CHAPTER – I
INTRODUCTION

INTRODUCTION

Queueing theory suggests mathematical methods of analyzing the congestions and delay of waiting in line. The theory is used to develop more efficient queueing systems that reduce customers' waiting time and increase the rate of service of servers. Queueing theory was born in the early 1900's with the work of A.K.Erlang of the Copenhagen Telephone Company, who derived several important formulae for teletraffic engineering that bear his name today. The range of applications has grown to include not only telecommunications and computer science but also manufacturing, air traffic control, military logistics, management science, industrial engineering and many other areas that involve service systems whose demands are random. In terms of the analysis of the queueing situations, the researchers are typically concerned with measures of the system performance and might include :

- The average length of a queue.
- The probability that the queue will exceed a certain length.
- The expected utilization of the server and the expected time period during which the server will be fully occupied.
- The expected waiting time of customers in the queue before they are served.
- If cost can be associated with the factors such as customers waiting time and server idle time then the designing of the system at minimum total cost can be investigated.

1.1 CHARACTERISTICS OF QUEUEING PROCESSES

A queueing system is composed of customers or units, who arrive at a service facility and join a queue, if service is not immediately available and evidently leave the system after receiving service.

The principal characteristics, which describe a queueing system are the following :

1.1.1 The Input Process or Arrival Pattern of Customers

The arrival pattern describes the manner in which the arrivals occur. It is specified by the inter-arrival time between any two consecutive arrivals or by the mean arrival rate. The input pattern also indicates whether the arrivals occur single or in batches of fixed or random sizes. The inter-arrival time may be deterministic or stochastic. When it is stochastic the probability distribution associated with it is required. In case of bulk arrivals, not only the time between successive arrivals may be probabilistic but also the number of customers in a batch.

1.1.2 The Service Pattern

Service pattern can be measured by the number of customers served per some unit of time or the time taken to complete a service. The service time may also be constant (deterministic) or stochastic. If it is stochastic, the probability distribution associated with it will be required. The service can be provided in single or batch.

1.1.3 Queue Discipline

It refers to the manner by which customers are selected for service when a queue has formed. The most common discipline that can be observed in everyday life is first come, first served (FCFS) or it is sometimes called first in, first out (FIFO). Another important discipline, which is very common in the inventory system, is the last in, first out (LIFO). Besides these two, there are other queueing disciplines such as random selection of service (RSS) and a variety of priority schemes (very common in hospital causality), where customers are selected for service on the basis of their priorities.

1.1.4 System Capacity

The number of customers in the queue and in service together is called the system capacity. The system may have a queue of finite capacity or effectively infinite capacity.

1.1.5 The Number of Service Channels

A system may have a single server or a number of parallel or series of channels. In parallel channels, each and every channel provides an identical service facilities, so that several customers may be served simultaneously. In case of series of channels, a customer must pass successively through the ordered channels before service is completed. Also a queueing system may have only a single stage of service or it may have several stages operated by a single server. An example of a multi-stage queueing system is the physical examination procedure in a hospital.

1.1.6 Kendall's Notation

Any queueing system is presented by the notation introduced by Kendall (1951) : A/B/C/Y/Z, where A represents interarrival time distribution of the customers, B denotes the service time distribution, C is the number of parallel servers, Y represents the capacity of the system and Z denotes the queue discipline. If the queueing system has infinite capacity and the queue discipline is FIFO, then the system is denoted as A/B/C without mentioning Y and Z.

For example, the notation M/G/1 denotes a queueing system with Poisson (Markovian) input, generally distributed service time with single server following FCFS queue discipline and the system capacity is infinite.

The following are some of the special characteristics of queueing systems considered in the thesis.

1.1.7 Bulk or Batch Arrival Queueing Models

A queueing system where arrivals or service or both takes place in batches of fixed or random sizes is called a bulk queueing system. Batch arrival queueing models can be used in many practical situations such as the analysis of message packetization in data communication systems.

The basic queueing system considered in the present work is $M^X/G/1$. In this queueing system, the customers arrive in batches and are served individually by a single server. The batches arrive according to a time-homogeneous Poisson process with rate λ. The number of customers in the

batches are independent identically distributed positive random variables. The number of customers in a batch is denoted by X and the probability distribution of X is given by $Pr(X = k) = g_k$, $k = 1, 2, 3, \ldots$ That is, the probability that a batch of k units arrives in an infinitesimal interval (t, t+h) is $\lambda\, g_k\, h + o(h)$. The probability generating function of g_k is $X(z) = \sum_{k=1}^{\infty} g_k\, z^k$, with mean $E(X) = X'(1)$. The arrival process is also called compound Poisson process (Medhi, 2003), with mean arrival rate $\lambda\, E(X)$.

1.1.8 N-Policy Queueing Models

In classical queueing systems, service is assumed to initiate as soon as an arrival occurs. But optimal control models suggest rules for turning the server on and off. There are many threshold policies to determine the queue levels at which service should start or stop. But the emphasis was mainly on the N-policy introduced by Yadin and Naor (1963). A stationary N policy is defined as one where the server is turned on when the number of units in the queue reaches the value N and turned off when the system becomes empty. This policy is designed to minimize server switchovers and to avoid excessive frequent uses of setups.

1.1.9 Queues with Server's Vacation

In classical queueing models, servers are always available in the service facility. However in many practical situations servers may become unavailable for a period of time due to a variety of reasons. This period of server absence is called server vacation i.e., vacation in queueing models represents the period of temporary server absence.

