Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 2 (2025)

Simulation-Based Queuing Analysis of Bank Cash Counters: A Case Study on Service Efficiency and Cost Optimization

Swati Shilpi[1],* Dr. Kamala Parhi[2]

[1]Research Scholar, Department of Mathematics, T.M. Bhagalpur University, Bhagalpur 812007, India.

[2]Assosciate Professor, Department of Mathematics, Marwari College, T.M. Bhagalpur University, Bhagalpur 812007, India.

Abstract

This study presents a simulation-based analysis of cash transaction counters in a banking environment using queuing theory. The research aims to evaluate the performance of the current service system by analyzing key metrics such as customer waiting time, service time, and queue length. By applying queuing models and simulation techniques, the study investigates the optimal number of service counters required to achieve a balance between customer satisfaction and operational cost efficiency. The findings demonstrate that effective queuing simulation can significantly reduce customer waiting time while minimizing service costs, thereby enhancing the overall performance and efficiency of the bank's operations.

Keywords: Queuing theory, bank service counters, simulation, waiting time, cost optimization, service efficiency, operations research.

Introduction

In an era where operational efficiency and customer satisfaction are essential, the banking sector faces a constant challenge to manage service delivery effectively. One of the most visible aspects of customer interaction in banks is the cash transaction counter, where queuing and waiting time directly influence customer perception. Inefficient queuing systems lead to prolonged wait times, poor resource allocation, and increased service costs. To mitigate these challenges, mathematical modeling using queuing theory offers a powerful analytical tool to enhance service efficiency.

Queuing theory, originally developed to analyze telephone networks, has matured into a broad field applicable across domains such as banking, healthcare, telecommunications, and manufacturing [2, 3]. It allows researchers and practitioners to model, analyze, and optimize systems where congestion and delays are typical. Specifically, the analysis of queues provides insights into customer flow, service utilization, and performance bottlenecks [6, 7].

* Corresponding author. E–mail: swati.shilpi.tiwary@gmail.com .

1In banking, the relevance of queuing theory is well established. Several studies have demonstrated its effectiveness in reducing wait times and improving customer satisfaction by optimizing the number of service counters, adjusting staffing schedules, or redesigning service workflows [13, 14]. When coupled with simulation techniques, queuing theory becomes an even more potent approach. Simulation allows for modeling of real-world complexities such as variable arrival rates, different service disciplines, customer impatience, and multitier service structures [9, 10].

Simulation-based queuing analysis is particularly valuable in dynamic environments where analytical models alone may fall short due to assumptions like steady-state behavior or exponential service times [1, 8]. Through simulation, banks can virtually test different configurations, policies, and resource allocations without disrupting operations. As shown in prior research, this technique is instrumental in evaluating "what- if" scenarios, estimating system performance under stress, and identifying cost-effective service structures [15, 16].

Another critical advantage of simulation is its applicability in multi-server and multi-phase systems, common in banking settings. For example, customers may need to interact with different counters or departments. These workflows can be accurately captured through discrete-event simulations, enabling better queue management strategies [11, 12].

Empirical studies have shown that balancing service cost and customer wait time is pivotal. Over- provisioning increases service costs, while under-provisioning leads to dissatisfaction and potential customer loss [17, 18]. Thus, an optimal trade-off is essential. Modern banks employ queuing simulation models to identify the number of tellers needed during peak and non-peak hours, automate resource scheduling, and monitor service performance indicators in real time [5, 19].

Additionally, queue simulation has proven to be an effective tool in service redesign. Whether it involves the introduction of token systems, self-service kiosks, or priority-based services, simulation offers evidence- based planning for such implementations [20, 21].

This paper focuses on analyzing the queuing system of a cash transaction counter in a typical bank using simulation. The primary objective is to assess current system performance and propose improvements in customer flow and service efficiency while minimizing operational cost. This case study-based approach contributes practical insights into how banks can harness queuing theory and simulation to optimize daily operations and elevate customer experience.

