CHAPTER – I

INTRODUCTION

1. Introduction

"MATHEMATICS is a great motivator for all humans.

Because its career starts with "ZERO" but it never end (INFINITY)"

-Vignesh. R

Queueing theory is the mathematical study of "queues" or "waiting lines". A queue is formed whenever the demand for service exceeds the capacity to provide service at that point in time. A queueing system can be described as customers arriving for service, waiting for service if it is not immediate and leaving the system after being served.

1.1 Characteristics of a Queueing System

The basic characteristics of a queueing system which provide an adequate description are arrival pattern, service pattern, queue discipline, system capacity and service channels.

Arrival Pattern

In a queueing system the process of arrival is stochastic. Arrival may be either single or batches of variable or fixed size. Thus it is necessary to know the probability distributions describing the times between successive arrivals and describing the size of the batch.

It is also necessary to know the reaction of the customer upon entering the system. If the queue is too long, a customer may decide not to enter it upon arrival. In this situation he is said to have balked. On the other hand, a customer may enter the queue, but after sometime lose patience and decide to leave. In this case, he is said to have reneged. In the event that there are two or more parallel queues, the customers may switch over from one to another. In this case he is said to jockey for position.

The arrival pattern of customers that does not change with time is called stationary arrival pattern otherwise it is called non-stationary.

Service Pattern

Service pattern describe the manner in which the service is rendered to the arrivals. Customers may be served either singly or in batches of variable or fixed size.

The service pattern of customers may be stationary or non-stationary with respect to time and state dependent or independent with respect to number of customers waiting for service. The time required for serving a unit is called service time. The service time may be deterministic or probabilistic.

Queue Discipline

Queue discipline refers to the manner in which customers are selected for service from the queue. The most common disciplines are first come first served (FCFS) and last come first served (LCFS) which are based on the arrivals of customers into the system. Customers may also be served randomly irrespective of their arrivals to the system called service in random order (SIRO).

Another discipline is priority queue discipline, which allows service to be offered to customer depending on their priority in relation to other customer. There are two types in priority discipline, that is preemptive priority and non-preemptive priority. In the preemptive case, the customer with high priority is allowed to enter service immediately suspending the service in progress to a customer with lower priority. In non-preemptive case the higher priority goes to the head of the queue but gets into service only after the completion of service in progress to the customer with lower priority.

System Capacity

In some queueing processes there is a physical limitation to the amount of waiting room so that when the line reaches a certain length, no further customers are allowed to enter until space becomes available as the result of a service completion. These are referred to as finite queueing situations. A queue with limited waiting room can be viewed as one with forced balking.

Service Channels

The number of servers in a queueing model may be finite or infinite. The number of servers may be arranged in series, parallel or a combination of both, depending upon the nature of the services required. In parallel channels, all the channels provide identical services so that several customers may be served simultaneously. In series channels, a customer must pass through successively in several ordered channels before service is completed.

1.2 Queue Notation

As shorthand for describing queueing processes, a notation has evolved, due to Kendall (1953), which is now standard throughout the queueing literature. A queueing process is described as A/B/X/Y/Z, where A indicates the inter arrival time distribution, B is the service pattern described by the probability distribution for service time, X is the number of servers, Y is the restriction on system capacity and Z is the queue discipline. Table 1 bellow summarizes some of the most common symbols.

Characteristic	Symbol	Explanation
Interarrival time	M	Exponential
Distribution (A)	D	Deterministic
and	E_k	Erlang type k (k=1, 2, …)
Service time	H_k	Mixture of k exponentials
Distribution (B)	G	General distribution
Number of servers (X)	1, 2…..∞	
Restriction on system capacity	1, 2…..∞	
Queue discipline	FCFS	First come first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

In many situations only the first three symbols are used. Current practice is to omit the service capacity symbol if no restriction is imposed (Y=∞), and to omit the queue discipline if it is first come first served (Z=FCFS). The symbol G represents a general probability distribution; all we know is that the inter arrival times are independent and identically distributed.

1.3 Retrial Queueing System

In conventional queueing theory it is usually assumed that an arriving customer who cannot get service immediately either joins the waiting line or leaves the system forever. Sometimes impatient customers leave the queue but it is also assumed that they are leaving the system forever. Usually such customers after some random period of time return to the system and try to get service. The standard queueing models do not consider the phenomenon of retrials and therefore cannot be applied in solving a number of practically important problems. Retrial queues have been introduced to meet this inadequacy.

Retrial queueing systems are characterized by the feature that arriving customers who cannot receive service immediately may join a virtual queue called orbit to try their request after some random time. General structure of a retrial queueing system is presented in Fig. 1.1.

