Chapter 1
General Introduction

1.1 Motivation

The queue is a congestion situation, wherein the customers arrive for service. In queueing system, the customers wait for service when service system is busy and leave the system after being served. The performance of the queueing system depends upon various characteristics of the system, e.g. the manner in which customers arrived, the performance of the service facility, service discipline, number of service station, etc. A queue is formed when the number of customers is more than the number of servers in the system. Consider a hospital, wherein all outdoor patients are checked by a number of doctors; the patients have to stand in a queue if the number of patients is more than the number of doctors. If the number of patients is less than the number of doctors then some doctors will be idle. When patients have to stand in a queue they face the inconvenience; on the contrary, if doctors stay idle, it is the wastage of resources. In such situation, it is an endeavor of the hospital management to find the optimum number of doctors to be appointed in the hospital. The queueing theory deals with the queueing system in order to predict its performance indices and also sometimes to optimize the resources of the system. Along with simple applications of queueing theory, there are special applications in predicting the performance and designing of manufacturing and production systems, multi-user time shared computer systems, communication networks, and data processing, etc.

There are two main approaches for performance evaluation of a system, viz. measurement approach and mathematical modelling approach. In measurement approach, the study of the performance of the system is possible only when a system is designed and is in working condition, on the other hand in mathematical modelling approach, the study of the performance of the system is carried out in advance before actually developing it by analysing a suitable mathematical model for the system. The information obtained from the analysis of the mathematical model can be used for increasing the efficiency of the system by improving the design of the system. In general, the arrival process, i.e. the pattern, in which the customers join the system, and the service process, i.e., the way customers are being served in the system, follow
random behaviour and thus stochastic models are being used to study the queueing system. Further, in some situations, the arrival rate and the service rate remain constants, i.e. does not depend on the state of the system and in some situations arrival rate and service rate depend on the current state of the system. Mathematical models of the queueing system also depend on the situation, wherein at a time either one customer joins the system or the customers join the system in bulk. When the customer joins the system in bulk, then queueing system is called as bulk arrival queueing system where bulk size may be fixed or may be random in nature.

In many congestion situations of the queueing system, the arrival rate and service rate depend on the present state of the system. For example, the queue size affects the efficiency of the server; server with good experience may act fast, in case of long queue that result in increase of service rate. Similarly, some servers act slowly under the pressure of the long queue size which decreases the service rate. There are two types of queueing systems; in some systems, the service terminates at any specified time whereas in the other, service continues till the queue becomes empty. The speed of service providing by the server may depend on the present load with a burden of the arrived units. This criterion is very much applicable for the service system having a human being as server and can be seen in the production system, where the service rate of the server is relatively less when there is much burden of the workload or when there is very slow rate of the work. Another example of the state dependent system is commonly realized when the decision-maker provides additional service facility on observing long queue to reduce the congestion in the systems.

Due to the complexity of the analytical solution, the queue with the state dependent arrival of units and general service time has not received the considerable attention so far. The applications of state dependent inter-arrival time of the units and state dependent general service time can be observed in various daily life congestion situations. In the present thesis, we have presented the stochastic modelling of state dependent queueing systems with different features namely general service time distribution, unreliable server, server vacation, $N$-policy, bulk arrival, service/repair in different phases and discouragement factors balking/reneging. The layout of this chapter is as follows: In section 1.2, we discuss the various characteristics of the queueing system, basic concepts of queueing theory and some classical queueing systems. The methodology used to investigate the stochastic models is discussed in section 1.3.
Review of relevant literature is given in section 1.4. Finally, the section 1.5 deals with the chapter wise summary of the present thesis.

1.2 Some Aspects of Queueing Theory

1.2.1 Characteristics of queueing system

The queueing system is described by the basic characteristics as follow:

(i) Arrival process

The arrival process depicts the manner in which customers arrive and join the system. Usually the inter-arrival times are assumed to follow a common distribution and are independent of each other. If customers arrive at the instants \( t_1, t_2, \ldots, t_n, \ldots \) then the inter-arrival times \( u_n = t_{n+1} - t_n \forall n = 1, 2, \ldots \) are normally assumed to be independent and identically distributed positive random variables.

(ii) Service process

The service process is the manner in which the service is rendered. It is normally assumed that the service times follow a common distribution which is independent of each other and independent of the inter-arrival times. In queueing system, the service times are usually modelled by exponential distribution/general distribution with constant mean service time or state dependent mean service time.

(iii) Service discipline

The service discipline specified the manner in which the units are picked from the queue for service. The commonly used disciplines are:

FIFO (First in, first out): In the first in first out discipline, the customer who enters first in the system will be served first.

LIFO (Last in, first out): In this discipline, the customer who arrives last in the system is served first, given that no further customers arrive in the system before his service get started.

SIRO (Random service): The customers in the queue are served according to some random phenomena.
Priority disciplines: In many queueing situations, some units get priority in service over others units irrespective of their order of arrival to the system. There are two broad categories of priority disciplines (i) preemptive priority according to which the unit with highest priority is allowed to join the service immediately by stopping the service of the lower priority unit and (ii) non-preemptive priority according to which the highest priority unit is allowed only to go to the head of the queue.

(iv) Number of servers

A system may have a single server or multi-server for providing the service to the customers. In a multi-server system, it is generally assumed that servers operate in parallel and independently to each other. If all the servers are busy at the time of arrival of a customer then he has to join the common queue for all the customers.

(v) System capacity

It is the maximum number of the customers that can be accommodated in the queue at any time. A system may have an infinite or finite capacity. In the infinite case, the queue in front of the server(s) may grow to any length and when there is a space limitation or other resources limitation; system can be considered with finite capacity. In the finite capacity system, a newly arrived customer can not join the system when the system has already attained its maximum capacity.

