Review of Literature

1.2. LITERATURE SURVEY

N- Policy Queueing models

Past works regarding queues may be divided into two categories: (i) the case of controlling service and (ii) the case of controlling the arrivals. In the case of controlling the service, the N- Policy applied to queueing system was originally considered by Yadin and Naor (1963). Since then, N- Policy queues have been widely used to provide stochastic modeling of many problems arising in production inventory system, in telephone switching system and in quality control problems. A classified bibliography with an extensive survey can be found in Grabill *et al.*, (1977) and in Teghem (1986). Applications can be found in Heymen (1977), Lee and Srinivasan (1987).

Analytical study of the N-Policy $M/E_k/1$ queueing system was first analysed by Wang and Huang (1995). The N policy M/G/1 queueing system was first studied by Heyman (1968) and was developed by several researchers such as Bell (1971), Teghem(1987), Wang and Ke (2000) and others. Ke and Wang (2002) presented a recursive method using supplementary variable technique to develop analytic closed form solution for the N policy GI/M/1 queueing system with finite capacity.

The analysis of N policy queueing systems mentioned above discusses the cases where requests for service arrive in units, (i.e.,) one at a time. But in many real queueing systems, however, requests for service usually arrive in batches. For example, in manufacturing systems of the job-shop type, each job order often requires the manufacture of more than one unit; and in computer communication systems, where messages that are to be transmitted could consist of a random number of packets. The first study of $M^x/G/1$, N- policy was done by Lee and Srinivasan (1989). They developed control policies for queueing systems with batch arrivals and presented a procedure to find the optimal

stationary operating policy under a linear cost structure. Later Lee *et al.*,(1994a) studied the batch arrival system 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 Probability Generating Function (PGF) of the customers in the system when the server is idle. This paper, also concentrated on the development of the queue waitingtime transforms and the probabilistic interpretations of the terms involved in it.

N policy queueing models with setup time

In queueing situations, the setup operations correspond to the preparatory work of the server before starting the service. In some actual situations, the server often requires a setup time before starting his each service period. Examining queueing systems which combine the N policy with setup time, Baker (1973) first proposed the N policy M/M1 queueing system with an exponential startup time. Later, Borthakur *et al.*, (1987) extended Baker's results to the general startup time. The N policy M/G/1 queueing system with setup time was first studied by Minh (1988) and was investigated by several researchers such as Medhi and Templeton (1992) , Takagi (1993) and so on. Later Hur and Paik (1999) examined the operating characteristics of the N policy M/G/1 queueing system with server startup and explained how the systems optimal policy and cost structure behave for various arrival rates. Batch arrival queue with general set up time and general service time was analysed by Hur and Ahn (2005).

N policy with server Vacations

In recent years, queues due to server vacations have emerged as an important area of queueing theory and have been studied extensively and successfully in various applications such as production inventory systems, communication systems, computer net works etc., A wide class of policies for governing the vacation mechanism have been discussed in the literature.

Comprehensive surveys in this topic can be found in Doshi (1986), Takagi (1991) and Fuhrmann and Cooper (1985). One of the fundamental objectives of vacation models is to investigate the optimal control of a system in which a cost structure is assumed. Zhang *et al* .,(1997) and Kella (1986) studied the M/G/1 queue with N policy and vacation and derived the system size distribution and optimal threshold policies under an average cost criteria.

At present, most of the studies are developed for batch arrival vacation models under different vacation policies. Lee and Srinivasan (1989) first provided detailed discussions concerning N-policy $M^X/G/1$ queueing system with vacations. Later Lee *et al*., (1994b &1995) ,analysed in detail the batch arrival $M^X/G/1$ queueing system under N-policy with a single vacation and repeated vacations. Their results significantly confirmed stochastic decomposition property given by Fuhrmann and Cooper (1985). A number of researchers, including Ke (2003), Reddy *et al*., (1998) have also considered batch arrival queueing system under N policy with various vacation policies. Recently Yu *et al*., (2010) developed an approach for the queue length distribution of the N-policy to batch arrival queueing system with single vacation and setup time and obtained the additional queue length distribution using Leibnitz formulae for derivatives. Various authors Chae and Lee (1995), Arumuganathan and Jeyakumar (2005) have analysed queueing problems of server vacations with several combinations

N policy and breakdowns

In earlier studies of queues most of the authors focused on reliable servers. But in many queueing systems occurring in real life situations, particularly when the service facility consists of mechanically operated device, the service gets interrupted due to the occurrence of occasional random failures of the service device. Due to these interruptions in service, various parameters of queueing systems are affected. As a result of breakdowns, service facility becomes inoperative and the units demanding service can be served only when it is restored to operative state. Wang (1995) first proposed Markovian queueing

system under the N – policy and server break downs. Wang (1997) and Wang *et al.*, (1999) extended the model proposed by Wang (1995) to $M/E_k/1$ and $M/H_2/1$ queue system respectively. They developed analytic closed form solutions and provided a sensitivity analysis.

