International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 2 Issue 1 ǁ January. 2014 ǁ PP-14-21

Queuing Process and Its Application to Customer Service Delivery (A Case study of Fidelity Bank Plc, Maiduguri)

H. R. Bakari[1], H. A. Chamalwa[1] and A. M. Baba[2]

[1] *(Department of Mathematics and Statistics University of Maiduguri, Nigeria.)*
[2]*(Department of Mathematical Sciences Progamme, Abubakar Tafawa Balewa University, Bauchi, Nigeria)*

ABSTRACT: *Standing in line can cause extreme boredom, annoyance and even rage to customers. Customers are often forced to wait in line whenever the service facility is busy. Although, automatic teller machines (ATM) have been designed to provide efficient and improve services to customers at the shortest time possible, yet customers wait too long before they are finally serviced by the facility. This is principally due to variation in arrival and service time, which eventually leads to the formation of queue. Therefore, a systematic study of waiting line system will assist the management of the Bank in taking certain decisions that may minimize the length of time a customer spends in a service facility. Against this background that queuing process is employed with emphasis to Poisson distribution to assess the utility function of service delivery. The data for this study was collected from primary source and is limited to ATM service point of Fidelity Bank Plc located in West-End, Maiduguri. Data was collected by observation, in which the number of customers arriving at the facility was recorded, as well as each customer's arrival and service time respectively. The assistance of a colleague was sought in recording the service time while the researcher records arrival time. The period for the data collection was during busy working hours (i.e. 8:00am to 4:00pm) and for a period of ten (10) working days. The study reveals that the traffic intensity (ρ) is 0.96. Since we obtained the value of the traffic intensity, otherwise known as the utilization factor to be less the one (i.e. ρ<1), it could be concluded that the system operates under steady-state condition. Thus, the value of the traffic intensity, which is the probability that the system is busy, implies that 95% of the time period considered during data collection the system was busy as against 4% idle time. This indicates high utilization of the system.*

KEYWORDS: *Automatic teller machine, Poisson distribution, Queuing process and Traffic intensity*

I. INTRODUCTION

The financial sector in Nigeria has witness a significant reforms in recent times all in an effort to maximize profit, reduce cost ad satisfy customers optimally in the most generally acceptable international standard. Despite these entire sterling efforts one phenomenon remains inevitable: queue. It is a common practice to see a very long waiting lies of customers to be serviced either at the Automated Teller Machine (ATM) or within the banking hall. Though similar waiting lines are seen in places like; bus stop, fast food restaurants, clinics and hospitals, traffic light, supermarket, e.t.c. but long waiting line in the banking sector is worrisome being the public's most important units.

[1] Queue is a general phenomenon in everyday life. Queues are formed when customers (human or not) demanding service have to wait because their number exceeds the number of servers available; or the facility doesn't work efficiently or takes more than the time prescribed to service a customer. Some customers wait when the total number of customers requiring service exceeds the number of service facilities, some service facilities stand idle when the total number of service facilities exceeds the number of customers requiring service. [2] defines queue as simply a waiting line, while[3] put it in similar way as a waiting line by two important elements: the population source of customer from which they can draw and the service system. The population of customer could be finite or infinite.

Waiting line management has the greatest dilemma for managers seeking to improve the on investment of their operation; as customers don't tolerate waiting intensely. Whenever customer feels that he/she has waited too long at a station for a service, they would either opt out prematurely or may not come back to the station next time when needed a service. This would of course reduce customer demand and in the long run revenue and profit. Moreover, longer waiting time might increase cost because it equals to more space or facilities, which mean additional cost on the management [4].

Despite being in the technology era; line are experienced at within and Banks ATMs in developing nations than elsewhere. ATM are adopted so as to reduce waiting time., offers considerable ease to both the bank and their customers; as it enables customers to make financial transactions at more convenient times and locations, during and after banking hours. Most importantly, ATM, are designed to provide efficient and improved services to customers at the shortest possible time. Yet customers spend a considerable time before they are finally served. Businesses especially banks are striving very hard to provide the best level of service possible, minimizing the service time, giving the customer a much better experience. However, in situations where queue arises in a system, it is appropriate to attempt to minimize the length of the queue rather than to eliminate it completely; complete elimination may be infeasible

Therefore, a systematic study of waiting line system would assist the management of the Bank in making certain decisions in an effort minimize the time a customer spends in a service facility.