(vi) Service stages

In the queueing system, service may be rendered in single stage or in multi-stage in a sequential manner. In multi-stage service system, the customer enters a queue and waits for service at first stage, gets served and departs the service station to enter a new queue for second stage service, and so on. Such situations can be observed in manufacturing/production processes where defective parts are sent back for reprocessing unless that meets the desired quality standards.

1.2.2 Basic concepts of queueing theory

In stochastic modelling of the queueing systems, it is the common assumption that the inter-arrival times and service times of the customers follow the exponential
distribution. There are various situations that exist in the real life activities, wherein the arrival and service pattern of the customers follow the arbitrary distribution. In this section, some basic concepts of the queueing theory are briefly defined as follows:

(i) Stochastic process

A stochastic process is defined as a family of random variables \( \{X(t), t \in T\} \), where \( T \) is an index set. A stochastic process \( \{X(t), t \in T\} \) is said to be stationary if the entire family of its finite dimensional distribution is invariant under a translation in \( t \). i.e. for given \( t_1, t_2, ..., t_n \) time points, the distribution of \( X(t_1 + \tau), X(t_2 + \tau), ..., X(t_n + \tau) \) is independent of \( \tau \). Therefore,

\[
\text{Prob.}(X(t_1) \leq X_1, ..., X(t_n) \leq X_n) = \text{Prob.}(X(t_1 + \tau) \leq X_1, ..., X(t_n + \tau) \leq X_n),
\forall n, t_i \in T, t_i + \tau \in T, i = 1, 2, ..., n.
\]

(ii) Markov process

A stochastic process \( \{X(t), t \in T\} \) is said to be a markov process if for any set of \( n \) time points \( t_1 < t_2 < ... < t_n \) in the index set, the conditional distribution of \( X(t_n) \) for the given values of \( X(t_1), X(t_2), ..., X(t_{n-1}) \), depends only on \( X(t_{n-1}) \) i.e. immediately preceding value. More precisely for any real number \( X_1, X_2, ..., X_n \),

\[
\text{Prob.}\left\{X(t_n) \leq X_n | X(t_1) = X_1, X(t_2) = X_2, ..., X(t_{n-1}) = X_{n-1}\right\} = \text{Prob.}\left\{X(t_n) \leq X_n | X(t_{n-1}) = X_{n-1}\right\}
\]

(iii) Birth-death process

The birth-death process is a special case of a continuous-time markov process where the states represent the current size of a population and where the transitions are limited to births and deaths. In this process, the transitions can take place between neighbouring states only.

A continuous parameter homogeneous markov chain \( \{X(t), t \geq 0\} \) with the state space \( \{0, 1, 2, ...\} \) is known as the birth-death process if there exist constants \( \lambda_n \) and \( \mu_n \forall n = 0, 1, 2, ... \) such that the transition from state \( E_k \) to \( E_{k+1} \) describe the birth and from state \( E_k \) to \( E_{k-1} \) describe the death with rates \( \lambda_k \) and \( \mu_k \), respectively. Let \( P_k(t) \)
be the probability of the system is in state \( k \) at time \( t \). The system of differential equations describing the birth-death process is as follows (cf. Gross and Harris, 1985):

\[
\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \tag{1.3}
\]

\[
\frac{dP_k(t)}{dt} = (\lambda_k + \mu_k) P_k(t) + \lambda_{k-1} P_{k-1}(t) + \mu_{k+1} P_{k+1}(t), k \geq 1. \tag{1.4}
\]

(iv) **Poisson process**

The Poisson process with rate \( \lambda > 0 \) is an integer-valued stochastic process \( \{X(t), t \geq 0\} \) which satisfies the following rules:

(i) The process increments

\[X(t_i) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1}), t_0 = 0 < t_1 < t_2 \ldots < t_n,\]

are independent random variables.

(ii) The random variable \( X(s+t) - X(s), s \geq 0 \) and \( t > 0 \) has a Poisson distribution given by

\[P\{X(s+t) - X(s) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!}; k = 0,1,\ldots \tag{1.5}\]

(iii) \( X(0) = 0 \).

A random variable \( X \) having probability mass function \( f_X(x) \) and distribution function \( F_X(x) \) is said to have a Poisson distribution with parameter \( \lambda \), if

\[f_X(x) = P(X = x) = \frac{\exp(-\lambda) \lambda^x}{x!}; x = 0,1,2,\ldots \tag{1.6}\]

\[F_X(x) = P(X \leq x) = \sum_{i=0}^{x-1} \frac{e^{-\lambda} \lambda^i}{i!}. \tag{1.7}\]

The mean and variance of the Poisson distribution are same and given by

\[E(X) = \lambda = \text{Var}(X). \tag{1.8}\]

(v) **State probabilities**

The variation in the probability of state \( k \) is the difference between the probabilities of moving into and out of state \( k \). The probability of moving from one state to another state or remaining in the same state during a single time period is defined as transition probability. Mathematically, the transition probability can be defined as
\[ P_{X_n|X_{n+1}} = \text{Prob.}\{X(t_{n+1}) = X_{n+1} | X(t_n) = X_n\}. \quad (1.9) \]

It can be described as the conditional probability of the system which is in state \( X_{n+1} \) at time \( t_{n+1} \), provided that it was previously in state \( X_n \) at time \( t_n \). If the transition probability of the system does not depend on time \( t \), the system is said to be in steady state.