Wang (2003) later studied a single removable and non-reliable server in the N-policy M/M/1 queueing system and derived the steady state results and developed cost model to determine the optimal operating N-policy at minimum cost. The research works mentioned above do not investigate cases involving both server breakdowns and vacations together. Ke (2003a) considered the control policy for batch arrival $M^X/M/1$ queueing system under N- policy in which the server is characterized by breakdowns and multiple vacations. J.C.Ke and Pearn (2003) discussed the optimal management policy of heterogenous arrival M/M/1 queue with server's breakdowns and vacations. Later J.C.Ke(2004b) studied the operating system characteristics of $M^X/G/1$ queueing system under vacation policies with startup and break down times in which the server may breakdown according to a Poisson process while working and his repair time has general distribution.

Bi-level (m,N) policy queueing models

The bi-level control policy was first introduced by Lee and Park (1997) for an M/G/1 queueing system. They have used the decomposition property of a vacation queue to derive the distribution of the number of units in the system and developed a procedure to find the optimal values of (m, N) that minimize a linear cost. They have shown that the double-threshold policy is more beneficial than the conventional single threshold policy. Lee *et al.*, (1998) analysed $M^X/M/1$ queue with bi-level control and obtained the queue length and waiting time distribution.

Later Lee *et al.*, (2003) extended Lee and Park (1997) model to a non-Markovian batch arrival system with / without server's vacation. The works mentioned above focused on reliable servers with servers early setup. Comparable work on batch arrival queues where the server operates N-policy with vacations, startup and breakdowns is rarely found in literature. J.C. Ke (2004a) considered a bi-level control of batch arrival $M^x/G/1$ queueing systems in which the system is unreliable and is characterized by an early setup and multiple vacations. All these papers use the well known decomposition property of vacation queues directly to derive the PGF of the stationary queue length distribution.

Two phase service queueing models

The class of queueing systems, where the service discipline involving more than one service has been receiving a lot of attention recently. There have been several contributions, considering queueing systems in which the service is provided in two phases by a single server. Madan (2000) introduced the concept of second optional service to an M/G/1 type queueing model where the server provides the First Essential Service (FES) to all the arriving customers. As soon as the FES of a customer is completed, he may leave the system with probability (1-θ) or may immediately opt for a Second Optional Service (SOS) with probability θ. Later Medhi (2002) and Choudhury (2003) generalized the results of Madan by deriving the steady state queue size distribution at the stationary point of time for general SOS service time. Choudhury (2003) has also obtained the Laplace - Stieltjes Transform (LST) of the waiting time distribution and some important performance measures which lead to remarkable simplification when solving similar types of queueing models. Later Choudhury and Paul (2006) considered an $M^x/G/1$ queueing system with a second optional service channel under N-policy as a generalization of results obtained by Lee *et al.*, (1994a). Recently there have been several contributions for example Krishna kumar

et al.,(2002) and Atencia and Moreno (2006) considering retrial queueing systems in which the server may provide a second phase of service.

The papers mentioned above are characterized by a common feature namely, the second phase of service is provided only to a portion of the original incoming customers. A single server queue where the server provides two phases of heterogeneous service one after the other to all the arriving customers also attracts many researchers because of the application. Such systems have been discussed by Krishna and Lee (1990) and Doshi (1991). Such types of models with or with out Bernoulli schedule Vacation can be seen in Choudhury and Madan (2004), Madan and Choudhury (2005), etc.,

Bernoulli schedule vacation was proposed by Keilson and Servi (1986). It is characterized by the feature that if the queue is empty after service completion, then the sever becomes inactive and begins a vacation period. If the queue is not empty then another service begins with specified probability p or a vacation period begins with probability (1-p). The queueing model with vacation under Bernoulli schedule has received attention from many authors due to the applications in many real life situations. Various aspects of Bernoulli vacation models for single server queueing systems including M/G/1 queue have been studied by Ramaswamy and Servi (1993), Madan *et al.*, (2005). Choudhury and Madan (2004) allow the input process to be bulk instead of orderly. Such types of models can also be seen in Madan (2001), Choi and Kim (2003) Choudhury and Madan (2005). Moreover in all the above mentioned papers, one can find important applications of the two phase service models to computer communication, production, manufacturing systems, control processes and multimedia communications.

