European Journal of Business and Management
ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online)
Vol.6, No.39, 2014

Efficient Queue Management in Supermarkets: A Case Study of Makurdi Town, Nigeria

ANTHONY IGWE
Department of Management, University of Nigeria, Enugu Campus, Enugu, Nigeria

J. U. J ONWUMERE
Department of Banking and Finance, University of Nigeria, Enugu Campus, Enugu, Nigeria

OBIAMAKA P. EGBO
Department of Banking and Finance, University of Nigeria, Enugu Campus, Enugu, Nigeria

Abstract
Waiting lines or queues are a common phenomenon in life, especially in the province of organizations that are for profit making. Queues are common in such places as petrol or filling stations, supermarkets stores, clinics, hospitals, motor parks, manufacturing firms, to mention a but a few. An interesting aspect of queuing process resides in the measures of its system's performance, especially in terms of average service rate, systems, utilization and the costs implied for a given capacity level. This paper examines efficient queue management in Nigeria with Makurdi as case study. The "M/M/I" model is applied owing, inter alia, to its relative simplicity as well as its relevance to the firms under-studied: the City Company, Nobis Ltd, Dobbiac and Eke Ltd; particularly the service unit of each of the firms. The analysis reveals that on the aggregate, queue management in Makurdi town, the capital of Benue State is grossly inadequate, inefficient and ineffective for the most part. The paper has consequently made suggestions to help mitigate the prevailing queuing problems in Makurdi town, which will be relevant to many cities in developing countries having similar challenges.
Keywords: Queue Management, Supermarkets, Nigeria

1.0 Introduction

The thrust of this paper is the queing theory as it relates to waiting lines events in contemporary Benue State of Nigeria with a focus on the for-profit sector. The focal issues are: what are the main costs related to queuing process? Who are the principal actors with respect to waiting lines? What are the main categories of queuing models? In other words, what does the generic queuing framework consist of? What are the queuing parameters?

Research and experience in most cases have shown that customers will often leave a store without making a purchase rather than stand in a long or slow moving check-out line. Despite large advancements in technology designed to decrease wait times, queue management remains a challenge for every retailer and shop owner. The truth is that when queues are not well managed, it will definitely result to having unhappy customers, sales decrease, lower customer satisfaction, and loss of customer loyalty.

Owners of businesses especially retailers have been trying to measure and manage queues and customer wait times for years. Basically, a lot of methods have been used. Some have required the use of expensive, inaccurate, and unreliable methods such as expensive entrance/exit traffic counters, proprietary solutions that combine expensive dedicated hardware with proprietary software, industrial engineering studies, and customer satisfaction surveys. But in spite of all these efforts, today's retail shoppers remain largely dissatisfied with their shopping experiences.

The objective of this study was to identify the forces militating against efficient and effective queuing management in the for-profit oriented ventures in Benue State. Precisely, the study was concentrated on the queuing activities as they relate to supermarket organizations/businesses in Makurdi town, capital of Benue State of Nigeria.

The study is divided into the following sections: Section one is the introduction;, section two discusses the cost implications of waiting lines, section three deals with the issue of customer and the server in waiting lines, while section four looks at the queuing framework in general. Section five identifies the key queuing parameters.

Importantly, section six and seven are centered on the theoretical framework and empirical analysis/results respectively. Finally, the discussion of the findings and the attendant recommendations form the pivot of section

eight of the study.

2.0 COSTS ASSOCIATED WITH QUEUES

Perhaps one of the most problematic issues in queuing analysis is that of how to attain the very important goal of queuing, whch is essentially to minimize total costs. As Stevenson (1999) has pointed out there are two basic sets of costs in queuing. These are (1) costs associated with customers waiting for service, and (2) costs related to capacity. "Capacity costs are the costs of maintaining the ability (of the system) to provide service…." (Stevenson 1999:813). .