II. QUEUING THEORY

Queuing Theory is a collection of mathematical models of various queuing systems. It is used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times. It also provides the technique for maximizing capacity to meet the demand so that waiting time is reduced drastically.

2.1.1 Queuing Discipline:

It is obvious to notice or reason that ticket at service station or store such as grocery checkout in supermarket, gasoline, manufacturing plants, and banks e.t.c line as a typical example of queuing system. Meaning when any arrival occurs, it is added to the end of the queue and service is not rendered on it until all the arrivals that are there before it are attended to in that order. This is in fact a common way by which queue is being handled. The process whereby arrivals in the queue are being processed is termed queuing discipline. the example highlighted above is a typical example of first-come-first served discipline or FCFS discipline; other possible discipline include last-come-first- served or LCFS and service in random order SIRO. it is worth noticing that the particular discipline chosen will greatly affect the waiting time for particular customers; as no one would want to arrive early in an LCFS discipline; the discipline doesn’t generally affect the important outcome of the queue itself, as arrivals are constantly receiving service respectively [5].

2.1.2 Kendall-Lee Notation:

Kendall in 1953 and lee in 1966 came-up with a much simpler notation that describes the characteristics of a queue termed Kendall-Lee notation. The notation gives six abbreviations for characteristics listed in order separated by slash as: M/M/A/B/C/D. the first and second characteristics describes in the arrival and service time based on their respective probability distribution, where M represents exponential distribution, E stands for Erlang and G stands for a general distribution: uniform, normal e.t.c, the third character gives the number of servers working together at the same time, known as parallel servers; the fourth describes the queue discipline, the fifth gives the maximum number of customers the system can accommodate, while the sixth gives the population size of the customers from which the system can draw from for example: M/M/6/FCFS/30/∞

2.1.3 Little’s Queuing Formula:

It is pertinent to determine the various waiting times and queue size for particular components of the system in order to make judgment about how to run the system. Suppose L denote the average number of customers in the queue at any given time, assuming that the steady-state has reached. We can be able to break that into Lq; average number of customers waiting in the queue and Ls; average number of customers in the service; and since the customers in the system can only either be in the queue or in service, this implies that: L=Lq+Ls.

Moreover, we can say W denotes the average time a customer spends in the queuing system. Wq is the average time spends in the queue; while Ws is the average time spends in the service. Therefore, W=Wq+Ws

Let λ denotes the arrival rate into the system, meaning thereby, the number of customers arriving the system per unit time, thus,

$L=W\lambda$

$L=Ws\lambda$

$L=Wq\lambda$

2.2 Queue Characteristics

[6] Stated that queuing system can be characterized by four (4) components or four main elements. These are: the arrival, the queue discipline, the service mechanism and the cost structure. [7] On the other hand stated

that queuing systems are characterized by five (5) components: The arrival pattern of customers; the service pattern, the number of servers, the capacity of the facility to hold customers, and the order in which the customers are served.

2.2.1 The Arrival Pattern of Customers

The arrival is the way in which a customer arrives and enters the system for services. It is the system input process. It is how units (customer) joined the queues; which could be static or dynamic i.e control depends on the arrival rate or service facility and customer. Whenever customers arrive at a rate that exceeds the processing system rate, a queue will be formed. The arrival process of customers is usually specified by the inter-arrival time (the time between successive customer arrivals to the service facility). It may be deterministic (known exactly) or it may be a random variable whose probability distribution is presumed known. The arrival process can be;

- Regular arrival; that is it follows a Poisson distribution with average arrival rate λ. It may be
- In a completely random manner
- Singly or in batches
- Non-stationary arrival
- General independent arrival

2.2.2 The Service Mechanism

This means how the service is being rendered; it could be persons (bank teller, barber) a machine (gasoline pump, elevator) or a space (parking lot, hospital bed). [8] Viewed that the uncertainty involved in the service mechanism are the number of servers, the number of customers getting served at any time, and the duration and mode of service. This also depend on the configuration: single server-single queue, single server-several queues, parallel servers-single queue, several servers- several queues and service facilities in series (multiple servers) and speed i.e. service rate and service time which are inversely proportional to each other e.g. if a barber attends 2 customers in an hour, the service rate is 2customers per hour and the service time is 30 minutes per customer

2.2.3 The Service Pattern

The service pattern is usually specified by the service time (the time required by one server to completely serve one customer). The service time may be deterministic or it may be a random variable whose probability distribution is presumed known. It may depend on the number of customers already in the facility or it may be state independent. Also of interest is whether a customer is attended to completely by one server or the customer requires a sequence of servers. Unless stated to the contrary, the standard assumption will be that one server can completely serve a customer.

