CHAPTER – I

INTRODUCTION

1.1 Nature of Queueing Theory

The most fantastic growth in recent times has not been in the field of population, Inflation, crime, punishment, pride or prejudice but in the field of queues. Queues have become the most ubiquitous sight these days. In fact, the sight of a queue is almost always both amusing and appealing. But like many of unsavory phenomenon of life, queues simply cannot be questioned. Think for a moment, how much time is spent in one’s daily activities waiting in some form of queue: waiting for breakfast; Stopped at traffic light; slowed down on the highways and freeways; delayed at the entrance to one’s parking facility; queued for access to an elevator; standing in line for morning milk and so on. The list is endless and too often also is the queues, serpent-time queues that have become the warp and woof of our life.

In our increasingly congested and urbanized society, all of us have experienced the annoyance of having to wait in line. Unfortunately, this phenomenon is becoming more and more prevalent. This problem is basically because of demand exceeding available service facilities. There may be many reasons, for example, shortage of available servers for the particular job, or it may be economically infeasible for a business firm to provide the level of service required so as to prevent the waiting, or there may be space limit to the amount of service that can be provided. Thus, services of others became a critical function of survival. If man had followed a philosophy of “immediate service on demand”, the result would have been an uneconomical utilization of his total effort. Thus, queues are a part of everyday life.

Queueing Theory was originated in the early 20th century, because of the practical requirements of teletraffic physics, and rational organization of mass service, such as theatre agencies, stores, automatics machines, etc. The pioneer investigator was the well-known Danish Mathematician A.K.Erlang, who in 1909 published ‘The Theory of Probabilities and Telephone Conversations’.
In his later works he observed that a telephone system was generally characterized by either (i) Poisson input, exponential holding times and multiple channels or (ii) Poisson input, constant holding times and a single channel. Since then a number of mathematicians, engineers and economists have shown their interest in this area and developed a number of similar models. The models so developed have found their applications to various other areas such as the natural sciences, engineering, economics, transportation, military problems, and Internet trafficking and production assembly line etc.

Today queues (the waiting line) are a part of everyday life. We all wait in queue whether to buy a movie ticket or to make a bank deposit or pay for groceries or mail a package or obtain food in a cafeteria or start a ride in an amusement park or to withdraw money from an ATM counter etc., but still get annoyed by usually long waits. However having to wait is not just a petty personal annoyance. The amount of time that a nation’s populace wastes in waiting is a major factor in both the quality of life there and the efficiency of the nation’s economy. Before its dissolution, the USSR was notorious for the tremendously long queues that its citizens frequently had to endure just to purchase basic necessities, which is an example of a country in crisis. Today even in a developed country, like United States it has been estimated that Americans spend 37,000,000,000 hrs per year (as by F.S. Hillier and G.J. Lieberman) waiting in queues. Even this staggering figure does not tell the whole story of the impact of causing excessive waiting. Great inefficiencies also occur because of other kinds of waiting than people standing in line. For example, making machines wait to be repaired may result in lost production. Vehicles (including ships and trucks) that need to wait to be unloaded may delay subsequent shipments. Airplanes waiting to take off or land may disrupt later travel schedules. Delay in telecommunication transmissions due saturated lines may cause data glitches. Causing manufacturing jobs to wait to be performed may disrupt subsequent production. Delaying service jobs beyond their due dates may result in lost of future business.
In manufacturing economics, one of the problems is how to assign the correct ratio of man to machine. If a man has responsibilities for few machines then a considerable portion of his time is nonproductive and some service time is lost for the company. If he is assigned too many machines time may be lost because a man who is servicing one machine is unavailable for work on another. Thus, the total cycle time for one machine is extended because like others it must wait for service. This queue of machines, idle because of inadequate servicing, may become costly to a company especially when the cost per machine-hour is considerably larger than the cost per man-hour.

There are many other valuable applications of the theory, most of which have been well documented in the literature of probability, operations research, management science and industrial engineering. Examples include: traffic flow (vehicles, aircraft, people, and communications), scheduling (patients at the doctor, programs on a computer), and facility design (banks, post offices, fast-food restaurants). Queuing theory originated as a very practical subject but much of literature up to middle 1980’s was of little direct practical value. Since then the emphasis in literature on finding the exact solution of queuing problems with clever mathematical tricks is now becoming secondary to model building and the direct use of these techniques in decision making. Most real problems do not correspond exactly to a mathematical model and do not always have closed-form solutions, but most of the time we are able to conduct computational analysis and find approximate solutions. We have to thank for this to our every day companion, the computer.

1.2 A Historical Perspective

This very important subject has completed its 100 years of existence a few years back. The paper by Johannsen’s “Waiting Times and Number of Calls” which was published in 1907 and reprinted in Post as Office Electrical Engineers Journal, London, October, 1910 seems to be the
first paper on the subject. But the method used in this paper was not mathematically exact, for this reason only many mathematician had not recognized this as the official work on Queueing Theory, which now a day is purely mathematical in nature. And therefore, from the point of view of exact treatment, the paper that has historic importance is A. K. (Agner Krarup) Erlang’s, “The Theory of Probabilities and Telephone Conversations” which was published in Nyt tidsskrift for Matematik, B, 20 (1909), p. 33. In this paper he lays the foundation for the place of Poisson (and hence, exponential) distribution in queueing theory. His papers written in the next 20 years contain some of the most important concepts and techniques; the notion of statistical equilibrium and the method of writing down balance of state equations (later called Chapman-Kolmogorov equations) are two such examples. Special mention should be made of his paper “On the Rational Determination of the Number of Circuits” Brockmeyer etal (1948), in which an optimization problem in queueing theory was tackled for the first time. Agner Krarup Erlang is now known by many as the father of Queueing Theory.

Inspired by the work done by A. K. Erlang’s the twenties and thirties of the last century shows a major thrust on this subject also partly motivated by the practical problem of congestion. During the next two decades several theoreticians became interested in these problems and developed general models which could be used in more complex situations. Some of the authors with important contributions are Crommelin, Feller, Jensen, Khintchine, Kolmogorov, Palm, and Pollaczek. A detailed account of the investigations made by these authors may be found in books by Syski (1960) and Saaty (1961). Kolmogorov’s and Feller’s study of purely discontinuous processes laid the foundation for the theory of Markov processes as it developed in later years.