A wide class of policies for governing the vacation mechanism have been discussed in the literature. There are two major types, namely vacations with exhaustive and non-exhaustive services. With an exhaustive service, the server cannot take a vacation until the system becomes empty. On the other hand, in non-exhaustive system, the server can take a vacation between two services, during busy period. In either case, the rules for resuming services at

vacation completion instant are numerous. Based on these rules the two main vacation policies, framed under the exhaustive service discipline are single and multiple vacation policies. The policies (i) and (ii) given below are defined for N-policy queueing models.

(i) Single and Multiple Vacation Policy

In both single and multiple vacation schemes, the server takes a vacation of random duration, as soon as the queue becomes empty. When the server returns from the vacation and finds N or more customers in the system, he immediately starts his service.

In the case of **single vacation policy,** if the server returns from vacation and finds fewer number of customers than N, then he joins the system and waits until the system size reaches atleast N and then begins to serve exhaustively.

In **multiple vacation policy** the server takes repeated number of vacations (may be infinitely many times) until the system contains at least N customers at a vacation completion instant.

Multiple vacation models may correspond to an efficient utilization of servers for secondary jobs, while single vacation models can represent a server (machine) maintenance or post production operation.

(ii) (a) J-Vacation Policy

Whenever the system becomes empty, the server immediately takes a vacation of random length. Upon returning from a vacation, the server will be immediately activated, if atleast N customers are waiting for service in the queue. Otherwise, if the queue length is less than N at the end of a vacation, then the server takes another vacation and the process continues until either the queue length reaches at least N at the end of a vacation or the server takes at most **J**-vacations consecutively. However, the server takes a maximum number of **J** vacations and at the end of the $\mathbf{J}^{th}$ vacation the server returns to the system even if the quorum is not reached.

(ii) (b) Randomized J-Vacation or < p, J > -Vacation Policy

A variant of **J**-vacation policy is, if all the customers are served in the queue exhaustively, the server then immediately takes a vacation. Upon returning from a vacation if the queue does not contain required number of customers, then the server either joins the system and remains idle with probability p or leaves for another vacation with probability q (p + q = 1). This pattern continues until the number of vacations reaches **J** or the queue length reaches a threshold value (N). At the end of the $\mathbf{J}^{th}$ vacation the server necessarily joins the service facility. Thus < 1, **J** > vacation policy is simply **J**-vacation.

The classical single and multiple vacation policies are the two extreme cases of J-vacation policy (respectively **J** = 1 and J tends to infinity).

(iii) Multiple Adapted Vacation (MAV) Policy

At the end of each busy period the server either remains idle with probability $(1 - \gamma_0)$ or takes a vacation with probability γ_0. If, at the end of the vacation no customer has arrived then the server, independently of everything else, takes a new vacation with probability γ_1, or remains idle and available to serve the first customer that arrives, with probability $1 - \gamma_1$. The process is repeated. That is the policy is determined by the sequence of probabilities $\{\gamma_k\}$, k = 0, 1, 2, . . . Multiple Adapted Vacation Policy reduces to the randomized J-vacation policy with the following selection of γ_k :

$\gamma_0 = 1$, $\gamma_k = p$ for k = 1 to J−1 and $\gamma_k = 0$ for $k \geq J$.

(iv) Bernoulli Schedule (Non-Exhaustive Service) Vacation Policy (BSV)

In all the vacation policies mentioned above, servers take vacations only when the system becomes empty. But in some situations, especially when the service is done in two or more phases, the maintenance of the system may be required at the completion of each service and in such cases, the service may be stopped for maintenance and overhauling, or continued, if there is no fault in the system. The overhauling may be utilized as a vacation time. The vacation scheme with Bernoulli service discipline originated by

Keilson and Servi (1986) is characterized by the feature that, at the completion of each service, the server may take a vacation with probability p or may continue to serve next unit if any, with probability (1 – p). The motivation for these types of models comes from computer networks and telecommunication systems where messages are processed in two stages by a single server.

1.1.10 Queueing Systems with Server's Breakdown (or) Service Interruptions

In queueing situations, servers may breakdown while providing service and the service of the customer being served is then interrupted and cannot resume service until the server is repaired. The period during which the system is in breakdown state is termed as breakdown period.

Delay Repair

There are situations in which, once the system breaks down, the repair does not start immediately and the repair takes place only after some random time. In this case the repair is said to be delayed. The time between the instant of breakdown and the initiation of repair work is called delayed repair time.

1.1.11 Queues with Two Phases of Heterogeneous Service

Queueing models, where the service discipline involves more than one service have been receiving a lot of attention recently. From the practical point of view, in production process, the machines produce certain items which may require two or more phases of service such as preliminary check followed by the usual processes to complete the processing of the raw materials. Various scenarios have been considered in the literature and they include, additional service channel, feedback service, optional re-service and phases of two or more heterogeneous service. The scenarios adopted in the present work are listed below :

Scenario 1

Each customer undergoes the essential service in the first phase (FES) and may choose optional service in the second phase or depart the system. In

certain cases the second stage may consist of multi optional service facilities and the customers can choose anyone of them after completing the FES.

Scenario 2

Each customer receives two phases of heterogeneous service one after the other and then departs from the system. If, the second phase contains multi optional service facilities then the customers who complete the first phase service will necessarily choose any one of the optional service facilities available in the second phase before departing the system.

Scenario 3 (Feedback Policy)

A customer who is not satisfied with his service, may join the queue as feedback customer with probability **f** to have another service or leave the system with probability **(1 – f)**. Many studies on feedback queues have conceded that dissatisfied customers may join the tail-end of the queue to repeat the service. The present work assumes that the dissatisfied customers reiterate the service (if required) immediately, finite or infinite number of times before leaving the system.

