INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Modeling Customer Service Delivery in Banks Using Queuing Theory.

[1]Abubakar Muhammad Auwal., [1]Chaku Shammah Emmanuel., [1]Saleh Ibrahim Musa, [2] Tiletswen Timothy Terhemba

[1] Department of Statistics, Nasarawa State University, Keffi. Nasarawa State. Nigeria.

[2] Department of Statistics, Akawe Torkula Polytechnic, Makurdi. Benue State. Nigeria

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1412000047

Received: 16 December 2025; Accepted: 23 December 2025; Published: 02 January 2026

ABSTRACT

Efficient customer service delivery in banking operations is critical for maintaining customer satisfaction and resource utilization. Queuing theory is an approach to analyzing waiting lines, it provides a robust framework for modeling and improving service delivery in banks. This research explores the application of queuing theory to model customer service delivery processes in banks, focusing on service efficiency, customer wait times, and resource allocation. The data for this research was taken for a period of four weeks, covering all work days from Mondays to Fridays and hours from 8am to 4pm. The data analysis was done with the aid of EXCEL package for descriptive statistics and TORA for optimization system. From the performance measures researched, it shows that increasing the teller points to 3 would reduce the waiting time in the queue and system to 0.01481hours (53.32 seconds) and 0.05829hours (3.50 minutes) respectively as against the present situation where each customer has to wait in the queue and system for 0.30095hours (18.06 minutes) and 0,34443hours (20.67 minutes) respectively. This will result to each teller being busy for 62.3% of the time while remaining idle for 37.7% of the time.

Keywords: Multi-Server, Service efficiency, Waiting times, Banking Operations, Resource Optimization, Queue length and Arrival Rate.

BACKGROUND TO THE STUDY

A good number of establishments encounter queues in their effort to offer certain services to their customers. These establishments include large organizations such as banks, hospitals, post offices, super markets, fuel stations etc, to smaller scenarios such as students queuing up to submit assignments. The different establishments vary in scope and complexity but they all consist of a set of activities and procedures that require queuing, in which a customer or client must undergo in order to receive the needed services. The resources (or servers) in these systems (queuing system) include trained personnel and specialized equipment required for effective service delivery**.** Often, customers get to these servers to receive the needed services only to find that they are not attended to as soon as they get there due to one reason or the other. This causes the customers to wait for the service delivery for an unknown period of time.

According to David (1985), waiting is frustrating, demoralizing, agonizing, aggravating, annoying, time consuming and incredibly expensive. The truth of this assertion cannot be denied. There should be just a few consumers of services in modern society who have not felt, at one time or another, each of the emotions identified above. These experiences can also attest to the fact that the waiting line experience in a service facility significantly affects customers overall perceptions of the quality of service provided. Once customers are being served, their transaction with the service organization may be efficient, courteous and complete; but the bitter taste of how long it took to get attention pollutes the overall judgment that they make about the quality of service and leaves a very negative impression.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Increasing criticisms, cost pressure and increasing demand of quality and efficiency from highly aware and educated customers have started putting more pressure on the many organizations urging them to improve on the quality of service they offer. The urge to study queues is prompted by two obvious features. Owing to the ebb and flow of customers, there would be some occasions when the service facility is not fully employed, i.e where there are more servers and fewer customers such that the servers wait idly for a period of time and others where it is under continuous pressure, with a long queue of customers waiting to be attended to. Costs are involved when the service is under-employed (low productivity), and in the congested period, loss of productive time for queuing members. According to Singh (2006), if the organization decides to increase the level of service provided, cost of providing service would increase, if it decides to limit the same, costs associated with waiting for service would increase. So the manager has to balance the two costs and make a decision about the provision of optimum level of service. Hence it is one of the tasks of queuing theory to try to see how these costs can be reduced by modifications to the mechanics of the system.