In steady state, the differential equations of birth-death process become:

\[ -\lambda_0 P_0 + \mu_1 P_1 = 0 \quad (1.10) \]
\[ -(\lambda_k + \mu_k) P_k + \lambda_{k-1} P_{k-1} + \mu_{k+1} P_{k+1} = 0, \quad k \geq 1. \quad (1.11) \]

(vi) Laplace transform

Let \( f(t), \ t \geq 0 \) be continuous real-valued function such that \( |f(t)| \leq Ae^{Bt} \) with some constants \( A \) and \( B \). Laplace transform of \( f(t) \) is defined as the function of the complex variable \( s \) with \( \text{Re}(s) > B \) and given by

\[ \mathcal{F}(s) = \int_0^\infty e^{-st} f(t) dt. \quad (1.12) \]

For continuous and non-negative random variable \( X \) with density function \( f_X(x) \) and distribution function \( F(x) \), the \( n^{th} \) moment of the random variable \( X \) can be obtained as

\[ E(X^n) = (-1)^n \left[ \frac{d^n}{ds^n} \mathcal{F}(s) \right]_{s=0}. \quad (1.13) \]

In particular,

\[ E(X) = \left[ \frac{d}{ds} \mathcal{F}(s) \right]_{s=0} \quad (1.14) \]
\[ E(X^2) = \left[ \frac{d^2}{ds^2} \mathcal{F}(s) \right]_{s=0} \quad (1.15) \]
\[ \text{Var.} (X) = E(X^2) - [E(X)]^2. \quad (1.16) \]
1.2.3 Some classical queueing models

(i) Bulk queue

In some queueing problems, the customers at service station may join in bulk. For example, the passengers at a bus stand/railway station arrive in bulk/group; goods are stored in bulk in cold storage, etc. In many queueing systems, the service of the customers is rendered in batches/groups of fixed or random size. The queueing systems having arrival and/or service in bulk are called bulk queueing systems. The $M^X/G/1$ queueing model deals with a single server queue with the Poisson arrival process, wherein units join the system in batches of random size as independent and identically distributed (i.i.d.) random variable. The service time of each customer is arbitrary distributed. In the $M^X/G^{(a,b)}/1$ model, units are served by the single server in batches of size $X$, where the random variable $X$ takes value $x$ ($a \leq x \leq b$); other assumptions are same as for $M^X/G/1$ model (cf. Chaudhry and Templeton, 1983).

(ii) Queue with server vacation

In many queueing scenarios, the server becomes unavailable for the occasional intervals of time; the time interval when the server is unavailable for providing the service to the customers is known as vacation period and such systems are modelled as a queue with vacation. When the service system is empty or some servers are idle, the server may go for vacation. The stochastic models of queueing systems with server vacations have enormous applications in industrial systems such as in high speed telecommunication networks, manufacturing and production systems and processor schedules in the computer and switching systems, etc. (cf. Tian and Zhang, 2006).

(iii) Queue with unreliable server

The performance of any system is one of the important issues that affect the design, development, configuration of the system. In many real time systems, the unreliable server makes a negative impact on the performance of the system. Therefore some effective measures are to be taken to maintain a desired grade of service. In many manufacturing/production systems, the server may fail and need repair to restore its capability to serve jobs and to maintain the efficiency of the system. In many fields, the
role of servers is performed by the mechanical/electronic devices, such as computers, telecommunications, ATM machines, traffic lights, etc. wherein the service may suffer due to accidental/random failures until the failed server is repaired.

(iv) Phase service

In some queueing scenarios, the service of the units is performed in many phases. Such situations can be realized in various real life congestion problems wherein the service of the units is provided through a series of different channels at work stations; each channel performs the service as assigned task. For example, in clinical examination procedure at hospital/clinic, the patient goes through a series of service stages such as blood pressure, X-ray, ECG, blood test, etc. Other practical examples of phase service can be found in automobile repair service stations, wherein the service of the vehicles is done through different channels.

(v) Queue with discouragement

It is specified by the inter-arrival time between any two consecutive arrivals and discouragement factors, namely Balking, Reneging and Jockeying. Balking in the queueing system occurs when the customer does not join the queue if he thinks that he has to wait for a long time for service. The Reneging can be observed in the circumstances, where the customers join the queue only for short periods and feel impatient after some time and decide to leave the queue. In multi-server queueing system, the process of switching over from one queue to another queue is called Jockeying.

1.3 Queueing Methodology

In many congestion situations of real life problems, the importance of the state dependent inter-arrival of units and state dependent service can be realised in many day-to-day activities. Some methodological aspects to develop the state dependent models of various queueing systems by incorporating the more realistic features such as vacations, discouragement factors, unreliable server and service/repair in different phases are presented. We describe some important classical, analytical and numerical solution techniques such as generating functions, supplementary variable, Laplace transform, maximum entropy principle and Runge-Kutta method which are frequently used to
analyse the queueing models in different frameworks. In this section, we briefly explain commonly used techniques for the performance modelling and analysis of queueing systems.

(i) Probability generating function method

Let $X$ be a non-negative discrete random variable with probability function $P(X = n) = p(n), n = 0, 1, 2,\ldots$

The generating function $P_X(z)$ of $X$ is defined as

$$P_X(z) = E(z^X) = \sum_{n=0}^{\infty} p(n)z^n, \quad |z| \leq 1.$$  \hspace{1cm} (1.17)

From the definition of probability generating function $P_X(z)$, we have

$$P_X(0) = p(0), \quad P_X(1) = 1, \quad P_X'(1) = E(X), \ldots, P_X^{(k)}(1) = E\{X(X-1)\ldots(X-k+1)\}$$

where

$$P_X^{(i)}(1) = \frac{\partial^i}{\partial z^i} P_X(z) \bigg|_{z=1}. \hspace{1cm} (1.18)$$

(ii) Supplementary variable technique

In the markovian models of queueing situations, wherein the arrival of the customers has a Poisson process and the service time is exponentially distributed, the queue size process may be described by using markov process. In the non-markovian queueing models, where arrival process and/or service process are general (arbitrary) distributed; the supplementary variable technique introduced by Cox (1955) can be used. It is one of the simple and convenient ways to obtain the solution of the non-markovian queueing system. The supplementary variables can be introduced corresponding to either elapsed time or remaining time in case of both inter-arrival and service processes.