If the queueing system contains two phases of service in which the first phase consists of essential single service facility and second stage consists of multi-optional service facilities, then the customers after completing the first round of service may repeat any of the services from the second phase or repeat from first phase service followed by a second phase before leaving the system.

1.1.12 Queues with Server Setup or Startup Time

In some of the practical situations servers often require startup operations before starting each busy period. The server startup time thus corresponds to the preparatory work of the server before starting the service. For example, when a typical machine is to be turned on, a proper setup operation before the normal use may extend its function and the number of unsteady conditions of the machine tool may be reduced. Thus an additional amount of time of random length, in order to set the system into operation mode before actual service begins, is called setup period.

1.1.13 (m, N)-Policy (Bilevel Control Policy or Double Threshold Policy)

In N-policy queueing models with setup time, the server begins the setup work when the queue length reaches or exceeds **N**, where **N** controls the starting condition of service. And as soon as the setup operation is completed, the server begins to serve the customers exhaustively.

In (m, N) policy models, the server starts the setup operation earlier, when **m** (**m** ≤ **N**) customers accumulate in the system. At the end of the setup work if the queue length is greater than or equal to **N**, then the server begins to serve the customers exhaustively, or else the server remains dormant in the system waiting for the queue length to reach at least **N** to start a busy period. This policy is also referred as bilevel threshold policy with early setup. The (m, N) policy is more general than the usual N-policy. When **m = N** the (m, N) policy coincides with the single threshold N-policy. Moreover when **m = N =** 1, it corresponds to the classical queue.

1.2 REVIEW OF LITERATURE

1.2.1 Classical Queues

The pioneer investigator of queueing theory was the Denish Mathematician Erlang (1909) who published “The theory of probabilities and telephone conversations” and modeled the telephone traffic systems. Due to its wide applications in many areas, queueing theory has been one of the most active research topics in Operations Research and Management Science for the past several years. Kendall (1951, 1957) was the pioneer who viewed and developed queueing theory from the perspective of stochastic process. Some excellent books on classical queueing theory have been published ; they include Takacs (1962), Cooper (1981), Cohen (1982), Gross and Haris (1985), Satty (1983), Wolff (1989), Prabhu (1997), etc.

1.2.2 Vacation Queues

Server vacations are useful for the systems in which the servers want to utilize their idle time for different purposes. Vacation queues have attracted many attentions from numerous researchers since Levi and Yechalli (1975) introduced the two standard vacation policies (single and multiple). The

comprehensive survey on vacation queues can be found in Doshi (1986 and 1990), Takagi (1991), in the book by Tian and Zhang (2006) and in recent years Ke et al. (2010b). The amount of literature relating to queueing models with vacations is growing rapidly and the analysis of the queueing systems with vacations has been discussed through a considerable amount of work recently.

The early work on vacation queueing models focused on exhaustive service policy, where the server takes vacations if and only if the system becomes empty. However, multifarious vacation policies with non-exhaustive service have important application values in communication network and computer systems. To adopt diversified application background, some new vacation policies were presented. Keilson and Servi (1986) introduced the vacation policy between services, in which after completion of a service to a customer, the server may take a vacation with probability p or may continue to serve the next customer if any with probability (1 – p).

A wide class of policies for governing the vacation mechanism has been discussed in literature. Takagi (1991) first proposed the concept of variant vacation for the M/G/1 regular system which generalizes the single and multiple vacation policies. Zhang and Tian (2001) investigated a Geo/G/1 queue with modified vacation policy, where the server can take at most **J** vacations continuously. This modified vacation policy can be reduced to the classical single and multiple vacation policy by setting **J** to be one or infinite. Ke and Chu (2006) studied a batch arrival system under this modified vacation policy. The extensions and variations of these models can be referred to Ke (2007), Wang and Huang (2009). Ke et al. (2010) have examined an $M^X/G/1$ queueing system with a randomized vacation policy and at most **J** vacations and derived the distributions of important system characteristics. They have developed a cost model to determine the joint suitable parameters (p^*, J^*) at a minimum cost. Recently, Mytalas and Zazanis (2015) considered the Multiple Adapted Vacation (MAV) policy for a $M^X/G/1$ queueing system with disaster and repair. This policy is determined by the sequence of probabilities $\{\gamma_k\}$, k = 0, 1, 2, … which provides additional

flexibility and includes, the cases of other vacation policies in the same framework. The vacation policy considered by them (MAV) was first introduced by Takagi (1991) and termed by Tian and Zhang (2006).

1.2.3 Queueing Models with Server Breakdowns

The study of queueing systems, wherein the server (service channel) is subjected to unpredictable breakdowns is a most popular subject that has received a lot of attention for the past 60 years. Queueing models with service interruptions have proved to be a useful abstraction in situations, where breakdowns occur while providing service and the service of the customer being served cannot resume service until the server is fixed (repaired). Service interruptions in queueing systems have been investigated by several researchers. Gaver (1962) seems to be the first to analyse a queueing process with service interruptions in which, he studied the effect of interruptions on the distributions of important measures such as busy period queue length and waiting time for M/G/1 queueing model. The study of service interruptions mainly assumes that the interruptions occur only when the server is busy and the server in an interrupted state will not be affected by the process of further interruptions. Regarding the resumption of services, there are several possible scenarios to restoring an interrupted service. They include (i) starting service from the very beginning independent of the earlier service (ii) starting a service from where it got interrupted and (iii) denying service to the one whose service got interrupted.