Noting the inadequacy of the equilibrium theory in many queue situations, Pollaczek (1934) began investigations of the behavior of the system during a finite time interval. Since then and throughout his career, he did considerable work in the analytical behavioral study of queueing
systems Pollaczek (1965). The trend towards the analytical study of the basic stochastic processes of the system continued, and queueing theory proved to be a fertile field for researchers who wanted to do fundamental research on stochastic processes involving mathematical models. A concept that plays a significant role in the analysis of stochastic systems is statistical equilibrium.

Even though Erlang did not explicitly state his results in these terms, he used this basic concept in his results. To this day a large majority of queueing theory results used in practice are those derived under the assumption of statistical equilibrium. Nevertheless, to understand the underlying processes fully a time dependent analysis is essential. But the processes involved are not simple and for such an analysis sophisticated mathematical procedures become necessary. And hence the growth of queueing theory can be traced on using existing mathematical techniques or developing new ones for the analysis of the underlying processes; and incorporating various system characteristics to make the model closely represent the real world phenomenon.

In 1953, David G. Kendall introduced the A/B/C type queueing Notation. The theory of Continuous time storage models were initiated by P.A.P. Moran, J. Gani and Nu Prabhu during 1956-63. The book “queues Inventory and Maintenance” by Philip M. (Mc Co0rd) Mor,e was published in 1958 and is considered the first text book on queueing. Other early researchers on server interruptions include Cohen (1958) Miller (1960), Keilosn and Kooharian (1960)Avai-Itzhak and Naor (1961), Gaver (1962), seem to have been the first to consider retrial queues. The proof of little’s formula (so called because it was first proved by Hon little) was published in 1961. Mandelbaum and Avi-Itzhak (1961) introduced the concept of the split and match queue. The queues with rest to the server were analyzed firstly by Maryan Ovich (1962). The Muller (1964) is concerned with an M/G/1 priority queue with rest to server. D.R. Cox (1965) introduced the “Supplementary variables technique” (1965) to analyze the queue.
Aggarwal (1965) consider a queueing system in which the interruption in service, although random does not occur during the period of service but occurs only after the service in hand’s completed. The serve so provided has been named as intermittently available service. The use of queueing for computer performance evaluation began around 1970.

Cooper (1970) was the first to define the reaction type of disciplines of exhaustive service, here the exhaustive server means the server returns from vacation to find one or more customers waiting; he works until the system is empty. Robert cooper did some of the earliest work (1969, 1970) on polling models. In particular, the “vacation models” was introduce in and a special case of decomposition theorem was given. Yecliali (1975) and Shanta Kumar (1980) investigate M/G/1 queueing systems with exhaustive service. An inituitive explanation for M/G/1 queueing systems with exhaustive service was first given by Fuhrmann (1983). M/G/1 queues with variations are also analysed by authors such as Cooper (1981) Ali and Neuts (1984), Fuhrmann are Cooper (1984) and Shogan (1971). The Daniel (1985) also treats some queueing models with vacations to the server with the assumption that the serve takes rest either after serving K consecutive units or whenever the system becomes empty, which occurs first, in contract to exhaustive service rule.

Here we mention only a few papers and books that have made a profound impart in the direction of research in queueing theory.

1.3. System’s Characteristics

A Queueing system may be described as one having a service facility at which the customers arrive for service, whenever the demand for service exceeds the capacity of service, a queue, or a waiting line develops. The whole mechanism of a Queueing System can be described by the figure below.
In the figure above we can clearly see that customer(s) come from some source which may or may not be finite and forms a queue/ goes directly to the server, depending on whether the server(s) is (are) free. After getting serviced from any of the parallel servers the customer leaves the system. These systems are determined by the organization and performance of the facility on one hand and the arrival and behavior of the customers on the other. There are six basic characteristics, which described a queueing System as given by;

(i) The input or arrival pattern of the incoming units.

The input pattern means 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 groups or batches of fixed or random sizes. The inter-arrival time may be deterministic, so it is the same between any two consecutive arrivals, or
it may be stochastic, in this case 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. Sometimes an arrival may not join the queue being discouraged by the length of the queue or being debarred from joining the system because of space constraint. Again, the arrivals may occur from an infinite source or sometimes from a finite source, with the system units circulating in the system that is, machines coming for repair whenever they fail. The nature of arrival pattern is general stochastic and can be described in terms of probability distribution. Assuming that a successive customers or batches arrive at a time.

\[ 0 \leq t_0 < t_1 < t_2 < t_3 \ldots < t_n < \ldots \ldots \]\n
the inter arrival time \( U_r = t_r - t_{r-1} \), where \( r = 1, 2 \ldots \) are mutually independent identically distributed random variables. Then \( A_n = P_r(U_r \leq X) \) is called the inter-arrival time distribution is exponential. In this case

\[ A(X) = 1 - e^{-\alpha x} \]

Where \( \alpha \) is mean arrival rate. Another type of input process is what Cox and Smith (1961) refer as renewal process in which the inter-arrival times are independently, identically distributed with probability density function \( f(x) \). Again, input distribution may depend on output distribution. There are two types of arrival process:

(a) Static arrival process

(b) Dynamic arrival process

(a) Static Arrival Process

The control depends upon nature of arrival rate (random or constant). Thus to analysis the queueing system it is necessary to attempt to describe the probability distribution of arrivals. From such distributions we obtained average time between successive arrivals, also called inter-
arrival time (time between two consecutive arrivals) and the average arrival rate (i.e. number of customers arriving per unit of time at the service system).

(b) **Dynamic Arrival Process**

It is controlled by both service facility and customers. The service facility adjusts its capacity to match changes in demand intensity, by either varying the staffing levels at different timings of service, varying service charges (such as telephone call charges at different hours of the day or week) at different timings or allowing entry with appointments.

(ii) **Service pattern of servers.**

Service pattern can be measured by the number of customers served per some unit of time or the time taken to complete the service. Here again this time may be constant (deterministic) or it may be stochastic. If it is stochastic, the probability distribution associated with it will be required. The service can be provided in single or batch. If it is batch, as in the case of arrival the batch size can be fixed or random. Service can again be state-dependent and also stationary and non-stationary with respect to time. Service facility may consist of one server or several ones. The capacity of server may be one or a précised signed number $b$ (i.e. it takes for service $b$ units or the whole queue 1 length while ever is less; whenever the server becomes free) or may be governed by the probability distribution, prob (capacity is $j$) = $p_j$, 

$$\sum_{j=1}^{c} p_j = 1$$

Due to uncertainly involved in the length of the service, we can consider the successive service time as random variables, having a certain distribution. Thus if $V_1, V_2, \ldots$ be the successive service and assuming that service time are mutually independent identically distributed random variable then

$$B(X) = P[V_t \leq X]$$

is known as service time distribution.
(iii) **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 first in, first out, as it is sometimes called. 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 discipline 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.