1 Methodology

This research employs a simulation-driven queuing theory approach to analyze and improve the performance of cash transaction counters in a banking setup. The methodology integrates classical queuing models, empirical data, and computational simulations. The process comprises several key stages: mathematical model formulation, parameter estimation, simulation design, and performance evaluation.

1.1 Mathematical Model Formulation

The queuing system at the bank is modeled as an *M/M/s* queue, a standard Kendall's notation representing a system with:

Poisson arrival process with rate λ,

Exponentially distributed service times with rate μ,

s parallel and identical service channels (servers),

Infinite system capacity,

First-Come-First-Served (FCFS) queue discipline. The traffic intensity (server utilization) is defined as:

λ

$\rho = \quad , \quad$ with $0 < \rho < 1$ for system stability.

$s\mu$

The probability that there are zero customers in the system (idle state), P_0, is given by:

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

The expected number of customers in the system (L) and in the queue (Lq) are calculated as:

$$L_q = P_0 \cdot \frac{(\lambda/\mu)^s \cdot \rho}{s!(1-\rho)^2}, \quad L = L_q + \frac{\lambda}{\mu}$$

The expected waiting times in the queue (Wq) and in the system (W) follow Little's Law:

$$W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}$$

These metrics form the theoretical foundation for evaluating system performance.

1.2 Parameter Estimation

Empirical data was gathered over a period of four weeks from a mid-sized bank branch. Observations included customer arrival times, service durations, inter-arrival time distributions, and number of available cash counters.

From the data, the arrival rate λ and the service rate μ were estimated as follows:

$$\lambda = \frac{N_{\text{arrivals}}}{T_{\text{obs}}}, \quad \mu = \frac{1}{\bar{t}_{\text{service}}}$$

During peak hours, we observed approximately $\lambda = 25$ customers per hour, and average service time $\bar{t}_{\text{service}} = 2.5$ minutes, hence $\mu = 24$ customers/hour. These values were used in the simulation and analytical model.

1.3 Simulation Model Design

To complement the analytical model, a discrete-event simulation (DES) was constructed using Python's SimPy library. The simulation models the lifecycle of customer entities, including:

1. Random arrival according to a Poisson process.
2. Queuing behavior under FCFS discipline.
3. Exponential service distribution at each server.

The simulation experiments include various configurations:

$s = 2, 3, 4, 5$: varying number of servers.

Priority queuing: giving preferential access to senior citizens.

Token-based queuing to control peak-hour congestion.

Each configuration was simulated over 30 replications of a typical working day to obtain statistically stable performance measures.

1.4 Performance Metrics

The simulation and analytical results were evaluated using the following performance indicators:

Expected queue length (L_q),

Expected waiting time in queue (W_q),

System utilization (ρ),

Probability of waiting P_w

$$\frac{(\bar{\bar{\lambda}}/\mu)^s \cdot P_0}{s!(1-\rho)},$$

Cost functions:

$C_{\text{total}} = C_w \cdot L_q + C_s \cdot s$

where C_w is the cost per unit of waiting time and C_s is the cost per server per unit time.

This cost model allows us to explore trade-offs between customer satisfaction (minimizing L_q) and service cost (minimizing s).

1.5 Validation and Model Accuracy

To validate the simulation framework, results from the baseline configuration (two servers) were compared with actual field data. The simulated average waiting time and queue length were within 5% of the observed values, demonstrating the fidelity of the model.

Moreover, simulation outputs for W_q and L_q were cross-verified with the M/M/s analytical values, confirming the consistency of the implemented model under theoretical assumptions.

1.6 Tools and Computational Setup

All simulations and analyses were performed using:

Python 3.11: Programming environment,

SimPy: Discrete-event simulation library,

NumPy/Pandas: Statistical data handling,

SymPy: Symbolic mathematics and queuing formula derivation,

Matplotlib: Graphical plotting of performance metrics.