Madan *et al.*, (2005) considered an extended $M^X/G/1$ queue with two types of general heterogeneous service and modified Bernoulli schedule vacation in which each customer, just before a service starts, has the option to choose one

of the two kinds of services with probability θ_i i=1,2. As soon as the service of the customer is completed the customer leaves the system and the server may take Bernoulli vacation with probability p. Such a model may find applications in many day to day real life queueing situations encountered in post offices, banks or computer centers and so on.

Working vacation queueing models

Classical server vacation represents a period of time of not attending the queue or doing other non queue jobs (i.e.,) servers stop primary service during vacation. Servi and Finn (2002) extended the classical vacation queueing system to the working vacation system where the server serves customers at a lower service rate instead of completely stopping service. Obviously the working vacation queue is a generalization of the classical vacation queueing system.

Servi and Finn (2002) first studied the M/M/1 queue with working vacations. They obtained the stationary queue length and waiting time using the quasi – birth and death model. Liu *et al.*, (2007) used the matrix geometric solution to show the stochastic decomposition in performance measures. Tian *et al.*, (2008b), Li and Tian (2007a), Xu *et al.*, (2009b) considered M/M/1 queue with different working vacation policies. Baba (2005), Li *et al.*, (2008), Banik *et al.*, (2007) studied the GI/M/1 type working vacation queues. Using different methods, Kim *et al.*, (2003), Wu and Takagi (2006), Li *et al.*, (2009) discussed several M/G/1 type working vacation queues. Due to the wide applications in the performance analysis of communication and computer systems, the discrete–time queues with various working vacation policies have been considered in Tian *et al.*, (2008a) Li and Tian (2007b), Li *et al.*, (2007) Li and Tian (2008) and in Yi *et al.*, (2007). Recently Tian *et al.*, (2009) provided a survey of the results of working vacation queues and demonstrated that the matrix analytic methods developed by Neuts (1981,1995 & 1999) are powerful tools for analyzing the working vacation queues and can be considered as a unified approach to this class of queueing models.

However, the bulk input working vacation queues are not involved in the literature mentioned above. Hence Xu *et al.*, (2009a) studied the results of Liu *et al.*, (2007) to bulk input model $M^X/M/1$ MWV. They have formulated the model as two dimensional Markovian chain and obtained the PGF of the stationary queue length and its stochastic decomposition result using the matrix analysis method.

Madhu and Agarwal (2007) motivated by the work of Servi and Finn (2002) extended the results to $M/E_k/1$ queue with server breakdown and working vacations. Sensitivity analysis is also carried out in order to obtain the effect of various system parameters on system performance characteristics for the model.

Restricted admissibility with/without random setup time

The vast literature in queueing theory abounds in results of considerable theoretical elegance and significances. Nevertheless, it is felt by users of various results in different situations that the theory has to a large extent, still remained behind the control of service process or arrival process.

However, Madan and Choudhury (2004) studied a model which deals with the aspects concerning the control of the arrival process as well as the service process itself. In existing queueing literature one finds some papers such as Rue and Roshen (1981), Stidham (1985), Neuts (1984) and Huang and Mc Donald (1998) which deal with control policies of arrivals into queue and queueing network. But these papers neither deal with batch arrival nor deal with server vacations. Madan and Abu (2003) analysed the steady state behavior of a single server (Bernoulli schedule) vacation queue, in which arriving batch may or may not be allowed to join the system at all times and obtained the PGF of the number of customers in the system. Later Madan and Choudhury (2006) considered a two stage heterogeneous service batch arrival queue with a vacation and random setup time.

1.3 THESIS ORGANIZATION

As far as the bi-level threshold queueing models with reliable single server are concerned, the batch arrival queueing system studied by Lee *et al.*, (2003) with double thresholds **m** and **N** along with server vacations is the most general one and the work includes many previous works as special cases. But this paper is focused only on reliable or perfect server. Later J.C.Ke (2004) had analysed the bi-level control batch arrival queueing model with an unreliable server and derived the probability generating function of the number of customers present in the system. In most of the bi-level threshold queueing models including Lee *et al.*, (2003) and J.C.Ke (2004), it is assumed that the customers' arrival is homogeneous and the arrival rate does not depend on the status of the system. Moreover, the decomposition property of vacation queues is used to derive the queue length PGF of the models. And also in their models, only one type of service is provided by the server present in the system. In chapters II to V of the present work, more general bi-level control policy queueing models with server startup, server vacations and server breakdown are analysed.