The costs of customer waiting include the salaries paid to employees while they wait for service. Other examples include the time lost as carpenters remain idle waiting for tools to be made available by the employer, the fuel consumed by cars or lorries waiting to park, the loss of any business due to customers refusing to wait and possibly going elsewhere next time around (Adam and Ebertm,2000).

3.0 CUSTOMER AND SERVICES

Weiss and Gershon (1989) have identified a number of situations and operations in which waiting lines can be ensured. These, according to them include the barbershops, gasoline or petrol stations, tollbooths, banks, airports and hospital emergency rooms. Others are the library, offices, computer centres and parking lots. Each of these locations has its corresponding "customer". Thus, the "customer" for the parking lots or garage is a car, that of mass transportation is a "commuter" and the one for the hospital emergency room is known as a "patient" and so on. (Jhingan, 2003; FS 2004).

Another interesting aspect of the waiting lines, according to Weiss and Gershon (1989, is the fact that they all have servers. For instance, the server in case of the hospital is the doctor, the server for airport is the runway and the server for the library is a book and the one for the banks is a teller and so on. In a barbershop, the server is, of course, the barber. In short, the fact that queuing system do inevitably engage the services of server implies the need for correct staffing levels in organizations or firms where queuing takes place. Under normal circumstances queuing systems are designed to optimize the use of server, as we shall see shortly (Sun, 2004: Thisday, 2004).

JUDICIOUS INPUTS UTILIZATION IN QUEUES

The basic questions that spring to mind when dealing with the staffing levels include what staffing level shall we maintain per week, month and year? How can we increase the number of servers without simultaneously enlarging the wage bills or salaries of the staff? in sum, the problem is largely skewed on how optimality can be attained under the prevailing scarce resources. In other words, the question is essentially: what are the focal parameters to the optimum utilization of limited facilities in organization where queues are not just desired but also where such waiting lines have, as it were become an imperative determinant which must be judiciously managed in order to attain the corporate goals fruitfully?

4.0 QUEUING FRAMEWORK IN GENERAL

Weiss and Gershon (1989) and Slack, et. al. (1995) have identified three main broad categories of models, namely; descriptive queuing models, prescriptive queuing models and analytical queuing models. Most of the early works on queuing models was based on descriptive models, whch tended to determine the effect of having two tollgates rather than three. Prescriptive queuing, a fairly recent waiting line model, places emphasis on how the service facility can be properly controlled in order to improve the general waiting situation.

Slack, et. al. (1998) have suggested the use of the analytical queuing models other than the purely descriptive and prescriptive models. They opine that analytical queuing might be proper in certain instances especially where concern is for predicting the behaviour of units when the arrivals are in a random fashion. The analytical queuing models, they assert, have the capacity to help predict the steady state behaviour of different types of queuing systems. The analytical systems or models are most useful for capacity management purposes. Perhaps one of the reasons why some tend to be less receptive to the application of analytical queuing models resides in the fact that these models portray a matrix of complicated mathematical formulae (Bunday, 1986; Mason, 2003).

5.0 THEORETICAL FRAMEWORK

Queuing Theory

Delays and queuing problems are the most common features not only in our daily-life situations such as at a

bank or postal office, at a ticketing office, in public transportation or in a traffic jam but also in more technical environments, such as in manufacturing, computer networking and telecommunications. They play an essential role for business process re-engineering purposes in administrative tasks. "Queuing models provide the analyst with a powerful tool for designing and evaluating the performance of queuing systems"(Bank, Carson, Nelson & Nicol, 2001). Whenever customers arrive at a service facility, some of them have to wait before they receive the desired service. It means that the customer has to wait for his/her turn, may be in a line. Customers arrive at a service facility (sales checkout zone in ICA) with several queues, each with one server (sales checkout counter). The customers choose a queue of a server according to some mechanism (e.g., shortest queue or shortest workload) (Adan, 2000). Sometimes, insufficiencies in services also occur due to an undue wait in service may be because of new employees. Delays in service jobs beyond their due time may result in losing future business opportunities. Queuing theory is the study of waiting in all these various situations. It uses queuing models to represent the various types of queuing systems that arise in practice. The models enable us into finding an appropriate balance between the cost of service and the amount of waiting.

