CHAPTER I

INTRODUCTION

The study of queueing systems has increasingly occupied the attention of researchers, since it has wide range of applications in the fields like assembly lines in production systems, communication systems, computer systems and even in day to day affairs.

The first work on waiting line (queue) situation was "The theory of probabilities and telephone conversations" by A.K. Erlang [12] in 1909. This was devoted for a study of telephone traffic congestion.

1.1 Basic Structure of Queueing Models

A queue, or a waiting line, involves arriving customers who wait to be served at the service station. The term "customer" may refer, for example, a machine arriving at an inspection centre or a person arriving at a booking counter in a railway station. Customers are selected for service by certain rule, known as queue discipline. The term "queueing system" is defined as queue consisting of customers and the service station put together. The required service for the
1.2 Characteristics of Queueing Models

The basic characteristics of a queueing system are listed below:

i. The arrival pattern of customers
ii. The service pattern of servers
iii. The queue discipline
iv. The system capacity
v. The number of servers

1.2.1 The arrival pattern of customers

In most of the queueing models, the arrival pattern is probabilistic or stochastic, whereas in some cases the arrival of customers may be deterministic. If more than one arrival enter 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 interarrival time of the batches, may be probabilistic, but also the number of customers in the batch. 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, may 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 The 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 service system is called as "bulk service system". In bulk service system, the batches may be of fixed size or variable size. Sometimes, the service rate may depend on the number of customers waiting for service.

The "general bulk service rule" was introduced by Neuts [39]. According to this rule, the server will start a service, if there are at least 'a' customers in the queue. If there are fewer, he waits for the queue length to reach 'a' and then serves a group of size 'a'. If after a departure, the queue length \( \xi \) is at least 'a', a group of size \( \min(\xi, b) \) (where \( b \geq a \)) is served. Service is independent of group size.
1.2.3 The queue discipline

The queue discipline is the procedure by which a customer is selected for service, from among customers who are waiting for service. The most commonly used queue discipline is "first in first out" (FIFO). Another queue discipline is "last in first out" (LIFO). Such a queue discipline can be observed in inventory systems, where there is no obsolescence of stored units, as it is easier to reach the nearest items, which arrived last. In some cases, customers are selected at random for service, independent of the time of arrival to the system, and such a queue discipline is known as "service in random order" (SIRO). Yet another discipline is "priority" (PRI), where customers are given priorities upon entering the system and the ones with higher priority are selected for service ahead of those with lower priority, irrespective of their time of arrival to the system.

1.2.4 System capacity

The number of customers, those can be accommodated both in the queue and service put together in the system, is called as system capacity. It may be finite or infinite.

1.2.5 Number of servers

The number of servers in a queuing model may be finite or infinite. When the number of servers is more than one, the servers may be arranged in
series or parallel or a combination of both, depending upon the nature of the service required. If the queueing system has only one server, it is called single server model, when the system has a number of parallel servers, it is termed as multiserver model and when the system has many servers in series, it is called multistage queueing system.

1.3 Kendall's Notation

A convenient and abbreviated form of notation for queueing systems is proposed by Kendall [23], popularly known as Kendall's notation. According to this notation any queueing system is simply denoted as A/B/C/X/Y, where 'A' stands for interarrival time distribution, 'B' represents the service time distribution, 'C' is the number of servers, 'X' is the system capacity and 'Y' stands for the queue discipline. In practice, if the system capacity is infinite and the queue discipline is FIFO the system is denoted as A/B/C, without mentioning X and Y. For example, a queueing system with Poisson arrival process and general service distribution with general bulk service of minimum capacity 'a' and maximum capacity 'b' and single server with infinite system capacity and FIFO queue discipline is denoted as M/G (a,b)/1.
1.4 Queueing Models with Server Failures

In many practical situations, the server may fail to render service to the incoming customers. This may mostly happen when the server is a machine. Server failures can be classified into operation and time dependent failures. The operation dependent failures can occur only when the server is in operation and the time dependent server failures can occur at all-time, independent of whether the server is rendering service or not. In queueing models with server failures, once failure occurs, immediately the repair process starts and after completion of the repair, the service is continued according to the "repeat" or the "resume" rule.