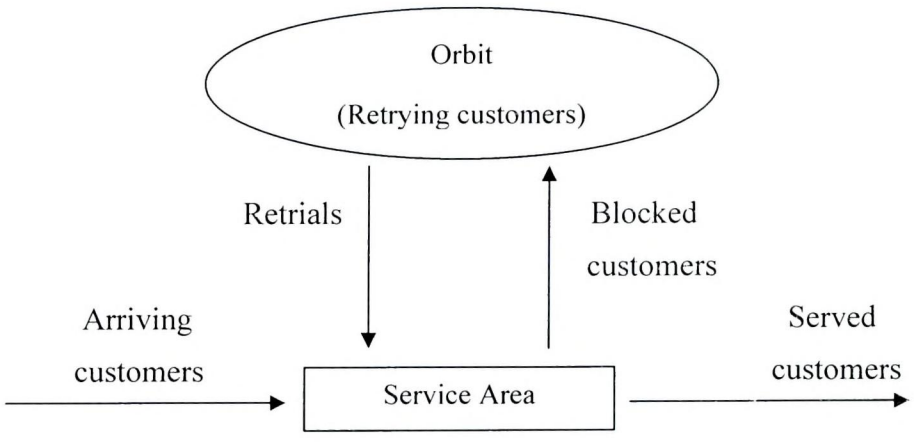

Fig . 1.1 General Structure of a Retrial Queueing System

1.4 Review of literature

The theory of queues was initiated by the Danish mathematician Erlang, who in 1909 published "The theory of Probabilities and Telephone Conversation". He observed that a telephone system was generally characterized by either (1) Poisson input, exponential holding time, and multiple channels, or (2) Poisson input, constant holding time and a single channel. Erlang was also responsible in his later works for the notion of stationary equilibrium and for the first consideration of the optimization of a queuing system.

In 1927, Molina published "Application of the Theory of Probability to Telephone Trunking Problems", and one year later Thornton Fry printed "Probability and its Engineering Uses" which expand much of Erlang's earlier work. In the early 1930's Felix Pollaczeck did some further pioneering work on Poisson input, arbitrary output, and single and multiple channel problems. Other names working in the same field during that period included Kolmogorov and Khintchine in Russia, Crommelin in France and Palm in Sweden.

Queueing systems with repeated attempts are found suitable for modelling the processes in telephone switching systems, digital cellular mobile networks, packet switching networks, local area networks, stock and flow etc. Review of **retrial** queueing literature can be found in the survey papers of Yang and Templeton (1987) and Falin (1990), the bibliographies of Artalejo (1999a, 1999b) and the books by Falin and Templeton (1997) and Artalejo and Gomez Corral (2008). The applications of retrial queues in science and engineering are given in Kulkarni and Liang (1997).

The batch arrival process is a useful mathematical model for describing busty traffic in modern communication networks. Batch arrival retrial queueing model was introduced by Falin (1976). Kulkarni (1986) and Falin (1988) analysed multiple classes of customers with batch arrivals. Chakravarthy and Dudin (2002) presented an article on multi-server retrial queue with BMAP arrivals and group services. A detailed study on batch arrival queue under both classical and constant retrial policies was done by Jain et al. (2008). Yamamuro (2012) analysed an M/G/1 retrial queue with batch arrivals and obtained the decomposition of expected queue length as the sum of two independent random variables, one corresponds to the queue length of a

standard M/G/1 batch arrival queue and another is compound Poison distributed. Ayyappan et al. (2013) analysed $M^{[X]}/G/1$ queue with two types of service subject to random breakdowns, multiple vacation and restricted admissibility.

Queues with server subject to **breakdowns and repairs** are often encountered in many practical applications. It is of basic importance to study the reliability of retrial queues with server breakdowns and repairs because of limited ability of repairs and heavy influence of the breakdowns on the performance measures of the system. Aissani (1988) and Kulkarni and Choi (1990) considered retrial queueing systems with server breakdowns and repairs. Aissani and Artalejo (1998) studied a single server retrial queueing system subject to active independent breakdown. Wang et al. (2001) obtained explicit expressions of availability, failure frequency and reliability function of the server for M/G/1 retrial queue with server breakdown. Li et al. (2006) provided BMAP/G/1 retrial queue with server breakdowns and repairs considering both from queueing view point and reliability view point. Gharbi and loualalen (2006) gave a detailed analysis of unreliable retrial system using generalized stochastic petrinets model.