Basic Queuing Process

Customers requiring service are generated over time by an input source. The required service is then performed for the customers by the service mechanism, after which the customer leaves the queuing system. We can have the following two types of models: One model will be as Single-queue Multiple-Servers model (fig.1) and the second one is Multiple-Queues, Multiple-Servers model (fig.2) (Sheu and Babbar, 1996).

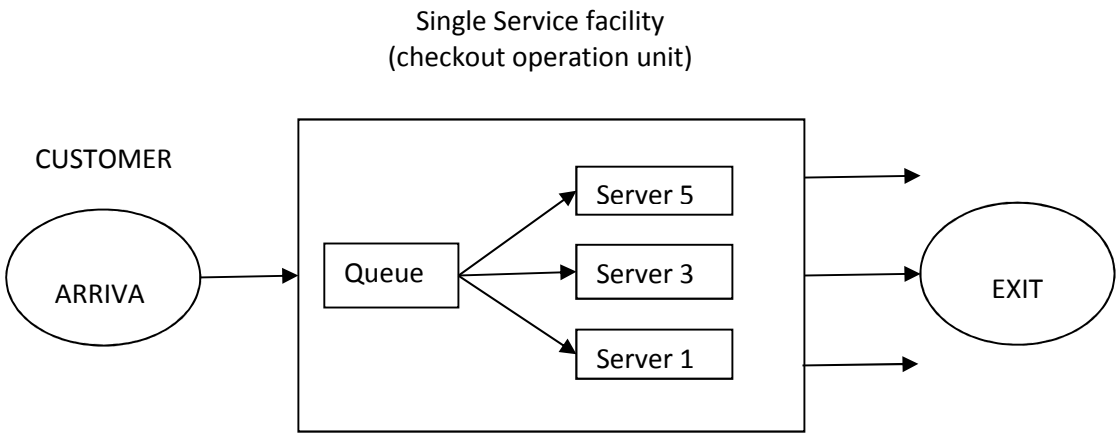

Fig. 1: Single Stage Queuing Model with Single-Queue and Multiple Parallel Servers

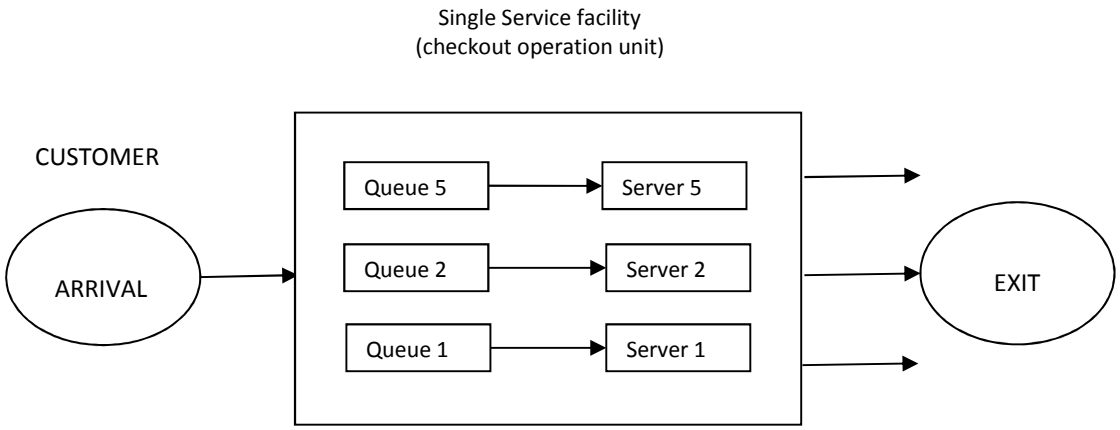

Fig. 2: Single Stage Queuing Model with Multiple Queues and Multiple Parallel Servers