In chapter II , single server batch arrival Markovian queueing systems along with server breakdowns and vacations are analysed under bi-level threshold policy for service and restricted admissibility policy for arrivals. Unlike the usual batch arrival queueing system, it is assumed that not all batches are allowed to join the system at all times. Regarding admissibility of batches, different policies are assumed. The probability that an arriving batch is allowed to join the system varies according to the system state. $r_i(i=1,2,3)$ respectively denotes the probability with which an arriving batch joins the system during the idle, busy and breakdown periods of the server.

A cycle starts as soon as the system becomes empty and the server leaves the system for a vacation of random length V (vacation period). After returning from the vacation, if the server finds less than **m** customers waiting in

the system then he remains idle (build up period) in the system until the queue length reaches at least **m**, to start the setup operation, that is, the server takes single vacation. On the other hand, if the server comes back from the vacation and finds **m or more customers**, then he starts preparatory work immediately, called startup or setup operation which takes a random length. At the end of the set up, if the queue length is greater than or equal to **N,** then the server begins to serve the customers exhaustively, otherwise he remains dormant in the system waiting for the queue length to reach or exceed **N** before he starts a busy period.

The server will serve only one unit at a time and the server is subject to break down at any time when working. The break downs are assumed to follow Poisson process of rate α. Whenever the server fails, he is immediately sent for repair. A customer under service during the failure waits in the queue until the server, returns from the repair facility and then completes the service. In this model, the first threshold **m** is used to control the starting condition of a setup time, while the second threshold **N** is used to control the starting condition of service. The vacation period, buildup period, setup period, dormant period, busy period and break down period constitute a cycle. The single vacation model is analysed in section 1 of chapter II and the service time, vacation time, setup time and repair time are random variables which follow the exponential distribution and are independent of each other and also independent of the arrivals. In section 2 of Chapter II, the same model is considered under repeated vacation, that is, if the server returns from vacation and finds less than **m** customers waiting in the system, he takes another vacation and repeats his vacation until he finds at least **m** customers in the system. In this case build up period is zero.

Chapter III deals with non –Markovian single server batch arrival queueing systems in which the server provides two types of heterogeneous service facility to all the arriving customers. It is assumed that all the arriving customers require the First Essential Service (FES) and only some of them demand the Second

Optional Service (SOS) with probability r. Each cycle is made up of buildup period, setup period, dormant period, and busy period. The vacation time, setup time and service time of both phases are assumed to be independent of each other and follow general distributions with finite moments. Remaining vacation time, setup time and service times are introduced as supplementary variables to analyze the model using supplementary variable technique. Section 1 of chapter III deals with single vacation policy and multiple vacation policy is considered in section 2 of this chapter.

In chapter IV, a more general batch arrival queueing system in which the server provides c - types of general heterogeneous service and the arriving customers have the option of choosing any one of the c- kinds of service is analysed. The customers are served on FCFS basis and each customer chooses the i^{th} type of service with probability r_i where $\sum r_i = 1$. It is assumed that the i^{th} type service follows an arbitrary distribution $S_i(t)$, $(1 \leq i \leq c)$ with finite moments. The customers entering the queue undergo only one type of service according to their choice and leave the system when their service is complete. The server is subject to breakdowns at any time while doing service of any kind. The break downs occur according to Poisson process with rates α_i (i= 1 to c). The rate indexed by i is to denote that the breakdown occurs at the ith type of service. Immediately the breakdown server is sent to the repair facility and the repair time follows a general distribution (heterogeneous) with finite moments. If the server breaks down during the service, the customer who is just being served waits in the service channel for the server to return back from the repair facility to complete his remaining service. It is assumed that the service time for the customer is cumulative and after repair, the server returns back as good as new. The other assumptions of the model are similar to that of the model of section 3.1. Only single vacation policy is considered.

Generally vacation models deal with the case of exhaustive policies, which means that the server leaves for a vacation of random length only when the system becomes empty. In contrast to this, Keilson and Servi (1986) have introduced a class of vacations with Bernoulli schedule. The Bernoulli schedule vacation models vary from the rest and they are general in nature up to some degree. The model considered in section 3.1 is analysed under Bernoulli schedule single vacation policy in chapter V. According to this policy, the server after completing service for each customer, takes a vacation of random duration with probability p or continues service to the next customer if exists or stays idle in the system if the system is empty, with probability (1-p). On completion of a vacation period, the server returns back to the system even if there is no customer and stays idle until the system size reaches at least m.