In queueing systems with arbitrary service time distribution, the number of customers in the system at time $t$ and the time spent by the system are considered to determine the future stochastic properties. For the analysis of such stochastic modelling of queueing problem, the supplementary variables are taken corresponding to remaining or elapsed service time because the queue size process cannot be described easily in the absence of markov property. Similarly, for general distributed arrival process, the
supplementary variables are introduced corresponding to the inter-arrival time of the customers (cf. Chaudhry and Templeton, 1983).

(iii) Maximum entropy principle

The maximum entropy principle (MEP) can be used to estimate the probability distribution of the system states in explicit form. The result of a probability distribution is consistent with the known constraints expressed in terms of expected values of one or more quantities, but is otherwise as unbiased as possible. Consider the discrete case of entropy for finite observations discussed by Shannon (1948). The system entropy function $Z(p)$ measures the amount of uncertainty contained by the probability distribution $p = (p_1, p_2, ..., p_n)$ (cf. Kapur and Kesavan, 1987; Kapur, 1989) and defined as

$$Z(p) = Z_n(p_1, p_2, ..., p_n) = -\sum_{k=1}^{n} p_k \log p_k$$

with following property:

$$Z_n(p_1, p_2, ..., p_n) \leq Z_n\left(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right) \text{ with equality iff}$$

$$p_k = \frac{1}{n}, k = 1, 2, ..., n.$$  \hspace{1cm} (1.20)

In the field of queueing theory, the maximum entropy principle has been applied to study the explicit results of the queue size distribution for the complex queueing models. The maximum entropy principle can be used to determine appropriate probability distributions of the waiting time distribution in terms of some moments of the inter-arrival time or service time or average queue size. In general, the maximum entropy principle is employed to obtain the probability distribution only based upon the available information. For stochastic modelling of complex queueing systems, maximum entropy principle provides a self consistent method to determine the unknown distribution subject to the mean value of constraints. The solution can be expressed in terms of normalizing constant along with the product of Lagrangian coefficients corresponding to the constraints in terms of many known performance indices. To illustrate the problem, we describe the maximum entropy principle for analysing the unreliable server queueing model with vacation. Consider the system entropy function defined as (cf. El-Affendi and Kouvatsos, 1983; Wang et al., 2007)
\[ H = - \sum_{n=0}^{\infty} P_0(n) \ln P_0(n) - \sum_{n=1}^{\infty} P_1(n) \ln P_1(n) - \sum_{n=1}^{\infty} P_2(n) \ln P_2(n) - \sum_{n=1}^{\infty} P_3(n) \ln P_3(n) \]  
\hspace*{2cm} (1.21)

with constraints based on different states

(i) Normalizing condition \[ \sum_{n=0}^{\infty} P_0(n) + \sum_{n=1}^{\infty} P_1(n) + \sum_{n=1}^{\infty} P_2(n) + \sum_{n=1}^{\infty} P_3(n) = 1. \]  
\hspace*{2cm} (1.22)

(ii) The probability that the server is on busy state \[ \sum_{n=1}^{\infty} P_1(n) \equiv a_1. \]  
\hspace*{2cm} (1.23)

(iii) The probability that the server is on breakdown state \[ \sum_{n=1}^{\infty} P_2(n) \equiv a_2. \]  
\hspace*{2cm} (1.24)

(iv) The probability that the server is on vacation state \[ \sum_{n=1}^{\infty} P_3(n) \equiv a_3. \]  
\hspace*{2cm} (1.25)

(v) The expected number of customers in the queue

\[ L_q = \sum_{n=0}^{\infty}nP_0(n) + \sum_{n=1}^{\infty}nP_1(n) + \sum_{n=1}^{\infty}nP_2(n) + \sum_{n=1}^{\infty}nP_3(n). \]  
\hspace*{2cm} (1.26)

(iv) Runge-Kutta method

Due to the complex nature of the differential equations governing the queueing models, most often it is not easy to find the analytical solution in particular when the equations having the transient probabilities of the system states. In such cases, the numerical techniques can be used to find the solution of the set of differential equations. The method developed by Runge and Kutta (cf. Iwaarden, 1985) is one of the techniques to provide the numerical solution of a set of differential equations. The fourth order Runge-Kutta method is the most popular one and many researchers have used it to obtain the solution of a set of differential equations governing the queueing model.

Various computer softwares namely MATLAB, Maple, Mathmatika, etc. are available for computational purposes to facilitate the numerical solutions of differential equations. For the modelling of some queueing systems, MATLAB with routines ode45 based on 4th and 5th order Runge-Kutta method is commonly used to find the numerical solution of the system of differential equations (cf. Ingolfsson et al., 2007).
1.4 Review of Literature

Due to enormous applications of the queueing theory in real world problems, it has attracted many researchers, since the pioneer work of Erlang (1909). Most of the researchers considered the markovian queueing problems wherein inter-arrival and service times were modelled by using the exponential probability distribution. The queueing models dealing with varying rates along with other features such as arbitrary distributed service time, bulk arrival queue with vacation, queues with unreliable server and service/repair in different phases etc. have not received much consideration so far due to the complexity of the solution procedure. These congestion situations play important role in various daily life activities and also have great applications in production/manufacturing systems, computer/communication systems and machine repair model, etc. Based on the investigations carried out in the present thesis, for the purpose of review, we classify the relevant literature which is closely related to present study in the following groups:

1.4.1 Historical development

The pioneer work in the queueing theory was done by Danish mathematician Erlang (1909, 1917). He discussed the performance prediction of a telephone system under the assumptions of Poisson input, exponential service time and multiple/ single server. He described the importance of Poisson distribution in queueing theory for modelling of the arrival process of the customers. After the pioneer work of Erlang, many researchers studied different queueing systems by discussing several variations in arrival process, service process, service discipline, etc. Molina (1927) and Fry (1928) presented the queueing models for practical problems in Erlang’s work. Theoretical analysis of queueing system started significantly with the development of Operations Research after World War-II and became the standard tool in various areas. Many researchers discussed the fundamental models in queueing theory with Poisson arrival process and exponential service time with a single server and provided a solution for limiting the queue size by solving the balance equations of the system’s model using recursive arguments. Bailey (1954) provided the state dependent solutions using the generating function for the differential equation governing the fundamental model. The systematic study of the queueing process by using the notations $A/B/C$ was started with the notable work of Kendall (1953). In queueing theory literature $A$ defines the
distribution of inter-arrival time, $B$ defines the distribution for service time and $C$ denotes the number of servers. In his investigations, Kendall considered some problems of queues and proposed the mathematical models for stochastic process and obtained various system characteristics to predict the behaviour of queueing system. Since then many researchers contributed to queueing theory by analysing queueing models with different variations in arrival process and service system, and discussed steady state and transient state solutions of queueing models. For comprehensive developments in the queueing theory, one may refer to Medhi (1982), Chaudhry and Templeton (1983), Gross and Harris (1985), Kashyap and Chaudhry (1988), Takagi (1991), etc.

1.4.2 State dependent queueing models

The queueing models with state dependent arrival and service rates have great applications. However, the queueing problems with state dependent rates have not received much attention because the solution of the models becomes cumbersome in nature. A few researchers have analysed the queueing models with state dependent rates under different assumptions. Harris (1967) modelled production system as the queueing system with state dependent service rates which find applications in many real life problems. Cohen (1969) developed the model with the markov chain approach in single server queue. The state dependent Erlangian queue was discussed by Conolly (1974). Hadidi (1974) discussed the busy period distribution for a general birth and death queueing model. The multi-server queues with Poisson arrivals were discussed by Cosmetatos (1978). Jackson and Asden (1980) explained multi-stage Poisson queueing system with infinite servers. Albin (1984) investigated a single server model with perturbed Poisson arrival process to predict the operating characteristics of queues which find applications when inter-arrival process is slightly different from the Poisson process. Gross and Harris (1985) provided a comprehensive development of the literature for the birth-death process for single and multi-server models. Knessl et al. (1987) considered a state dependent arrival single server queueing model and assumed that arrival rate is a function of the backlog work in the system and obtained asymptotic approximations of both the busy period distributions as well as the residual busy period distribution. Smith (1994) discussed the applications of state dependent models and presented the techniques to design the models. The stochastic model with state dependent service rates was analysed by Wang (1996) and probability generating...
function technique was applied to obtain the various performance measures. Whitt (1999) investigated a time dependent markovian queueing system and discussed the decomposition approximations to solve the time dependent ordinary differential equations. Boxma and Perry (2001) considered a queueing model with dependence between a service request and subsequent inter-arrival time to find a direct link between the workload distributions of the model. To discuss the impact of time dependent behaviour on the performance of the system, a model with time varying rates was analysed by Massey (2002) and illustrated the application of the model in telecommunication system. Jain (2005) has analysed a model of multi-server queueing system by using recursive method to solve the steady state equations and obtained various performance indices. Guan et al. (2006) presented an algorithm for new approach to control the delay in service time with the adjustment of threshold in terms of arrival rates of the units. To compute the performance measures, the multi-server model with a finite capacity model was studied by Cruz and Smith (2007). They described the methodology for the analysis of approximate results. Winands et al. (2009) considered a single server model with priority and obtained the queue length distribution for non-markovian model under the different assumptions about customers with low and high priority queues. To analyse the multi-server queue, a system optimum model for state dependent non-markovian system was presented by Cruz et al. (2010). Lee (2011) provided state dependent stochastic networks to establish the different moment stability properties of the system by using the concept of birth-death process. Banerjee and Gupta (2012) investigated a single server bulk queueing system to reduce the congestion by controlling the arrival and service rates using specific threshold policy and obtained various performance measures. Recently, Claeys et al. (2013) have analysed the stochastic model of the queueing problem for a batch service by including the timing mechanism wherein service time is dependent on the number of customers within a batch and presented the numerical illustration to validate the analytical results.