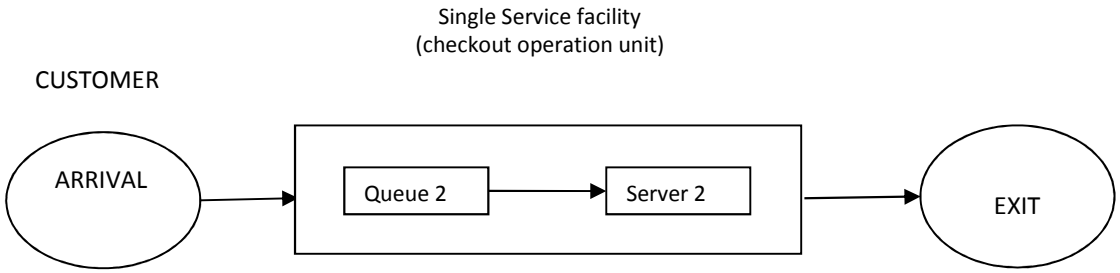

Fig. 3: Single Stage Queuing Model with Multiple Queues and Multiple Parallel Servers

In these models, three various sub-processes may be distinguished:

- **Arrival Process**: includes number of customers arriving, several types of customers, and one type of customers' demand, deterministic or stochastic arrival distance, and arrival intensity. The process goes from event to event, i.e. the event "customer arrives" puts the customer in a queue, and at the same time schedules the event "next customer arrives" at some time in the future.
- **Waiting Process**: includes length of queues, servers' discipline (First In First Out). This includes the event "start serving next customer from queue" which takes this customer from the queue into the server, and at the same time schedules the event "customer served" at some time in the future.
- **Server Process:** includes a type of a server, serving rate and serving time. This includes the event "customer served" which prompts the next event "start serving next customer from queue" (Troitzsch, 2006).

The Queuing model is commonly labeled as M/M/c/K, where first M represents Markovian exponential distribution of inter-arrival times, second M represents Markovian exponential distribution of service times, c (a positive integer) represents the number of servers, and K is the specified number of customers in a queuing system. This general model contains only limited number of K customers in the system. However, if unlimited number of customers exist, which means K = Q, then our model will be labeled as M/M/c (Hillier & Lieberman, 2001.)

Parameters in Queuing Models (Multiple Servers, Multiple Queues Model)

n = Number of total customers in the system (in queue plus in service)
c = Number of parallel servers (Checkout sales operation units in ICA)
λ = Arrival rate (1 / (average number of customers arriving in each queue

in a system in one hour))

μ = Serving rate (1 / (average number of customers being served at a server per hour))

cμ = Serving rate when c > 1 in a system

r = System intensity or load, utilization factor (= l/(cμ)) (the expected factor of time the server is busy that is, service capability being utilized on the average arriving customers)

Departure and arrival rates are state dependent and are in steady-state (equilibrium between events) condition.

Notations & their Description for single queue and parallel multiple servers model (fig.1) assuming the system is in steady-state condition

Po Steady-state Probability of all idle servers in the system, i.e.

$$Po = \left[\sum_{n=0}^{c-1} \frac{\gamma^n}{n!} + \frac{\gamma^c}{c!(1-\rho)}\right]^{-1} \qquad \text{where } \gamma = \frac{\lambda}{\mu}$$

Pn Steady-state Probability of exactly n customers in the system

$$P_n = \frac{\lambda^n}{C!C^{n-c}\mu^n} Po \quad n > c$$

Lq Average number of customers in the waiting line (queue) $= \frac{\lambda^n}{(C)!(1-\rho)^2} X\ Po$

Wq Average waiting time a customer spends waiting in line excluding the service time $= \frac{L_q}{\lambda}$

There are no predefined formulas for networks of queues, i.e. multiple queues (fig.2). Complexity of models is the main reason for that. Therefore, we use notations and formulas for single queue with parallel servers. In order to calculate estimates for multiple queues multiple servers' model, we may run simulation.

