$$

Therefore,

$$P_0 = [41.647 + 84.21]^{-1} \tag{5.81}$$

$$P_0 = (125.857)^{-1} = 0.00795 \tag{5.82}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.83}$$

$$P(\text{wait}) = 84.21 \times 0.00795 \tag{5.84}$$

$$P(\text{wait}) = 0.669 \tag{5.85}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.86}$$

$$L_q = \frac{0.669 \times 0.857}{1 - 0.857} \tag{5.87}$$

$$L_q = \frac{0.573}{0.143} \tag{5.88}$$

$$L_q = 4.02 \tag{5.89}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.90}$$

$$W_q = \frac{4.02}{30} \tag{5.91}$$

$$W_q = 0.134 \text{ hours} \tag{5.92}$$

Converting to minutes,

$$W_q = 0.134 \times 60 \tag{5.93}$$

$$W_q = 8.03 \text{ minutes} \tag{5.94}$$

Fifth Analysis
Consider the parameters

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

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.95}$$

$$\rho = \frac{30}{5 \times 9} \tag{5.96}$$

$$\rho = \frac{30}{45} = 0.667 \tag{5.97}$$

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

$$\frac{\lambda}{\mu} = \frac{30}{9} = 3.333 \tag{5.98}$$

Probability of Zero Customers in the System

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.99}$$

$$\sum_{n=0}^{4} \frac{(3.333)^n}{n!} = 1 + 3.333 + 5.556 + 6.173 + 5.144 \tag{5.100}$$

$$= 21.206 \tag{5.101}$$

$$\frac{(3.333)^5}{5!(1-0.667)} = \frac{411.52}{120 \times 0.333} = 10.29 \tag{5.102}$$

Therefore,

$$P_0 = [21.206 + 10.29]^{-1} \tag{5.103}$$

$$P_0 = (31.496)^{-1} = 0.0318 \tag{5.104}$$

Probability that a Customer Waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.105}$$

$$P(\text{wait}) = 10.29 \times 0.0318 \tag{5.106}$$

$$P(\text{wait}) = 0.327 \tag{5.107}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.108}$$

$$L_q = \frac{0.327 \times 0.667}{1-0.667} \tag{5.109}$$

$$L_q = \frac{0.218}{0.333} \tag{5.110}$$

$$L_q = 0.65 \tag{5.111}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} \tag{5.112}$$

$$W_q = \frac{0.65}{30} \tag{5.113}$$

$$W_q = 0.0217 \text{ hours} \tag{5.114}$$

Converting to minutes,

$$W_q = 0.0217 \times 60 \tag{5.115}$$

$$W_q = 1.31 \text{ minutes} \tag{5.116}$$

Sixth Analysis
Consider the parameters

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

Traffic Intensity

$$\rho = \frac{\lambda}{c\mu} \tag{5.117}$$

Substituting the values,

$$\rho = \frac{33}{5 \times 7} \tag{5.118}$$

$$\rho = \frac{33}{35} = 0.943 \tag{5.119}$$

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

$$\frac{\lambda}{\mu} = \frac{33}{7} \approx 4.714. \tag{5.120}$$

Probability of zero customers in the system

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \tag{5.121}$$

$$\sum_{n=0}^{4} \frac{(4.714)^n}{n!} = 1 + 4.714 + 11.10 + 17.42 + 20.53 \tag{5.122}$$

$$= 54.76 \tag{5.123}$$

$$\frac{(4.714)^5}{5!(1-0.943)} = \frac{2298.7}{120 \times 0.057} \approx 336.9 \tag{5.124}$$

Therefore,

$$P_0 = \frac{1}{54.76 + 336.9} = \frac{1}{391.66} \approx 0.00255 \tag{5.125}$$

Probability that a customer waits

$$P(\text{wait}) = \frac{(\lambda/\mu)^c}{c!(1-\rho)} P_0 \tag{5.126}$$

$$P(\text{wait}) = 336.9 \times 0.00255 \approx 0.861 \tag{5.127}$$

Average Queue Length

$$L_q = \frac{P(\text{wait})\rho}{1-\rho} \tag{5.128}$$

$$L_q = \frac{0.861 \times 0.943}{1 - 0.943} \tag{5.129}$$

$$L_q = \frac{0.812}{0.057} \approx 14.20 \tag{5.130}$$

Average Waiting Time

$$W_q = \frac{L_q}{\lambda} = \frac{14.20}{33} \approx 0.430 \text{ hours} \tag{5.131}$$