1.4.3 Non-markovian queueing models

In some daily life congestion problems, the service time of the units may not follow an exponential distribution. Such situations can be noticed in the clinics performing X-rays and blood test etc. of patients, in a bank at cash counters and many
other places. In queueing systems with arbitrary service time distribution, the number of units in the system at time \( t \) and the length of time for which the unit is in service (if any) are sufficient to determine the future stochastic properties of these variables. Yadin and Naor (1963) investigated the important model for \( M/G/1 \) queueing system. Then after, many researchers contributed to the \( M/G/1 \) queueing models by discussing the optimal solution for the systems. In the recent past, a large number of queueing models with arbitrary service time for single arrival of the units or arrival in batches under different assumptions have also been developed by Levy and Yechiali (1975), Kimura (1981), Keilson ad Servi (1986), Choi and Park (1990), Takagi (1991) and Madan (1992) etc. and the references cited therein. Abolnikov et al. (1993) considered the single server model to study the general distributed state dependent service process with continuous time parameter. The finite buffer queueing system for retrial service was discussed by Bocharov et al. (1999). Wang and Ke (2000) analysed the model of non-markovian finite/infinite capacity system by using recursive method for arbitrary distributed service time to obtain the performance characteristics of the system. The queueing system with the general distributed arrival process was provided by Ke and Wang (2002). They studied the behaviour of the system under \( N \)– policy by providing the recursive method with the use of the supplementary variable technique. Chaudhry and Kim (2003) investigated the system for multi-server queueing problem with the deterministic service time to study the behaviour of the system. To study the performance measures, the \( M/G/1/K \) queue was considered by Smith (2004) by using embedded markov chain approach to demonstrate the applications of the model. Cruz et al. (2005) presented a multi-server system to discuss the problem of service and capacity allocation in the system and analysed the stochastic model to provide the computational results. Yin et al. (2007) considered the problem of sensitivity analysis to discuss the steady state performance measures for the \( M/G/1 \) queue and proposed the algorithm to provide the numerical illustration of the system. Single server finite capacity system was investigated by Wang et al. (2008). In this investigation, they presented the recursive method by using the supplementary variable technique to develop the steady state probability distribution under \( F \)– policy and studied the behaviour of performance measures. Kahraman and Gosavi (2011) proposed the system of queueing problem by developing the theory based on discrete-time markov chain for the situation of bulk arrival and bulk service in the single server arbitrary distributed service time. To examine the threshold policy with the arrival and service patterns of the
customers and to compute the fundamental formula for evaluation of the performance of the system, a queueing system was presented by Claeys et al. (2011). Recently, Arrar et al. (2012) have investigated a stochastic model of retrial queue with batch arrival of customers. They have obtained the steady state distribution by using the supplementary variable technique and discussed the asymptotic behaviour of the customers. Liu and Zhao (2013) have discussed the asymptotic properties of the loss probability for non-markovian \( M/G/1/N \) model with vacation and obtained the asymptotic rates of the loss probability.

### 1.4.4 Queueing models with vacation

The queueing systems with server vacations can be seen in many queueing scenarios and very useful to analyse the performance of real time queueing systems where the server may go for single/multiple vacations. Such systems have wide applicability in analyzing the various real life traffic situations of day-to-day as well as industrial queues. Numerous researchers have studied single arrival/batch arrival queues with vacation time, we refer the important contributions by Baba (1986), Doshi (1986), Takagi (1991), Zhang and Vickson (1993), Lee et al. (1995), Li and Zhu (1996), Borthakur and Choudhury (1997), Chao and Zhao (1998) etc. and the references cited therein. Madan (1999) considered an \( M/G/1 \) queueing model to study the steady state behaviour of the single server model with deterministic server vacations by using the supplementary variable technique. It was assumed that at the completion of each service, the server may decide to take a vacation of fixed period or may continue to be in system for the next service. The stochastic model for bulk arrival queueing system with setup and vacation period to study the behaviour of the queue length distribution was analysed by Choudhury (2000). Madan (2000) considered a model in which he has assumed that the units arrive one by one with the homogeneous arrival rate. He obtained the probability generating functions for queue length distribution with the provision of second optional service available in the system. Choudhury (2002) studied a system for batch arrival queue and discussed the behaviour of the system with a vacation time under single vacation policy and derived the expression of the queue length distribution. The stochastic model of queueing system by using the matrix geometric method was analysed by Zhang and Tian (2003). Choudhury and Madan (2005) analysed a model with batch arrival process and two stages of service with the Bernoulli schedule
under $N$–policy. It was further assumed that the server remains idle till the queue size becomes at least pre-specified value $N$ ($N \geq 1$). The queue size distribution and the existence of stochastic decomposition property of the queue size distribution at a random and departure epoch were derived. Wu and Takagi (2006) analysed the system for $M/G/1$ queue under multiple working vacation policy and exhaustive service discipline by using Laplace-Stieltjes transform. Choudhury (2006) derived the expressions for the queue size distribution at random epoch and at a departure epoch for a model. It was assumed that the arrival follows the Poisson process with a uniform rate at busy state, state of first regular vacation and second optional vacation state. Choudhury (2008) discussed the queue size distribution of the queue with a random setup time and Bernoulli vacation schedule under a restricted admissibility policy. A batch arrival queueing system with a randomized vacation policy with the assumption that the server takes vacation if the system is empty and immediate activate for service when at least one customer found for service, was discussed by Ke et al. (2010). Recently, Dimou et al. (2011) analysed the system of a single server queue for the geometric abandonment case wherein customers decide sequentially whether they leave the system or not without getting the service. Recently, an iterative method of boundary value was discussed by Chan et al. (2012). They obtained the transient solutions of queueing systems and demonstrated the efficiency by taking numerical illustration. Gao and Liu (2013) have considered an $M/G/1$ queue with the single working vacation and analysed the model by using the matrix analytic method. The supplementary variable method is also used to obtain the joint distribution to obtain the queue length distribution of the system.

