CHAPTER 1

INTRODUCTION

A queue is a waiting line of ‘customers’ requiring service from one or more servers. A queue forms whenever existing demand exceeds the existing capacity of the service facility; that is, whenever arriving customers are unable to receive immediate service due to busy servers. Delays and queuing problems are most common features not only in our day to day life situations, but also in more technical environments, such as, in manufacturing systems, computer networking and telecommunication systems. Queueing theory provides a mathematical basis for understanding and predicting the behaviour of such situations. As a stochastic process, queueing theory contributes vital information required in decisions associated with waiting lines, by predicting various performance measures such as mean queue length, mean busy period, etc,. Therefore, the ultimate goal is to achieve an economic balance between the cost and the cost associated with waiting time for that service.

The Danish Mathematician A.K. Erlang an employee of the Danish telephone company was the pioneer investigator of queueing theory who investigated the theory of probabilities and telephone conversations in 1909. Later he observed that the telephone system was generally characterised by either

1. Poisson input, exponential service times, and multiple channels (servers) or
2. Poisson input, constant holding times, and single channel.

He is also responsible for the notion of stationary equilibrium, balance of state equations and optimization of queueing system. Many researchers motivated by his work have developed this queueing theory for practical applications. Also, queueing theory combines the concept of service capacity, customers arrival intensity and a quality of service in an elegant and simple way such that their dependencies are expressed with equations.

1.1 TYPES OF CLASSICAL QUEUEING MODEL

A classical queueing system can be described as customers arriving for service, if service is not immediately provided, having to wait for service and leaving the system after being served. A classical queueing model consists of three parts namely, arrival process, service process and queue discipline. The term customer may refer to, for example, data packets waiting for transmission, calls arriving at a telephone keyboard or a programme waiting for a command to run in a computer, a machine arriving at an inspection centre, etc. For a detailed study on classical queueing models, one can refer the contributions made by Neuts (1989), Yechiali (1993), Takagi (1988, 1991) and others. The general structure of a classical queue is shown in Figure 1.1.

![Figure 1.1 General structure of a classical queue](image.png)

In recent years, queueing theory has been extensively applied as a model for different computer networks and telecommunication systems which
provides service for randomly arriving customers with a specified quality of service. These arrivals are combined with the variety of queueing phenomena leading to a large number of variants. Few such variants are described as follows:

1.1.1 Continuous and Discrete Time Queueing Models

The standard models in classical queueing theory are systems operating in continuous time which were first developed to analyze telephone networks. In continuous time, arrivals and departures take place at any time instant. But in practice, there are many systems which show an inherent generic slotted time scale. That is why discrete-time queueing theory is distinguished from its more developed continuous-time counterpart by the synchronous nature of its service. Discrete-time models are very popular studies of computer and telecommunications systems because in some cases, time is divided into fixed length intervals called time-slots and packets of information called cells are of fixed length, such that exactly one cell can be transmitted during a time-slot. Examples of such cases include technologies, such as digital computers and communication systems including mobile and B-ISDN networks based on Asynchronous Transfer Mode (ATM) technology and the IEEE 802.6 Metropolitan Area Network (MAN) standard.

It should be noted that the probability of queueing activities occurring simultaneously is not zero in discrete-time. Thus, to solve the conflict of the simultaneity, it is necessary to establish the order in which the queuing activities take place. There are two different policies, namely, if the arrivals precede the departures (Late Arrival System (LAS) or Arrival First (AF) policy) and if the departures precede the arrivals (Early Arrival System (EAS) or Departure First (DF) policy). That is, if upon arrival when the server is idle, then the service of that customer starts immediately (Immediate Access -IA) or in the following slots (Delayed Access - DA). If these
concepts are combined with LAS, then the discipline is known as Late-Arrival System with Immediate Access (LAS-IA or EAS) and Late-Arrival System with Delayed Access (LAS-DA), respectively. For more details on these concepts one can refer to Hunter (1983) and Gravey & Hebuterne (1992).

1.1.2 Queueing Models with Multi Server

By and large, systems that have more than one server are much more difficult to analyse than those with the single server. Queues in which service can occur simultaneously over more than one channel are known as multi-channel queues.

1.1.3 Queueing Models with Two Phases of Service

In day-to-day life all arriving customers are in need of a main service and only few of them may require the subsidiary service provided by the server. More specifically, all customers are processed in the first phase and only the customers who are in need can proceed to the second phase. Such type of service is called queueing models with two phases of service.