Statement of the Problem

This study is important because Benysta microfinance Bank, Makurdi has always experienced failure in terms of customer satisfaction due to the population of customers who constantly need different services to be rendered. This sometimes leads to chaos in the banking hall. Therefore, modeling customer service delivery in banks using queuing theory is relevant.

Aim and Objectives of the Study

The aim of this dissertation is to model customer service delivery using queuing theory to the operations of Benysta Microfinance Bank, Makurdi. The specific objectives of this dissertation are to:

1. Determine the average service rate and intensity of the bank per unit time.
2. Determine the average number of transactions in a given time and compute the average time a customer spends for a transaction.
3. Determine the average waiting time a customer waits in the system and the utilization services of the bank at different number of server.
4. Find the optimal number of server for the transactions in the system.

Significance of the Study

Overcrowding and long waiting lines in many organizations are a common occurrence. Any arrangement that alters the system and reduces waiting time would go a long way to reduce customer's anxiety as they wait to get services. The findings would be useful to systems which make use of queue in a bid to reduce their waiting times. e.g. Banks, IT Centres, Hospitals, telecommunication service providers etc. In addition to this, if this arrangement helps in minimizing cost to the establishment, it would have improved on the smooth running of the organization and increase system's revenue. Thus results of this research would provide innovations that would assist organizations which make use of queue to improve on their services while minimizing costs.

Conclusively, the study will guide Benysta Microfinance Bank and related areas in making policies that affect the management of queues especially when customers are in season. This will in turn create efficiency and effectiveness in services rendered to customers.

Scope of the Study

This work focuses on the operations of Benysta Microfinance Bank, Makurdi. Customers come in and queue up to make transactions. There is usually more than one server available and customers queue up waiting to go in and consult on a first come first served (FCFS) basis. The data is a primary data and was collected over a period of four weeks and covers all the working days from Mondays to Fridays during the peak farming season. The

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

scope of this project is limited to Benysta Microfinance Bank and any other related banking institution in terms of services; the data is of a primary source for an observational period.

METHODOLOGY

Research Design

The Poisson and Exponential processes are used to model the arrival rates and service rates. This is due to the following reasons:

1. There is a certain amount of regularity in the arrivals of customers in the Bank. Although individual arrivals are impossible to predict, there is perhaps some statistical regularity in the sense that when we observe customers arrivals during a period of say one month, without knowing the absolute time frame, then we have no way to decide whether we observe the time period of January, February or June. In probability terms, the process is stationary in time. In other words, the course of time should not change the probability properties of the process.

2. The fact that there is an occurrence at the particular time, says nothing about the probability of an occurrence at or around a later or earlier time. In other words, there seems to be some kind of independence with respect to various occurrences.

3. The next occurrence cannot be predicted from past or current information. In other words, the process of occurrences seems to have no memory. The fact that something happened in the past has no effect on the probabilities for future occurrences.

4. There is no accumulation of occurrences at anytime. That is, in each finite time interval, there are only finitely many occurrences.

Data Collection

For the purpose of this research work, the primary source is used. The data was obtained from Benysta Microfinance Bank, in Makurdi, Benue State, Nigeria. The Bank opens for services at 8:00am and closes at 4:00pm from Mondays to Fridays.

The data that is used for this research work is of primary source, it was obtained for four weeks and the collections were done from 8:00am to 4:00pm each day. The arrival time and the number of customers serviced per period were duly recorded by observations.

Models and Technique for Analysis

Queuing Model Specification

According to John (2010), Queuing theory modelling is classified by using special or standard notations described by D.G. Kendell in 1953 in the form of (a/b/c). Later, A.M. Lee added the symbols d and e to the Kendell notation. In the literature of queuing theory, the standard format used to describe queuing models is as follows:

Fig 2.1: Component of a basic queuing system

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

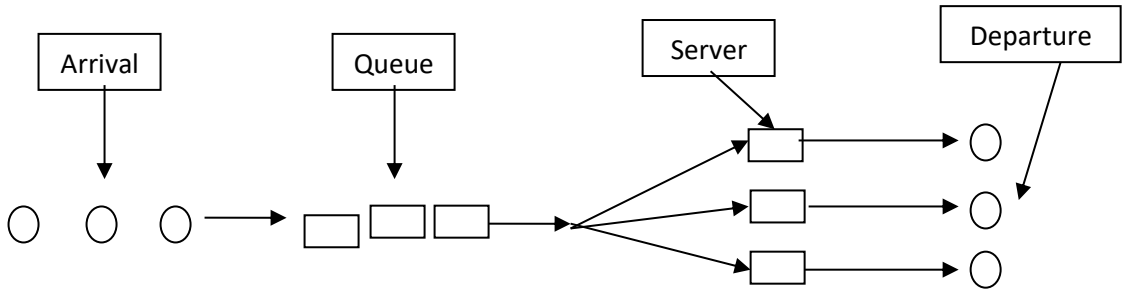

Fig 2.2: A multiple server queuing system, M/M/C

{(a/b/c): (d/e)}

Where a = arrival distribution

b = service time distribution

c = number of servers (service channels)

d = capacity of the system (queue plus service)

e = queue (or service discipline)

In place of notation a and b, the following descriptive notations are used for the arrival and service times distribution:

M = Markovian (or exponential) inter-arrival time or service-time distribution.

D = Deterministic (or constraint) inter-arrival time or service topic.

G = General distribution of service time, i.e. no assumption is made about the type of distribution with means and variance.

Gl = General probability distribution normal or uniform for inter-arrival time.

E_k= Erlang-k distribution for inter-arrival or service time with parameter k (i.e. if k = 1, Erlang is equivalent to exponential and if k = ∞, Erlang is equivalent to deterministic)

Method of Analysis

The analysis is done with the aid of EXCEL Package for descriptive statistics (averages) and TORA for Optimization System.

$$L_s = \frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)^k}{(k-1)!\,(k\mu-\lambda)^2}P_0 + \frac{\lambda}{\mu} \qquad (2.1)$$

$$W_s = \frac{L_s}{\lambda} \qquad (2.2)$$

$$L_q = \frac{\lambda\mu\left(\frac{\lambda}{\mu}\right)^k}{(k-1)!\,(k\mu-\lambda)^2}P_0 \; or \; L_q = L_s - \rho \qquad (2.3)$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

$$W_q = \frac{\mu\left(\frac{\lambda}{\mu}\right)^k}{(k-1)(k\mu-\lambda)^2} P_0 \; or \; W_q = W_s - \frac{1}{\mu} = \frac{L_q}{\lambda} \qquad (2.4)$$

$$\rho = \frac{\lambda}{\mu c} \qquad (2.5)$$

Where,

L_s is the expected number of customers in the system at any point in time

W_q is the average time spent on queue

L_q is the average queue length

W_s is the expected time spent in the system

ρ is the utilization of server in the bank

λ is the average arrival

μ is the average service rate

P_0 is the Initial Probability i.e the probability that no customer at the system

ρ is the rho (utilization)

Economic Analysis

The two basic types of costs associated with queuing systems are the costs involved in operating each service facility like the costs for equipment (including maintenance), materials, labor, etc. These cost increases as the number of service facilities put into operation increases and the opportunity costs associated with causing customers to wait in the system. . The total of these two basic types of costs goes to a minimum at some specific number of facilities. This then is the optimum number of service facilities which should be operated by the manager- optimum because it minimizes the total cost of both operating the service facilities and waiting time in the system. The total cost model includes the cost of waiting and the cost of service:

$$TC = C_w L_s + C_s k$$

where:

Cw= the waiting cost per time period for each customer

Ls=average number of customers in the system

Cs= the service cost per time period for each channel

k = the number of channels

TC= the total economic cost per time period

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

RESULTS AND DISCUSSION

Summary of inter-arrivals time of customers for four weeks between the intervals of one hour.