1.4.5 Unreliable queueing models

It is realized that in real time system, no server can always work perfectly, i.e. it is subject to failure during busy state. In manufacturing/production systems, the failure of machines occurs randomly so as to stop the service. It is noticed that during repair of the machines or during a delay in repair due to unavailability of the repairman, the system does not work smoothly and affects the arrival rates of the customers also. Several researchers have studied different models in this field, that describe the failure of a server and its repairs as well as the rules of servicing a customer who finds a server in a broken down condition, we refer the contributions of Wartenhorst (1995), Li et al.
(1997), Tang (1997), Gray et al. (2000) etc. and the references cited therein. Ke (2003) investigated the queueing system with general distributed service time under $N$–policy for unreliable server. It was assumed that there is a provision of vacation and startup time before providing the service. Wang (2004) developed a model with second optional service and unreliable server to discuss various performance measures. He considered that after availing first essential service, some of the customers may demand for second optional service. Lam et al. (2006) presented a $M/M/1$ model with geometric process to obtain the performance characteristics by using the supplementary variable technique. The stochastic model of the batch arrival queueing system with server breakdowns was presented by Ke (2007). In this system, he investigated the situation with general distributed startup/closedown times to obtain the system characteristics. Choudhury and Tadj (2009) considered an $M/G/1$ queueing model with an additional second phase of the optional service subject to breakdowns of the server. This model generalizes both the classical $M/G/1$ queue subject to random breakdown and delayed repair as well as $M/G/1$ queue with second optional service and server breakdowns. Jain and Jain (2010) provided the stochastic model with multiple types of server breakdowns wherein it is assumed that the arrival and service rates are not uniform and depend on the different system states. In this study, they analysed the system by using matrix-geometric approach to compute the performance indices. An $M^{[X]}G/1$ queueing system with the provision of randomized vacation policy was considered by Ke et al. (2011). This investigation examined the system for the unreliable server with the assumption that the server may go for vacation if the system is empty and immediately activate for providing the service when found the customer in the system waiting for service. Recently, Choudhury and Deka (2012) discussed the single server unreliable non-markovian system with two phases of heterogeneous service and derived the joint distribution of state of the server and obtained the probability generating function of the queue size distribution. Hassan and Ibrahim (2013) have analysed the multi-level queueing system by using the recursive solution technique and discussed the steady state behaviour of the system. Various performance measures of the system are also obtained.
1.4.6 Queueing models with discouragement

Sometimes, because of the shortage of time or other reasons, the customer may drop the idea of getting the service or may join the other service station to get the service immediately. In the busy schedule of day-to-day activities, the situations of queueing systems with discouragement factors can be noticed in all spheres of life and should be incorporated in order to predict the performance indices of queueing problems. Takacs (1974) analysed a model with discouragement factor due to the busy period with the assumption that if a customer arrives at any instant of busy state then it may or may not to join the queue. Haghighi et al. (1986) derived the steady state probabilities for $M/M/c$ queue under assumptions of balking and reneging. Abou-El-Ata and Shawky (1992) proposed a single server queueing model with balking, reneging and an additional server for long queues and presented the analytical explicit solution to study the behaviour of the system. Thomo (1997) considered a $M^X/G/1$ queueing model with balking and multiple vacations. A customer in call centre, who cannot be helped immediately by a human server might be told to wait for some time, then the customer might hang up (i.e. balk) or decide to hold. This balking behaviour arises when a customer refuses to enter the queue if the wait is too long. Gans et al. (2003) have presented the significant works on customer’s impatience in practice when studied the performance modelling of the call centers. Garnett et al. (2002) and Whitt (2005) have obtained the important results on the models of customer’s impatience behaviour to examine the congestion with different variations. Altman and Yechiali (2006) have studied the behaviour of the customer due to unavailability of the server upon arrival of a customer. Singh et al. (2007) discussed a single server queueing model with controllable arrival rate with reneging and obtained the performance indices to characterize the model. Liu and Kulkarni (2008) extended the balking and reneging concepts in multi-channel service queueing model. Chakravarthy (2009) studied the behaviour of the queue with disasters failures in which the customer becomes impatient during repair period of the failed server. The $G/M/1$ queue with a constant patient time of the customers and level crossing argument was considered by Bae and Kim (2010). They studied the characteristics for the distribution of the workload of the server, and the virtual waiting time. Heavy traffic analysis for Earliest-Deadline-First queues was presented by Kruk et al. (2011) and analysed the model under an EDF-policy to obtain the performance measures of the system for impatient customers wherein the
performance is measured by minimum reneged work by EDF-policy. In multi-server
markovian queueing system, the discouragement factors balking and reneging were
studied by Choudhury and Medhi (2011). They derived the explicit closed form
expressions for the queue length distribution. Recently, Boudali and Economou (2012)
have considered the optimal and equilibrium strategies in queue, wherein all customers
are forced to abandon the system. It was also assumed that the decision of the customer
whether to join the system or not, depends on natural reward-cost structure. The balking
behaviour of the customers in the $Geo/Geo/1$ queueing system under multiple vacation
policy was discussed by Ma et al. (2013). In this investigation, the equilibrium
behaviour of the customer is discussed under different cases and provided the numerical
results to demonstrate the effects of parameters.

