International Journal of Scientific and Management Research
Volume 8 Issue 6 (June) 2025
ISSN: 2581-6888
Page: 176-189

Assessment of Single and Multi-Server Exponential Queuing Models in Banking System

Kolawole Daramola[1], Ajeka Friday[2], & Enoch Yabkwa Yanshak[3]
[1]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, **Nigeria.**
[2]Department of Computer Systems Technology, North Carolina A&T State University, **USA**
[3]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, **Nigeria.**
DOI - http://doi.org/10.37502/IJSMR.2025.8611

Abstract

Queuing occurs when the number of customers awaiting service exceeds the system's service capacity, often leading to extended wait times and congestion. The banking sector in Nigeria is facing challenges related to prolonged queues, adversely impacting the nation's economic growth. This article assessed both single and multi-server exponential queuing models. Performance indicators for both single and multi-server queuing models, such as utilization factor, average queue length, average system length, average queue waiting time, and average system waiting time, were computed and analysed. The result revealed that the (M/M/S): (FCFS/∞/∞) model outperforms the (M/M/1): (FCFS/∞/∞) model by minimizing customer waiting time from approximately 2.0 minutes to 0.03 seconds. The findings emphasized the efficiency of employing multiple servers; this shows that introducing more servers reduces the workload per server, potentially attracting more customers. Furthermore, a comprehensive analysis of cost implications and utilization factors served as a target for achieving a balance between minimizing cost and ensuring an optimal server level at the customer service of Access Bank Plc. The results indicated that for an optimal balance between service level and total cost, adopting the (M/M/4): (FCFS/∞/∞) model is recommended, as it results in a lower cost on the maintenance and servicing of queuing facilities (N9,104.99) compared to (N10,793.43).

Keywords: Queue, Arrival rate, Service rate, Single-Server, Multi-server model

1. Introduction

Queuing is the process of lining up customers to place orders, with the term "queue" originating from the French word queu, meaning "to line up." Queuing can involve human customers or physical entities waiting for service in places such as petrol stations, banks, mechanic workshops, airports, car parks, and goods for shipment. In Nigeria, queuing is a common experience, especially in banks and fuelling stations (Nsude et al., 2017). Queuing occurs when the number of individual items waiting in line exceeds the maximum capacity of the system. This happens when the line grows longer than the available servers. The line is made up of individuals who are waiting to complete a task, process or receive a service, and the number of

people waiting goes beyond the limit set by the system's capacity. The main causes of long queues include inadequate service systems and low service quality, which increase the waiting time. In service-related industries, reducing waiting time and providing prompt service are critical factors in enhancing customer satisfaction, which leads to improved service quality (Nsude et al., 2017).

Queuing theory is a mathematical study of waiting lines and their associated problems. Service operates on a First-In-First-Out (FIFO) basis, with customer being served one at a time. The queuing theory is a quantitative analysis technique used to predict the characteristics of a waiting line. It enables the mathematical analysis of queuing behaviour, including customer arrival time and the amount of time a customer waits in the system in a real-world queuing situation (Yakubu & Ussiph, 2014). By applying the queuing theory, companies can estimate and improve their service capacity, which helps them provide better service quality and reduce waiting time (Amit & Nurdia, 2018).

Waiting time refers to the amount of time that a customer spends in a queue before receiving service (Yakubu & Ussiph, 2014). Waiting for service is a common occurrence in various customer-centric settings, including gas stations, banks, hospitals, restaurants, and departmental stores (Adewole, 2016). When customers visit a service facility, they may need to wait for their turn, which can be frustrating if the waiting time is long. Such situations can lead to customer dissatisfaction, as they are unable to receive the level of service they desire (Nkrumah, et al., 2015). Moreover, some customers may faint or even die while waiting in the queue. Service delays occur when the need for a service surpasses the existing capacity, resulting in the formation of a queue (Nkrumah, et al., 2015).

Service time in a queuing model refers to the amount of time it takes to serve a customer or process a task at a service point, such as a checkout counter or a customer service desk. It is a critical factor in analyzing and modelling queues or waiting lines. Service time can vary based on factors like the complexity of the task, the efficiency of the service, and the nature of the customers or tasks in the queue. In queuing theory, understanding service time distribution helps in predicting waiting times, queue lengths, and overall system performance.

The banking sectors faces challenges related to prolonged queues due to inadequate queue management. This issue has detrimental effects on the country's economic growth and development. The extended waiting times lead to customer dissatisfaction and often prompt them to leave the system prematurely, hindering productivity and complicating customer engagement. In the increasingly competitive banking environment, customers expect high-quality, cost-effective, and prompt service delivery. Waiting is perceived as inconvenient, and the associated time becomes a tangible cost to customers. Moreover, the extended waiting period incurs an economic loss for individuals in the queue, necessitating the minimization, if not the elimination, of challenges posed by long queues in the country's banking system. To ensure the efficient operation of banking services and enhance customer satisfaction, it is imperative to adopt strategies that reduce waiting times. To address these concerns, this study concentrates on customer services, utilizing single and multi-server exponential queueing models.