Expected length of each Queue

Besides service time, it is important to know the number of customers waiting in a queue to be served. It is possible that any customer would change his queue and choose another if find a shorter queue in another parallel server. In general, variability of inter-arrival and service time causes lines to fluctuate in length. Then the question arises, what could be the estimated length of the queue in any server? Some papers describe the general criterion for counting the number of customers in a queue. These counts are a combination of input processes, which are: arrival point process, Poisson counting process (which counts only those units that arrive during the inter-arrival time and these units are conditionally independent on Poisson interval), and counting group of units being served within the Poisson interval. The above mentioned formula of Lq is defined for average queue length of the queuing system but does not evaluate a length of parallel queues.

We are next concerned with how to obtain solution for a queuing model with a network of queues? Such questions require running Queuing Simulation. Simulation can be used for more refined analysis to represent complex systems.

Queuing Simulation:

The queuing system is when classified as M/M/c with multiple queues where number of customers in the system and in a queue is infinite, the solution for such models are difficult to compute. When analytical computation of T is very difficult or almost impossible, a Monte Carlo simulation is applied in order to get estimations. A standard Monte Carlo simulation algorithm fixes a regenerative state and generates a sample of regenerative cycles, and then use this sample to construct a likelihood estimator of state (Nasroallah, 2004). Although supermarket sales do not have regenerative situation but simulation here is used to generate estimated solutions. Simulation is the replication of a real world process or system over time. Simulation involves the generation of artificial events or processes for the system and collects the observations to draw any inference about the real system. A discrete-event simulation simulates only events that change the state of a system. Monte Carlo simulation uses the mathematical models to generate random variables for the artificial events and collect observations(Banks, 2001). Discrete models deal with system whose behavior changes only at given instants. A typical example occurs in waiting lines where we are interested in estimating such measures as the average waiting time or the length of the waiting line. Such measures occur only when the customer enters or leaves the system. The instants at which changes in the system occur, identify the model's events, e.g. arrival and departure of the customers. The arrival events are separated by the 'interarrival time' (the interval between successive arrivals), and the departure events are specified by the service time in the facility. The fact that these events occur at discrete points is known as "Discrete-event Simulation." (Taha, 1997)

When the interval between successive arrivals is random then randomness arises in simulations. The time t between customers' arrivals is represented by an exponential distribution; to generate the arrival times of the next customers from this distribution, we have $\left[\frac{1}{\mu}\right] 1n\,(1-R)$ where R = random number. (1 – R) is a compliment of R, so we can replace (1 – R) with R.

6.0 METHODOLOGY

The methodological framework used for this study was the M/M/I: pronounced "em em one". It means a memory less queuing system. It is the most basic queuing system. The first "M" means that the arrival is in consonant with the Poisson process i.e. units acting independently coverage at a point or spot. The second "M" implies memory less (exponential) service times. The "I" means only "one serve". The model is depicted below in six equations:

$$Lq = \frac{\lambda 2}{\mu(\mu-\lambda)} \qquad (1)$$

$$L = \frac{\lambda 2}{\mu-\lambda} \qquad (2)$$

$$Wq = \frac{\lambda}{\mu(\mu-\lambda)} \qquad (3)$$

$$W = \frac{1}{(\mu-\lambda)} \qquad (4)$$

$$P = \frac{\lambda}{\mu} \qquad (5)$$

$$Pn = \left(1-\frac{\mu}{\lambda}\right)\left(\frac{\mu}{\lambda}\right)^n \qquad (6)$$

where

λ= arrival rate
μ= service rate
Lq = average number of customers in the queue
L = average number of customers in the system.
Wq = average time spent waiting for service
W = average time spent in the system.

P = utilization i.e. the proportion or percentage time each server is busy
Pn = the probability that exactly n customers are in the system.