Keilson (1962) treats the first two cases of interruptions. Yue and Tu (2001) studied M/G/1 queueing system with server interruptions in which the interrupted service is repeated from the beginning and investigated the completion period of jobs. Wang (2004) analysed an M/G/1 queue with two phases of service and obtained the transient and steady-state solution for interesting system measures for the case where the interrupted service is resumed when the server is fixed. In a survey paper, Krishnamoorthy et al. (2012) give a detailed description of research on queueing models with interruptions that occur due to many reasons and highlighted the rules for

resumptions. Ke et al. (2011) examines a repairable $M^X/G/1$ queueing system with a randomized J-vacation policy.

Many researchers have studied queueing systems with interruptions wherein one of the underlying assumptions is that as soon as the service channel fails, it instantaneously undergoes repairs. But it is a common phenomenon that as a result of a sudden breakdown, the system has to wait for repair to start. This waiting time is termed as delay time. Choudhury and Tadj (2009), studied M/G/1 queueing system with interruptions and delayed repair. Khalaf et al. (2011a, b) have examined the effect of delay time on the $M^X/G/1$ queueing system with Bernoulli schedule server vacations and random system breakdowns. Ke and Huang (2010) and (2012) also analysed an unreliable $M^X/G/1$ queue under randomized vacation policy with delayed repair.

1.2.4 Queueing Models with Two Phase Service

Queueing systems with two phases of service have been extensively studied in the literature over the past two decades. Madan (2000) introduced a single server Poisson arrival queue with two phases of service, in which, the first phase of service is provided to all arriving customers, whereas only a few customers opt for a second phase of service. The essential service time is assumed to be generally distributed, whereas, the second phase of optional service time is assumed to be exponentially distributed. Medhi (2002), generalized the results of Madan (2000), by assuming that the service times in both the phases of service are generally distributed. Further generalizations of this model were provided by Krishna Kumar et al. (2002) and Madan et al. (2002), who incorporated the features of no waiting capacity, an unreliable server and Bernoulli schedule server vacations respectively. Choudhury (2003), has studied a similar type of queueing model in greater detail. Al-Jararha and Madan (2003) and Wang (2004) have examined the transient behaviour of the M/G/1 queueing system with two phases of service for reliable and unreliable servers respectively. Thangaraj and Vanitha (2010), have considered a similar model with compulsory server vacations and

random breakdowns. Choudhury and Madan (2004), Choudhury and Paul (2006a) and Choudhury et al. (2007) have studied queueing systems with two phases of service under Bernoulli vacation schedule, in which after two successive phases of service the server may go for Bernoulli vacation. A repairable M/G/1 queue with general retrial times, Bernoulli Vacation Schedule, setup times and two phase service is investigated by Wang and Li (2008).

1.2.5 Queueing Models with Feedback Policy

In feedback queueing models, if the service of a job of a customer is unsuccessful then, the customer tries the job again and again until a successful service is completed. In computer and communication networks with cyclic queueing systems, messages are processed in two stages and a fraction of the messages may re-enter the system. Queueing systems with various feedback policies have been investigated by many authors. Most feedback queueing systems have the Bernoulli feedback policy. In this policy the customers who completed their services, feedback instantaneously to the tail of the queue with probability p or leave the system forever with probability (1 – p). In queueing systems with this policy, the memory less property of the number of feedbacks of a customer makes it easy to analyze the system. The concept of feedback was introduced by Takacs (1963) and since then many papers have appeared about this topic. He considered an M/G/1 Bernoulli feedback queue with single class customers and obtained the distributions of queue size and the total response time of a customer. Disney and Konig (1985) have given an overview of the literature concerning Bernoulli feedback studies.

Fewer results are known for feedback queueing systems in which the feedback policy is not Bernoulli. Baskett et al. (1975) obtained the product form of the joint queue size distribution for the M/M/1 queueing system with several types of customers and general feedback policy. Simon (1984) considered an M/G/1 priority queueing system with several types of customers and general bounded feedback policy and obtained a system of linear equations for the mean sojourn times for each class of customer type.

Thangaraj and Vanitha (2010), Choi and Tae-Sung (2003), Badamchi Zadeh and Shahkar (2008) and Choudhury and Paul (2005) derived the queue size distribution for M/G/1 queue with two phases of heterogeneous services and Bernoulli feedback system. Li and Wang (2006) have considered an M/G/1 retrial queue with a second multi optional service, an unreliable server and feedbacks. Shahkar and Badamchizadeh (2006) and Salehirad and Badamchizadeh (2009) have studied queueing systems with k phases of heterogeneous service and random feedback. Saravanarajan and Chandrasekaran (2014) analysed M^X/G/1 feedback queue with two-phase service, compulsory server vacation and random breakdowns.

In these papers mentioned above the customers who wish to obtain another round of service (feedback) have to go to the tail of the queue and await their turn for the next round of service to be provided. And moreover, it is assumed that the customers may repeat their service to any number of times before leaving the system. Recently Kalidass and Kasturi (2013) have studied a two-phase service M/G/1 queue with a finite number of immediate Bernoulli feedbacks. The model is motivated by the working of an ATM machine. After performing one transaction, say withdrawing money, the customer may want a mini statement of accounts or may want to change the pin number. For this purpose the customer immediately again starts his operations by giving the pin number which is the first phase of service. The customers do not have to go to the tail of the queue for another transaction. Maragatha Sundari and Srinivasan (2012) also have analysed a multi phase M/G/1 queue with finite number of immediate Bernoulli feedback where the server takes multiple vacations.

1.2.6 The N-Policy Models

Controllable queueing systems aim to find the rules for turning the server on and off that result in the lowest long-run cost. Work on controllable queueing systems could be divided into service control and arrival control. Service control queueing systems include the N-policy, the T-policy and the D-policy, which are introduced by Yadin and Naor (1963), Heyman (1977) and

Balachandran (1973) respectively. The development and applications on service control queueing systems are rich and varied. Among these service control policies, the investigations of N-policy queueing models gained a lot of attention recently. The development and applications on the optimal control and management operating policies of queueing systems are listed in the survey conducted by Tadj and Choudhury (2005).