1.1 Theoretical Framework

The beginnings of queueing theory can be identified in the early 20th century, when Danish engineer (Erlang, 1909) extensively employed this theory to examine the dynamics of telephone networks. Recognized as the pioneer of queuing theory, Erlang formulated numerous queueing principles that continue to be fundamental in evaluating queueing performance today.

Queuing theory has its origin in research by Erlang (1909) when he created models to describe the Copenhagen telephone exchange. The ideas have since seen applications including telecommunications, traffic engineering, computing, and the design of factories, shops, offices, banks and hospitals.

1.2 Aim and Objectives

This study aims to assess single and multi-server exponential queuing models in banking system. The specific objectives of the study are to:

i. Determines the average waiting time of the customers on the queue
ii. Estimate the waiting cost of customers and the service cost of the bank facilities
iii. Investigate the suitable queuing model for bank services.

By focusing on these objectives, this study aims to provide insights into how long customers are likely to wait in a queue, how many customers are likely to be in a queue at any given time, and how many servers are needed to minimize queue length and wait times. findings from this study aim to help banks make informed decisions about their service delivery strategies, such as staffing and operational processes. the study aims to contribute to the body of knowledge in queuing theory. The study not only contributes to the global discourse on the queue model but also sheds light on the understanding of queuing theory and its practical applications.

2. Empirical Review

Khaskheli et al. (2020) conducted a study to enhance the efficiency of the existing queuing systems in the outpatient departments (OPDs). the study aimed to identify the optimal number of receptionists and doctors. They initiated their investigation by selecting the most congested OPD, specifically the medical OPD in the initial case hospital. Subsequently, they replicated the same study in another public sector hospital, referred to as case hospital 2, situated in Sindh, Pakistan. Data was collected over two weeks, with a focus on various parameters including the rate of patient arrivals, the rate at which patients are served, the number of available servers, the compensation of the staff, and the associated costs incurred by patients waiting for their turn. The distribution of patient arrivals and service times was validated using the input analyzer function of Rockwell Arena 14.5, based on the assumptions of the multi-server queuing model (M/M/c). To assess the performance of the queuing system, the research team computed performance metrics using TORA optimization software, while MS Excel was used for cost calculations and graph plots. The results indicated that increasing the number of receptionists and doctors by one at both OPDs would help minimize patient congestion and reduce waiting times.

Segun (2020) evaluated the effectiveness of healthcare service delivery using queuing theory. The primary objective was to assess the waiting, arrival, and service times of patients at the AAUA Health Centre and establish a suitable queuing system model through simulation techniques to validate its accuracy. The investigation was carried out at the AAUA Health Centre in Akungba Akoko and involved the use of both analytical and simulation methods to formulate an appropriate model. Data collection involved the use of a stopwatch to measure the time patients spent in different sections of the health centre, starting from the reception where patients arrived and received their hospital cards or registered, to the consulting room. The study gathered data on patient arrival times, waiting times, and service times on weekdays (Mondays through Fridays) over three weeks. Microsoft Excel was employed for data calculation and analysis. Subsequently, the current state of the patient queuing system was modelled and simulated using PYTHON software. The outcomes of the simulation model disclosed that, during the first week, the mean patient arrival rate on a Friday was lower than the mean patient service rate (i.e., 5.33 was less than 5.625 ($\lambda < \mu$)). This implied the formation of a waiting line, resulting in a continuous increase in waiting time, while the service facility remained consistently occupied. The comprehensive analysis of the entire AAUA health centre system indicated that queue length increased during peak periods. As a recommendation, the study underscores the significance of enhancing the quality of service provided to patients visiting the AAUA Health Centre.

Yuqi and Haoxuan (2020) conducted Bank Queuing Optimization using Markov Process. The researchers focused on analyzing the arrival of customers and the service provided by cashiers to create a birth-death process and its transfer rate matrix. They introduced the concepts of service intensity and customer satisfaction index and found that it is possible to calculate and predict the optimal number of open cashiers for different time periods.