Theoretical framework is based on the assumption that average arrivals and average service follow the Poisson discrete probability distribution. The average inter-arrival time and the average services time follow the exponential distribution.

DATA COLLECTION AND SAMPLING TECHNIQUE USED

The data were obtained via a direct observation by the researchers. The researchers personally visited four supermarkets stores or shops in the "Wurukum-High-Level" axis of the town. There were six of such visits in all. The choice of the area was informed by the fact that it was a highly populated and "business-activity-centered" area.

The researchers used a deliberate sampling technique in the choice of the part of the town studied. However, the selection of the four shops under reference was based on simple random sampling technique.

7.0 DATA ANALYSIS

The data (in table 1) shows that the mean arrival per hour for the four supermarkets ranges between 7 and 12 inclusive while the mean service per hour is relatively higher- it ranges between 9 and 15. The city company Ltd had both the highest mean arrival and mean service of 12 and 15 respectively. The city company seemed to enjoy the most strategic location via-avis other firms. It is very close to several motor parks including the Agesi Motor Park. The Nobis Ltd came in terms of mean arrival and mean service per hour, which were 8 and 11 respectively. This could be attributed to the fact that its location is comparatively better than that of Dobbiac Ltd with the mean arrival of 8 per hour and mean service of 10 per hour respectively. The Eke Ltd which is relatively obscure or hidden from the railway, enjoyed the least score in terms off mean arrival and mean service per hour,

respectively 7 and 9. The study showed that city Ltd was the most efficient in terms of mean service per hour of 15, a staggering difference of 6 when juxtaposed with Eke Ltd with the mean service per hour of 9. comparatively, there was a gross lack of efficiency in the use of service facilities, given that only about 11 customers could be serviced per hour on the average instead pf say, 14.

The data (in table 2) depicts the computed values in respect of the equations 1 to 5. Equations one shows that with the exception of Nobis Ltd, the average number of customers in the queue had tended to be 3. The average number of customers in the system gravitated around 4 generally as equation two has indicated. The average time spent waiting for service as shown by equation three was least for Nobis Ltd where a customer spent 0.24 hour (14.4 mins) waiting for service. The Dobbiac and Eke Ltd recorded 0.4 hour (24 mins) as being the time a customer spent waiting for service. The data shows hat both Dobbiac Ltd and Eke Ltd are less efficinet in terms of the average time spent in the system. In each of them, 0.5 hour (30 mins) was spent on the average in the system. Finally, equation five shows that the input utilization (i.e. the proportion of time each server is busy) was generally very high. The least utilization was recorded against Nobis Ltd where it stood at 0.73 (or 73%). The Dobbiac Ltd and City Ltd recorded utilization of 80% each, to top the list. This implies prima facie that servers were being over-stressed in these supermarkets. They hardly had enough time to rest once work got started on that day.

If utilization was defined as the percentage of capacity (other than human resources) employed, it would mean that the existing physical facilities were being over-stressed and such would lead to increased costs including the maintenance and replacements costs.

8.0 DISCUSSION ANDS RECOMMENDATIONS

The study has shown that the mean service rate was rather poor generally. The study has also shown that the average time a customer spent while waiting for service was too long. A situation where a customer has to wait up to nearly half an hour before he was serviced smacks of inefficiency on the part of the firm. It must be borne in mind that most of the customers are in a hurry for one thing or another. They demand to be treated or serviced with dispatch. The sense of urgency was clearly absent.

The study has shown that “utilization” is rather too high on the average. The input (labour) had been over-stressed given that there was no adequate time for the workers to rest. The service process seemed to be continuous one and this could lead to the breakdown of the human resources and even of inanimate inputs. In fact, there might be a tendency towards a declining or diminishing productivity of the labour if it is not given adequate period for relaxation after being used for a long time. There was also the tendency for the task to become unnecessarily burdensome and monotonous to the worker if the task did not permit one to rest at least briefly before he continued with. The study indicated a gross lack of queuing discipline in virtually all of these supermarkets. Many customers were allowed to jump the queue or renege.