1.5 Queueing Models with Server's Vacation

The term "vacation", related to queueing theory, refers the absence of the server or nonavailability of the service in the system. There are situations in which the server requires some rest due to continuous serving or fatigue. In some other cases, such as bulk service models, the required number of customers may not be available to render the service. And in such a situation, the server or the operator in the case of machine may be given some other job, other than the known as ancillary duties. We then say that the server or service station on regular work. This type of taking rest or doing some other job is called as "vacation". The vacation time in a queueing system is aimed to minimise the total some cost function.
average-cost. There are different types of vacation, some of the important types are discussed below:

1.5.1 Multiple vacations

In case of bulk service, after completing a service, if the queue length is less than a predetermined value say 'a', the server may leave for a vacation. When he returns, if the queue length is still less than 'a', he may leave for another vacation and so on, until he finally finds at least 'a' customers waiting for service. This type of server's vacation is known as multiple vacations.

1.5.2 Multiple vacations of alternate type

In some situations the server may take two different types of jobs other than the regular job, when he is idle. After a service, if the queue length is less than 'a', the server leaves for a vacation of type $V_1$ with cumulative distribution function $V_1(.)$. After returning from vacation type $V_1$, if still the queue length is less than 'a', he leaves for another vacation of type $V_2$ with cumulative distribution function $V_2(.)$. The server takes the vacations of types $V_1$ and $V_2$ alternately, until he finds, on returning from a vacation, at least 'a' customers waiting for service. This type of server's vacation is called multiple vacations of alternate type.
1.5.3 Single vacation

After completing a service, if the server finds the number of customers waiting for service is not to the required level, say 'a', for service, he will go for a vacation. When he returns, if the queue length is at least 'a', he starts service. Otherwise, he remains in the system till the queue length reaches 'a' and then starts a service. This type of vacation is called as single vacation.

1.5.4 Limited service vacation

Sometimes, the server may not be allowed to work more than a specified duration of time. In the case of a machine, it may not be allowed to work continuously for a longer period; since machine fault may occur due to increase in temperature after certain limit. In such a case, server may be allowed to take a vacation after the specified duration of time and it is known as limited service vacation.

1.5.5 Gated vacation

In this type, the server serves only those customers who entered the system by the time of his return from a vacation and takes another vacation without serving the customers who arrived during the service period. One can assume that when the server returns from a vacation, a gate closes behind the last waiting customer and the server serves only those customers in front of the gate.
and on completion of service, the server goes for another vacation after opening the gate.

1.6 Queueing Models with Setup Times

Before starting a service, the server may have to do some preparative work such as some alignment (setup) must be done in the case of certain machines. Such queueing models considering this aspect are called models with setup times.

1.7 Queueing Models with Threshold - Policy

In some queueing models, the server can start service if required number (threshold value) of customers, say N, available in the queue. Such queueing models are often called as queueing models with N-policy. In the case of general bulk service queueing models with multiple vacations, the server usually avails the multiple vacations until the queue length reaches the minimum service capacity, say 'a'. But, in the models with d-policy, the server avails multiple vacations until the queue length reaches d (d < a). On returning from a vacation, if the queue length is at least 'd', he remains in the system till the queue length reaches 'a' and then starts servicing.
1.8 Bulk Queueing Models

Queueing systems 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 broadly classified as Markovian queueing models and non-Markovian queueing models.

1.9.1 Markovian queueing models

If both interarrival time of customers and the service time follow exponential distributions, then the model is said to be Markovian queueing model.

Markovian queueing models are generally solved by using difference-differential equation method or Neuts matrix-geometric algorithm.

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. This procedure is discussed in detail by Gross and Harris [15], Kleinrock [25] and Saaty [44]. Steady state queueing models can be solved using matrix geometric algorithmic...
approach, which was developed by Neuts [40,42]. This method involves only real arithmetic and avoids the calculation of complex roots based on Rouche's theorem.