For the past three decades, the queueing systems with classical vacations have been well studied because of their applications in modeling the computer networks, communication and manufacturing service systems. In these studies it is assumed that the server stops the primary service completely during the vacations. In 2002, Servi and Finn analysed an M/M/1 queue with **working vacation** where the server works at a lower rate than completely stopping service during vacation period.

Thus working vacation period becomes an operation period of lower speed for the queueing system. The available research works on working vacations mainly concentrated on single arrival and single service queueing systems. Thus in chapter VI and VII , some bulk arrival and bulk service queueing models are analysed under working vacation policies.

In sections 1 and 2 of chapter VI the batch arrival $M^x/M/1$ queueing system is analysed in detail under both multiple and single **working vacations.** In section 3 of chapter VI the general bulk service queueing model M/M(a,b)/1 is considered under **multiple working vacations**. Chapter VII is devoted to a

Non-Markovian GI/M(a,b)/1 queueing system under server multiple **working vacations** . The bulk service rule followed in chapters VI and VII is the most general bulk service rule introduced by Neuts (1967). According to this rule the server takes all the k units in a batch for service, if he finds k ($a \leq k \leq b$) units in the queue and he takes the first **b** units , if he finds more than b units in the queue. On completing his service, if the server finds less than the quorum of **a** units he takes repeated **working vacations,** which is distributed exponentially. During working vacations, the server follows the same rule for bulk service with exponential service rate μ_v which is different from the regular service rate μ. When the vacation ends the server switches his service rate from μ_v to μ. The service rates are assumed to be independent of the size of the service batch. The results of M/M/1 working vacation queueing models obtained by Liu *et al.*, (2007) and Tian *et al.*, (2008a) and GI/M/1 working vacation queueing model of Baba (2005) are deduced from the results of the models of chapters VI and VII respectively as particular cases.

1.4 OBJECTIVES OF THE WORK

The prime objective of the thesis is to develop an analytical treatment of some general bi-level threshold queueing models and working vacation models and to obtain various performance measures. The proposed models are theoretically developed and numerically justified. The results obtained for the models of the present research work are listed below :

- Theoretical frame work for various bulk arrival bi-level threshold queueing models is developed.
- The steady state system size distribution is presented in a closed form so that numerical calculations can be done directly
- The existence of stochastic decomposition property for the proposed models is proved and it is shown that the system size distribution of

these models decomposed into the distributions of two or more independent random variables.

- The performance measures with numerical illustrations and graphical representations are analysed.
- Some interesting particular cases are deduced.
- The system size probabilities and the mean system size when the server is in different states are derived.
- A study on the cost analysis of the models is carried out by determining the total expected cost function per unit time by taking into account various cost elements. The optimal values of the double thresholds that minimize the long run average cost under a linear cost structure are determined.

1.5 METHODOLOGY

The models in chapter II and VI are Markovian. Chapter II and sections 6.1 & 6.2 consider single server queueing systems in which customers arrive in batches in accordance with a time–homogeneous Poisson process with parameter λ. They are non-birth and death processes but still Markovian. The model in section 6.3 deals with single arrival queue in which the service is done in batches according to the general bulk service rule of Neuts (1967). The Chapman –Kolmogrov balance equations satisfied by the steady state system size probabilities are derived for these models and the analysis is done in a straight forward manner similar to that of birth and death queueing processes.

The models of chapter III , IV and V are Non-Markovian, since the service time, setup time and vacation time follow general distributions. The models are analysed by the supplementary variable technique introduced by Cox in 1955. This technique introduces one or more random variables to convert a Non-Markovian process into Markovian. In the models of chapters III to V the

remaining service time, the remaining setup time, repair time and vacation time are used as supplementary variables. Kendall (1953), Keilson and Koharian (1960) and Henderson (1972) have indicated this technique.

The non-Markovian model of chapter VII GI/M(a, b)/1 is analyzed using embedded Markov chain technique. In this model except the inter arrival time the other distributions such as vacation time, service time during regular and vacation time are exponential. Hence the system is studied at the pre arrival epochs so that between any two such consecutive epochs, the only possible transitions that occur are services of customers or completion of vacations which follow exponential distributions. The time points namely the pre – arrival epochs are called regeneration points and the process considered at these epochs is called a Markov chain embedded in the total process, which is non-Markovian. Kendal (1951) uses the concept of regeneration point by suitable choice of regeneration points that involves extraction from the process $\{N(t), t \geq 0\}$ Markov chains in discrete time at those points. The detailed procedure of such embedded Markov chain technique can be seen in Gross and Harris (1985), Medhi (1981,2006) etc.,