Convert to minutes:

$$W_q = 0.430 \times 60 \approx 25.83 \text{ minutes} \tag{5.132}$$

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

System Parameters		**Performance Measures**			
λ (cust./hr)	μ (cust./hr)	ρ	$P(\text{wait})$	L_q (cust.)	W_q (min)
27	8	0.675	0.339	0.70	1.57
30	8	0.750	0.462	1.39	2.77
33	8	0.825	0.603	2.84	5.17
30	7	0.857	0.669	4.02	8.03
30	9	0.667	0.327	0.65	1.31
33	7	0.943	0.861	14.20	25.83

From Table 2, several key observations can be made. Firstly, increasing the arrival rate λ while keeping the service rate μ constant results in higher server utilization. For instance, when $\mu = 8$, increasing λ from 27 to 33 customers per hour increases the utilization ρ from 0.675 to 0.825. Consequently, the probability that a customer waits, the average queue length, and the average waiting time also increase. This shows that as the system becomes busier, more customers experience delays, and queues grow longer. Secondly, variations in the service rate μ significantly affect system performance. For an arrival rate of $\lambda = 30$, reducing the service rate from 9 to 7 customers per hour increases the utilization from 0.667 to 0.857. Correspondingly, the probability of waiting rises from 0.327 to 0.669, the average queue length from 0.65 to 4.02 customers, and the average waiting time from 1.31 to 8.03 minutes. This indicates that slower service leads to increased congestion and longer customer delays. Furthermore, the impact of high utilization is particularly notable. When the system operates near full capacity, such as for $\lambda = 33$ and $\mu = 7$ resulting in $\rho = 0.943$, the probability of waiting reaches 0.861, the average queue length grows to 14.20 customers, and the average waiting time extends to 25.83 minutes. This illustrates the exponential growth of queueing metrics as utilization approaches unity, a characteristic behavior of $M/M/c$ queueing systems. Conversely, low utilization corresponds to improved system performance. For $\lambda = 30$ and $\mu = 9$, the utilization is 0.667, yielding a low probability of waiting (0.327), a short average queue length (0.65 customers), and a minimal average waiting time (1.31 minutes). This emphasizes the importance of maintaining adequate service capacity relative to arrival demand to ensure smooth operation and customer satisfaction. Overall, the table shows the sensitivity of queueing performance to both arrival and service rates. Maintaining moderate utilization ensures acceptable waiting times and manageable queue lengths, while high utilization can lead to significant congestion and long delays. These insights are crucial for designing efficient service systems and allocating adequate resources to meet customer demand.

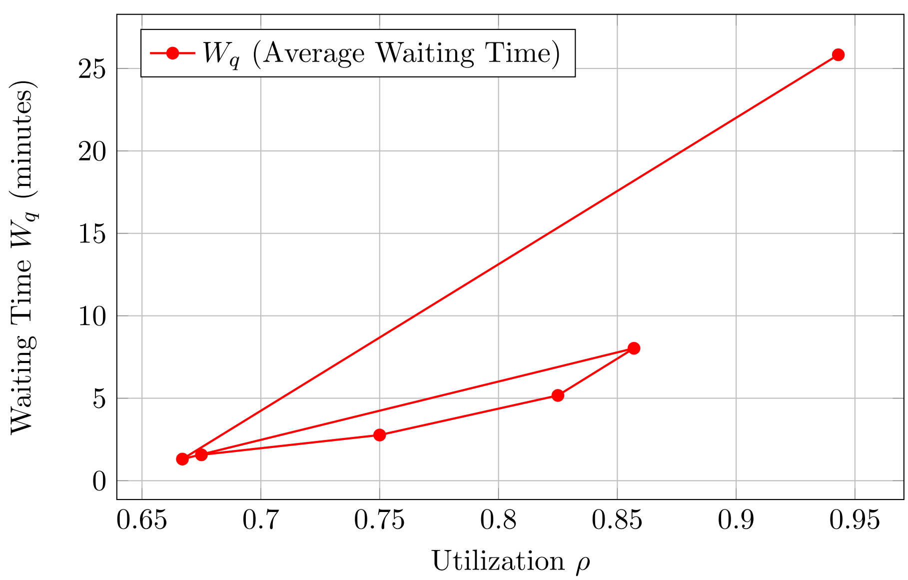

Figure 2: Relationship between utilization ρ and average waiting time W_q.

Figure 2 illustrates the waiting time (W_q) of the M/M/c model. Due to the presence of multiple tellers, waiting times are minimal at low to moderate utilization. However, as ρ approaches high values, W_q grows quickly, indicating that high utilization greatly exacerbates customer waits. This emphasizes how crucial it is to maintain appropriate service levels by striking a balance between teller availability and customer arrival 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.