	Inter-Arrival Time/hour	**Days/week**						
		1	**2**	**3**	**4**	**5**	**Total**	**Mean**
Week One (1)	8:00 – 9:00	74	78	103	126	58		
	9:00 – 10:00	108	67	68	72	109		
	10:00 – 11:00	87	86	86	64	76		
	11:00 – 12:00	63	52	78	72	74		
	12:00 – 1:00	61	67	72	31	44		
	1:00 – 2:00	57	63	27	39	55		
	2:00 – 3:00	39	40	32	31	52		
	3:00 – 4:00	38	40	32	38	68		
	Total						**2527**	**63.18**
Week One (2)	8:00 – 9:00	82	107	100	88	63		
	9:00 – 10:00	89	88	78	73	100		
	10:00 – 11:00	75	57	121	77	96		
	11:00 – 12:00	87	97	67	58	65		
	12:00 – 1:00	53	84	53	68	63		
	1:00 – 2:00	73	0	65	69	67		
	2:00 – 3:00	46	54	51	52	0		
	3:00 – 4:00	19	0	13	48	28		
	Total						**2574**	**64.35**
Week One (3)	8:00 – 9:00	13	18	29	13	78		
	9:00 – 10:00	2	16	47	50	183		
	10:00 – 11:00	21	0	26	2	112		
	11:00 – 12:00	16	17	62	19	72		
	12:00 – 1:00	11	11	0	6	24		
	1:00 – 2:00	0	4	17	16	7		

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

	2:00 – 3:00	6	1	22	0	69		
	3:00 – 4:00	1	5	1	12	105		
	Total						**1114**	**27.85**
Week One (4)	8:00 – 9:00	16	1	12	0	1		
	9:00 – 10:00	38	3	47	21	26		
	10:00 – 11:00	22	28	0	53	35		
	11:00 – 12:00	19	44	31	20	14		
	12:00 – 1:00	1	53	3	24	0		
	1:00 – 2:00	13	21	4	24	0		
	2:00 – 3:00	18	25	6	22	3		
	3:00 – 4:00	12	16	6	74	0		
	Total						**756**	**18.9**
	Mean arrival rate/hour (λ) = 43 customers							

Summary of inter-service time of customers for four weeks between the intervals of one hour.

	Inter-Arrival Time/hour	**Days/week**						
		1	**2**	**3**	**4**	**5**	**Total**	**Mean**
Week One (1)	8:00 – 9:00	26	33	0	36	40		
	9:00 – 10:00	35	42	31	0	32		
	10:00 – 11:00	42	34	57	27	41		
	11:00 – 12:00	46	59	53	42	60		
	12:00 – 1:00	30	44	33	24	35		
	1:00 – 2:00	22	31	14	37	36		
	2:00 – 3:00	32	25	31	15	43		
	3:00 – 4:00	11	25	16	23	17		
	Total						**1280**	**32**
Week One (2)	8:00 – 9:00	36	28	3	22	13		
	9:00 – 10:00	29	19	8	34	2		
	10:00 – 11:00	41	33	38	46	47		

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

	11:00 – 12:00	48	38	61	53	25		
	12:00 – 1:00	53	51	47	57	53		
	1:00 – 2:00	48	0	44	72	41		
	2:00 – 3:00	33	0	42	47	21		
	3:00 – 4:00	31	0	15	21	27		
	Total						**1327**	**33.18**
Week One (3)	8:00 – 9:00	29	21	6	13	28		
	9:00 – 10:00	0	0	25	13	14		
	10:00 – 11:00	0	22	24	17	22		
	11:00 – 12:00	0	26	23	19	22		
	12:00 – 1:00	26	0	0	6	12		
	1:00 – 2:00	15	0	17	16	12		
	2:00 – 3:00	14	14	12	0	32		
	3:00 – 4:00	1	5	1	12	105		
	Total						**602**	**15.05**
Week One (4)	8:00 – 9:00	6	15	0	0	15		
	9:00 – 10:00	13	18	0	13	19		
	10:00 – 11:00	21	16	0	22	2		
	11:00 – 12:00	14	0	0	26	14		
	12:00 – 1:00	17	0	15	12	0		
	1:00 – 2:00	21	8	21	16	0		
	2:00 – 3:00	3	13	29	0	3		
	3:00 – 4:00	22	13	21	0	0		
	Total						**428**	**10.7**
	Mean service rate/hour (μ) = 23 customers							