2.2.4 System Capacity

[7] Stated that the system capacity is the maximum number of customers, both those in service and those in the queue(s), permitted in the service facility at the same time. Furthermore; whenever a customer arrives at a facility that is full, the arriving customer is denied entrance to the facility. Such a customer is not allowed to wait outside the facility (since that effectively increases the capacity) but is forced to leave without receiving service. A system has an infinite capacity i.e. no limit on the number of customers permitted inside the facility ha, while a finite system has limited capacity.

2.2.5 Calling Population

This simply means the set of potential customers that are expected to receive or require the services; which could be finite or infinite depending on the customer source. In this research work, the inter-arrival time of the customers follows an exponential distribution. The number of arrivals over a specific time interval follows a Poisson distribution with mean λ. That is,

$P_n = \lambda^n e^{-\lambda}/n! \qquad n = 0,1,2,……..$

III MATERIAL AND METHODS

The purpose of this study is to examine the performance characteristics of the Fidelity bank Plc West-End, Maiduguri ATM service point. The system's characteristics of interest that will be examined in this research work include; number of arrivals (number of customers arriving to the service point at a given time), service time (the time it takes for one server to complete customer's service), the average number of customers in the system, and the average time a customer spends in the system. The results of the operating characteristics will be used to evaluate the performance of the service mechanism and to ascertain whether customers are satisfied with the banks' services. This is essential since a customer's experience of waiting can radically influence his/her perception of service quality of the bank.

The data for this study was collected from primary source and is limited to the ATM service point of Fidelity Bank Plc located in West-End, Maiduguri. Data was collected by observation, in which the number of customers arriving at the facility was recorded, as well as each customer's arrival and service time respectively. The assistance of a colleague was sought in recording the service time while the researcher records arrival time. The period for the data collection was during busy working hours (i.e. 8:00am to 4:00pm) and for a period of ten (10) working days. Based on the system's arrival and service pattern, and the assumptions made during data collection, the M/M/1 queuing system was used to analyze the data collected using Micro Soft computer package.

3.1 Queuing Models

Model as an idealized representation of the real life situation; in order to keep the model as simple as possible however, some assumptions need to be made [9].

3.1.1 Assumptions Made on the System

1) Single channel queue.
2) There is an infinite population from which customers originate.
3) Poisson arrival (Random arrivals).
4) Exponential distribution of service time.
5) Arrival in group at the same time (i.e. bulk arrival) is treated as single arrival.
6) The waiting area for customers is adequate.
7) The queue discipline is First Come First Served (FCFS).

Although several queuing models abound; designed to serve different purposes; [5] highlighted the following:

The M/M/S/GD/∞/∞; there is s servers to serve from a single line customer, if the arrival is less than or equals to s server every customer is being attended to; if j arrival is greater than the s servers, then j-s customers are waiting in the line.

M/G/∞/GD/∞/∞ queuing system, there e is infinite servers: meaning customer need not to wait for their service to begin; one good example of this is the self-service system such as shopping on the internet.

The machine repair model M/M/R/GD/K/K, where R is the number of servers and K both stand for population of customers and the maximum number allowed to the system, in other words; there are K machines that each break down at rate λ and R repair workers who can fixed a machine atv rate μ. Where λ and μ are dependent on either how many machines are remaining in the population or how many repair workers are in service.

The M/G/s/GD/s/∞ queuing system, here when customers arrives and found that all the servers are busy he simply walks away without being served. No queue is actually formed. Since no queue is formed $L_q=W_q=0$

If λ is the arrival rate and 1/ μ is the mean service time, the $W=W_s= 1/\mu$

Thus, the model considered in this study is the single server queuing systems.

3.2 M/M/1 Systems

An M/M/1:(∞/FCFS) queuing system: here the arrival and service time both has an exponential distribution, with parameters λ and μrespectively, 1 server, FCFS is the queue discipline and infinite population size from which the system can draw from.