1.9.2 Non-Markovian queueing models

Though it is convenient to assume the exponential probability distributions, it may not always be realistic. There are situations in which Markov assumptions fail. Queueing models having interarrival times and/or service times which are not exponentially distributed are known as non-Markovian queueing models.

The techniques generally used in studying non-Markovian queues are:

(a) **Embedded Markov chain technique**

This technique, introduced by Kendall [24], is commonly used when one or the other among the service times and interarrival times is exponentially distributed, while the other is not.

(b) **Supplementary variables 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 variables technique. Cox [9] and
Cox and Miller [10], have analysed non-Markovian stochastic processes by the inclusion of supplementary variables. Lee [32] developed an efficient technique to compute the steady state probabilities of vacation models using supplementary variables technique.

1.10 Relevant Literature Survey

1.10.1 Bulk queueing models

The single server bulk service queues have been analysed by many researchers, following the pioneering work of Bailey [4]. The steady state probabilities for M/M(1,b)/C model is obtained by Ghare [14]. Neuts [41] has studied the transient state distribution of the number of customers in the system for the M/G(a,b)/1 model. Borthakur [5] and Medhi [35] 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. Medhi and Borthakur [37] have latter extended these results to the M/M(a,b)/2 model. Borthakur and Medhi [6] have studied a queueing system with arrival and service in batches of variable size. They derived the queue length distribution for the M'/G(a,b)/1 model, using the supplementary variables technique. Arora [1] has analysed two server bulk service problem. The model M/M(a,b)/N has been
studied by Neuts and Nadarajan [43]. The steady state probability vector and waiting time distribution have been obtained using matrix - geometric approach.

Krishna Reddy et.al [28,29] analysed multiserver bulk service queueing models and obtained steady state probability vector, using matrix - geometric approach. Characteristics of $M^{n}/G/1$ system with $N$ - policy is analysed by Ho Woo Lee et.al [16]. Sharma [48] has studied a queueing system with arrival in batches of variable size and correlated departures. Detailed analysis of some bulk queueing models is found in the monographs by Chaudry and Templeton [8] and Medhi [36].

Sim and Templeton [51] have obtained the steady state probability distribution of the waiting customers for the model $M/M(a_1,b_1), M(a_2,b_2)/2$, the two heterogeneous server Markovian queue with general bulk service, using difference - differential equations method. Love [34] has developed the steady state solution for the $E_k/M(a,b)/N$ system, using semi Markov process technique.

The $GI/M(1,b)/N$ system has been analysed by Shyu [50], using embedded Markov chain technique. Jaiswal [18,19] has obtained the transient and steady state distribution of the queue and the waiting time distribution in steady state for the system $M/E_R(1,b)/1$, where $E_R$ denotes modified Erlangian distribution having
a random number $R$ of exponential phases. Queueing systems with general bulk service have also been studied by Audsin Mohana Dhas [2], Jayaraman [20], Kandasamy [21] and Sivasamy [52].

1.10.2 Queueing systems with server failures

Some queueing models with the service station subject to breakdown are analysed by Avi-Itzhak and Naor [3]. Buzacott and Hanifin [7] classified the machine failures as operation dependent failures and time dependent failures. A time dependent failure model with service and repair times limited to Erlang distribution and arrival rate dependent on server's position is analysed by Shogan [49]. Shanthikumar [47] analysed a $M/G/1$ queueing system with time and operation dependent server failures, in which service and repair times follow general distribution.

1.10.3 Queueing systems with vacation

An $M/M/S$ queueing system with server's vacation is analysed by Levy and Yechiali [33]. The distribution of number of busy servers and the mean number of units in the system have been obtained by considering repeated vacations and single vacation.
Vinod [59] has analysed a queueing system in which the server may take a vacation which is either independent of the number of customers in the system or dependent on the number of customers in the system.