The N-policy introduced by Yadin and Naor (1963) for M/M/1 queueing systems, turns the server on when N($\geq$ 1) or more customers are present, and turn the server off only when none is present in the system. The extensions of their basic model can be found in several research papers. For a reliable server the N-policy M/G/1 queueing system was first studied by Heyman (1968) and was developed by Bell (1971), Teghem (1987) and others. Analytic steady state solutions of the N policy M/E_k/1 queueing system were first obtained by Wang and Huang (1995). Later, Artalejo (1998) proposed a new stochastic decomposition property for the waiting time in the N policy M/G/1 queue. Wang and Ke (2000) analysed the N policy M/G/1 queueing system using supplementary variable technique.

The first study of batch arrival queue with N-policy was done by Lee and Srinivasan (1989). They presented a procedure to obtain the optimal stationary policy under a suitable cost structure. Later, Lee et al. (1994a) studied the M^X/G/1 queueing system under N-policy and proved that the system size is decomposed into two random variables, one is the system size of the classical M^X/G/1 queue and the other is the PGF of the number of customers in the system when the server is idle. For unreliable server, Wang (1995), Wang (1997), Wang et al. (1999), Wang and Ke (2002) analysed control policies for M/M/1, M/E_k/1, M/H_2/1 and M/G/1 models respectively.

In a few real time scenarios, the server often requires a start-up time before starting each busy period. Combination of N-policy and setup in a queueing system was first considered by Baker (1973) who proposed the N-policy M/M/1 queueing system with exponential start-up time.

Borthakur et al. (1987) extended Baker's result to the general start up time. The N-policy M/G/1 queueing system with start-up time was first studied by Minh (1988) and then investigated by many researchers such as Medhi and Templeton (1992), Takagi (1993) and Hur and Paik (1999) and so on. Batch arrival queueing system with general setup time and general service time was analysed by Hur and Ahn (2005).

1.2.7 N-Policy Queueing Models with Server's Vacations

The optimal control and management of vacation models have also received considerable attention in the literature. In N-policy queueing systems with server's vacations, the server becomes unavailable at the end of each busy period and resumes service instantaneously, when the queue length reaches a critical number N. This type of policy is called a threshold policy or N-policy with server vacations.

Kella (1989) studied the M/G/1 queue with N-policy and vacations. Batch arrival queues with thresholds and with/without multiple vacations were first studied by Lee and Srinivasan (1989). Later Lee (1991) developed a procedure to calculate the system size probabilities for a batch arrival queueing model with server vacations under control operational policy. Lee et al. then (1994b) studied the same queueing systems and found that the system size decomposes into two random variables : one is the system size of the classical $M^X/G/1$ queue and the other is the PGF of the conditional system size distributions due to vacations and threshold. They also derived the optimal threshold N^*, under a linear cost structure. While this paper concentrated on the development of system measures and waiting time transform solutions for multiple vacation models, Lee et al. (1995) derived the probability distributions and waiting time transform solutions for $M^X/G/1$ single vacation queueing model and obtained decomposition property. Ke et al. (2010a) investigated the threshold models (N-policy) with the modified vacation policies. Ke (2003) considered the control policy for $M^X/M/1$ queueing system in which the server is characterized by breakdowns and multiple vacations.

The single server queueing models with N policy and Bernoulli schedule vacations have many applications in flexible manufacturing systems, telecommunication systems, transport systems, etc. Tadj et al. (2006a) and (2006b) analysed N policy for batch service queueing system under Bernoulli schedule vacation. Most of the N policy queueing systems with BSV deal with two phase service facilities.

1.2.8 N-Policy Queueing Models with Two Phase Service Facilities

Two phase service rendered by a single server has been found to be useful to analyse many practical situations, arising in packet transmissions of communication networks, multimedia communications, central processors etc. Recently there have been considerable attention paid to the study of N-policy queueing models with two phases of service. The optimal control of the N-policy for a reliable server queueing system with two phases of service under Bernoulli vacation schedule was investigated by Choudhury and Madan (2005). They have highlighted the applications of two stage batch arrival queues in digital communication system and production system, where the concept of Bernoulli Schedule along with a vacation time is introduced under N-policy. Jain and Agarwal (2010) have made an attempt to generalize a batch arrival two stage service system with a modified Bernoulli vacation schedule by including state dependent rate and ℓ-stages of service. Choudhury and Paul (2006) analysed N-policy for a batch arrival queueing system where all customers receive batch mode service in the first phase, followed by individual service in the second phase. Choudhury et al. (2009) examined the N-policy for an unreliable server with delaying repair and two phases of service. Choudhury and Tadj (2011) studied the optimal control of an unreliable batch arrival Bernoulli vacation queueing system with two phases of service. Choudhury et al. (2011) dealt with M/G/1 queueing system with two phase of service and Bernoulli vacation schedule for an unreliable server, which consists of breakdown period and delay repair, under N policy and random setup time. These works unify several cases of related batch arrival queueing systems. Recently Afthab et al. (2014a) studied N-policy for

repairable bulk arrival queueing model with setup, second multi optional service facility under J-vacation policy.