The expected inter-arrival time and the expected time to serve one customer are (1/ λ) and (1/μ) respectively. An M/M/1 system is a Poisson birth-death process. The probability, $P_n(t)$ i.e. the system has exactly n customers either waiting for service or in service at time t satisfies the Kolmongorov equation with $\lambda_n= \lambda$ and $\mu_n= \mu$, for all n.

The steady state probabilities for a queuing system are

$$P_n = \text{Lim } P_n(t) \quad \text{as } t \dashrightarrow \infty \qquad (n= 0, 1, 2, 3,\ldots)$$

If the limit exist.

For an M/M/1 system, we define utilization factor (traffic intensity) as $\rho= \lambda/\mu$

And steady-state probabilities as:

$$P_n= \rho^n(1- \rho)$$

If $P < 1$.

But,

if $P > 1$, i.e the arrival comes at a faster rate than the server can accommodate.

3.2.1 Measure of Effectiveness

L= the average number of customers in the system
Lq= the average length of the queue
W= the average time a customer spends in the system
Wq = the average time a customer spends in the queue
W(t) = the probability that a customer spends more than t units of time in the system.
Wq(t) = the probability that a customer spends more than t unit of time in the queue.
For an M/M/1 system $\hat{\lambda} = \lambda$ and the six (6) measures are explicitly:

$$L = \frac{\rho}{(1-\rho)}$$

$$Lq = \rho^2/(1-\rho)$$

$$W = \frac{1}{(\mu-\lambda)}$$

$$Wq = \frac{\rho}{(\mu-\lambda)}$$

$$W(t) = e^{-t/w} \quad (t \geq 0)$$

$$Wq(t) = \rho e^{-t/w} \quad (t \geq 0)$$

IV. RESULTS AND DISCUSSION

The tables below show a summary of frequencies for the inter arrival time and service time from the data collected as depicted in appendix B and C.

Table 1: Frequencies for Inter Arrival time

X	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11
Y	94	89	94	23	15	10	4	5	2	0	2

At X=0-1 implies that 94 times, customers arrived at an inter arrival time between 0 to 1minute.

Table 2: Frequencies for Service Time

P	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9
Q	23	171	91	17	16	10	6	0	3

At P=0-1 implies that 23 times, customers were served at a period of not more than 1 minute.

4.2 Arrival Rate

The arrival rate is given by $\frac{1}{\lambda} = \frac{\sum_{i=1}^{11} XiYi}{\sum_{i=1}^{11} Yi}$

Where $\sum_{i=1}^{11} XiYi$ =1042, $\sum_{i=1}^{11} Yi$ =338

Therefore, $\frac{1}{\lambda} = \frac{1042}{338}$ =3.0832

Thus, the Average arrival rate per hour

$\lambda = \frac{60}{3.0832}$ =19.46 per hour

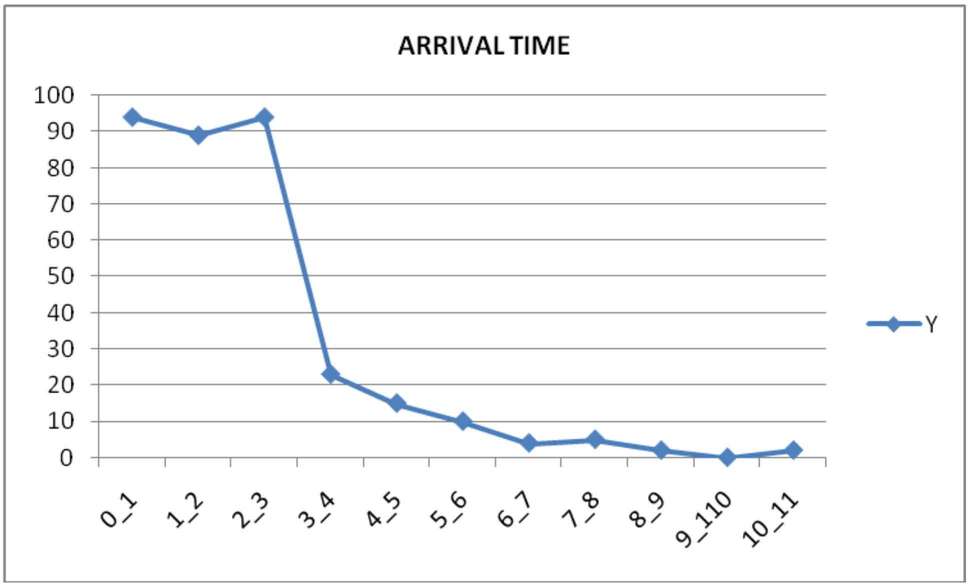

FIGURE A

4.3 Service Rate

The average service rate is given by $\frac{1}{\mu} = \frac{\sum_{j=1}^{9} PjQj}{\sum_{j=1}^{9} Qj}$