Burodo et al. (2021) assessed queue management practices and patient satisfaction at specific hospitals in Northwestern Nigeria. Their primary objective was to gauge the impact of queue management practices on customer satisfaction. The study utilized a descriptive research design and randomly selected 2,850 registered patients from these hospitals. Questionnaires were distributed to these patients, but only 2,793 were returned. The investigation focused on aspects such as waiting times for service, conditions in the waiting area, and service quality concerning their effect on customer satisfaction. The results revealed that a notable portion of patients expressed dissatisfaction with the queue management practices in the chosen hospitals. Furthermore, a regression analysis demonstrated that all three dimensions of service quality had a significant impact on patient satisfaction. Service quality and the waiting environment showed a positive association with overall patient satisfaction, while service time had a negative influence on patient satisfaction. The research also put forth recommendations, suggesting that hospital management and staff should maintain up-to-date technological practices, undergo regular training in patient care, and consistently be reminded of their core values, mission, and vision in serving patients.

Samuel et al. (2021) applied a mathematical model to examine the waiting times at two particular banks in the Sekondi-Takoradi Metropolis. The goal of the study was to compare the average waiting times between these banks and estimate waiting times using a stochastic

model. To accomplish this, the study utilized a case study and observational research approach, gathering firsthand data. The two chosen banks were selected through purposeful sampling, with the target population consisting of customers intending to conduct transactions between 11 a.m. and 12 p.m. The first bank had sample sizes of 28, 17, and 20 on the first, second, and third days, respectively, with three servers each day. On the other hand, the second bank had sample sizes of 20, 9, and 17 on the first, second, and third days, respectively, with two servers each day. The analysis involved employing a multiple server (M/M/s) model, and the statistical tool used was Tora Software. The study's results indicated that the second bank had a higher utilization factor compared to the first bank. Additionally, the second bank consistently had a greater number of customers in the banking hall throughout the observation period, in contrast to the first bank. Finally, customers at the first bank required less time to complete their transactions when compared to the second bank.

Limlawan and Anussornnitisarn (2021) used the queuing management approach to investigate the unpredictable arrival pattern of customers at a service system. To prevent customers from abandoning the queue, they employed an Artificial Neural Network-based waiting time predictor. This method enables the system to anticipate and generate the expected waiting time for each customer. Consequently, customers have the option to engage in other activities instead of waiting in line until it's their turn.

Idigo et al. (2021) analyzed patient flow system in the radiology department of a tertiary hospital and made improvements to the scheduling of patients. They performed a cross-sectional study using data from 768 patients who arrived at the hospital via ambulance. They utilized MATLAB software to analyze various parameters related to the queue. The study revealed that the arrival pattern of patients was characterized by a high level of randomness. The researchers were able to estimate the peak arrival time and service rate based on their observations.

Shamsuddeen et al. (2022) examined the implementation of single and multi-server exponential queuing models in several selected hospitals located in the North-Western region of Nigeria. Primary data was gathered through an observational approach, where the researchers observed the time intervals between patient arrivals and the time taken for service in the hospitals included in the study. These observed time data were then used to calculate the queuing parameters for all the hospitals. They performed queue analyses based on the observational data collected from the eight hospitals that were part of the study. The results from the multi-server-single channel queuing models indicated that General Hospital Hunkuyi was identified as the most congested hospital. This conclusion was drawn based on its highest utilization rate and the largest number of patients waiting in the queue. In contrast, Ahmadu Bello University Teaching Hospital, Zaria, was found to be the least congested facility, and the Federal Medical Centre in Katsina had the fewest patients waiting in the queue. In light of these findings, the researchers recommended the allocation of additional medical staff to the hospitals, especially in cases where only one doctor is available at a time. This suggestion is aimed at converting the single-channel queuing units into multi-channel queuing units.

While previous studies have explored queuing models in healthcare and banking environments, this study extends the analysis by integrating cost optimization with queuing performance

metrics in a real-world Nigerian banking setting. Unlike earlier works that focused primarily on queue length and waiting time, our research incorporates a detailed cost-benefit assessment, enabling the identification of an optimal server configuration (M/M/4) that minimizes both customer wait time and operational expenditure. Moreover, by drawing data directly from a functioning bank branch and employing R-based simulation for multiple scenarios, this study provides a more granular and practical contribution to service optimization literature in developing economies.

3. Methodology

In this study, Primary data were collected through observations at the customer service unit of Access Bank Plc, Anyigba, Nigeria. The data collected was the customer's time of arrival, time spent waiting in the queue, time spent receiving services, and the time of departure from the bank. Customers arrive randomly as units and form a single file in the waiting line until a server serves them. Customers are served individually in parallel, according to the order in which they arrived. This helps with the estimation of the average number of customers that entered the bank per hour (α) and the average number of customers that were attended to per hour (μ). The data collected covers a period of two weeks, in which five days of the week from Monday to Friday (8:00 a.m. to 4:00 pm) were considered. A customer is considered to have arrived when he or she joins the queue at the customer service unit. The waiting time in the queue ended immediately after the customer gained access to the bank service. Also, the service time was recorded from the time the customer gained access to the bank service to the time of exit. The queue discipline observed is First Come, First Serve (FCFS). The data analysis is carried out using R software. Time records obtained in the process of data collection were entered into Microsoft Excel for conversion of the recorded time into interval time. The time interval is computed to get the mean arrival rate α and the mean service rate μ, which will then be entered into R software to create scenarios different from the real-life observations obtained from the process of data collection. The software, as stated earlier, is used to assess the performance measures of the queueing system.

