CHAPTER I

INTRODUCTION

Queueing theory is an important branch of applied probability and it involves the mathematical study of "queues" or "waiting lines". The formation of waiting lines is a common phenomenon which occurs whenever the current demand for a service exceeds the current capacity to provide that service. Decisions regarding the amount of service to be provided are made frequently in industry and elsewhere.

Study of queueing theory is an attempt to predict the fluctuating demands from observational data and to enable an enterprise to provide adequate service for its customers with tolerable waiting. However, this theory also basically improves the understanding of queueing situations, enabling better control.

Therefore, the ultimate goal is to achieve an economic balance between the cost and the cost associated with waiting time for that service. Queueing theory, as a stochastic process contributes vital information required in decisions associated with waiting lines, by predicting various performance measures such as mean queue length, mean busy period etc.,
The pioneering work on waiting line, "The theory of probabilities and telephone conversions", was done by A.K. Erlang [37] in 1909. The study was to provide solution to congestion arising in telephone traffic.

The objective of studying queueing theory is to optimize the behaviour of the system as a whole.

1.1 BASIC STRUCTURE OF QUEUEING MODELS

A queueing system can be described as customers arriving for service - if not immediately provided and if having waited for service - leaving the system after being served. The term customer may refer to, for example, calls arriving at a telephone exchange or a computer programme waiting for a command to run. The required service to the customer is usually performed as per queue discipline by the service station, after which the customers leave the system.

1.2 CHARACTERISTICS OF QUEUEING SYSTEM

The basic structure of a queueing system is given as follows:

i) arrival pattern of customers

ii) service pattern of servers

iii) number of servers

iv) system capacity and

v) queue discipline
1.2.1 Arrival Pattern of Customers

In most of the queueing models, the arrival pattern is probabilistic or stochastic, where as, in some cases, the arrival of customers may be deterministic. If more than one arrival enters the system simultaneously, the input is said to occur in group or bulk or batch. In such a situation, not only the time between successive arrivals, called inter arrival time of the batches, but also the number of customers in the batch, may be probabilistic. If the probability structure of the arrival process does not vary with time, then it is said to be stationary; otherwise, it is called as non-stationary. In some models, the arriving customer, on seeing the existing position in the system, may decide not to enter the system and this is called as “balking”. In some other models, a customer, who enters the system, will wait for some time for service, becomes impatient and leaves the system without being served. In such a situation, it is referred that the customer has “reneged”.

1.2.2 Service Pattern of Servers

The service times may be deterministic or probabilistic. Customers may be served individually or in batches. In case of batch service, the system is called as “bulk service system”. In bulk service system, the
batches may be of fixed size or variable size. Some times, the service rate may depend on the number of customers waiting for service.

According to Neuts [91], the general bulk service rule is defined as follows:

At a service completion epoch, if there are $\xi$ customers waiting in the queue for service, then the following rule is adhered to:

(i) for $0 \leq \xi < a$ no service

(ii) for $a \leq \xi < b$ serving for a batch of $\xi$ customers

(iii) for $\xi \geq b$ serving for a batch of $b$ customers and the remaining $\xi - b$ customers are kept waiting in the queue

1.2.3 Number of Servers

The number of servers in a queueing model can be finite or infinite. In case of more than one server, depending upon the nature and requirement of the service, the servers may be arranged in series or parallel or combination of both. In a single server model, there will be only one server, whereas in a multiple server model, there will be more than one parallel server. The queueing system, which has many servers in series, is called “multistage queueing system”.

4
1.2.4 System Capacity

The System capacity refers to the physical limitation of the system i.e., the number of customers, who can be accommodated in the queue and at the service centre put together, is called the system capacity. It may be finite or infinite.