Where $\sum_{j=1}^{9} PjQj$ =1012, $\sum_{j=1}^{9} Qj$ =343

$\frac{1}{\mu} = \frac{1012}{343}$ =2.9513

The average service rate per hour

μ= $\frac{60}{2.9513}$ =20.33

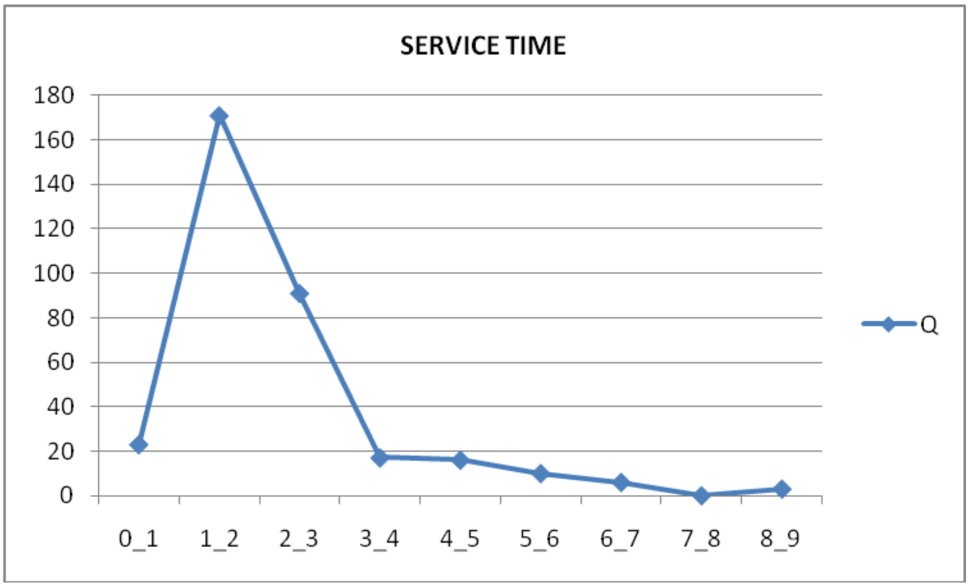

FIGURE B

Both Fig. A and B indicate an decrease in the number of arrival and service time, which is attributed to limited operation period of the bank (8:00am to 4:00pm)

4.4 Traffic Intensity and Measures of Effectiveness

The traffic intensity and measures of effectiveness are calculated using an MS60 software package. The output is given below:

```
--------------------------------------------------------------------------------
Summary of A One Channel Waiting Line With
Mean Number of Arrivals Λ=19.46
Mean Number of Services μ=20.33
 L=       22.3678          Lq=      21.4106
 W=       1.1494                    Wq=      1.1002
 W(t)=    0.0428                    Wq(t)= 0.9572
--------------------------------------------------------------------------------
```

From the results above, the average number of customers in the system, L =22

The average length of the queue, Lq = 21
The average time a customer spends in the system, W = 1.1494
The average time a customer spends in the queue, Wq = 1.1002
The probability that a customer spends more than t units of time in the system, W(t) = 0.0428
The probability that a customer spends more than t unit of time in the queue, Wq(t) = 0.9572
Results from the study revealed that the queue is quite long and as such customers that come to the ATM would have to wait too long before they are serviced by the ATM. Observations also showed that due to network complexities the service rate is relatively slow.

While this may seem to confirm the fact that the excessively long queue and lengthy service time could be due to the influx of customers and shortage of service mechanism, the opinion of management regarding the inadequacy of the service mechanism in meeting the needs of customers was sought.
Asked if management would be willing to reduce the waiting time of customers by deploying more ATM's to the banks' premises, they answered in the affirmative. The bank however added that it could only afford one additional ATM for the time being.