1.2.9 Queueing Models with Double Threshold Policies

Many authors tend to combine various features in a single threshold queueing model. But still, there are some more kinds of classification of research exist on the optimal control of queues under N-policy. Lee and Park (1997) considered an Poisson arrival M/G/1 queueing model with an early setup. The early set up N-policy is termed as (m, N)-policy which is more general than the single threshold (N, N)-policy. The decomposition property of vacation queues is used to derive the distribution of the system size. A cost model is developed and a procedure, to find the optimal values of m and N that minimize the average cost of the system, is presented. It is illustrated that the double threshold policy is more beneficial than the conventional single threshold N-policy when the setup cost is excessively high compared with the inventory holding costs. Lee et al. (1998) extended the bilevel control policy to a batch arrival queueing system $M^X/M/1$ and obtained queue length and waiting time distribution for the model. Later Lee et al. (2003) have analyzed a batch arrival system $M^X/G/1$ under bilevel threshold with early setup and with/without server vacation and used the decomposition property of vacation queues to derive directly the system size PGF and mean queue lengths of the system. These models focused on reliable server. Using the stochastic decomposition approach, Ke (2004) has studied the probability generating function of the number of customers for two different models of (m, N) policy $M^X/G/1$ queueing system with an unreliable server.

Afthab and Fijy (2014) and Afthab et al. (2014b) analysed the (m, N) policy for a batch arrival two phase service queueing systems under J-vacation for unreliable and Bernoulli schedule single vacation for reliable cases respectively.

1.3 THESIS ORGANIZATION

The contribution of the author is schematically represented in Figure 1.

Figure 1. Schematic Representation of various Models Developed

CH - Chapter

BSV – Bernoulli Schedule Vacation

Lee et al. (2003) analyzed the queue length of $M^X/G/1$ systems under bilevel threshold control and early setup with or without server vacations. Later Ke (2004) obtained the Probability Generating Functions (PGF) of the number of customers in the system for an unreliable case using the stochastic decomposition approach. Recently, Julia Rose Mary et.al. (2013) investigated the (m, N)-policy for $M^X/G/1$ queues where the service discipline involves more than one service. Nevertheless, these authors derived the steady state results by considering the classical single and multiple vacations as two different cases. Ke and Huang (2010) studied the N-policy for M/G/1 queue with at most J-vacations and developed the joint optimum thresholds that minimize the cost function. In Chapters II, III and V the author examines the (m, N)-policy for more general $M^X/G/1$ queues under J-vacation policy, so that the corresponding results for single and multiple vacation models can be deduced as special cases.

In Chapter II, it is assumed that a cycle starts whenever the system empties and the server is deactivated and leaves the system immediately for vacation of random length. The server operates (m, N)-policy with at most J-multiple vacations. During busy period, the server provides single FES to all arriving customers in the first phase and C-optional heterogeneous services in the second phase. As soon as the FES is completed, each customer may either opt for a certain (i^{th}) service in the second phase (with probability r_i, $0 \le r_i \le 1$) or may leave the system (with probability $(1-\sum_{i=1}^{C} r_i)$). The server is subject to random breakdowns during busy period and the server's lifetime follows the exponential distribution in the first phase with parameter $\mathbf{a}$. In the second phase, the server fails at an exponential rate $\mathbf{a_i}$ ($1 \le i \le C$). Whenever breakdowns occur, it is assumed that the server is sent for repair immediately. The customer, just being served just before breakdown, waits for the server to return from the repair facility to complete the remaining service. As soon as the server is fixed, the service resumes for the waiting customer. The vacation period, buildup period, setup period, dormant period, busy period and breakdown period constitute a cycle for the model.

The model analyzed in chapter III differs from the model of chapter II, in service facility and the customers' behaviour during breakdown period. The author scrutinizes the (m, N)-policy by considering three possible scenarios to restoring an interrupted service together in a single $M^X/G/1$ queueing system. They include, if the server fails, then the customer in service may join the head of the queue and opt for a new service (with probability q_1) or may leave the system without completing the service (with probability q_2) or else stay in the service facility (with probability $q_3 = 1 - (q_1 + q_2)$) to complete the remaining service. The repair times of the server follow distinct general distributions of finite moments. The hypothesis regarding (m, N)-policy with at most J-vacations is as in Chapter II.

In Chapters II and III, it is assumed that the server takes multiple vacations only when the system becomes empty. In some situations, especially when the services are executed in two or more phases, maintenance of the system is required at the completion of each service. In Chapter IV, the (m, N)-policy is inspected for an $M^X/G/1$ queueing system with the provisions of various types of generally distributed long and short vacations. It is assumed that at the end of busy period, the server becomes free and takes a single vacation (VI) (with probability p′) or stays idle in the system (with complementary probability). During busy period, after completing a service to a customer, the server may choose (i^{th}) type of vacation VB_i (with probability p_i, $(1 \le i \le M)$) or continue to serve the next customer (with probability $(1-\sum_{i=1}^{M} p_i)$). The other assumptions regarding (m, N)-policy, busy period, breakdown period and repair facility are as in Chapter II with C = 1 and J = 1.

In Chapter V, the (m, N)-policy is studied for an $M^X/G/1$ queue, in which the unsatisfied customers, after completing the regular service, may demand a re-service with probability **f** or leave the system with probability **(1–f)** is admissible. It is also assumed that the server may take Bernoulli Scheduled Vacations in between services. Two different vacation distributions, $VB_1(t)$ and

$VB_2(t)$ are considered. The server, after completing each service, either takes a vacation of type VB_1 (with probability p_{b_1}) or VB_2 (with probability p_{b_2}), depending on whether the customer leaves the system or opts for feedback. Otherwise, the server continues to serve with probability $(1 - p_{b_1})$ in case 1 or $(1 - p_{b_2})$ in case 2.