1.2.5 Queue Discipline

From the customers who are waiting for service, the procedure by which the customer is selected for service is termed as queue discipline. The “first in first out” (FIFO) is the most commonly used service discipline following the natural law. “Last in first out” (LIFO), is also followed in many occasions. Such a queue discipline is followed in inventory systems, where there is no obsolescence of stored units, since the last arrived unit is easier to reach. In a “service in random order” (SIRO), the customers are selected at random for service. In the “priority” (PRI) queues, the customers are assigned priorities while they enter the system. The customers with higher priority are selected for service before those with lower priority, whatever may be their time of arrival to the system.
1.3 KENDALL’S NOTATION

Any queueing system is represented by the notation introduced by Kendall [60] as follows: $A/B/X/Y/Z$, where $A$ represents inter arrival time distribution of the customers, $B$ denotes the service time distribution, $X$ is the number of parallel servers, $Y$ represents the capacity of the system and $Z$ denotes the queue discipline.

In practice if the system capacity is infinite and the queue discipline is FIFO, then the system is denoted as $A/B/C$ without mentioning $X$ and $Y$.

For example, a queuing system with Poisson bulk arrival and general service distribution with general bulk service of minimum capacity ‘a’ and maximum capacity ‘b’ and single service with infinite system capacity and FIFO queue discipline, is denoted as $M^X/G(a,b)/1$.

1.4 QUEUEING MODELS WITH SERVER’S VACATION

The absence of the server or nonavailability of the service in the system can be termed as “vacation”. The following are the situations where the vacation can be popularly referred to:
(i) the fatigue resulted because of continuous service rendered by the server.

(ii) the non-availability of the minimum number of customers to start the service in case of bulk queueing models.

In the above cases, the server or the operator in the case of machines may be given some secondary job, other than the normal work. The vacation time in a queueing system is aimed at minimizing the total average cost. There are different types of vacation and some of them are discussed below.

1.4.1 Multiple Vacations

At a service completion epoch, if the queue length is less than ‘a’ (fixed before the start of the service), then the server may leave for a vacation of random length. When the server returns from the vacation, again if the queue length is less than ‘a’, then the server may leave for another vacation and so on, until the server finds at least ‘a’ customers waiting for service. This type of server’s vacation is known as multiple vacations.

1.4.2 Single Vacation

After completing a service, if required number of customers, say, ‘a’ is not available in the queue, then the server leaves for a vacation of random length. After completing this vacation, upon returning, if required number of customers is available, the server will start the service;
otherwise, the customer will remain in the system till the queue length reaches the required number of customers. This type of vacation is called a single vacation.

1.4.3 Modified Vacation

After completing a vacation, if the server finds inadequate number of customers waiting in the queue then he avails another vacation and continues in this manner until either he finds required number of customers or completes at most the specified number of vacations, say, ‘M’. After ‘M’ vacations, upon returning, if adequate number of customers in the queue is available, then he will start the service, otherwise he will remain in the system till the queue length reaches the required size. By this assumption, the maximum number of vacations that a server can avail is restricted to M, during an idle period. This type of vacation is called "modified vacation".

1.5 QUEUEING MODELS WITH SETUP TIME

Before starting a service, the server may have to do some preparatory work or some alignment (setup) in the case of certain machines. Queueing system considering this aspect is called "queueing system with setup time".
1.6 QUEUING MODELS WITH CLOSEDOWN TIME

After completing a service, the server has to do some closing work like shutting down the machine, removing the tools etc., in certain machines. Queueing system considering this aspect is called "queueing system with closedown time".

1.7 QUEUING MODELS WITH THRESHOLD POLICY

In some queueing models, the server starts the service only if he finds required number (threshold value) of customers, say, N, available in the queue. Such a queueing system is said to operate under the threshold policy.

1.8 BULK QUEUEING MODELS

Queueing models in which arrivals occur in bulk and/or service done in bulk are known as "bulk queueing models".

1.9 TECHNIQUES FOR SOLVING PROBLEMS OF QUEUEING MODELS

Queueing models are classified as Markovian queueing models and non-Markovian queueing models. The techniques generally adopted to solve these types of queueing models are explained below.
1.9.1 Markovian Queueing Models

Queueing models with exponential inter arrival time and exponential service time are called “Markovian queueing models”. Some of the techniques used to solve Markovian queueing models are:

(i) Difference–differential equation method

(ii) Neuts Matrix-geometric algorithm

(iii) Continued fraction method.