Atencia et al. (2008) studied a batch arrival retrial queue subject to breakdown where the retrial time is exponential and independent of the number of customers applying for service. Choudhury and Deka (2008) discussed a Poisson input queueing system wherein the server delivers a second phase of optional service and server subject to breakdown and repair. Jain and Charu Bhargava (2008) analysed the waiting time distribution and sensitivity analysis for a batch arrival retrial queueing model with priority subscribers and unreliable server. Wang (2008) derived the transient and steady state solutions for reliability measures of M_1, M_2/G_1, $G_2/1$ retrial queues with server breakdown. Choudhury and Deka (2009) obtained the limiting distribution of the number of customers in the system at departure epoch and idle period completion epoch for an M/G/1 retrial queue with two types of heterogeneous service subject to random breakdown and repair under linear retrial policy. Zhou et al. (2009) discussed an M/G/1 retrial queue with repairable server and exhaustive vacation and derived the necessary and sufficient condition for system stability and queue indices using supplementary variable technique. Falin (2010) investigated an M/G/1 retrial queue with an unreliable server and general retrial time with the help of

embedded Markov chain. Using matrix geometric method, Kalyanaraman and Seenivasan (2010) analysed a multi-server retrial queueing system with unreliable server in which service time distribution is negative exponential.

In most of the papers with unreliable server it is assumed that whenever the system breaks down the repair process starts instantaneously. However, this is not the case in many real life situations. The system has to wait for repair to start. Recent work on unreliable retrial queueing system with delayed repair includes Prakash Rani et al. (2011), Choudhury and Ke (2012), Ebenesar Anna Bagyam and Udaya Chandrika (2012), and Ayyappan and Shyamala (2014).

In the retrial setup, each service is preceded and followed by the server's idle time because of the ignorance of the status of the server and orbital customers by each other. Server's idle time is reduced by the introduction of **search of orbital customers** immediately after a service completion. Artalejo et al. (2002) considered a retrial queue in which immediately after a service completion the server searches for customers from the orbit or remains idle. Dudin et al. (2004) extended the model to a batch arrival retrial queue and performed the steady state analysis of the queueing system. Krishnamoorthy et al. (2005) analysed M/G/1 retrial queue with non persistent customers and orbital search using supplementary variable method and discussed the structure of the busy period and its analysis in terms of Laplace transform. Chakravarthy et al. (2006) studied a multi-server retrial queueing model with orbital search using direct truncation and matrix geometric approximation. Sumitha and Udaya Chandrika (2011) discussed a single server batch arrival retrial queueing system with additional optional service and orbital search and obtained the steady state distributions of the server state and the number of customers in the orbit. Sumitha and Udaya Chandrika (2012) investigated a repairable M/G/1 retrial queue with Bernoulli vacation and orbital search and derived the queueing and reliability indices to predict the system behaviour. Deepak et al. (2012) obtained expected queue length of a batch arrival retrial queueing system with two types of search of customers from the orbit.

Queues with negative arrivals called **G queues** were first introduced by Gelenbe (1989) with a view to modelling neural networks. In recent years, a variety of

industrial applications have created interest in the modelling of reliability in G queues. Liu et al. (2009) analysed an M/G/1 retrial G queue with preemptive resume and feedback under N policy vacation, where the negative customers not only remove the customer in service from the system but also causes the server breakdown. Wang and Zhang (2009) considered a discrete time retrial queue with negative arrivals. Aissani (2010) obtained the generating function of the number of primary customers in the stationary regime of an M/G/1 retrial queue with negative arrivals and unreliable server. Wu et al. (2011) studied the MAP/PH/N retrial queue with finite number of sources and MAP arrivals of negative customers and provided an elaborate algorithm for calculating the stationary state probabilities. Wu and Yin (2011) considered an unreliable M/G/1 retrial G-queue with non exhaustive random vacations and derived steady state solution for both queueing measures and reliability quantities. Wu and Lian (2013) discussed an M/G/1 retrial G queue with priority resume, Bernoulli vacation and server breakdown. The authors obtained the necessary and sufficient condition for ergodicity of embedded Marko chain with the help of Lyapunov functions. Peng et al. (2013) suggested an M/G/1 retrial G- queue with preemptive resume priority and collisions under linear retrial policy subject to server breakdowns and delayed repairs. Kirupa and Udaya Chandrika (2014) analysed Batch Arrival Retrial G-Queue and an Unreliable Server with Delayed Repair.

1.5 Profile of Present Work

The main objective of the dissertation is to analyse the steady state behaviour of retrial queueing systems with server breakdown and orbital search.

The concept of the thesis is presented in five chapters.

- Chapter one gives the preliminary results and review of literature.
- Single server retrial G queue with server breakdown is analysed in chapter two.
- Batch arrival retrial G queue with server breakdown is considered in chapter three.

- In chapter four, M/G/1 retrial G queue with server breakdown and orbital search is discussed.

- In the final chapter an unreliable batch arrival retrial queue with positive and negative customers and orbital search is investigated.

In all the above models, the retrial time, service time and repair time are assumed to follow general distribution. All the models are formulated mathematically and analysed at equilibrium state using supplementary variable technique. The explicit expressions of expected number of customers in the orbit, expected number of customers in the system, availability and failure frequency of the server are derived. Stochastic decomposition property is verified. Special cases are discussed and numerical results are presented.