During idle period, a more general J-vacation policy controlled by the parameter **p** is considered. That is, when the system becomes empty, the server immediately takes a vacation. Upon returning from the vacation, the server is activated if there are at least **m** customers waiting in the queue. On the other hand, if the number of customers waiting in the queue is less than **m**, the server either joins the system with probability (1 – p) or leaves for another vacation with probability **p**. The pattern continues until the number of vacations reaches **J**. At the end of the $\mathbf{J^{th}}$ vacation if the queue length is still less than **m**, the server joins the system and stays idle until the queue length reaches at least **m** to start the setup work. The process after setup period regarding (m, N)-policy is as in Chapter II.

Most of the queueing models with service interruptions assume that interrupted server is sent for repair immediately. On the contrary, this is possible only when the repair facility is available on a permanent basis in the service station. In chapter VI, the author analyses a repairable batch arrival queueing model with two phases of heterogeneous service where the repair work is delayed. The life time of the server follows exponential distributions with parameters a_i, i = 1, 2 in phase 1 and 2 respectively. The customer whose service has been interrupted will repeat the service from the beginning. After the first round of service, the customer who wants to repeat the service will immediately reiterate service with feedback probability **f**. Bernoulli Schedule Vacation policy controlled by the parameter **p** is considered during busy period and MAV policy is adopted during idle period. A setup time is also introduced at the beginning of every cycle.

Chapters VII and VIII examine $M^X/G/1$ queueing system with immediate feedbacks and multi second optional service facilities. The server operates single service in the first phase and different kinds of heterogeneous services in the second phase. A customer is said to complete the first round service if he undergoes the first phase service and any one of the second phase services. After having completed the first round service, the customer is permitted to repeat services from the second multi-optional services which may be different from the one chosen earlier. Kalidass and Kasturi (2013) analyzed a reliable Poisson arrival M/G/1 queue with two phases of heterogeneous service and a finite number of immediate Bernoulli feedbacks before leaving the system. In Chapter VII of the present work, the author studies the $M^X/G/1$ queueing system where customers can feedback finitely many times, the server undergoes unpredictable breakdowns and takes optional vacations between services. When the server fails, the customer in service may resume or repeat service. These two cases are considered separately in sections 7.1 and 7.2. The corresponding infinite feedback cases are investigated in sections 8.1 and 8.2.

The random variables service times (iid), setup times, vacation times (iid) and repair times (iid) considered in the present work follow general distributions with finite moments and are independent of each other.

1.4 OBJECTIVES AND SCOPE OF RESEARCH WORK

- The steady state system size equations governing the mathematical models are derived using the supplementary variable technique. The probability generating functions of the system size at arbitrary epoch are obtained.
- The performance indices namely the probability that the server is in different states and the mean queue length when the system is in different states are also established.
- A procedure to find the optimal threshold values under a linear cost structure is developed for the models from Chapters II to V.

- The effects of various system parameters on the system performance are numerically analyzed and graphically depicted to interpret the data more comprehensively.
- The stochastic decomposition property has been verified.
- The systems studied are among the most general queuing systems and include many previous works as special cases.

1.5 METHODOLOGY

1.5.1 Transient and Steady-State Solution :

Let N(t) denote the number of customers in the system at time t and its probability distribution be denoted by $P_n(t) = Pr\ (N(t) = n / N(0) = \bullet)$. For a complete description of the queueing process, the transient or time-dependent solutions are necessary. But it is often difficult to obtain such solutions. Further in many practical situations, we need to know the behaviour in steady-state, i.e., when the system reaches an equilibrium state, after being in operation for a pretty long time. The time-dependent solutions are called transient solutions. And the solutions obtained as $t \to \infty$ are called steady-state solutions. Throughout the present work, the queueing systems are analysed at steady state, assuming that the system size probabilities are independent of time t, as $t \to \infty$.

1.5.2 Markovian and Non-Markovian Queueing Models

The exponential distribution is the only continuous distribution which satisfies the Markov property. The queueing models in which all the continuous time random variables involved, follow exponential distribution are said to be Markovian queueing models. In non-markovian queueing models atleast one random variable follows distribution other than exponential.

The first step in analyzing a queueing system is to set it up as a Markov process. There are a number of techniques or approaches that are used for this purpose. In most practical queueing systems, supplementary variables are usually needed to achieve this. The alternative to that is an embedded Markov chain method.

1.5.3 Supplementary Variable Technique (SVT) (Alfa and Srinivasa Rao, 2000)

In queueing literature, there exists two kinds of supplementary variables, in general. They are elapsed time and the remaining time of random variables. In both the cases, the approaches of deriving the queueing characteristics are different. The main reason for adding supplementary variables to a stochastic process variable is to make the system Markovian. Cox (1955) considered elapsed service time as supplementary variable to study the M/G/1 queueing system. The use of the supplementary variable technique (SVT) in queueing was introduced by Kosten (1973). Later, the technique became popular for most stochastic models. The technique of remaining time as the supplementary variable is simple and elegant. The supplementary variable technique analysis for the queueing problems by considering the remaining time as supplementary variable involves, probability density function and partial differential equations.

The SVT applied to the M/G/1 model :

Consider a single-server queue wherein the inter-arrival times is exponential with parameter λ and service times (S) are independent and identically distributed (i.i.d.) random variables having density functions s(x) ($x \geq 0$) and distribution function $S(x) = \int_0^x s(y)\,dy$.