Nadarajan and Subramanian [38] analysed a single server queueing system with general bulk service rule. Both multiple and single vacation of server are considered and the steady state probability vector of the number of customers in the system and the stability condition are obtained, using matrix - geometric method.

Krishna Reddy et.al [26,27] analysed two general bulk service queueing models with vacations. The first one is a Poisson arrival queueing model with additional server. Whenever the number of customers in the queue is less than the quorum, the regular server takes an exponential vacation. An additional server is introduced, when the queue length exceeds a pre-assigned number. The second one is a tandem queueing model with vacation. The steady state probability vector of the number of customers is obtained, using matrix-geometric technique.

Soon Seok Lee et.al [54] considered an M*/G/1 queueing system with N-policy and single vacation. Using supplementary variables technique, system size distribution and waiting time distribution of an arbitrary customer are
obtained. A procedure to find the optimal stationary operating policy under a linear cost structure also presented.

Ho Woo Lee et. al [17] have analysed an M^\infty/G/1 queueing system with N-policy and multiple vacations, using supplementary variables technique. The system size distribution is obtained. The system size is decomposed into three random variables one of which is the system size of ordinary M^\infty/G/1 model. The optimal stationary operating policy is achieved under a linear cost structure.

Application of vacation models to polling systems is presented by Takagi [55]. The time dependent analysis and process of M/G/1 vacation models with exhaustive service is obtained by Takagi [56,58]. The Laplace transforms of the joint distribution of server state, queue size and elapsed time in that state are obtained. Exhaustive service M/G/1 systems with multiple vacations, single vacation, an exceptional service time for the first customer in each busy period and a combination of N-policy and setup time are considered. The decomposition property in the steady state, joint distribution of the queue size and remaining service time is demonstrated.
A comprehensive survey on vacation queueing systems is found in the monographs by Doshi [11], Takagi [57], Scholl and Kleinrock [45] and Fuhrmann and Cooper [13].

1.11 Author's Contribution

This dissertation presents the analysis of some single server queueing models with certain combinations of single arrival, bulk arrival, single service, bulk service, server failures and server vacations.

Chapter two is devoted for the investigation of "Poisson bulk arrival queueing model with time and operation dependent server failures". The server is subjected to time and operation dependent failures and arrival rate depends on the up and down state of server. The time to failure is exponentially distributed and repair times follow general distribution. The failure rate depends on whether the server is idle or busy, and the repair time depends on whether the failure occurred during the idle or busy period of the server. The probability generating function of the steady state system size is obtained at customer departure epoch, using embedded Markov chain technique. The probability generating function of the steady state system size at an arbitrary time is obtained, using level crossing method. Expressions for expected system size, expected length of busy and idle periods are derived. Numerical illustration is presented.
Chapter three contains the analysis of "M/G(a,b)/1 queueing system with time and operation dependent server failures". The customers arrive singly and service is done according to general bulk service rule. As per this rule, the server will start a service, if there are at least 'a' customers in the queue. If there are fewer, he waits for the queue length to reach 'a' and then serves a group of size 'a'. If after a departure or repair completion epoch during idle period, the queue length is at least 'a', a group of size min(ξ, b) (where b ≥ a) is served. Service time is independent of the group size. The server is subjected to time and operation dependent failures and arrival rate depends on the up and down state of the server. The time to failure is exponentially distributed and repair time follows general distribution. The failure rate depends on whether the server is idle or busy and the repair time depends on whether the failure occurred during the idle or busy period of the server. The probability generating function of the steady state system size at customer departure epoch is obtained, using embedded Markov chain technique. The probability generating function of the steady state system size at an arbitrary time is obtained, using level crossing method. Expressions for expected system size, expected length of busy and idle periods are derived. Numerical solution for particular values of parameters is also presented.
In chapter four, "Poisson bulk arrival queueing model with general bulk service and multiple vacations" is considered. After completing a service, if the queue length is less than 'a', the server leaves for a vacation. When he returns, 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 for service. After a vacation, if the server finds 'a' or more customers in the system, he serves a batch of min (ξ, b) customers, where b ≥ a. Using supplementary variables technique, the probability generating function of the steady state queue size at an arbitrary time is obtained. Expressions for expected queue length, expected length of busy and idle periods are derived. A cost model for the queueing system is discussed. Numerical illustration is presented.