Thus, with the bank agreeing to deploy one additional service point, and considering a service pattern of single queue, multiple servers in parallel, the output of the M/M/S :(∞/FCFS) where S=2 is given as:

```
-----------------------------------------------------------------------------------
Summary of a Two Channel Waiting Line With
Mean Number of Arrivals λ=19.46
Mean Number of Services μ=20.33
L=        5.2236          Lq=                4.1865
W=        0.0638          Wq=                0.0146
W(t)=     0.5512          Wq(t)=   0.4598
```

V. CONCLUSION

Results from the analysis showed the traffic intensity (ρ) to be 0.96. Since we obtained the value of the traffic intensity, otherwise known as the utilization factor to be less the one (i.e. $\rho<1$), it could be concluded that the system operates under steady-state condition. Thus, the value of the traffic intensity, which is the probability that the system is busy, implies that 95% of the time period considered during data collection the system was busy as against 4% idle time. This indicates high utilization of the system.
Results from the MS60 software package on the measures of effectiveness for an M/M/1:(∞/FCFS) shows the average number of customers in the system to be 22, the average number of customers in the queue are 21. A customer spends an average of 1.15 hours before he/she is serviced by the system while a customer is likely to spend an average of 1.10 hours in the queue waiting for service.
Comparatively, the measures of effectiveness for an M/M/2 :(∞/FCFS) puts the traffic intensity at 45%, average number of customers in the system at approximately 5 while the average number of customers in the queue is 4. Similarly, a customer spends an average of 0.063 hours in the system while queue waiting time for a customer is 0.014 hours on the average.

According to [10], in designing queuing systems we need to aim for balance between service to customers (short queues implying many servers) and economic considerations (not too many servers). Though, the provision of an additional service mechanism may be capital intensive, it would pay the bank more since the primary aim of every business organization besides profit making is customer satisfaction.

The conclusion was reached without considering cost models for the system (i.e. the cost of deploying an additional ATM and cost implication resulting from the banks' inability to provide additional service point). However, the investigator strongly recommends that management should make provision for one additional ATM so as to enable her minimizes customer waiting time and improve service rate.

5.1 RECOMMENDATIONS

[11] In his study on waiting lines highlighted that to contain queue length, utilization (i.e. traffic intensity) must be less than one, the server must have unused capacity and the server must at times be sitting idle. An average of one entity is not uncommon per queue. This also corresponds to 50% channel utilization.

The findings of this study have revealed that the system is highly, if not over utilized. This implies that arrival comes at a faster rate than the system can accommodate. The following recommendations if accepted and implemented by the bank management may help in tackling these problem.

The need for the management of the bank to deploy another ATM (i.e. an M/M/S with S=2 or more) within the bank's premises as this will minimize the waiting time of customers and hence reducing the inconveniences and frustrations associated with waiting.

The bank should review its maintenance policy so as provide timely and periodic maintenance on these machines. This will drastically reduce machine or server complexities while at the same time increasing service efficiency.

REFERENCES

[1] J. K. Sharma, Operations Research: Theory and Application, 3rd Ed. (Macmillan Ltd., India 2007)

[2] A. H. Taha, Operations Research: An Introduction, 7th Ed. (Prentice Hall, India, 2003)

[3] J. Hiray, Waiting Lines and Queuing System, Article of Business Management, 2008

[4] E. Anderson, A note on managing waiting lines. UT McCombs School of Business, 2007.

[5] Ryan Berry (2006), Queuing Theory,

[6] R.A.Nosek Jr, and J.B Wilson, Queuing Theory and Customer Satisfaction: a review of terminology, trends and applications to pharmacy. Hospital pharmacy, 36, 2001, 275-279

[7] H. A. Taha, operation research. An Introductory 2nd Ed,(Macmillan, New York, 1976)

[8] B. F. Adam, I. J. Boxma, and J. A. C. Resing, Queuing models with multiple waiting lines queuing systems, 37, 2001, 65-9

[9] D.S Hira and P.K.Gupta, Simulation and Queuing Theory Operation Research, (S Chand and company ltd, new Delhi, India, 2004)

[10] J. E. Beasley, Queuing Theory, Operational Research Notes, 13, 2002, 1-5.

[11] J. B. Atkinson, Some Related Paradoxes of Queuing Theory: New cases of unifying explanation. Journal of Operational Research Society, 51, 2000, 8-11.