3.1 The Single Queue Single-Server (M/M/1) Model

This model has Poisson arrival, Poisson service, and exponential inter-arrival /service time, including single channel, infinite system capacity and First-Come- First -Serve queue discipline. The letter "M" here is used to honour Russian mathematician Andreyevich Markov, who has published extensively on queuing theory (Varma, 2016). It is used to represent both the inter-arrival and service time distributions in Kendall's notation, as presented in Fig. 1

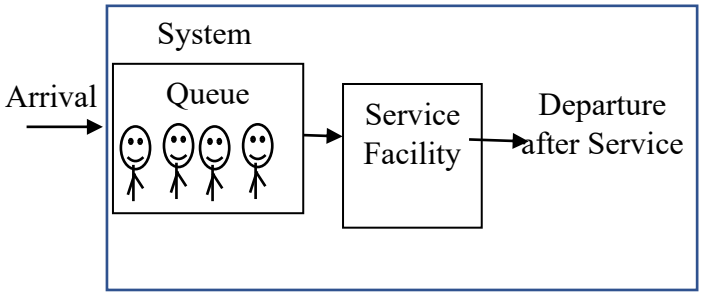

Fig. 1. A typical single-server queuing process

3.2 Characteristics of M/M/1 Queue Model

The M/M/1 queuing model means that the arrival and service time are exponentially distributed (Poisson process). In most cases, it is useful to determine the various waiting times and queue sizes of the queuing system in order to determine if the queue is in a good state or suggest better models for the queue. Little's formula is an equation that shows that the average number of customers in a queuing system is equal to their average arrival rate multiplied by the average amount of time spent in the system (El-Taha & Stidham, 1999).

The average server utilization, ρ is given by:

$$\rho = \frac{\alpha}{\mu} \tag{1}$$

Equation (1) is the proportion of time that a server actually spends with a customer where, α is the average number of customers arriving per unit of time and μ is the average number of customers completing the service per unit time.

$$P_O = \frac{\mu-\alpha}{\mu} \tag{2}$$

Equation (2) is the probability of having no customer in the system.

$$P_k = \left(\frac{\alpha}{\mu}\right)^k P_O \tag{3}$$

Equation (3) is the probability of having k customers in the system.
Where P_O is as defined in equation (2), k is the number of customers.

$$L_S = \frac{\alpha}{\mu-\alpha} \tag{4}$$

Equation (4) is the average number of customers in the system.

$$L_q = \frac{\alpha^2}{\mu(\mu-\alpha)} \tag{5}$$

Equation (5) is the average number of customers in the queue

$$W_q = \frac{\alpha}{\mu(\mu-\alpha)} \tag{6}$$

Equation (6) is the average waiting time in the queue

$$W_S = \frac{1}{\mu-\alpha} \tag{7}$$

Equation (7) is the average time spent in the system, including the waiting time.

3.3 The Single Queue Multi-Server (M/M/S) :(FCFS/∞/∞) Model

The (M/M/S) :(FCFS/∞/∞) model describes a scenario where there are multiple service facilities operating simultaneously, all providing the same service. In this format, several customers in the waiting line can receive service from any of these stations. It is assumed that customer arrivals and service times are distributed according to the Poisson and exponential distributions, respectively. Multiple channels, infinite system capacity, and First-Come-First-Serve queue discipline are applicable in Fig. 2.

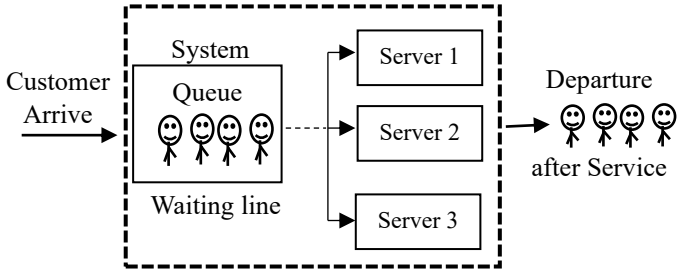

Fig. 2. A typical multi-server queuing process

3.4 Characteristics of (M/M/S) :(FCFS/∞/∞) Model

(i) If $k \geq S$, every one of the servers are occupied, and the maximum number of customer waiting in the queue will be $(k - S)$, then $\mu_k = S\mu$. Additionally, $\alpha_k = \alpha$ for $k = 0, 1, 2, 3, ...$