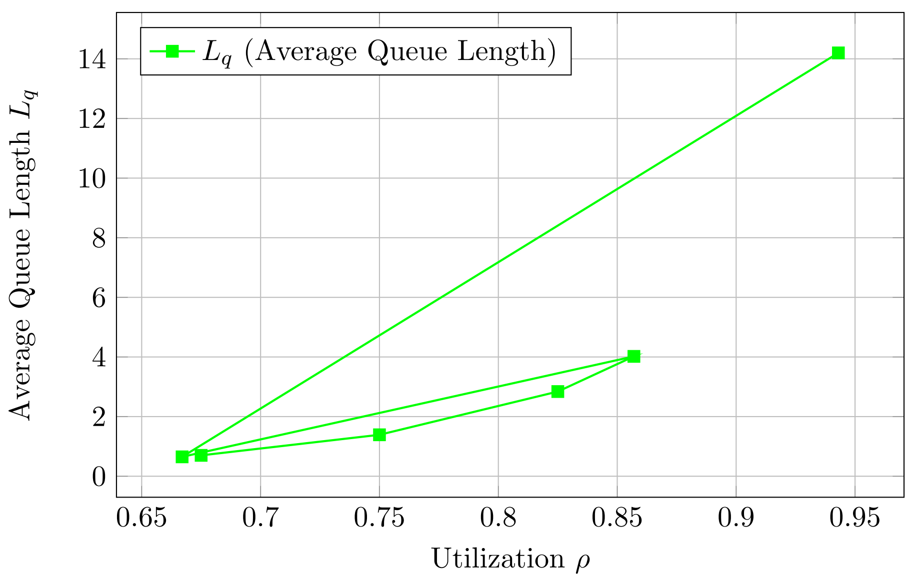

Figure 3: Relationship between utilization ρ and average queue length L_q.

Figure 3 shows how the average queue length (L_q) and system utilization (ρ) are related of the $M/M/c$ model. The average queue length climbs rapidly as ρ approaches high levels, indicating that the bank becomes increasingly congested near full capacity, although it stays relatively low for moderate values of ρ as usage increases. This pattern illustrates the crucial effect of getting close to the maximum system load on performance by reflecting the nonlinear rise of queue length at high utilization leading to potential customer satisfaction.

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.

Conclusions

This study evaluated peak-hour teller staffing in a retail banking system using the M/M/c queuing model. The analysis showed that congestion becomes severe when the number of tellers is insufficient relative to customer arrivals. At four tellers, the system operated close to full utilization, resulting in a long expected queue and an average waiting time of 25.96 minutes. Increasing teller capacity produced clear improvements in service performance. With five tellers, the average waiting time declined to 2.80 minutes, and further increases to six and seven tellers reduced waiting time to below one minute. The sensitivity analysis also showed that system performance is strongly influenced by both arrival rate and service rate. Higher arrival rates and lower service rates increased utilization, queue length, and waiting time, particularly when utilization approached one. These results indicate that moderate utilization should be maintained during peak periods to avoid excessive customer delay. Overall, the study supports the use of queuing models as practical tools for staffing decisions, capacity planning, and service improvement in retail banking operations.

Limitations

This study is limited by the scope and detail of the available operational data. The model uses average arrival and service rates, but the analysis does not report the full number of observation days, total customer count, variability of service times, or confidence intervals for the estimated parameters. The M/M/c framework assumes Poisson arrivals, exponential service times, parallel identical tellers, and first-come-first-served service discipline. These assumptions were not statistically tested in the present analysis and may not fully represent all banking conditions. The study also focuses on teller staffing during peak salary-payment hours and does not include off-peak periods, customer abandonment, priority service, self-service channels, or the cost of additional staffing. Therefore, the findings should be interpreted as model-based estimates for the observed setting rather than universal staffing rules. Future work should use larger datasets, validate distributional assumptions, and include cost-service trade-off analysis.

Declaration of AI Use

This manuscript was prepared through the combined contributions of all author(s), including contributions to the study design, data, content development, results, interpretation, and related scholarly work. The author(s) acknowledge the use of Grammarly and ChatGPT to assist with grammar checking, language refinement, reference formatting. These AI-assisted tools were not used as authors and did not replace the intellectual contributions or scholarly judgment of the author(s). All AI-assisted outputs, including content, references, and interpretations, were carefully reviewed, revised, verified, and approved by the author(s). The author(s) accept full responsibility for the accuracy, integrity, and final content of the manuscript.

References

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

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.

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.