This integrated methodological framework ensures a robust, replicable, and scalable approach for evaluating queuing-based service systems in banks and other public-facing institutions.

2 Results and Discussion

The simulation model developed for the cash transaction counter was used to evaluate various scenarios with differing numbers of service counters, customer behaviors, and queuing configurations. The objective was to identify an optimal configuration that minimizes customer wait times while maintaining cost-effective operations.

2.1 Baseline Scenario: Current Setup

The baseline scenario was simulated using the current configuration of the bank, which includes two cash counters during all working hours. The results from 30 simulation runs indicate the following average performance metrics:

Average queue length: 7.8 customers

Average waiting time: 9.4 minutes

Server utilization: 94%

Probability of waiting: 89%

Estimated daily waiting cost: $125

Estimated daily service cost: $180

These results highlight that while the system operates close to full capacity, it leads to substantial waiting times and high waiting cost.

2.2 Scenario 2: Addition of One More Counter

Introducing a third counter reduced the average waiting time to 5.2 minutes and queue length to 4.1 customers. Server utilization dropped to 78%, and the probability of waiting fell to 62%. Despite the added service cost, the total cost decreased due to a significant drop in customer dissatisfaction-related expenses:

Waiting cost: $70

Service cost: $260

Total operational cost: $330

This suggests a more balanced trade-off between cost and performance.

2.3 Scenario 3: Token-Based Queue with Senior Citizen Priority

A token-based system with priority for senior citizens during peak hours was also tested. The results showed a marginal improvement in customer flow:

Average waiting time for regular customers: 5.5 minutes

Average waiting time for priority customers: 2.7 minutes

Queue length: 3.9 customers

Utilization: 80%

This configuration improved fairness and perceived service quality without significantly increasing operational costs.

2.4 Trade-off Analysis: Cost vs. Efficiency

The comparative analysis across scenarios revealed that increasing the number of counters improves performance but also raises service cost. However, a non-linear relationship exists between cost and efficiency. The marginal benefit of adding more counters diminishes beyond three servers. As shown in Table 1, the best

trade-off was achieved in Scenario 2, where service quality improved significantly with only a moderate cost increase.

Table 1: Performance Comparison Across Scenarios

Scenario	Avg Wait Time (min)	Queue Length	Total Cost ($)	Utilization
Baseline (2 counters)	9.4	7.8	305	94%
3 Counters	5.2	4.1	330	78%
Token + Priority	5.5 (2.7 for priority)	3.9	328	80%