1.4.7 Maximum entropy principle in queueing models

To study the steady state probabilities and to find the approximate values of
unknown probability distribution and corresponding performance measures in many
complex situations, the maximum entropy principle is a powerful technique. In the field
of queueing theory, the maximum entropy principle has been applied to study the
explicit results of the queue size distribution for the complex models. Shannon (1948)
introduced the principle of maximum entropy as a measure of an amount of uncertainty
and this concept with some more assumptions was extended by Jaynes (1968). Several
researchers have used the maximum entropy principle to analyse the various models for
different assumptions in queueing scenarios. Some notable works in this regard were
al. (1991), Wang et al. (2002), Wang et al. (2005) etc. and the references cited therein.
Wang et al. (2007) discussed the maximum entropy principle for the analysis of a
queueing system with multiple vacations and server breakdowns. Wang and Huang
(2009) discussed the $M/G/1$ queueing model with single removable and unreliable
server to obtain the approximate formulae for the probability distributions of the number
of customers and the expected waiting time in the system under $(p,N)$-policy. Yang et
al. (2010) dealt with the $(N,p)$-policy $M/G/1$ queue with server breakdowns and
general startup times, where customers arrive to demand the first essential service and
some of them further demand a second optional service. The method of maximum
entropy principle is used to develop the approximate steady state probability distribution
of the queue length in the system. Wang et al. (2011) analysed a single removable and unreliable server in an \( M/G/1 \) queueing system operating under the \((p,N)\)-policy. They employed an improved maximum entropy method with several well-known constraints to estimate the probability distributions of the system size and the expected waiting time in the system. To obtain the expectation of waiting cost and other characteristics of the queueing system, the entropy maximization principle was considered by Guo et al. (2011). By using this principle, they discussed the optimal decision to join the queueing system to get the service. Recently, Wang et al. (2012) have described the analytical method for \( M/G/1 \) queue with unreliable server, second optional service and general startup times. In this investigation, they presented the approximate results by using the maximum entropy principle and compared the results with the exact results of performance measures.

### 1.5 Chapter wise Summary

In the past, many researchers have considered the various queueing models with the constant arrival and service rates by using different probability distributions for modelling of the queueing problems in different frameworks. The queueing problems with state dependent rates have drawn a little attention of queue theorists because of the complexity of the solution. The queues with varying rates have great applications in many daily life congestion situations arising in flexible manufacturing systems, telecommunication systems and computer networks, etc.

In some systems, the arrival of the units occurs according to a Poisson process or general distributed fashion with different arrival rates depending on the states of the system. Besides it, the arrival rates of the units in the service system may also depend on the queue length in the system. The queueing models with variable arrival rates of the units can observe in many day-to-day queueing scenarios, e.g. in healthcare systems. In some cases, the arrival rates may depend upon the time but independent of the system state. The service rate of the server may be dependent on the workload with a burden of the arrived units. Motivated by the above mentioned factors, the objective of the present work is to investigate some stochastic models of queueing problems with state dependent arrival rates and to predict the performance behaviour of such queueing systems.
The brief summary of the chapters of the present thesis are as follows:

The ongoing **chapter 1** deals with the introductory aspects of queueing systems. It provides an overview of stochastic modelling of queueing problems in different scenarios. It also covers the brief knowledge of the techniques used to analyse the different models in the present thesis. As there is a vast literature on various aspects of queueing theory, we have reviewed the topics related to the present research work. The developments on these topics in recent past have also been reported.

In **chapter 2**, we extend the model of Madan (1999) by incorporating the balking behaviour of the units with bulk arrival and provision of optional deterministic vacation. By using the basic assumptions of probability reasoning and supplementary variable technique, the steady state behaviour of the system is studied and various performance measures are obtained. In order to obtain the approximate values of system probabilities for queue length distribution, the principle of maximum entropy is also employed. To verify the tractability of performance measures obtained, the numerical illustrations are provided. The results of this chapter are to be published in International Journal of Operational Research (Singh et al., 2013).

In **chapter 3**, we generalize the model of Choudhury (2006) for the bulk arrival queueing system by incorporating varying arrival rates during the busy period, regular vacation period and optional vacation period. We employ the supplementary variable approach to obtain the queue size distribution and other performance measures. The results of this chapter have been published in International Journal of Mathematics in Operational Research (Singh et al., 2012).

In **chapter 4**, we extend the model of Madan (2000) for the situation wherein the units arrive in bulk with different arrival rates depending upon server status. There is a provision of second optional service as per demand; the second optional service is rendered by the same server to the unit after completing the essential service of the unit who requests for it. The results of this chapter have been published in International Journal of Mathematics in Operational Research (Singh et al., 2011).

**Chapter 5** dealt with single server model based on the work of Madan (2003), wherein arrival of units are in bulk and follows the Poisson process with varying arrival rates depending upon the server status that may be in an idle state, operating state and repair state. It is also assumed that the system is subject to breakdown and the failed server immediately joins the repair facility, which takes constant duration to repair the server. By using the supplementary variable technique, we analyse the model to obtain
the expression of some performance indices of the system. The results of this chapter have been communicated for publication.

**Chapter 6** is devoted to the study of a single server stochastic model of queueing system. The investigation extends the work of Thangaraj and Vanitha (2010), wherein the arrivals of units are in batches and follow the Poisson process with state dependent arrival rates. There is a provision of two stages of heterogeneous service with arbitrary distributed different service times. The server may take optional vacation after the completion of the both stages of service of each unit. The server may fail at any instant of service and requires repair. The results of this chapter are to be published in International Journal of Industrial and Systems Engineering (Singh and Kumar, 2013).

**In chapter 7**, we propose the extension of the model studied by Choudhury et al. (2009) for two phases of service and different phases of repair under $N$ – policy with the assumption that the units arrive in bulk; the arrival rates of units depend on the current status of the server who may be in an idle state, busy state, waiting for repair, under repair state. The server may breakdown at any instant during any stage of the service. There is a provision of essential and optional repair in $m$ phases. The results of this chapter are to be published in International Journal of Services and Operational Management (Singh et al., 2013).

**In chapter 8**, the model of Perel and Yechiali (2010) is discussed and propose a single server markovian queueing model for finite capacity/finite population under the assumption that the customer can leave the system (reneging) due to slow speed of the server. We obtain the transient solution of queue length distribution. Various performance measures are obtained in terms of transient probabilities of the system states. The numerical illustrations are provided to validate the tractability of performance measures and facilitate the sensitivity analysis. The results of this chapter have been communicated for publication.

**Research Publications:**

The Research papers based on investigations done in this thesis are as follows:


