CHAPTER TWO

**2.0 LITERATURE REVIEW**

**2.1 Introduction**

Waiting in lines seems to be a component of a human daily life. Queues form when the demand for a service exceeds its supply. In Filling stations, customers can wait a certain period of time (minutes, hours, days) to receive products or service. For many customers, waiting in lines or queuing is annoying or negative experience. The disagreeable experience of waiting in line can often have a negative consequence on the rest of a customer’s experience with a particular firm. The way in which managers address the waiting line issue is critical to the long term success of their firms.

Literature on queuing models indicates that waiting in line or queue causes problem to economic expenses to persons and institutions. Hospitals, banks, airline companies, industrialized firms etc., attempt to decrease the total waiting price, and the cost of service provided to their customers. Therefore, speed of service is increasingly becoming a very important competitive parameter. Davis (2003) assert that providing ever-faster service, with the ultimate goal of having zero customer waiting time, has recently received managerial attention for several reasons. First, in the more highly developed countries, where standards of living are high, time becomes more valuable as a commodity and consequently, customers are less willing to wait for service. Second, this is a growing realization by organizations that the way they treat their customers today significantly impact on whether or not they will remain loyal customers tomorrow. Finally, advances in technology such as computers, internet etc., have provided firms with the ability to provide faster services.

For these reasons managers and service providers are always finding way to deliver more rapidly services, believing that the waiting will negatively affect the organization performance evaluation.

Researchers have argued that service waits can be controlled by two techniques: operations management or perceptions management (Hall, 2006). The operation management feature deals with the organization of how customers, queues and servers can be coordinated towards the goal of rendering efficient service at the minimum cost. The act of waiting has significant impact on patients’ satisfaction. The amount of time customers must spend waiting can significantly influence their satisfaction. Additionally, research has demonstrated that customer satisfaction is affected not just by waiting time but also by customer expectations or attribution of the causes for the waiting. Consequently, one of the issues in queue management is not only the actual amount of time the customer has to wait, but also the customer’s perceptions of that wait. Clearly, there are two approaches to increasing customer satisfaction with regard to waiting time: through decreasing actual waiting time, as well as through enhancing customer’s waiting experience (Singh, 2011).

**2.2 Theoretical Framework**

**2.2.1 Queuing Theory**

Queuing theory is the mathematical study of waiting lines, or queues (Sundarapandian, 2009). In queuing theory a model is constructed so that queue lengths and waiting times can be predicted (Sundarapandian, 2009). Queuing theory is generally considered as branch of operations research because the results are often used when making business decisions about the resources needed to provide a service. Queuing theory has its origins in research by Agner Erlang when he created models to describe the Copenhagen telephone exchange (Sundarapandian, 2009). The ideas have since seen applications including telecommunication, traffic engineering, computing and the design of factories, shops, offices and hospitals (Schlechter, 2009).

**Key Variables**

• Τ: interarrival time

• λ: mean arrival rate = 1/E[Τ]

• s: service time per job

• µ: mean service rate per server = 1/E[s]

• n: number of jobs in the system(queue length)

• nq: number of jobs waiting

• ns : number of jobs receiving service

• r: response time = time waiting + time receiving service

• w: waiting time = time between arrival and beginning of service

Typically we can talk of this individual sub-system as dealing with “customers” queuing for “service”. To analyze this sub-system we need information relating to arrival process, service mechanism, and queue characteristics.

**Essential components to describe a phenomenon of waiting line**

The following components are essential to describe a phenomenon of waiting line: the population Source, the arrival, queues, queue discipline, service mechanism, departure or exit.

a. Population source

The population source serves as where arrivals are generated. Arrivals of customers in a filling station may be drawn from either a finite or an infinite population. Arrivals of customers in Bovas filling station, Opopogboro Ado Ekiti are drawn from an infinite population. A finite population source refers to the limited size of the customer pool. Alternatively, an infinite source is forever.

b. Queue discipline

The queue discipline is the sequence in which customers are processed or served. The most common discipline is first come, first served (FCFS). Other disciplines include last come, first served (LCFS) and service in random order (SIRO). Customers may also be selected from the queue based on some order of priority (Taha, 2005).

c. Service mechanism

The service mechanism describes how the customer is served. It includes the number of servers and the duration of the service time, both of which may vary greatly and in a random fashion. The number of lines and servers determines the choice of service facility structures. The common service facility structures are: Single-channel, Single-phase; Single-channel, Multiphase; Multi-channel, Single phase and Multi-channel, Multiphase.

d. Departure or Exit

The departure or exit occurs when a customer is served. The two possible exit scenarios as mentioned by Davis (2003) are:

(a) The customer may return to the source population and immediately become a competing candidate for service again;

(b) There may be a low probability of re-service.

**Birth and Death Process Queuing Models**

A number of important queuing theory models fit the birth-and-death process. A queuing system based on the birth-and–death process is in state En at time t if the number of customers is then n, that is, N (t) =n. A birth is a customer arrival, and a death occurs when a customer leaves the system after completing service. Thus, given the birth rates {λn} and death rate {µn}, and assuming that

S=1 + C1 + C2 + C3 + … < ∞ (1)

Where

Cn = , n= 1,2,3,… (2)

We calculate

P0 = 1/S (3)

and

Pn= P [N=n] = CnP0, n= 1, 2, 3,… (4)

From the probabilities calculated by (4) we can generate measures of queuing system performance (Nosek, 2001).

**M/M/1 Queuing System**

****

**Figure 2.1: M/M/1 Queuing System.**

This model assumes a random (Poisson) arrival pattern and a random (exponential) service time distribution. The arrival rate does not depend upon the number of customers in the system and the probability of an arrival in a time interval of length h > 0 is given by

(5)

Thus, we have

λn = λ n = 0,1,2,... (6)

By hypothesis, the service time distribution is given by

(7)

Then, when a customer is receiving service, the probability of a service completion (death) is a short time interval, *h*, is given by

(8)

(Here we have used the memoryless property of the exponential distribution in neglecting the service already completed.) (Little, 1961).

Thus,

n = 1, 2, 3, (9)

Thus the state-transition diagram for the M/M/1 queuing system is given by Figure 2 and therefore,

by (1), since λ/µ=ρ, and each *Cn* is equal to ρn,

(10)

Hence,

n=0, 1, 2, 3… (11)



**Figure 2.2: State-transition diagram of the M/M/1 queuing system.**

But (11) is the pmf for a geometric random variable, that is, N has a geometric distribution with p = 1- ρ and q = ρ.

Hence,

L = E [N] = q/p = ρ / (1 – ρ), (12)

And

σ2 N = ρ / (1 – ρ )2 (13)

By Little’s formula,

W = E[q] = W – E[s] = ρ E[s] / (1 – ρ), (14)

Since ρ = λE[s].

Now,

Wq = E[q] = W – E[s] = ρ E[s] / (1-ρ) (15)

Applying Little’s formula, again, gives,

Lq = E [Nq] = λWq = ρ2  / (1-ρ) (16)

By (11) we calculate

P [server is busy] = 1 – P[N=0] = 1 – (1-ρ) = ρ.

By the law of large numbers this probability can be interpreted as the fraction of time that the sever is busy; it is appropriate to call ρ the “server utilization”.

We now have the four parameters most commonly used to measure the performance of a queuing system, *W*, Wq, L and Lq as well as the pmf q*n* , of the number in the system.( Allen, 1978)

2.4 Summary

References