(ii) If $k < S$, every one of the customer might be served concurrently and there will be no queue. $S - k$ number of servers is likely to be idle and then $\mu_k = k\mu$ for $k = 0, 1, 2, 3, ... k$

$$\rho = \left(\frac{\alpha}{S\mu}\right) \tag{8}$$

There are S servers in parallel, i.e. an M/M/S system but the expected capacity per time is the $S\mu$ in equation (8)

$$P_k = \begin{cases} \frac{\rho^k P_0}{k!}; & k < S \\ \frac{\rho^k P_0}{S^{k-S} S!}; & k \geq S \end{cases} \tag{9}$$

Where P_k is the probability of having k unit (customer) in the system.

$$P_0 = \left[\sum_{K=0}^{S-1} \frac{1}{k!}\left(\frac{\alpha}{\mu}\right)^k + \frac{1}{S!}\left(\frac{\alpha}{\mu}\right)^S \frac{S\mu}{S\mu - \alpha}\right]^{-1} \tag{10}$$

Equation (10) is the probability of having no customer in the system.

Expected number of customers in queue (L_q)

$$L_q = \left[\frac{1}{(S-1)!}\left(\frac{\alpha}{\mu}\right)^c \frac{\alpha\mu}{(S\mu - \alpha)^2}\right] P_0 \tag{11}$$

Expected number of customers in the system (L_S)

$$L_S = L_q + \rho \tag{12}$$

The average time a customer waits for service (W_q)

$$W_q = \frac{L_q}{\alpha} \tag{13}$$

The average time a customer spent is in the system, (W_S)

$$W_S = \frac{L_S}{\alpha} \tag{14}$$

3.5 Introduction of Cost into the Queuing Model

To determine the optimal service level and cost implications of a given queuing system, Murugan & Saratha (2017) stated that we have to contend with two cost, namely:

i. Customers' Waiting Costs (Cost of delay in offering service)

ii. Service cost (Cost of offering service)

Let the Expected Waiting Cost E(WC) per unit time be defined as:

$$E(WC) = \alpha W_s C_w \quad (15)$$

Where W_S is as defined in equation (7), C_w is the opportunity cost for a customer waiting in the system.

Also, Let the Expected Service Cost E(SC) be defined as:

$$E(SC) = C_0 + SC_S \quad (16)$$

Where C_0 is the fixed cost per unit time, S is the number of servers, C_S is the cost per server.

In order to obtain the total cost of the system, we add the waiting and service cost in equations (15) and (16). Hence, the E(TC) is defined as:

$$E(TC) = \alpha W_S C_w + C_0 + SC_S \quad (17)$$

The Total cost model in equation (17) attempts to create a balance between the two conflicting costs namely the waiting and service cost, because an increase in one cost automatically causes a decrease in the other (Ikotun et al., 2016).

4. Result and Discussion

The data was obtained from the customer service unit of Access Bank Plc, Anyigba, Kogi State, Nigeria, over the period of two weeks.

Table 1: Condensed Overview of Data Collection Spanning Two-Week Period.

Day	Week 1		Week 2	
	α_1	μ_1	α_2	μ_2
Monday	10.097	10.71	8.899	9.697
Tuesday	9.144	10.81	9.21	9.647
Wednesday	9.737	8.849	9.019	9.402
Thursday	9.599	10.12	9.441	10.242
Friday	9.267	8.905	9.322	10.109
Total	**47.844**	**49.394**	**45.8913**	**49.097**
Average	**9.5688**	**9.8788**	**9.17826**	**9.8194**

Table 1 offers a comprehensive breakdown of data collected over a two-week timeframe, with a distinction between Week 1 and Week 2. The collected data was subsequently scrutinized to derive crucial input parameters, namely the arrival rate and service rate. These parameters are instrumental in assessing the performance of both single-server and multi-server scenarios by employing an appropriate queuing model.

For the entire 10 days considered in this study, the mean of arrival rate of customers (α) is obtained thus: α=9.3734. The mean of service rate (μ) is obtain thus: α =9.8491.

4.1 Service Utilization Factor (ρ)

This performance metric signifies the proportion of time during which an equipment or system is actively operating in relation to the total time it could potentially be in use. ρ=0.9517, this indicates that the queuing system is operational for about 95.17% of the time and remains inactive merely for 4.83% of the time.