(iv) **System capacity.**

It concerns with the physical limitation of the waiting room such as when the line reaches a certain queue length, no further customers are allowed to enter until space becomes available by a service completion. These are referred to as finite queueing situations. A system which has a limit to its capacity is also sometimes referred to as a **loss system**; as any incomer after the limit of the system will be considered being loss.

(v) **Number of service channels.**

A system may have a single server or a number of parallel servers. In the case of parallel servers an arrival who finds more than one free servers may choose at random any one of them for receiving his service. If he finds all the servers busy, he joins a queue common to all the servers. There can be single queue for multi-channel system or each channel having a separate queue for service. In the latter case, the system can be considered as a number of parallel single channel systems.

(vi) **Number of Service Stages.**

A queueing system may have only a single stage of service or it may have several stages. An example of a multi-stage queueing system is the physical examination procedure in a hospital.
**Notation**

The basic representation widely used in queueing theory is made up of symbols representing three elements: input/service/ number of servers. i.e. \( A/B/C: u/v/w \) where symbols are as follows:

A – Arrival time distribution

B – Service time distribution

C – No. of parallel service channel

u – Queue discipline

v – System capacity

w – Size of customer population

For instance, using \( M \) for Poisson or exponential, \( D \) for deterministic (constant), \( E_k \) for the Erlang distribution with scale parameter \( k \) and \( G \) for general (also \( GI \), for general independent). The table below explains some of the common notations that are being followed:

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Symbol</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interarrival-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>M</td>
<td>Exponential</td>
</tr>
<tr>
<td>distribution</td>
<td>D</td>
<td>Deterministic</td>
</tr>
<tr>
<td></td>
<td>E_k</td>
<td>Erlang type k (k= 1,2,3,..)</td>
</tr>
<tr>
<td></td>
<td>H_k</td>
<td>Hyperexponential type k</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>Phase type</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>General</td>
</tr>
<tr>
<td>Service-time distribution</td>
<td>M</td>
<td>Exponential</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>Deterministic</td>
</tr>
<tr>
<td></td>
<td>$E_k$</td>
<td>Erlang type $k$ ($k=1,2,3,...$)</td>
</tr>
<tr>
<td></td>
<td>$H_k$</td>
<td>Hyperexponential type $k$</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>Phase type</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>General</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of parallel servers</th>
<th>1,2,3,----</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restriction on system capacity</td>
<td>1,2,3,----</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Queue discipline</th>
<th>FCFS</th>
<th>First come first serve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LCFS</td>
<td>Last come first serve</td>
</tr>
<tr>
<td></td>
<td>RSS</td>
<td>Random selection of service</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>Priority</td>
</tr>
<tr>
<td></td>
<td>GD</td>
<td>General discipline</td>
</tr>
</tbody>
</table>

These symbolic representations are modified when other factors are involved. The above symbolic representation of the queueing system was introduced for the first time by David G. Kendall in 1953.

1.4. Operating Characteristics of a Queueing System.

Queueing models enable the analyst to study the effects of the decision variables on the operating characteristics of a service system. This operating characteristic of the system also
defines the measures of performances of the system. Some of the more common operating characteristics of the system are:

(i) The Expected Length of the System/Queue

The expected number of customers in the system/queue reflects one of the two conditions. Short queues could mean either good customer service or too many servers. Similarly long queue could also mean either low server efficiency or the need to increase the number of servers.

(ii) Expected Waiting Time in the System

A Long queue does not necessarily mean long waiting times always, if the service rate is fast. However, when waiting time seems long to customers, they perceive that the quality of service is poor. Long waiting time is an indication of a system in problem, such a system require an immediate attention to avoid the unnecessary crisis.

(iii) Busy/Idle Period of the Servers

Busy period of the system starts with the arrival of a customer in an empty system and end when the system becomes empty again for the first time. Long duration of busy period is an indication of shortage of server or the inefficiency of the existing server. In either case an immediate attention is required to address the existing problem. Whereas long durations of idle period are an indication of excess server, in this case an action from the system manager is required.

Symbols:

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Random Variable</th>
<th>Expected Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.5. Transient and Steady-State Behavior

The solutions of queueing system are obtained under two different assumptions. Firstly one assumes that the probabilities of interest are functions of the system time, such a solution is termed as the transient solution of the system, e.g. computation of \( P_n(t) \) the probability that there are ‘n’ units in the system at time \( t \). Assuming that the system is being exposed for a sufficiently long period of time and the system starts behaving as if it is independent of the system time, the solution so derived under this condition is termed as steady state solutions, e.g. \( P_n \), the probability that there are \( n \) units in the system regardless of the system time.

Given the transient solution of the system the steady state solution of the system can easily be derived by taking \( \lim_{t \to \infty} \), i.e.

\[
P_n = \lim_{t \to \infty} P_n(t)
\]

Either type of solution has its own significance, e.g. consider a CPU which allocates buffer in real time. We need the steady state solution of buffer contents to find the number of buffers required. And the transient solution will give the load on CPU for buffer allocation.

The overwhelming majority of work in queueing theory is concerned with the steady state probabilities rather than the time dependent behavior. These steady state solutions are simple in
form and easy to compute as compared to the time dependent solutions which are notoriously difficult.

In queueing theory, one typically concentrates on steady-state solutions, yet real-life systems are almost never in steady-state. In this sense, transient solutions are more realistic. One reason for using steady state results is that in many cases, they are easier to obtain than transient results. Also, the initial probabilities of being in the different states are often unknown. On the other hand, if these initial probabilities are known, and if the transient probabilities are not too difficult to obtain, one should use them rather than equilibrium probabilities. In particular, if rates are piecewise constant when considered as functions of time, and this is a reasonable approximation (Ingolfsson et al., 2007), then transient solutions should be considered. The distribution of the state variables after the rate change then provides the initial distribution for the new period. The main focus of our present works is on the transient solution of the Queueing system.