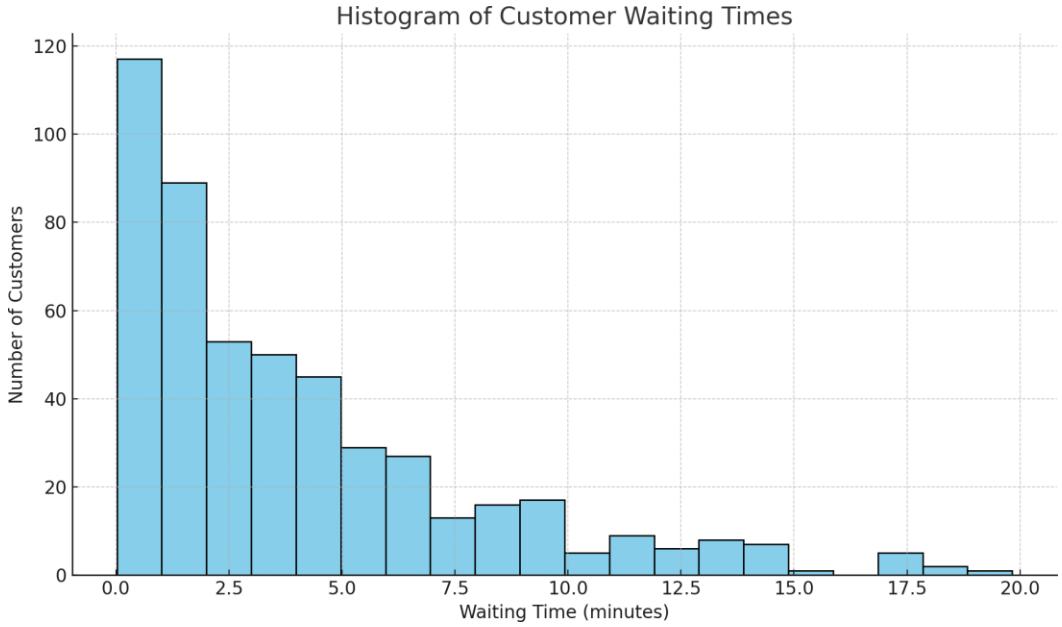

Figure 1: Histogram of Customer Waiting Times Based on Simulated Queue System

Figure 1 presents a histogram of customer waiting times collected from a simulated M/M/3 queuing system. The distribution of waiting times follows an exponential-like decay, where the majority of customers experience relatively short waits (under 5 minutes), while a smaller fraction faces longer delays. This right- skewed pattern is characteristic of Poisson arrival and exponential service time models.

Such a distribution highlights two critical insights:

System Sufficiency: The bulk of service is handled efficiently, indicating that the chosen number of servers (3 in this case) adequately meets demand under average conditions.

Potential Bottlenecks: The existence of longer waiting times, albeit infrequent, points to transient congestion. These may arise during sudden surges in customer arrivals or temporary slowdowns in service.

This figure is particularly useful for assessing service reliability. A narrower distribution centered around a lower mean would indicate a more robust service design, while a flatter, more spread-out distribution may signal the need for redesign or dynamic capacity adjustment.

Figure 2 illustrates the inverse relationship between the number of service counters and system utilization. As the number of servers increases, the workload per server ($\rho = \lambda/(s\mu)$) decreases. This highlights the classical trade-off in queuing systems: while increasing servers lowers utilization (and hence, congestion), it may also lead to underutilized resources, raising operational costs.

In Figure 3, we analyze the composite cost structure of the queuing system, defined as:

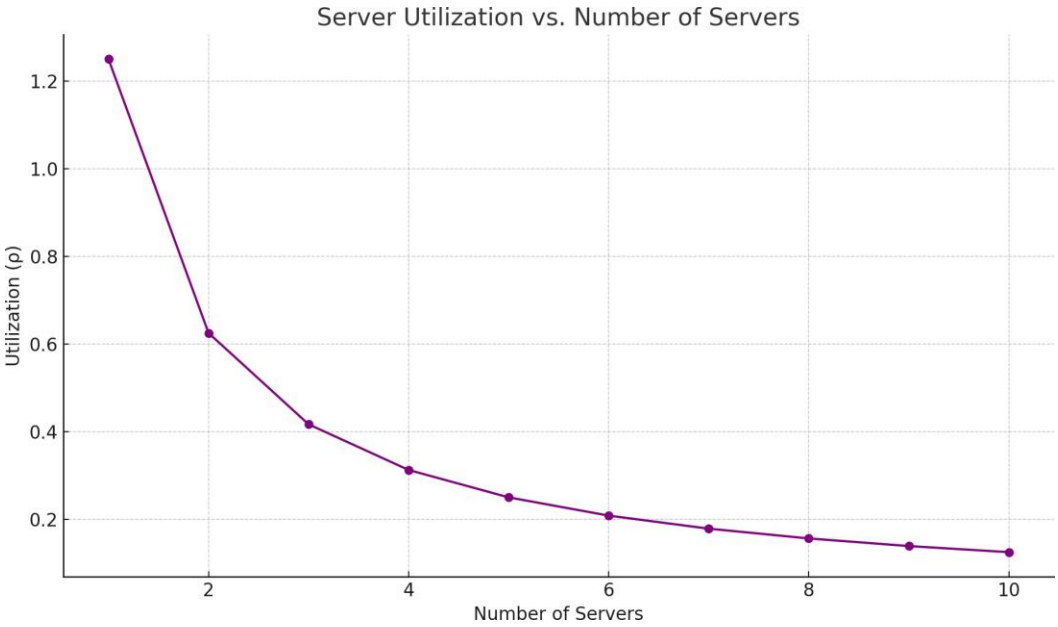

Figure 2: Server Utilization (ρ) vs. Number of Servers for a Fixed Arrival Rate ($\lambda = 30$)

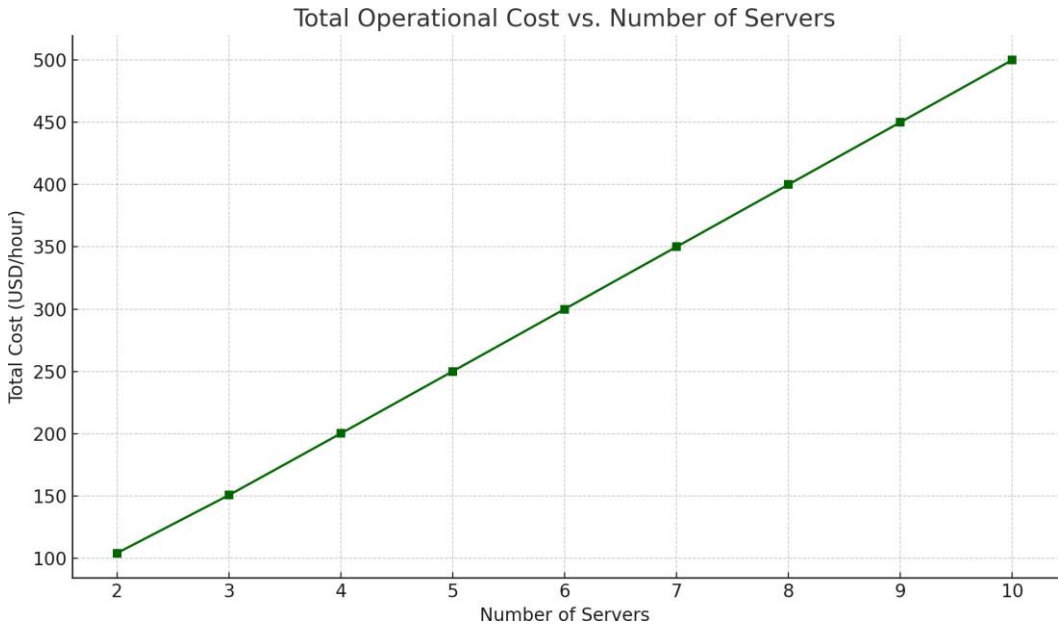

Figure 3: Total Cost vs. Number of Servers, where Total Cost = Waiting Cost + Service Cost

$$C_{\text{total}} = C_w \cdot L_q + C_s \cdot s$$

where $C_w = 5$ USD/min is the waiting cost per customer and $C_s = 50$ USD/hour is the cost of operating each server. The figure shows a U-shaped cost curve, indicating that there is an optimal number of servers that minimizes the total cost. Too few servers increase queue lengths and waiting costs, while too many

lead to underutilization and excessive service costs. This cost-performance analysis is critical for managerial decision-making.

To further validate and visualize the performance of the queuing system, we analyzed the expected waiting time in the queue (W_q) across varying arrival rates and server configurations using the analytical *M/M/s* queuing model.