In the light of the above, it might be pertinent to proffer the following recommendations to help mitigate the adverse effects these pitfalls might bring to bear on the firms in the long run. In the first place, there is a need to improve on the service facilities in these supermarkets. The maintenance and replacement costs were rapidly skyrocketing. More modern, sophisticated equipment and facilities including modern computers rather than manual and obsolete facilities should be provided to enhance labour productivity. Such would also help to drastically reduce the time the customers spend on the average waiting for service. There should also be dynamic and efficient mechanism by management of these supermarkets to seek for optimal system utilization. This would help to decrease the tendency for costs associated with system utilization to rise over time. There should be some optimal number of hours that a given server (worker) can perform. The tendency to over-labour or over-stress an input including human resources should be eliminated in order to motivate the workers to increase their productivity in the place. It is our view that these recommendations will go a long way in ameliorating the queuing problems of these organizations and by extension, to organizations in developing countries having similar challenges.

APPENDEXES

Table 1: Supermarkets Data Concerning the Mean Arrival and Man Service

(1) (2) (3)

Name of supermarket	Mean arrival per hour	Mean service per hour
1. Dobbiac Ltd	8	10
2. City Ltd	12	15
3. Nobis Ltd	8	11
4. Eke Ltd	7	9

Table 2: Supermarket Data: Computed values for Equation 1 to 5.

Supermarket:	Computed Vales of the Queuing Equations (1) Lq	(2) L	(3) Wq	(4) W	(5) P
DOBBIAC Ltd	3.0	4	0.40	0.50	0.80
CITY Ltd	3.0	4	0.27	0.33	0.80
NOBIS Ltd	2.0	3	0.24	0.33	0.73
EKE Ltd	3.0	4	0.40	0.50	0.78

REFERENCES

Adam, E.E. and Ebert R.J. (2000) Production and Operations Management (New Delhi Prentice-Hall)

Adan, I.J.B.F., Boxma1, O.J., Resing, J.A.C. (2000), "Queuing models with multiple waiting lines," Department of Mathematics and Computer Science, Eindhoven University of Technology,

Banks, J., Carson, J. S., Nelson, B. L., Nicol, D. M. (2001), Discrete-Event System Simulation, Prentice Hall international series, 3rd edition, p24–37

Bunday; B.D. (1986) Basic Queuing Theory (London: Edward Arnold Publishers Ltd).

Foote B.L. (1976) A Quickening Case Study of Drive in Banking in Interface 6. N0 4, August.

FS (2000): Financial Standard Weekly Paper 2nd August 204.

Hillier, F. S., Lieberman, G. J. (2001), Introduction to Operations Research, McGraw-Hill higher education, 7th edition, p834–8

Jhingan, M.L. (2003) Modern Macro-Economic (Delhi: Vrinda Publications Ltd)

Lee A. (1996) Operations Management (London: Pitman Publishing Company Ltd)

Nasroallah, A. (2004), "Monte Carlo Simulation of Markov Chain Steady-state Distribution," Extracta Mathematicae, Vol. 19, No. 2, p279-288

Sheu, C., Babbar S. (Jun 1996), "A managerial assessment of the waiting-time performance for alternative service process designs," Omega, Int. J. Mgmt Sci. Vol. 24, No. 6, pp. 689-703

Slack et al (1995) Operations management (London: Pitman Publishing Company Ltd)

Stevenson, W.J. (1999) Production and Operations Management (New York: Irvin McGraw-Hill).

Taha, Hamdy A. (1997), Operations Research an Introduction, PHIPE Prentice, 6th edition, p607–643

Troitzsch, Klaus G., Gilbert, Nigel (Sep 2006), "Queuing Models and Discrete Event Simulation," ZUMA Simulation Workshop 2006

Weiss H.J. and Gershon, M.E. (1989) Production and Operations Management (London: Alllyn & Bacon Incorporated.)

.