1.6 Different Method for finding the Transient Solution of Queuing System

The following are some of the main methods use by different mathematician for finding a transient solution to any queuing system;

i. Analytical Method.

ii. Continue fraction

iii. Uniformization Method and

iv. Combinatorial Method.

(i) Analytical Methods

A common approach in most of these methods is to formulate a set of differential difference equations governing that system and then solve the resulting system of equations through some technique. The following is the set of differential difference equations for M/M/1 queueing model or forward Kolmogorov equations for birth and death processes;
\[ P'_n(t) = - (\lambda + \mu) P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t); \quad n \geq 1 \]  \hfill (1.1)

\[ P'_n(t) = -\lambda P_0(t) + \mu P_1(t); \quad n = 0. \]  \hfill (1.2)

Where \( P_n(t) \) denote the number of units in the system at time \( t \). Solution (1.1) was for the first time given by Ledermann and Reuter (1956) through the spectral theory and by Bailey (1954) through the method of generating function. Morse in (1955) obtained \( P_n(t) \) as a correction term to the steady state probability Later using the difference equation technique this system was explicitly solved by Conolly (1958) and Karlin and McGregor (1958). Alternative method for solving (1.1) was proposed by Cox and Smith (1961). Further, Neuts (1964) developed a new approach avoiding the use of generating function and Rouches theorem. Further, Pegden and Rosenshine (1982) suggested the use of bivariate probabilities and Sharma (1990) extended this approach to finite capacity queue etc. Parthasarathy (1987) using a simple transformation had given another expression.

(ii) **Continue fraction**

Continue fraction approximation often provide good representations for transcendental functions that are much more generally valid than the classical representation by power series. They are employed in control theory to decide whether a polynomial with real coefficients is stable. All analytic functions have continued fraction representations. Among those functions which have fairly simple expansions are many of the special functions of mathematical physics. Other applications deal with analytic continuation, location of zeros and singular points, stable polynomials, acceleration of convergence, simulation of divergent series, asymptotic expansion and birth and death processes. A systematic study of the theory of continued fractions with stress on computation can be found in Jones and Thron (1980). Its application to the study of birth and death processes was initiated by Morphy and O’Donohoe (1975).

In particular, applications employing continued fractions have occupied a conspicuous place in mathematical literature due to their interesting convergence properties and also due to their connections with many branches of mathematics like number theory, special functions,
differential equations, moment problems, orthogonal polynomials and other. On account of their algorithmic nature they are used in numerical analysis, computer science, automata, electronic communication, etc., Their importance has grown further with the advent of fast computing facilities.

A continued fraction is denoted by

\[ f_0 = \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \cdots + \frac{a_n}{b_n} + \cdots \]

Where \( a_n \) and \( b_n \) are real or complex numbers. This fraction can be terminated by retaining the terms \( a_1, b_1, a_2, b_2, \ldots, a_n, b_n \) and dropping all the remaining terms \( a_{n+1}, b_{n+1} \). The number obtained by this operation is called the \( n^{th} \) convergent or \( n^{th} \) approximant and is denoted by \( \frac{A_n}{B_n} \). Both \( A_n \) and \( B_n \) satisfy the recurrence relation

\[ U_n = a_n U_{n-2} + b_n U_{n-1} \]  \hspace{1cm} (1.3)

With initial values \( A_0 = 1, A_1 = a \) and \( B_0 = 1, B_1 = b \).

(iii) **Uniformization Method:**

The uniformization technique, also known as the randomization technique and Jensen's method, is widely used to obtain \( \pi(t) \), transient probabilities for continuous-time Markov chains. This technique has also been used to calculate a variety of transient measures. It was first suggested by Jensen (1953), and for this reason it is often called Jensen’s method. In a Markovian event system, one has a number of variables, called state variables, which jointly determine the future of the system. A state is a particular assignment of values to all state variables. At certain rates, the state changes, and such changes are brought about by events. Markovian event systems are special cases of discrete event systems, with the restriction that they are Markovian. Specifically, the rate of an event depends only on the present state, and not on the past of the system. This means that there is no need to schedule events, which makes the logic easier. It is relatively easy to convert Markovian event systems into continuous-time Markov chains.
For finding the transient probabilities of a Markov chain with state space $S$, we need the initial probability vector $\pi(0) = [\pi_i(0), i \in S]$ as well as the transition rates $q_{ij}, i, j \in S, i \neq j$. Also, $q_{ii}$ can be obtained by multiplying $q_i$ of leaving state $i$ by -1, that is, $q_{ii} = -q_i$. Let $Q = [q_{ii}, i, j \in S]$. The first step of uniformization is to find the maximum leaving rate $q = \max(q_i)$, and Uniformize $Q$ according to the formula

$$P = \frac{Q}{q} + I$$

Since the matrix $P$ is the transition probability matrix no elements of the matrix has negative entries, and the sum of each row of $P$ is equal to 1, at least provided the rows of $Q$ all have a sum of 0. Let $\pi^n = [\pi_i^n]$ denote the probabilities of being in state $i$ after $n$ steps of this discrete Markov chain. We can find $\pi^n$ recursively as;

$$\pi^0 = \pi(0)$$

$$\pi^{n+1} = \pi^n P, \ n \geq 0$$

For finding $\pi(t)$, we use the concept of Poisson distribution with parameter $qt$, given by

$$p(n; qt) = \frac{e^{-qt}(qt)^n}{n!}.$$ The transient solutions for time $t$ are now:

$$\pi(t) = \sum_{n=0}^{\infty} \pi^n p(n; qt)$$

(iv) Combinatorial Methods

A different class of methods namely combinatorial methods were also proposed for finding solutions of queuing systems in finite times. In these methods sequence of arrivals or departures or alternatively inter-arrivals or service times are considered. Chmpernowne (1956) had used combinatorial method (Random Walk Method) for the first time to find the solution of the single server Markovian Queueing system, where he considered the virtual departures time and used the
reflection principle. It was because of the contribution by Takacs (1962), this method has gain its popularity among the researchers. He then employed this method to investigate queueing system the use of the well-known ballot theorem Takacs (1962, 1967).

1.7 Various queueing system are given below without being categorized

(i) Queueing networks.