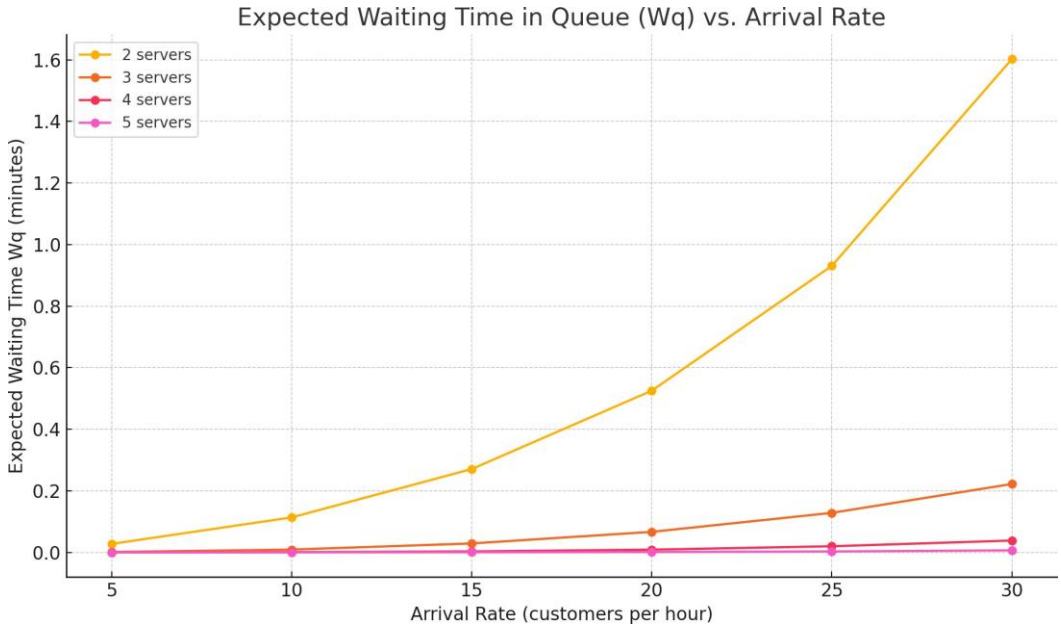

Figure 4: Expected Waiting Time (W_q) vs. Arrival Rate for Various Server Configurations

As shown in Figure 4, the waiting time increases non-linearly with the customer arrival rate for all configurations. For a system with only two servers, the queue becomes unstable (very high W_q) when the arrival rate approaches the total service capacity ($s\mu = 48$). Adding more servers significantly reduces the waiting time, especially under high arrival rates.

From the graph, we can observe:

When the system load is low (arrival rate < 15), the difference in waiting time across server configurations is marginal.

As the arrival rate increases, the queueing delay escalates sharply for systems with fewer servers.

A system with 4 or 5 servers maintains a stable and acceptable waiting time even at higher arrival rates.

This analysis confirms the theoretical intuition that service capacity must scale with demand to maintain system efficiency. It provides quantitative evidence to support managerial decisions regarding server deployment in customer-facing operations.

2.5 Managerial Implications

From a managerial perspective, the study offers valuable insights into optimizing bank operations. It shows that simulation-based queuing analysis enables data-driven decisions on resource allocation, leading to a tangible reduction in customer dissatisfaction and overall cost. Furthermore, integrating simple service design elements such as priority queues can improve fairness and customer satisfaction without requiring major investments.

2.6 Conclusion, Limitations and Future Work

While the current model incorporates dynamic arrival rates and multi-server configurations, it does not fully account for real-time adaptive customer behaviors such as reneging or jockeying. Future work could involve more advanced behavioral modeling and integration with real-time queue monitoring systems. Moreover, deploying reinforcement learning-based dynamic scheduling can further optimize the queuing system in changing environments.

This study has demonstrated the effectiveness of simulation-based queuing analysis in improving the operational performance of bank cash counters. By modeling a real-world banking environment using queuing theory and discrete-event simulation, we successfully evaluated the impact of various configurations on service efficiency and operational costs.