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Performance Measures of Multiserver Queuing Model at the Benysta Microfinance Bank

Scenario	C	Lambda	Mu	L'da eff	P_0	Ls	Lq	Ws	Wq
1	2	43.00000	23.00000	43.00000	0.03371	14.81061	12.94104	0.34443	0.30095
2	3	43.00000	23.00000	43.00000	0.13320	2.50628	0.63671	0.05829	0.01481
3	4	43.00000	23.00000	43.00000	0.15010	1.99546	0.12589	0.04641	0.00293
4	5	43.00000	23.00000	43.00000	0.15339	1.89741	0.02785	0.04413	0.00065

Summary analysis of the Multiserver Queuing Model of Benysta Microfinance Bank

Performance Measure	**2 Tellers**	**3 Tellers**	**4 Tellers**	**5 Tellers**
Arrival rate (λ)	43	43	43	43
Service rate (μ)	23	23	23	23
System Utilization (*p*)	93.5%	62.3%	46.7%	37.4%
L_s	14.81061	2.50628	1.99546	1.89741
L_q	12.94104	0.63671	0.12589	0.02785
W_s(Hours)	0.34443	0.05829	0.04641	0.04413
W_q(Hours)	0.30095	0.01481	0.00293	0.00065
P_0	3.371%	13.32%	15.01%	15.339%
Total System Cost/hr	₦1681.06	₦550.63	₦599.55	₦689.74

From the queue performance measures, increasing the number of teller points to 3 indicates that the waiting time in the queue and system would reduce to 0.01481hours (53.32 seconds) and 0.05829 hours (3.50 minutes) respectively as against the present situation where each customer has to wait in the queue and system for 0.30095 hours (18.06 minutes) and 0.34443 hours (20.67 minutes) respectively. As a result of this, each teller will be busy for 62.3% while the remaining 37.7% of the time would be idle. Furthermore, the total economic cost will also decrease from ₦1681.06 with 2 tellers to ₦550.63 with 3 tellers which is economically optimal.

From the analysis, it can be observed that the number of tellers necessary to serve customers in the case study of Benysta Microfinance Bank, Makurdi is 3 teller points. This has been proven in the tables above, it is the appropriate number of tellers that can serve the customers as at when due without waiting for long before customers are been served at the actual time necessary for the service.

FINDINGS

From the results, the Bank is expected to increase the service facility because the utilization factor is very high at 2 servers. The results further showed that:

1. The mean rate is 43 customers/hour and the mean service rate is 23 customers/hour.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

2. The average number of customers in the system at any point in time in the scenario 1, 2, 3 and 4 are 14.81061, 2.50628, 1.99546 and 1.89741. i.e 15 customers, 3 customers, 2 customers and 2 customers respectively.

3. The average queue length for the scenario 1, 2, 3 and 4 are 12.94104, 0.63671, 0.12589 and 0.02785. i.e there will be 13 customers in the queue at first scenario, 1 customer at the second scenario while no queues at the third and fourth scenario.

4. The average time spent in the system at scenario 1, 2, 3 and 4 are 0.34443x60 = 20.67 minutes, 0.05829x60 = 3.50 minutes, 0.04641x60 =2.78 minutes and 0.04413x60 = 2.64 minutes.

5. The expected time spent in the system by customer for the scenario 1, 2, 3 and 4 are 0.30095x60 = 18.06 minutes, 0.01481x60 = 53.31 seconds, 0.00293x60 = 10.54seconds and 0.00065x60 = 1.04 seconds respectively.

6. The probabilities that there are no customers in the system are 0.03371, 0.13320, 0.15010 and 0.15339 respectively for the scenarios.

7. The utilization at scenario 1, 2, 3 and 4 was calculated to be 0.93478 = 93.48%, 0.62319 = 62.32%, 0.46739 = 46.74% and 0.15339 = 15.34% respectively; this showed that the Bank was below efficiency at the first scenario and also showed that when there is increase in the server there will be increase in the work efficiency and satisfying the customer's needs.

8. The total system cost for the scenarios are ₦1681.06, ₦550.63, ₦599.55 and ₦689.74 respectively. That indicates that, the optimal server for the Bank is 3.

SUMMARY AND CONCLUSION

Summary

The situation or conditions experienced in this bank during the period of study are similar to what operates in other banks in the country. Excessive waste of time in the banking hall would have a negative impact on the economy of the country in terms of the opportunity cost, which is the excess time that would have been used elsewhere. Also from the M/M/C model, we see that there is a drop in waiting time as more servers are added. There is also a drop in queue length, this would probably come with some degree of happiness and satisfaction to the customers, but heavy cost would be incurred by the Bank if they have to employ more servers for the operation unit of the bank. For this study, given the queue characteristics, the optimal number of servers which would minimize the operational cost is five for this particular establishment with time saving and the reduction of operational cost as the only factors considered.

Conclusion

From the findings, we conclude that the number of customers patronizing the bank was very higher at first and second weeks; this is because of the planting season during which some customers need money to make certain purchases and others come in to deposit money from sales made. Providing customers with timely access to the needed services while minimizing cost to the establishment is one of the major goals of every organization. Generally, an excess of demand over supply i.e. more customers than servers causes waiting. Since we obviously have more customers than servers, the complete elimination of waiting line is impossible; we only try to minimize the waiting time. With the results obtained from this study, we see that the use of three servers can help to improve the operations of Benysta Microfinance Bank. Finally, we advice that a similar study should be carried out for systems which may have different service times or service rates in order to ascertain if pooling could also be beneficial to such systems as this work is limited only to the Operation Unit of Benysta Microfinance Bank, Makurdi, Benue State, Nigeria.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

REFERENCES

1. Dakingari, U.M., Burodo, M.S., and Shehu, S. (2024). Application of Queuing Theory: A Tool for Minimizing customer Waiting Time with ATM Servives in Selected Deposit Money Banks in Gusau Metropolis. Abuja Journal of Business and Management. 2(4).
2. Chaku S.M., Suleiman S.C. and Awigbo E.B.(2024). Modeling Queuing Operational Characteristics at Automated Teller Machine Points. International Journal of Research and Innovation in Applied Science (IJRIAS). 9(6), 446-461. doi: 10.51584/IJRIAS.2024.906040
3. Kupolusi, J. (2022). Queuing Modelling to Customer Management at a commercial Bank: a Case Study of UBA Branch, Ondo. SSRN.
4. Paveun, P.F., and Danyaro, M.L. (2025). Optimal Queuing model to Optimize Banking Service Performance: A Study of Zenith Bank Plc, Bwari Branch, FCT Abuja, Nigeria. International Journal of Research and Innovation in Applied Svience (IJRIAS).
5. Qi-Ming, Jinguixie and Xiaobo Zhao, (2009). Stability conditions of a preemptive repeat priority queue with customer transfers. Institute of Mathematics and Informatics, Vilnius: 463-467.
6. Shanmugasundaram, S. and Umarani, P. (2015). “Queuing Theory Applied in Our Day to Day Life”, 6(4), 533-541.
7. Yifter, T. (2023). Modelling and Simulation of a Queuing System to Improve Service Quality at a Commercial Bank in Ethiopia. Cogent Engineering/Taylor&Francis.