Let N(t) denote the number of units present in the system at time t and let $S^o(t)$ denote the remaining service of the unit in service at time t. Then the state of the system at time t defined by $\{N(t), S^o(t); t \geq 0\}$, is Markovian in continuous time. Furthermore, let $P_0(t)$ be defined as the probability that at time t, the system is empty (i.e., idle), and let $P_n(x, t)\,dt$ ($n \geq 1$) be defined as the joint probability that at time t, the number of units in the system is n and the remaining service time of the unit in service lies in the interval (x, x + dt) ; that is,

$P_0(t)$ = Prob $(N(t) = 0)$,

$P_n(x, t)\,dt$ = Prob $\{N(t) = n, x < S^o(t) \leq x + dt\}$, $x \geq 0$, $n \geq 1$

Then relating the states of the system at time t and t + dt, we obtain the following Chapman-Kolmogorov forward equations :

$$P_0(t + \Delta t) = P_0(t)\,(1 - \lambda\,\Delta t) + P_1(0, t)\,\Delta t$$

$$\Rightarrow \frac{d}{dt} P_0'(t) = -\lambda\,P_0(t) + P_1(0, t)$$

$$P_n(x - \Delta t, t + \Delta t) = P_n(x, t)\,(1 - \lambda\Delta t) + \lambda\,(1 - \delta_{1,n})\,\Delta t\,P_{n-1}(x, t) + P_{n+1}(0, t)\,s(x)\,\Delta t$$

These equations respectively imply :

$$\frac{d}{dt} P_0(t) = -\lambda\,P_0(t) + P_1(0, t) \tag{1.1}$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) P_n(x, t) = -\lambda\,P_n(x, t) + (1 - \delta_{1,n})\,\lambda\,P_{n-1}(x, t) + P_{n+1}(0, t)\,s(x), \quad n \geq 1 \tag{1.2}$$

where $\delta_{1,n}$ is the Kronecker delta function.

Assuming that at steady state, as $t \to \infty$ the system size probabilities are independent of time t, the steady state equations obtained from (1.1) and (1.2) are given by

$$\lambda\,P_0 = P_1(0)$$

$$-\frac{d}{dx} P_n(x) = -\lambda\,P_n(x) + (1 - \delta_{1,n})\,\lambda\,P_{n-1}(x) + P_{n+1}(0)\,s(x), \quad n \geq 1$$

The SVT will give results in a unified way to obtain the distributions of units in the system of departure, pre-arrival and arbitrary epochs.

1.6 PRELIMINARIES

There are several methods of solving the difference-equations and presenting the probability distributions of the system size to calculate important performance measures of the queueing models. Very often the transforms such as the probability generating function $P(z, t) = \sum_{n=0}^{\infty} P_n(t)\,z^n$

and the Laplace transform of the function $L(P_n(t)) = \int_0^\infty e^{-\theta t} P_n(t)\, dt$ are used to solve the difference equation.

1.6.1 Probability Generating Function (PGF) of the Random Variable X :

In probability theory, the probability generating function of a discrete random variable is a power series representation of the probability mass function of the random variable. The probability generating functions are often employed for their succinct description of the sequence of probabilities Pr (X = i), and to make available the well-developed theory of power series with non-negative coefficients.

Definition

Suppose that X is a random variable that assumes non-negative integral values 0, 1, 2, 3, … and that Pr {X = k} = p_k, k = 0, 1, 2, … with $\sum_{k=0}^{\infty} p_k = 1$, then the corresponding generating function $P(z) = \sum_{k=0}^{\infty} p_k z^k$ of the sequence of probabilities { p_k } is known as the probability generating function (PGF) of the random variable X. Also called as the z-transform of the random variable X.

We have p(1) = 1 ; the series P(z) converges for atleast $-1 \le z \le 1$ and is infinitely differentiable. The function P(z) is defined by { p_k } and in turn defines { p_k } uniquely, i.e., a PGF determines a distribution uniquely.

The k^{th} factorial moment of X is given by

$$E(X\,(X-1), \ldots, (X-k+1)) = \left[\frac{d^k}{dz^k} P(z)\right]_{z=1}, \text{ for } k = 1, 2, \ldots$$

1.6.2 Laplace Stieltjes Transform (LST)

Laplace transforms serve as very powerful tools in many situations and provide an effective means for the solution of many problems arising in queueing theory. The transforms are very effective for solving linear differential equations and reduce a linear differential equation to an algebraic

equation. In the study of some probability distributions, this method could be used to find the Laplace transform of a probability distribution rather than the distribution itself.

The Laplace-Stieltjes transform of a non negative random variable X with distribution function F(.), is defined as $F^*(\theta) = \int_{x=0}^{\infty} e^{-\theta X} dF(x)$, $\theta \geq 0$. When the random variable X has a density f(.), then the transform simplifies to

$F^*(\theta) = \int_{x=0}^{\infty} e^{-\theta X} f(x)\, dx$, $\theta \geq 0$. Note that $|F^*(\theta)| \leq 1$ for all $\theta \geq 0$. Further F(0) = 1.

The k^{th} factorial moment of X is given by $\left[\frac{d^k}{d\theta^k}(F^*(\theta))\right]_{\theta=0} = (-1)^k E(X^k)$.

The property of Laplace transform of derivatives namely

$L(f'(t)) = \theta\, L(f(t)) - f(0)$ is used in all the chapters. Whenever more than one continuous random variables say (X and Y) exist, the LST corresponding to the random variable Y may be denoted by :

$$F^{*}_{1}(\theta_1) = \int_{0}^{\infty} e^{-\theta_1 y} dF(y), \quad \theta_1 \geq 0.$$

For numerical study, the algorithms were implemented in computer programmes written in C^{++}, using objective oriented tools. The graphical representations were constructed by means of softwares origin 16-bit 32 and matlab 2013 for 3-D figures and Microsoft Excel 2013 for 2-D line graphs.