Some queueing systems are studied analytically by deriving the corresponding difference-differential equations and solving them by applying Rouche’s theorem through suitable generating functions. The first method is discussed elaborately by Gross and Harris [44] and Saaty [103]. Neuts [92] developed the matrix-geometric algorithmic approach to study the steady state queueing models. This method involves real arithmetic and avoids the calculation of complex roots based on Rouche’s theorem.

1.9.2 Non-Markovian Queueing Models

The exponential assumption on probability distribution, although very convenient, is not always realistic. There is practical need for models that do not depend on strict Markov assumptions. Queueing
models having the inter arrival times and/or service times which are not exponentially distributed are known as non-Markovian queueing models.

The techniques generally used to study non-Markovian queueing models are:

(a) **Embedded Markov Chain Technique**

This technique, introduced by Kendall [61], is commonly used when one among the service times and inter arrival times is exponentially distributed, while the other is not.

(b) **Supplementary Variable Technique**

Some non-Markovian models can be analysed by converting them into Markovian models through the introduction of one or more supplementary variables. This is known as “supplementary variable technique”. Cox [32] and Cox et al. [33] analysed non-Markovian stochastic processes by the inclusion of supplementary variables. Lee [72] developed an efficient technique to compute the steady state probabilities of vacation models using supplementary variable technique.

### 1.10 LITERATURE SURVEY.

In this section, the relevant literature survey is provided for batch arrival queues, bulk service queues and queues with vacation policies.
Lee et al. [73, 76] considered two queueing systems with N-policy, with and without vacation. He derived the system size distribution and waiting time distribution of an arbitrary customer using supplementary variable technique. A procedure to find the optimal stationary operating policy under a linear cost structure was also presented. Characteristics of $M^X/G/1$ system with N policy was analysed by Lee et al. [74]. Hysteretic and Heuristic control of queueing system was discussed in detail by Nobel [95]. Nobel [94, 96] analysed two $M^X/G/1$ queues with two service nodes for the first one he applied regenerative approach and the for the second he has derived an optimal control. Borthakur [16] and Medhi [85] have obtained the steady state probabilities for the number of customers in the queue and the waiting time distribution for the $M/M(a,b)/1$ queueing system. Borthakur and Medhi [17] have studied queueing system with arrival and service in batches of variable size. They derived the queue length distribution for the $M^X/G(a,b)/1$ model, using the supplementary variable technique. Arora [3] has analysed a two server bulk service queue model and derived various performance measures. Wang [115] analysed a single server queue with optional second service and server break downs.
Krishna Reddy et al. [67, 68] analysed some Markovian bulk service queueing models with delayed vacations and obtained the stationary distribution of the number of customers in the queue and the waiting time distribution of an arriving customer. They also studied a $M^X/G(a,b)/1$ queue with different Laplace transform of the joint distribution of the queue length and the remaining service time or the remaining vacation time depending on the state of the server. Ghare[41] has analysed a multi channel queueing system with unlimited Poisson input and exponential service, which is considered under a bulk service discipline.

A single server bulk service queue has been analysed by many researchers following the pioneering work of Bailey[13,14]. Jaiswal [50] analysed bulk service queueing problems Zhu (80) analysed an $M/G/1/N$ queue with generalized vacation and exhaustive service, where arrival rates depend on the number of customers in the system. They developed an efficient recursive algorithm with overall computational complexity of $O(N^2)$ for computing the exact stationary queue length distribution.

Krishnamoorthy and Deepak [64] analysed an $M/G/1$ queue with modified N-policy. According to this policy, all ‘N’ units are served at a
time; then subsequent arrivals are served one at the time. They obtained steady state distribution and numerical results. A finite buffer $M/ G (a,b)/1$ queueing system with general bulk service rule was analysed by Gupta et al. [46, 47].

Detailed analysis of some bulk queueing models can be seen in the studies of Chaudhry and Templeton [24] and Medhi [86,87]. Queueing systems with general bulk service were also studied by Anitha [2], Audsin Mohana Dhas [11], Jayaraman [57], Kandasamy [58] and Tackacs [110]. A comprehensive study on bulk queueing models were also analysed by Ross [102].

A class of single server vacation queues which has single arrival and non-batch server in discrete time was studied by Alfa [1]. Nadarajan and Subramanian [90] analysed a single server queueing system with general bulk service rule. They considered both multiple and single vacation of server and by using Matrix-geometric method, they obtained the steady state probability vector of the number of customers in the system and the stability conditions were obtained. Jau-Chuan Ke [56] analysed an $M^X/G/1$ system with startup server and J-additional options for service in which they obtained the steady state results, including system size distribution at a random epoch and at a departure epoch.
They also obtained the distributions of idle and busy periods and waiting time distribution in the queue. Krishnakumar et al.[65] analysed on $M/G/1$ feedback queue with varying arrival rates with N-policy. Madan et al.[83] studied a single server queue with batch arrivals and two types of heterogeneous service with different general service time distributions with optional reservice. They obtained steady state results in explicit and closed form in terms of the probability generating functions for the number of customers in the queue and the system.

Fuhrmann and Cooper [40] analysed an $M/G/1$ queue with generalized vacations, they illustrated that the decomposition property holds for a very general class of $M/G/1$ queueing models. Minh [89] obtained the transient solutions for some exhaustive $M/G/1$ queue with generalized independent vacations. Takine and Hasegava [113] studied an $M/G/1$ queue with server’s vacation. They exhibited the decomposition property of the $M/G/1$ queue in a generalized way. A batch arrival queue with a vacation time under single vacation policy was discussed by Choudhury [27]. Chuoudhry [26] discussed an $M/G/1$ queue with two different vacation times under multiple vacation policy. Choudhury [28] generalised the results of Madan [81] and obtained various performance measures using supplementary variables. Choudhury and Lotfi tadj [30]
analysed a bulk service queueing system with unreliable server. They obtained stability condition and steady state system size distributions.

Rosenberg and Uri Yachiali [101] analysed a bulk queue with single and multiple vacations under LIFO regime. Li and Zhu [79] did an analysis of M/G/1 queues with delayed vacations and exhaustive service discipline. Jau-Chuan Ke [52, 53] discussed the optimal control of an M/G/1 queueing systems with server vacations, startup breakdowns and two vacation types. Nishianura and Jiang [93] analysed an M/G/1 vacation model with two service nodes. A queue with starter and vacations delay was analysed by Levy and Kleinrock [77]. Madan et al. [82] analysed a single server queue with optimal phase type server vacations based on exhaustive deterministic service. They obtained explicit steady state results for the probability generating functions of the queue length, expected number of customers in the queue and expected waiting time of the customers. Tadj [111] analysed an M / G / 1 quorum queueing system under T-policy.

Chang and Choi [22] did a performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and multiple vacations. Sikdar and Gupta [106] exhibited the analytic and numerical results for batch service queue with single vacation. Using
supplementary variable technique, they obtained probability generating function at various epochs. Arumuganathan et al. [7,12] studied two non-Markovian bulk queueing system with control policy and different types of vacations, in which they obtained the probability generating function of number of customers in the queue at an arbitrary time epoch and other measures as well.

Levy and Yechiali [78] studied an M / M / S queueing system with server's vacation. They derived, a formulae for the distribution of the number of busy servers and the mean number of units in the system. An M / G / 1 queue with exceptional first vacation was analysed by Lee [71]. He concentrated on a vacation system in which the first vacation was differently distributed from the subsequent vacations. He derived the transform solution of the system size distribution by defining supplementary variable. An M / G / 1 queue with D-policy was discussed by Artalejo [4]. He has obtained steady state probabilities of the queue size.

Doshi [36] made a survey of queueing systems with vacations, in which he attempted to provide a methodological overview with the objective of illustrating how the seemingly diverse mix of problems is closely relayed in structure and can be understood in a common
framework. A comprehensive survey on vacation queueing systems is found in the monographs of Takagi [112] and Kleinrock [62]. Lee [70] analysed a $M / G (a,b) / 1$ queue with single vacation. Lee et al.[75] analysed on $M^X / G / 1$ queueing system with N-policy and multiple vacations, using supplementary variable technique. The system size was decomposed into three random variables one of which was the system size of ordinary $M^X / G / 1$ model. The optimal stationary operating policy was achieved under a linear cost structure.

Choudhury and Paul [29] considered a batch arrival queue with N-policy. They have considered the concept of additional service Harold J-Kushner and Ramachandran [48] discussed the approximately optimal control problem for tandem queueing or production networks under heavy traffic. A computational analysis of steady state probabilities of bulk and non-bulk queues was discussed by Chaudhry et al.[23]. Choudhury [25] analysed single server Poisson bulk arrival general service queue with a set up period and a vacation period. They obtained the system state probabilities and some performance measures of the system.
Arumuganathan et al. [8,10] studied some bulk queues with multiple vacation with fast and slow service rates and state dependent arrivals, in which they obtained the probability generating function of the number of customers in the queue at an arbitrary time epoch and other measures as well. Perry et al.[100] considered two models of M / G / 1 and G / M / 1 type queueing systems with restricted accessibility and obtained closed form expressions for the Laplace transform of the lengths of the busy periods. Martin and Artlejo [84] discussed an M / G / 1 with two types of impatient units. An identity on the moments of the transient delay of M / G / 1 queue was done by Wang [114]. Wang and Jau-Chuan Ke [116] studied optimal control of an M / G / 1 queueing system with finite and infinite capacity. Dharmaraja [34] discussed the transient solution of a two processor heterogeneous system with Poisson arrival of jobs having exponentially distributed execution times.

Dharmaraja et al. [35] did performance modeling of wireless networks with generally distributed handoff inter arrival times. A combinatorial technique for M / G / 1 type queues was discussed by Mercankosk et al. [88]. Bansal [18] analysed an M / G / 1 processor sharing queue with bulk arrivals. Boxma and Takine [20] discussed an M / G / 1 FIFO queue with several customer classes and they obtained joint queue length distributions for the queueing system. Chydzinski [31]
analysed the distribution of the remaining service time upon reaching a target level in M / G / 1 queueing system. Wang et al. [117,118] studied single removable server N-policy M / G / 1 queueing systems in steady state. They demonstrated that the maximum entropy approach is a useful method for solving a complex queueing system.

Bischof [19] analysed an M / G / 1 queue with setup times and vacations under six different service disciplines. He discussed that the performance of the system depended on the customer waiting time, the queue length and the cycle duration. Frans [39] discussed the inter departure time distribution of each class of the customers in $\sum M^X_i/G_i/1$ queue with setup time and multiple vacation policy.

Guan et al. [45] analysed control queue delay in buffer with time-varying arrival rate. They developed an algorithm that controls the delay by dynamically adjusting the threshold which, in turn, controls the arrival rate. Jaiswal [51] discussed the time-dependent solution of the bulk-service queueing problem.
1.11 OBJECTIVES OF THE WORK

The main objective of this research is to develop analytical treatment of some bulk queueing models to obtain some important performance measures. This research proposes different bulk queueing models arise in many practical situations. The analytical treatment of all these bulk queueing models is obtained using supplementary variable technique. The proposed models are theoretically developed and numerically justified.

Main objectives of this research are:

- to develop the theoretical framework for various bulk queueing model.
- to be motivated through many practical applications.
- to obtain some important performance measures.
- to analyse the performance measures with numerical illustrations.
- to obtain some interesting particular cases.

1.12 AUTHOR'S CONTRIBUTION

A steady state analysis of a Non-Markovian Bulk Queueing System with Overloading and Multiple Vacations is considered in Chapter II. It is considered that the server may adjust the service rate and the service capacity depending upon the queue length. After completing a
service, if queue length is less than a threshold value ‘a’, then the server leaves for a vacation of random length. When the server returns from a vacation, and if the queue length is still less than ‘a’, then the server leaves for another vacation and so on until the server finds at least ‘a’ customers waiting for service. After a service or a vacation, if the server finds at least ‘a’ customers waiting for service (say) $\xi (a \leq \xi < N)$, then the server serves a batch of min $(\xi, b)$ customers, where $b \geq a$. If $\xi \geq N$, then the server increases the service capacity and serves a batch of $N$ customers with a different service rate. Thus, the general bulk service rule is modified with variable service capacity. The probability generating function of number of customers in the queue at an arbitrary time epoch is obtained, using supplementary variable technique. Expressions for the expected length of queue and expected length of idle and busy periods are derived. A cost model for the queueing system is discussed and illustrated numerically.

In Chapter III, a Steady State Analysis of a Bulk Arrival General Bulk Service Queueing System with Modified M-Vacation Policy and Variant Arrival Rate is considered. After a service completion, if the number of customers waiting in the queue is less than a threshold value ‘a’, then the server avails a vacation of random length. After a vacation, if the queue length is still less than ‘a’, then the server
avails another vacation and so on until he finally finds at least 'a' customers waiting in the queue or he completes 'M' number of vacations, consecutively. After completing the Mth vacation, if the queue length is still less than 'a', then the server remains in the system till the queue length reaches 'a' (this period is known as dormant period). At a vacation completion epoch or a service completion epoch or during the dormant period, if the queue length $\xi$ is at least 'a' then the server serves a batch of min ($\xi, b$) customers, where $b \geq a$. It is more realistic that the arrival rate may not be constant all the time. Addressing this, it is considered that the arrival rate is dependent on the state of the server. That is, the arrival rate is $\lambda$ when the server is busy and the arrival rate is $\lambda_0 (\lambda_0 < \lambda)$ when the server is on vacation or in the dormant period. Using supplementary variable technique, the steady state queue size distribution at an arbitrary time is obtained. Performance measures like the expected queue length, expected length of busy and idle periods are derived. Expected waiting time in the queue is also obtained. A cost model is derived. Numerical illustration is also presented.

In Chapter IV, a Steady State Analysis of a Poisson Bulk Arrival Single Service Queueing Model with Threshold Policy on Number of Primary services for Secondary Jobs is considered. In most of the
queueing systems with vacations, the operator of the service station will be allotted to secondary jobs (vacations) only if, the number of waiting customers is zero. But it may not be the case always; there are situations in which the operator has to shutdown the machine after some finite number of processes. Addressing this, it is assumed that after rendering M consecutive services, the operator of the service station closes the service and avails a vacation of random length. At a vacation completion epoch, if the number of waiting customers in the queue is less than a threshold value N (N > M), the operator waits in the system (dormant period) till the queue length reaches N. At a vacation completion epoch or during a dormant period, if the queue length is at least N, then the server starts the single service. Using supplementary variable technique, the probability generating function of the number of customers in the system at an arbitrary time is derived. Expected length of the system, busy period and idle period are derived. A cost model is developed. Numerical illustrations are provided.

In Chapter V, a Steady State Analysis of a Non-Markovian Bulk Queueing Model with Multiple Vacations, Accessible Batches and Closedown times is considered. The service starts only if, minimum of 'a' customers available in the queue. At the service completion epoch if the number of customers waiting in the queue is ξ, where
a \leq \xi \leq d - 1 \quad (d \leq b), \text{ then the server takes the entire queue for batch service and admits the subsequent arrivals for service till the service of the current batch is over or till the accessible limit 'd' is reached, whichever occurs first. At the service initiation epoch if the number of customers waiting in the queue '\xi' is at least 'd', then the server takes min \((\xi, b)\) customers for service and does not allow further arrival into the batch. On completion of a service, if the queue length is less than 'a', then the server performs a closedown work such as shutting the machine, removing the tools etc., . Following the closedown work, the server leaves for multiple vacations of random length irrespective of queue length. When the server returns from a vacation and if the queue length is still less than 'a', he leaves for another vacation and so on until he finally finds at least 'a' customers waiting in the queue. Using supplementary variable technique, the probability generating function of the number of customers at an arbitrary time is derived. Expected length of the queue, busy period and idle period are derived. A cost model is developed. Numerical illustrations are presented.

In Chapter VI, an overview of all the proposed bulk queueing models and their scope for future enhancements are presented to conclude the thesis.