4.2 Performance Measure of M/M/S:FCFS/ ∞ / ∞

The performance measure for the various queuing model are presented in Table 2

Table 2: Performance Indicators for Single and Multi-Server Queuing Models

Performance measures	M/M/1	M/M/2	M/M/3	M/M/4	M/M/5	M/M/6
α	9.3734	9.3734	9.3734	9.3734	9.3734	9.3734
μ	9.8491	9.8491	9.8491	9.8491	9.8491	9.8491
ρ	95.17%	47.49%	31.66%	23.74%	18.99%	15.83%
P_0	0.0502	0.5251	0.6834	0.7626	0.8100	0.8417
Lq	18.9066	0.9043	0.4632	0.3114	0.2345	0.1881
Ls	19.7044	1.3792	0.7798	0.5488	0.4244	0.3464
Wq	2.0006	0.0965	0.0494	0.0332	0.0250	0.0201
Ws	0.2134	0.1471	0.0832	0.0585	0.0453	0.0369

Table 2 presents a comparison of different queuing systems, ranging from M/M/1 to M/M/6, and assesses their performance using various metrics such as utilization factor (ρ), Probability of having no customer in the system (P0), average number of customer in the queue (Lq), average number of customer in the system (Ls), average waiting time in the queue (Wq) and average time spent in the system (Ws) for both single-server and multi-server queuing models as presented in the methodology. The result revealed that the average customer waiting times in the system (Ws) per minute decrease as the number of servers increases: 0.2134, 0.1471, 0.0832, 0.0585, 0.0453, and 0.0369 for 1, 2, 3, 4, 5, and 6 servers, respectively. This can be generalized and applied to other banks or sectors facing similar challenges to optimize customer service and operational efficiency. The results demonstrate that as the number of servers increases, the average customer waiting time per minute significantly decreases from 0.2134 minutes with one server to 0.0369 minutes with six servers indicating that additional servers lead to fewer customers waiting in line. Concurrently, there is a notable decrease in the service utilization rate, dropping from 95.17% with one server to 15.83% with six servers, which suggests that each server becomes less busy as more servers are added. This reduction in both waiting times and server utilization implies that increasing the number of servers can enhance customer satisfaction by minimizing wait times and improving the overall customer experience, potentially leading to higher customer patronage. Consequently, other banks and sectors with high customer traffic can adopt similar strategies, adjusting their number of service points or servers to balance customer wait times and server utilization, thereby improving service quality and operational efficiency.

4.3 Cost Analysis on the Maintenance and Servicing of Queuing Facilities

To determine the marginal cost, hourly charges were employed to maintain consistency in the calculations.

i. For server cost (Cs) per hour, the monthly earned allowances for two staff are at a flat rate of N200,000. The server cost (Cs) per hour = N200,000/30 days/8 hrs = N833.33.
ii. The server attends to the customers between 08:00 a.m. and 04:00 p.m. (total of 8 hours in a day).

iii. The waiting cost per customer per hour is N2,479.7.
iv. Bank management claimed to spend on an hourly basis N5,000 on the maintenance and servicing of queuing facilities, e.g., desktops, cooling systems, sanitation, and other equipment.

Table 3: The Computation of Cost Analysis

S	α	W_S	αW_S	C_0	C_S	C_w	CC_S	$E(W_c)$	$E(S_c)$	$E(TC)$
1	9.3734	0.2134	2.000284	5000	833.33	2479.7	833.33	4960.103	5833.33	10793.43314
2	9.3734	0.1471	1.378827	5000	833.33	2479.7	1666.66	3419.078	6666.66	10085.73766
3	9.3734	0.0832	0.779867	5000	833.33	2479.7	2499.99	1933.836	7499.99	9433.825902
4	9.3734	0.0332	0.311197	5000	833.33	2479.7	3333.32	771.6749	8333.32	9104.994903
5	9.3734	0.0453	0.424615	5000	833.33	2479.7	4166.65	1052.918	9166.65	10219.56787
6	9.3734	0.0369	0.345878	5000	833.33	2479.7	4999.98	857.6748	9999.98	10857.65482

Table 3 provides a breakdown of the cost analysis for both single-server and multi-server queue models. The computation involved considering server (staff) salary structure, waiting cost, and service cost to identify the model with the lowest overall cost. In the single-server model (M/M/1) :(FCFS/∞/∞), the total cost is ₦ 10,793.43, resulting in 95.17% server utilization, indicating a high degree of overutilization at 95%. For multi-server models (M/M/s): (FCFS/∞/∞), the cost for (M/M/2) is ₦ 10,085.74 at a 47.49% utilization level. Subsequent models, (M/M/3), (M/M/4), (M/M/5), and (M/M/6), incur costs of ₦ 9,433.83, $ 9,104.99, ₦ 10,219.57, and ₦ 10,857.65, respectively, with decreasing utilization levels.