Chapter five is devoted for the analysis of "M/G(a,b)/1 queueing model with multiple vacations and setup time and threshold - policy". Two queueing models with different threshold policies to start the setup process are considered. In Model I, after completing a service, if the queue length is ξ, where ξ < a, then the server takes the vacation according to the following procedure. If the queue length ξ is less than the threshold value 'd', he leaves for a vacation. When he returns from the vacation, if the queue length is still less than 'd', he leaves for another vacation and so on, until he finally finds at least 'd' customers waiting for
service. After a service, if the queue length $\xi$ is such that $d \leq \xi < a$, then the server avails a single vacation. After a vacation, if the queue length is at least $d'$, then the server requires a setup time $R$ to start the service. After setup time, if the server finds $a'$ or more customers, then he serves according to general bulk service rule, otherwise he remains in the system till the queue length reaches $a'$. Model II differs from Model I in that, after a service if the queue length is less than $a'$, the server avails multiple vacations till the queue length reaches $N$ ($N \geq b \geq a$). After a vacation, if the queue length is at least $N$, then the server requires a setup time $R$ to start the service. After the setup time, he starts service with a batch of $b'$ customers, since at least $b'$ customers will be waiting in the system. The probability generating function of the steady state queue size at an arbitrary time is obtained, using supplementary variables technique. Expressions for expected length of queue, busy period and idle period are obtained. A cost model is discussed. Numerical illustration is presented. Corresponding results for the "$M^a/G(a,b)/1$ queueing model with multiple vacations and setup time" are deduced as a particular case of the Model I by substituting $d'$ equal to $a'$.

Chapter six is devoted for the analysis of "$M^a/G(a,b)/1$ queueing system with multiple vacations and repair of service station on request by a leaving batch of customers". It is assumed that whenever a leaving batch of customers is
dissatisfied with the low quality of service rendered, they may complain and request that the station be adjusted to its proper level of service. The leaving batch of customers may request for repair of service station with probability $\pi$ and it is assumed that more than one request will never be made by the same batch. After the repair time or service completion without request for repair, if the queue length is less than 'a', the server avails multiple vacations till the queue length reaches 'a'. After a vacation or a service completion without request for repair or a repair completion, if the server finds at least 'a' customers waiting for service, he serves according to general bulk service rule. Using supplementary variables technique, the probability generating function of the steady state queue size at an arbitrary time is obtained. Expressions for expected length of queue, busy period and idle period are obtained. A cost model is developed. Numerical illustration is presented.

Chapter seven contains the analysis of "M$^a$/G(a,b)/1 queueing system with multiple vacations of alternate type". After a service, if the queue length is less than 'a', the server leaves for a vacation of type $V_1$ with cumulative distribution function $V_1(.)$. After returning from vacation type $V_1$, if still the queue length is less than 'a', he leaves for another vacation of type $V_2$ with cumulative distribution function $V_2(.)$. The server takes vacations $V_1$ and $V_2$ alternately, until
After a vacation, if the server finds \( \xi \) (\( \xi \geq a \)) customers in the system, then he serves a batch of \( \min(\xi, b) \) customers, where \( b \geq a \). Using supplementary variables technique, the probability generating function of the steady state queue size at an arbitrary time is obtained. Expressions for expected queue size, expected length of busy and idle periods are derived. A cost model is developed. A numerical solution for particular values of parameters of the model is presented.

In chapter eight, an application for a production line system is discussed. A numerical model is analysed to illustrate how the management of flow line system can use the results discussed in chapters four, five, six and seven to make the decision regarding the threshold value 'a' of bulk service rule to minimise the total average cost.

In the last chapter, a brief discussion about the work done; what is achieved, what is not possible to achieve and suggestions for further work is presented.