Networks of queues are systems which contain an arbitrary, but finite, number m of queues. Customers, sometimes of different classes, travel through the network and are served at the nodes. The state of a network can be described by a vector, where \( k_i \) is the number of customers at queue \( i \). In open networks, customers can join and leave the system, whereas in closed networks the total number of customers within the system remains fixed. The first significant result in the area was Jackson networks, for which an efficient product form equilibrium distribution exists. Traffic processes in computers and computer networks have necessitated the development of mathematical techniques to analyze them. The first article on open queueing networks is by J.R. Jackson (1957). Complex queueing network problems have been investigated extensively since the beginning of the 1970’s. Several survey papers and books summarize the major contribution made in this area. These include Baskett (1973), Kelly (1979), Disney and Kiessler (1987), Ananthram et al. (1993), and Dai (1998).

(ii) Polling models

These models represent systems in which one or more servers provide service to several queues in a cyclical manner. i.e. the server visits the \( M \) classes in a fixed order, which can be labeled 1, 2, ............... \( M \), without loss of generality. Application of polling system occurs in many situations. The term “polling system” apparently originated in the telephone industry where several applications are. eg. A switch may poll each of input channels to see if there is incoming traffic or a telephone handset may pole to line to determine whether the number is being dialed. Based on variations on the system structure and queue discipline a large number of models
emerge. Beginning in the early 1980’s there was an explosive growth of research on polling systems, motivated apparently by the increasing number of applications to computer and communications system. For research on polling models see a special issue of queueing Systems edited by Boxma and Takagi (1992), as well as Takagi (1997) and Hirayama et al. (2004), all of which provide excellent bibliography on the subject.

(iii) Vacation model

Vacation queueing theory was developed in the 1970's as an extension of the classical queueing theory. In a queueing system with vacations, other than serving randomly arriving customers, the server is allowed to take vacations. The vacations may represent server's working on some supplementary jobs, performing server maintenance inspection and repairs, or server's failures that interrupt the customer service. Furthermore, allowing servers to take vacations makes queueing models more flexible in finding optimal service policies. Therefore, queues with vacations or simply called vacation models attracted great attentions of queueing researchers and became an active research area. The method of analysis are similar to those used for polling system, but because of the focus on one class, none general results have been obtained. Balachandran (1973, 75), Heyman and Sobel (1982) introduced N-policy. Here each time when system becomes empty, the server wait until exactly N-customers are accumulated. One of the most significant results of the research on vacation systems has been the discovery that the waiting time in the queue M/G/1 queue with vacations is distributed as the sum of two independent components, one distributed as the waiting time in queue in the corresponding M/G/1 queue without vacations and other as the equilibrium residual time in vacation. Many studies on vacation models were published from the 1970's to the mid 1980' and were summarized in two survey papers by Doshi and Teghem, respectively, in 1986. Stochastic decomposition theorems were established as the core of vacation queueing theory. In the early 1990' Takagi published a set of three volume books entitled Queueing Analysis.
(iv) Bulk Queues

A queueing system where arrivals or service or both takes place in batches of fixed or random sizes is called a bulk queueing system. The literatures on bulk Bailey (1954) introduced the concept of bulk queues; he considered a situation where service can be given in a batch of up to X customers—that is, all waiting customers up to a fixed capacity X are taken for service in a batch. Graver (1959) introduced bulk-arrival queues, where arrival could be in bulk or batch. The literature on bulk queues with bulk arrival and/or with bulk service is now quite vast. Chaudhry and Templeton (1983) discussed this subject at great length.

A batch may contain a minimum number of ‘a’ units and a maximum of ‘b’ units. If immediately after completion of service of a batch, the server finds less than ‘a’ units present, he waits till there are ‘a’ units, he finds more than ‘a’ units waiting, he takes in the batch of ‘b’ units for service. This is known as general bulk service rule. Neuts (1967), Kumar and Tuteja (1984) and other considered this type of queue.

(v) Non-poisson Queues

When either the inter-arrival time or the service time or any other distribution involved in the system is non-exponential, then the queues so developed are called an non-Poisson queues.

(v) Phase Technique

This technique is essentially due to Erlang, although modification has been made by various authors. We assume that service on a customer consist of k imaginary phases, which are mutually independent and exponentially distributed with the common expected sojourn time 1/μ in any of the phase. A customer on arrival phases in sequence though all the k-phases before it is discharged. After a customer leaves the sever, a new customer, is taken up instantaneously if one is waiting in the queue, otherwise the server remain idle. It has been demonstrated by Gaver (1954) and latter by Luchak (1956) that it is possible to obtain a wide class of service time distributions of practical interest by varying Cj.
(vii) Imbedded Markov Chain Technique

Consider a physical system which is observed at a discrete set of points 0, 1, 2, ..... Let the successive observations be \( N_0, N_1, N_2, \ldots \), where \( N_n \) is a random variable. Further assume that each of these variable is capable of taking the value 0, 1, 2, ..... The sequence \( \{N_n, n \geq 0\} \) is said to form a Markov chain. If for all \( n, n = 1, 2, 3, \ldots \), the conditional distribution of \( N_n \), for given values of \( N_0, N_1, \ldots N_{n-1} \), depends only on \( N_{n-1} \), the most recent known value. In symbols these becomes:

\[
P\{N_n = k / N_0 = i_0, N_1 = i_1, \ldots N_{n-1} = i_{n-1}\} = P\{N_n = k / N_{n-1} = i_{n-1}\}
\]

(1.5)

where \( i_0, i_1, i_2, \ldots, i_{n-1}, k \) takes the value 0, 1, 2, 3,....

The system is said to be in state \( k \) at the \( n \)th point if \( N_n = k \). Also if \( N_{n-1} = i \) and \( N_n = k \), the system is said to have made a transition \( i \rightarrow k \) at the \( n \)th time point. The conditional probability \( P\{N_n = k / N_{n-1} = i\} \) are called one-step transitional probabilities. These probabilities may depend on \( i, k, n \) and \( n-1 \). However, if they are independent of \( n \), the discrete – time variable, then the Markov chain is said to be homogeneous. In this case, the transitional probability may be denoted by

\[
P_{ik} = Pr\{N_n = k / N_{n-1} = i\} \quad i, k \geq 0.
\]

In this case of a homogeneous, Markov chain, the \( n \)-step, transitional probability are denoted by

\[
P_{ij}(n) = P\{N_{m+n} = k / N_m = i\}
\]

(1.6)

and are independent of \( m \).

This method was developed by Kendall (1951). He extracted a Markov Chain from the non-Markovian queueing process \( M/G/1 \) by taking the epoch of departures as the regration points. This was the first systematic approach in the direction of solving queueing problems with general distribution.

(viii) Supplementary Variable Technique
This technique was used in (1942) by Kostern. The name appears to be due to Cox (1955), who used a supplementary variable to study the queueing system M/G/1. Kendall (1953) considered this technique, which he labelled “augmentation” but preferred the use of an impelled Markov Chain as leading to simpler calculation. In spite of this, supplementary variable have been used by many authors to solve number of queueing problems. Jaiswal’s (1968) used this technique in his book, while kendall’s imbedded Markov chain technique is very powerful and elegant. It gives only approximate solution to queueing.

Problems considered in continuous time. Even in steady state case, many problems are more readily treated by supplementary variable technique then by the imbedded Markov chain technique. Regies (1973) obtain waiting time distribution in queue with poisson input, by the supplementary variable technique.

(ix) Cost Analysis and Optimal Control

In any system where the service facility has a large capacity that queue forms, then the facility is likely to remain idle and so the unused capacity could exist. On the other hand, the unending long queues will lead to the loss of goodwill of customers. Queueing theory method help the executives to maintain a suitable balance between these two extremes by providing a proper aim can be achieved by first studying the model analytically and then optimizing the objective function by superimposing a suitably chosen cost or profit function. The validity of the results can further be judged by empirical studies.

Earlier work in this direction was carried but by Morse (1958) who considered M/M/1 and M/M/C models with limited and unlimited space. Hiller and Lieberman (1967) analyzed the optimization problem for the multi-channel in finite source of the systems. Taha (1971), Katiah and Slater (1973), Balachandran and Schaefer (1980), Shantikumar (1983) and others have considered various cost structure.

(x) Control Policies in Queues

In general the following two assumptions are made 3 for a queueing system:
(1) The idleness of the server will be interrupted as soon as a unit arrive at a facility.

(2) The server becomes idle only when the system is empty.

According to Yadin and Naor (1963), in the single server case with stationary position arrival and generally distributed service times expected idle fraction is of order $30 - 40$, if the queue length is small as one. Hence the utilization of server fraction is considered a very important concept.

Considering these facts and relaxing the assumption (i) they also studied a queueing system wherein the idleness of the server terminates only when a queue of certain size is built up and a service facility. Heyman (1968), Sobel (1969), Bell (1971), Jaiswal and Simha (1972) and others also studied many queueing system under the doctrine dismantle the service station when the system becomes empty on a service completion and set up it again when N units form a queue.

The on-off type policy in a different form was considered by Balachandran was the total work load –D. Kumar and Tuteja (1984) studied a policy which is independent of arrival pattern and named it as Random Idle Policy. The doctrine is-the server goes to idle star for a randomly distributed period. By taking idle period distribution as exponential/general, they studied M/M/1 and M/G/1 queueing systems considering the start up cost, shut down cost and holding cost. They found optimal value of the average idle period. Gupta and Kumar (1984) also studied this queueing system for general bulk service rule. Erlangian input, queue was studied by Chitkara and Gupta (1984). Other queueing system with this policy and involving the concepts of an additional channel, priority rule and bulking have been studied by Gupta (1984) and Chitkara (1985).

Another policy called Random-N-policy was studied by Yadav and Kumar (1980), Arora and Tuteja (1989). Here the idleness of the server interrupted when a queue of fix size built or the ransom time was elapsed whichever occur earlier. This policy was further studied by Arora and Tuteja (1990) for Erlangian input and additional channel queue. In all these policies, it is
assumed that the service facility is available as soon as the decision to switch on the server is taken. Assuming a time gap between decision of the management and actual starting of service.

**(xi) M/M/1 Queue**

The equation of M/M/1 system may be obtained from the equation of birth and death process. The Kolmogorov equations governing the system are

\[
\frac{d}{dt} P_n(t) = -(\alpha + \beta) P_n(t) + \alpha P_{n-1}(t) + \beta P_{n+1}(t), \quad n > 0
\]

\[
P_0(t) = -\alpha P_0(t) + \beta P_1(t)
\]

where \( \alpha, \beta \) are the mean arrival and service rates and \( P_n(t) \) is the probability that there are \( n \) units in the systems at time \( t \). The steady states solution is given by

\[
P_n = (1-\rho)^n \rho^n, \quad \rho < 1, \quad n = 0, 1, 2, 3, \ldots
\]

where \( \rho = \alpha/\beta \), is the traffic intensity.

The time dependent solution of M/M/1 model was first given by Bailey (1954), Clarke (1956) solved this system may by the method of generating functions. Conolly (1958) used different equation technique. Heathcote and Moyal (1959) gave the transient solution of M/M/1 queue when the waiting space is limited. For M/M/1/N system, Scott and Ulmer (1972) Use the Laplace transform for obtaining the generating function for the time dependent, joint distribution of the number of customers in the system at time \( t \) and the number of customers served in the time interval \((0, t)\). Conolly (1975), Conolly and Chan (1977) dealt with the generalized birth and death queueing models with parameters \( \alpha_n \) and \( \beta_n \). Eisen and Tuteja (1979) studied heterogeneous queue on the assumption that the arrival and service intensities are subject to Poissonian jumps between two different states. Kashyap (1965a, b; 1966), Srivastava and Kashyap (1982) considered the double-ended queue of customers (passengers) and server (taxis) at a taxi stand.

**1.8 Stochastic Process**

A Stochastic process \( N(t) \) has three elements
(i) The state space
(ii) The parameter set T
(iii) The dependent relations among the random variables N(t) when t varies in T

The state space s, which is the set of values which may be taken by N(t) for typical values of t \( \in \) T, could be discrete or continuous, scalar or vector. This parameter T could also be could be discrete or continuous, scalar or vector. While in queueing theory t is normally a time parameter. The arrival process, the distribution of system length and so on are further examples of stochastic process.

1.9 Markov Processes

A stochastic process is called Markov process if the future development of the process depends present and is independent of the past history of the process: More precisely, if for \( n > 0 \) and any set of n points \( 0 < t_0 < t_1 < t_2 \ldots < t_n \) in T,

\[
P_r \{ x(t_n) = X_n / x(t_{n-1}) = X_{n-1}, x(t_{n-2}) = X_{n-2} \ldots \ldots \ldots \ldots x(t_1) = X_1 \} = P_r \{ x(t_n) = X_n / x(t_{n-1}) = X_{n-1} \} \tag{1.10}
\]

For a discrete – state Markov process, often called a Markov chain, the departure take the form of transition probabilities \( P_{ij}(t, s) \) defined by

\[
P_{ij}(t, s) = P\{ (N(t + s) = j / N(s) = i) \quad i, j \in S, \, s \geq 0 \} \tag{1.11}
\]

We will usually consider the special case of a homogeneous Markov chain, in which \( P_{ij}(t, s) \) does not depend on the initial epoch \( s \in T. \) The transition probabilities \( P_{ij}(t, s), \) then may be written as \( P_{ij}(t). \) \( P_{ij}(t) \) satisfies the following conditions

1) \( P_{ij}(t) \geq 0 \quad t > 0 \)

2) \( \sum_{j=1}^{n} P_{ij}(t) = t \quad t > 0 \)

3) \( \sum_{k=1}^{n} P_{i}(t)P_{kj}(h) = P_{ij}(t + h) \quad t, h > 0 \)
Parzen (1962), Kerlin and Taylor (1976), Cinlar (1975), have treated stochastic process.

### 1.10 Poisson Process

A stochastic process in continuous time and with discrete state space is said to be poisson process under the following postulates:

(i) The probability that an arrival occurs between time \( t \) and \( t + \Delta t \) is equal to \( \lambda \Delta t + o(\Delta t) \).

(ii) \( P(\text{more than one arrival between } t \text{ and } t + \Delta t) = o(\Delta t) \)

(iii) The numbers of arrivals in non-overlapping intervals are statistically independent; i.e. the process has independent increments.

Under the above assumptions, if \( X(t) \) denotes the number of arrivals in the time interval \([0, t]\), then \( X(t) \) follows Poisson distribution with parameter \( \lambda t \), so

\[
P(X(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k = 0, 1, 2, 3, \ldots
\]
gives the probability of \( k \) arrivals taking place in the time interval \([0, t]\). The poisson process is widely used in the development of queueing models. This is due to the fact that it enjoys many marvelous probabilistic properties. In addition, there are many real life situations which can be exhibited by this process.

**Birth and Death Process**

The birth and death process can always be viewed as the generalization of the Poisson process. Assume that for a small interval of time \( \Delta t \), we need only consider that at most one unit enters or leaves the system and that

\[
P_{n,n+1}(\Delta t) = \alpha_n \Delta t + o(\Delta t), \quad n = 0, 1, 2, \ldots
\]

\[
P_{n,n-1}(\Delta t) = \beta_n \Delta t + o(\Delta t), \quad n = 1, 2, 3, \ldots
\]

so that

\[
P_{n,n}(\Delta t) = 1 - P_{n,n-1}(\Delta t) - P_{n,n+1}(\Delta t)
\]

\[
= 1 - (\alpha_n + \beta_n) \Delta t + o(\Delta t), \quad n = 1, 2, \ldots
\]

\[
P_{0,0}(\Delta t) = 1 - P_{0,1}(\Delta t)
\]
\[= 1 - \alpha_0(t) + o(\Delta t) \ldots \] (1.15)

where \(\alpha_n, \beta_n \geq 0\) for all values of \(n \geq 1\).

Forward Kolmogonov equations becomes

\[
P'_0(t) = -\alpha_0 P_0(t) + P_1(t) \tag{1.16}
\]

\[
P'_n(t) = -(\alpha_n + \beta_n) P_n(t) + \alpha_{n-1} P_{n-1}(t) + \beta_{n+1} P_{n+1}(t), n \geq 1 \tag{1.17}
\]

The steady state solution of forward Kolmogorove equations (1.16) and (1.17) is given by

\[
P_n = P_0 \prod_{i=1}^{n} \frac{\alpha_{i-1}}{\beta_i}
\]

Since \(\sum_{n=0}^{\infty} P_n = 1\), the necessary and sufficient condition for the existence of steady state solution is the convergence of infinite series \(\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\alpha_{i-1}}{\beta_i}\) by giving different values to \(\alpha_n\) and \(\beta_n\), birth death process, a special type of Markov chain in continuous time gives rise to various Markovian queueing models.

1.11 Mathematical Concepts

The following are the some of the mathematical functions and their properties that will be utilized in the coming chapters.

**Random Variable**

Random variables are denoted by capitals, \(X, Y\), etc. The excepted value or mean of \(X\) is denoted by \(E(X)\) and its variance by \(\sigma^2(X)\) where \(\sigma(X)\) is the standard deviation of \(X\). This random variable plays a very important role in the study by applied probability theory.

**Generating function**

Let \(X\) be a non-negative discrete random variable with \(P(X = n) = p(n), n = 0, 1, 2, \ldots\). Then the generating function \(P_X(z)\) of \(X\) is defined as;
\[ P_X(z) = E(z^X) = \sum_{n=0}^{\infty} z^n p(n) \]

Note that \(|P_X(z)| \leq 1\) for all \(|z| \leq 1\). Further \(P_X(0) = p(0), P_X(1) = 1\) and \(P'_X(1) = E(X)\), more general, \(P_X^k(1) = E(X(X-1)(X-2)\ldots(X-k+1))\)

where the superscript \((k)\) denotes the \(k^{th}\) derivative. For the generating function of the sum \(Z = X + Y\) of two independent discrete random variables \(X\) and \(Y\), it holds that \(P_Z(z) = P_X(z) P_Y(z)\).

**Exponential Distribution**

A continuous random variable, \(X\) is said to follow an exponential distribution with parameter, \(\lambda > 0\) if its probability density function is given by,

\[ f(x) = \lambda e^{-\lambda t}, t > 0 \]

then

\[ \text{Mean} = E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Variance} = \frac{1}{\lambda^2} \]

An important property of an exponential random variable \(X\) with parameter \(\lambda\) is the memoryless property. This property states that for all \(x \geq 0\) and \(t \geq 0\),

\[ P(X > x + t| X > t) = P(X > x) = e^{-\lambda t} \]

Exponential distribution is the only continuous distribution with this property. There is a very important relationship between Poisson distribution and exponential distribution in the queueing context, which states that, “if the arrival processes follows Poisson distribution then the inter-arrival time follows exponential distribution and vice versa”.

**Laplace Transform:**
The Laplace transform (L.T.) of a function $f(x)$, $t \geq 0$, will be denoted by $\hat{f}(s)$ and is defined as,

$$L(f(x)) = \hat{f}(s) = \int_0^\infty e^{-st} f(t) \, dt$$

A sufficient condition for the existence of the above integral is that $f(x)$ must be of bounded exponential growth, which means that there are numbers $\lambda$ and $\mu$, such that $\forall \ t > 0, \ |f(t)| < \lambda e^\mu t$.

If two function $f_1$ and $f_2$ have the same values for all $t$ except at most on a set of points with no finite point of accumulation.

The following are some properties of Laplace transforms:

<table>
<thead>
<tr>
<th>Expression</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L(af(t) + bg(t))$</td>
<td>$\hat{a}f(s) + \hat{b}g(s)$ (additive property)</td>
</tr>
<tr>
<td>$L(f'(t))$</td>
<td>$sf(s) - f(0)$ (derivative)</td>
</tr>
<tr>
<td>$L\left(\int_0^t f(u) , du\right)$</td>
<td>$\frac{1}{s} \hat{f}(s)$ (Integral)</td>
</tr>
<tr>
<td>$L\left{\int_0^t f(t-u)g(u) , du\right}$</td>
<td>$\hat{f}(s)\hat{g}(s)$ (Convolution)</td>
</tr>
<tr>
<td>$L\left{e^{at}f(t)\right}$</td>
<td>$\hat{f}(s-a)$</td>
</tr>
<tr>
<td>$L\left{\frac{1}{t}f(t)\right}$</td>
<td>$\int_0^\infty \hat{f}(y)dy \quad \text{x}$</td>
</tr>
<tr>
<td>$L{tf(t)}$</td>
<td>$-\hat{f}(s)$</td>
</tr>
</tbody>
</table>

**Thesis at Glance**

A queue forms and grows when the units demanding service arrive at a service facility which provides the service they seek with a rate greater than the number of customers being served and leaving the queue at that point of time. A system, in which arrivals place demands upon a finite capacity, may be termed as a queueing system. If the arrivals times or the size of these demands
are unpredictable then the conflicts for the use of resources will arise and queue of waiting customers will form. If the rate of arrival is less than the queue will be reduced or may disappear altogether. The units demanding service are called customers. The service facility at which or the person by whom these customers are served is called “channel” or servers. Thus a queue can be regarded as a buffer between an input process (arrival) and an output process (service). Today, Queueing theory, the mathematical study of waiting lines (or queues) is a very important branch in applied stochastic process. The theory enables mathematical analysis of several related processes, including arriving at the queue, waiting in the queue, and being served by the server(s) at the front of the queue. The theory permits the derivation and calculation of several performance measures including the average waiting time in the queue or the system, the expected number waiting or receiving service and the probability of encountering the system in certain states, such as empty, full, having an available server or having to wait a certain time to be served.

Furthermore, allowing servers to take vacations makes queueing models more flexible in finding optimal service policies. Therefore, queues with vacations or simply called vacation models. In our present work we investigate some of queueing models with system breakdown for non-reliable server. A non-reliable server means that a server is typically subject to unpredictable breakdowns. The joint distribution of arrival and service is studied for the multichannel queueing models with controllable arrival rates. The behavior of many characteristics is studied numerically and graphically.

The present thesis is divided into six chapters and some chapters are subdivided into sections depending on the diversity of its subject matter.

**Chapter – I**

Chapter – I is introductory in nature and includes the fundamental concepts, origins, history, definitions and various tools with emphasis on Queueing Theory and also explains various terms related to the subsequent chapters of the thesis. Some of the mathematical concept such as
Poisson process, stochastic process, Laplace Transformation etc. that have been utilized in the present work is also given.

**Chapter – II**

In this chapter, we discuss about the optimal operation of a single removable and non-reliable server in a markovian queueing system under steady state condition. In general the service facility in a queuing system is assumed to provide uninterrupted service to the customer unit. There are however, many real-life situations where the service gets interrupted due to breakdown in service mechanism. Therefore, the breakdown period is the length of time per cycle when the server is turned on found to broken down and customer is waiting for their service. Such type of breakdown is common in industries, telephone traffic, communication system, etc.

**Chapter – III**

In previous chapter, we discussed about the optimal operation of a single removable and non-reliable server in a markovian queueing system under steady state condition. In this chapter we extend those results by incorporating the concepts of bulk arrival and service provided to them in batches. This chapter is divided into two sections.

In section-I, we consider that there is bulk arrival of the customer.

In section-II, there is bulk service queue with arbitrary service time is considered.

Interarrival time and the service time distribution of the customer are assumed to be exponentially distributed. Operational characteristics for number of customers in the queue/system when system is in idle state, working state, broken state are determined. Optimum N-policy is also determined for the busy period when system is in working condition.

**Chapter – IV**

We extend the result of chapter-III in the way; that the customers arrive in batches of variable sizes and also serve in random size. We have considered a bulk arrival, bulk service queueing model with removable and non-reliable server in a markovian queueing system. We derive the time dependent solution of mean number of arrivals and departure. Such situations are not
uncommon in our daily life like horses in their stable, students in their class rooms before starting the class and after ending the class. The total expected cost function per unit time is developed to obtain optimal operating policy at minimum cost.

**Chapter – V**

In this chapter, we have studied a multi – server finite capacity queueing models with controllable rates. The arrival and servive processes are correlated and follow a bivariate poisson process. If the arrival customer does not receive any service, he will wait only for some time or limited time, on the expiry of which it is lost. Here there will be two arrival rates; faster arrival rates or slower arrival rates. Probability for the faster rate and slower rate of arrival can be determined. Mean queue length, Expected waiting time for the customer in the system can also be determined.

**Chapter – VI**

In this chapter the result of the previous chapter can be extended by incorporating the concepts of bulk arrival and service in multi-server finite capacity queueing models. If an arrival customer does not receive any service, he will wait only for some time or limited time, on the expiry of which it is lost e.g. In network systems when packet arrives at a node it is buffered in a queue before being processed. All packets have got a field called time to live. When the field reaches zero value, the packet is ignored by the node so the sender should resend it again. Here the result of chapter-V, is extended for a bulk arrival queueing system all the other assumptions are same as chapter V. We investigate the Mean queue length of system, Expected waiting time, Conditional probability of the system. Such situations are held in exams. Such system must have practical example in sugar cane industry.

**Bibliography**

This section contains a list of all the references which has been used in writing the thesis.