The results highlight that, at the time of this study, the single-server model (M/M/1) (FCFS/∞/∞) has the highest total cost at ₦ 10,793.43. As the number of servers increases, the system's cost decreases, reaching ₦ 9,104.99 for the (M/M/4) :(FCFS/∞/∞) model, which achieved a 23.74% utilization level. Striking a balance between waiting time and service level, this model represents the optimal service level with the minimum total cost.

Fig. 3. The overall cost of the system

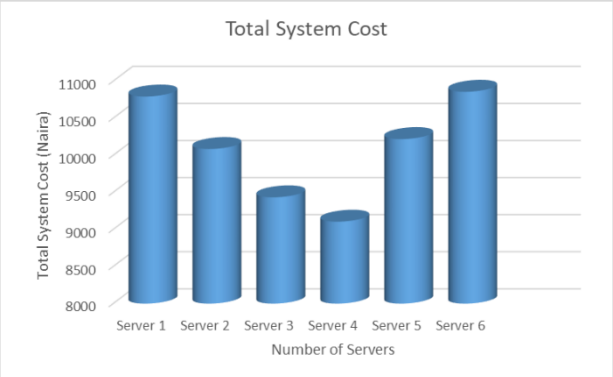

Fig. 3 illustrates the overall cost of the system for both single- and multiple-server models. The chart indicates that the most favorable service level is attained when employing the multi-server model with four servers.

5. Limitation

Despite the insightful results derived from the queuing models, certain assumptions inherent in this study may limit the generalizability of the findings. First, the models assume Poisson arrival and exponential service time distributions, which may not fully capture the actual variability in customer behavior, especially during peak or unpredictable hours. Secondly, the assumption of infinite queue capacity and constant arrival/service rates simplifies real-world conditions where space, time, and human factors affect queue dynamics. Finally, the First-Come, First-Serve (FCFS) queue discipline may not reflect priority-based or segmented services used in modern banking systems.

6. Conclusion

This study assesses the importance of utilizing single and multi-server exponential queuing models. It reveals that both inter-arrival and service times adhere to the (M/M/1) :(FCFS/∞/∞) general distribution model, employing the First-Come, First-Serve (FCFS) queuing discipline. The results indicate that the (M/M/s) :(FCFS/∞/∞) multi-server model outperforms the (M/M/1) :(FCFS/∞/∞) single-server model. This improvement is evident in average queue and system sizes, as well as average waiting times, particularly when there is an increase in the number of servers. The findings highlight the efficacy of employing multiple servers to enhance efficiency. Additionally, a thorough examination of cost implications and utilization factors serves as a benchmark for achieving a balance between cost minimization and ensuring an optimal service level. The study highlights the importance of considering both cost-effectiveness and service quality to enhance customer service delivery. Based on the research outcomes, the following recommendations are proposed: It is advisable to implement the (M/M/4): (FCFS/∞/∞) multi-server model, as it effectively reduces customer waiting time, improves service delivery, and significantly reduces operational costs. Promoting online banking is also recommended to alleviate pressure on customer care services.

Implementing the (M/M/4): (FCFS/∞/∞) multi-server model, while effective in reducing customer waiting times, improving service delivery, and lowering operational costs, may face several challenges and limitations. One potential challenge is the initial investment required to establish and maintain multiple servers, including the costs associated with infrastructure, technology, and staffing. Additionally, managing a multi-server system can be complex, requiring robust scheduling and coordination to ensure that servers are efficiently utilized without causing downtime or overstaffing during off-peak hours. There is also the risk of underutilization of resources if customer demand fluctuates significantly, which could lead to increased costs without proportional benefits. Furthermore, the model assumes an infinite queue capacity and steady arrival rates, which may not always align with real-world scenarios where customer flow can be unpredictable and vary widely. Lastly, customer satisfaction depends not only on reduced wait times but also on the quality of service provided, which requires continuous training and management oversight. Therefore, while the (M/M/4) model presents a promising strategy, its implementation must be carefully planned and monitored to

address these potential challenges and ensure sustainable improvements in service delivery and cost efficiency.

Recommendation

Based on the research outcomes derived from the conducted analyses, we recommend implementing the (G/G/4): (FCFS/∞/∞) multi-server model, as it effectively reduces customer waiting time, enhances service delivery, and notably decreases operational costs.
To address the limitations identified, future studies could explore non-Markovian models (e.g., G/G/s systems) that allow for more realistic arrival and service time distributions. Incorporating priority-based queuing disciplines and customer segmentation could also provide a more accurate reflection of banking operations.

Acknowledgements

Our profound gratitude to God who has given us the grace and wisdom for this research, we also appreciate the encouragement and invaluable contributions of Dr. Abubakar Yahaya, Dr. Umar Kabir Abdullahi, during this research and the management of Access Bank Plc, Anyigba who granted us access to gather data and provide necessary information needed to carry out this research.

References

1) Adewole, P. O. (2016) "Waiting Lines, Banks' Effective Delivery Systems and Technology Driven Services in Nigeria: A Case Study." *International Journal of Finance and Banking Research* Vol. 2, no. 6, pp 185-192. doi: 10.11648/j.ijfbr.20160206.11.
2) Amit, N. and Nurdia. A. G. (2018). "Using simulation to model queuing problem at a fast-food restaurant." *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016) Theoretical and Applied Sciences*. Springer Singapore. https://doi.org/10.1007/978-981-13-0074-5_104.
3) Burodo, M. S., Suleima, S., & Garba, Y. (2021). An assessment of Queue management and Patient Satisfaction of Some Selected Hospitals in North-Western Nigeria. *International Journal of Mathematics and Statistics Invention* (*IJMSI)*, 9(8), 14-24.
4) El-Taha, M. & Stidham, S. (1999). Little's formula and extensions. In Sample-Path Analysis of Queuing Systems., 10(5), 159–212. Springer.
5) Idigo, F. U., Agwu, K. K., Onwujekwe, O. E., Okeji, M. C., & Anakwue, A.- M. C. (2021). Improving patient flows: A case study of a tertiary hospital radiology department. *International Journal of Healthcare Management*, 14(1), 153–161.
6) Ikotun, D.O., Justus, A. A. & Festus, D. F. (2016) "Comparative Analysis of Customers' Queue Management of First Bank Plc. and Guaranty Trust Bank Plc, Isokun Ilesa, Nigeria." *IJ Mathematical Sciences and Computing, (November)* pp 1-11. DOI: 10.5815/ijmsc.2016.04.01
7) Khaskheli, S. A, Marri H. B., Nebhwani M., Khan M. A., & Ahmed M. (2020). Compartive Study of Queuing Systems of Medical Out Patient Departments of Two

Public Hospitals. *Proceedings of the International Conference on Industrial Engineering and Operations Management,*19(13), 2702–2720.

8) Limlawan, V., & Anussornnitisarn, P. (2021). Enhance system utilization and business revenue with ai-based queue reservation system. *International Journal of Machine Learning and Computing (IJMLC)*, 11(3), 252–256.
9) Murugan, A. N. & Saratha, S. V. (2017). "Minimizing the total cost in the Outpatient Department (OPD) of a multispecialty hospital". *World Journal of Research and Review (WJRR),* vol. 4, no. 3, pp. 50-53.
10) Nkrumah, S, Yeboah F. B., & Adiwokor E. (2015). Client Satisfaction with Service Delivery in the Health Sector: The Case of Agogo Presbyterian Hospital, *Int. J. Bus. Adm*., 6(4), 64–78.
11) Nsude, F. I., Elem-Uche, O., & Uwabunkonye, B. (2017). Analysis of Multiple-Queue Multiple-Server Queuing System: a case study of first Bank Nig. PLC, Afikpo branch, *International Journal of Scientific & Engineering Research,* 8(1), 29-55
12) Samuel, M. A., John, A.A., Frank B. K. T., & Emmanuel M. B. (2021). Stochastic Model of Waiting Time: A Case of Two Selected Banks in the Sekondi-Takoradi Metropolis, *Open Journal of Statistics*, 11, 906-924.
13) Segun, M. O. (2020). Performance Modelling of Health-care Service Delivery in Adekunle Ajasin University, Akungba-Akoko, Nigeria Using Queuing Theory. *Journal of Advances in Mathematics and Computer Science*, 35(3), 119-127.
14) Shamsuddeen, S., Muhammad, S. B., & Zubairu, A. (2022). An Application of Single and Multi-Server Exponential Queuing Model in Some Selected Hospitals of the North-Western Nigeria. *Asian Journal of Probability and Statistics,* 16(2), 1-9.
15) Varma, M. S. (2016). "Minimization of traffic congestion by using queueing theory". *IOSR Journal of Mathematics (IOSR-JM)*, vol. 12, no. 1, 116-122.
16) Yakubu, A.B.N., & Najim, U. (2014). An application of queuing theory to ATM service optimization: A case study, *Mathematical Theory and Modelling*, 5(6), 11-23.
17) Yuqi, H., & Haoxuan, L. (2020). Bank Queuing Optimization Based on Markov Process. 3rd International Symposium on Big Data and Applied Statistic. *Journal of Physics: Conference Series,* 8(12), 279-287.