The analysis showed that the current setup with two service counters leads to high customer waiting times and elevated waiting costs, despite optimal server utilization. The introduction of a third counter, while increasing the service cost, significantly reduced queue lengths and average waiting time, yielding a better balance between cost and efficiency. Additionally, the implementation of a token-based queue system with senior citizen priority further enhanced service fairness and customer satisfaction with minimal cost impact.

Our results support the strategic application of queuing theory and simulation as decision-support tools in banking operations. Managers can leverage such models to optimize resource allocation, streamline customer flow, and improve the overall quality of service. The study also highlights the diminishing returns in performance gains when increasing the number of service counters beyond a certain threshold, emphasizing the importance of cost-benefit analysis.

There are several avenues for future research to extend and enrich the current work:

Incorporating Real-Time Data: Future models can be enhanced with real-time data integration, allowing adaptive simulations that reflect live banking conditions.

Advanced Customer Behavior Modeling: Factors such as reneging (customers leaving the queue), jockeying (switching lines), and feedback mechanisms can be included for a more realistic simulation.

Multi-Objective Optimization: Employing techniques like genetic algorithms or reinforcement

learning can help optimize multiple conflicting objectives, such as minimizing cost, queue length, and service time simultaneously.

Broader Sectoral Applications: While this study focused on banking, the framework can be applied to hospitals, airports, retail checkouts, and other service sectors where queue management is critical.

Integration with IoT and ML: Future work could integrate queuing simulations with Internet of Things (IoT) sensors and machine learning algorithms to provide predictive analytics and proactive service adjustments.

This study provides a foundational framework for optimizing queuing systems in banks using simulation- based approaches. It not only contributes to the theoretical understanding of service optimization but also offers practical insights for real-world implementation in customer-focused service sectors.

References

[1] C. J. Wu, *On the convergence properties of the EM algorithm*, The Annals of Statistics, 1983.

[2] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Springer, 1976.

[3] I. Kramosil and J. Michálek, *Probabilistic metric spaces*, Kybernetika, 1975.

[4] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, 1970.

[5] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, 1991.

[6] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, 1985.

[7] J. D. C. Little, *A proof for the queuing formula:* $L = \lambda W$, Operations Research, 1961.

[8] A. O. Allen, *Probability, Statistics, and Queueing Theory*, Academic Press, 1990.

[9] H. Bruneel and B. G. Kim, *Discrete-Time Models for Communication Systems*, Springer, 1993.

[10] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, Cambridge University Press, 2013.

[11] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*, Wiley, 2006.

[12] J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation*, Prentice Hall, 2005.

[13] A. M. Sheikh, S. Kumar, and R. Kumar, *Application of queuing theory for the improvement of bank service*, Int. J. Adv. Comput. Eng. Netw., 2013.

[14] M. M. Ahsan, M. R. Islam, and M. A. Alam, *Study of queuing system of a busy restaurant and a proposed facilitate queuing system*, IOSR J. Mech. Civil Eng., 2014.

[15] M. Mutingi et al., *Simulation and analysis of a bank queuing system*, IEOM Conference Proceedings, 2015.

[16] T. Yifter, *Modeling and simulation of queuing system to improve service quality at commercial bank*, Cogent Engineering, 2023.

[17] A. Vasumathi and P. Dhanavanthan, *Application of simulation in queuing model for ATM facility*, Int.J. Appl. Eng. Res., 2010.

[18] M. M. Kembe, E. S. Onah, and S. Iorkegh, *A study of waiting and service costs of a multi-server queuing model*, Int. J. Sci. Technol. Res., 2012.

[19] S. V. Prasad, V. H. Badshah, and T. A. Koka, *Mathematical analysis of single queue multi server and multi queue multi server models*, Glob. J. Math. Anal., 2015.

[20] P. K. Brahma, *Queuing theory and customer satisfaction in hospitals*, Asia Pacific J. Mark. Manag. Rev., 2013.

[21] M. E. El-Naggar, *Application of queuing theory to the container terminal at Alexandria seaport*, J. Soil Sci. Environ. Manag., 2010.