1.1.4 Queueing Models with Closedown Times

In practice, after completing the service, the server may require some amount of time for closing the service job, such as making the templates for copy turning, checking the components etc., in a pump manufacturing industry. In such situations, operator always shuts down the machine and removes the templates before taking up other works. This aspect in queueing system is called the queueing models with closedown times.
1.1.5 Queueing Models with Vacations

During the idle period the server may be utilised for some other purposes such as maintenance work or serving secondary customers etc. In such situations, the server may shut down the service facility and may not be available to the next customer who arrives to the empty system. This is known as the queueing models with server vacation. The periods for which the server is unavailable is said to be 'server vacation periods’. The following are some types of vacations used in the thesis:

1.1.5.1 Single vacation

At a service completion epoch, if the required number of customer is not available in the queue to start a service, then the server may leave the system for a vacation of random length. After returning from this vacation, if the required number of customer is available in the queue, then the server begins the service. Otherwise the server will remain in the system till the queue size reaches the required level. Such type of vacation is called single vacation.

1.1.5.2 Multiple vacations

The server avails a vacation every time whenever the system becomes empty or the number of customers waiting for service is less than a predetermined value. After completing a vacation, if the server finds inadequate number of customers in the queue for service, then the server avails of another vacation and continues to do so until the server finds a sufficient number of customers to start a service. These types of vacations are called multiple vacations.
1.1.6 Queueing Models with Priorities

When some customers in the queueing system are given more preference for service they are called priority queueing systems and are of two types namely pre-emptive priority and non pre-emptive priority.

1.1.6.1 Pre-emptive priority

The arrival of higher-priority customer interrupts the service of the lower-priority customer. In such situations the service of the higher priority begins and the lower priority customer returns to the queue. The interruption of service can cause a loss of progress and customers have to start the service from the beginning once again. This is called pre-emptive-repeat. In a pre-emptive-resume scenario a customer can continue at the point of interruption.

1.1.6.2 Non pre-emptive priority

The arrival of higher priority customer may not interrupt the service of the lower priority customer and has to wait until the service completion of the lower priority customer. This is called non pre-emptive priority.

1.1.7 Queueing Models with Server Breakdown

In most of the practical situations, the server may breakdown and fails to render service to the arriving customers. Since the performance of a system may be heavily affected by the service station breakdowns, such systems with repairable server are worth investigating from the queueing theory view point. Such type of server failures can be classified into two types namely, operation dependent server failures and time dependent server failure. The operation dependent server failures can occur only when the server is in operation. The time dependent server failures can occur at any time, independent of whether the server is rendering service or not. In
Queueing models with server breakdown, failures can be repaired immediately and the service is continued according to the repeat or the resume rule.

1.1.8 Queueing Models with Starting Failures

A customer if upon arrival finds the server idle must turn on the service station. In such situations, if the server is activated successfully with certain probability, the service of that customer begins immediately and abandons the system forever after service completion. On the contrary, the server started unsuccessfully with the complementary probability the server is sent to repair station immediately and the customer joins the queue. Such type of failures is known as queuing system with starting failure.

1.1.9 Queueing Models with Different Types of Customers

The theory of queues reasonably considers several types of customers who will directly or indirectly affect the measures of the queueing system. The following are some types of customers used in the thesis.

1.1.9.1 Impatient customers

Waiting times in queues are most often taxing and customers may lose patience and leave the system without service from the server. Such types of customers are called impatient customers.

1.1.9.2 Negative customers

The arrival of negative customers in the queue removes (kills) the ordinary (positive) customer according to strategies such as individual removal, batch removal, disaster, triggered movement, and random work removal. For details one can refer the survey paper by Artalejo (2000). Negative arrivals to an empty queue depart immediately with no effect. Thus
a negative arrival represents some kind of work cancelling signal and does not require any service.

1.1.9.3 Feedback of customers

The queueing system which includes the possibility for a customer to return immediately to the counter for additional service with certain probability or departs permanently with compliment probability, are called queues with feedback. The concept of feedback is explicitly introduced by Finch (1959).

1.2 TECHNIQUES USED TO SOLVE QUEUEING MODELS

In general, queueing models are classified into Markovian and non-Markovian queueing models according to arrival and service processes. The techniques generally adopted to solve these types of queueing models are as follows.

1.2.1 Markovian Queueing Models

Poisson arrivals and exponential service makes queueing models Markovian that are easy to analyze and obtain usable results. In early stages these models were used to help decision-making in the telephone industry. Some of the techniques used to solve Markovian queueing models are:

(a) Difference – differential equation method
(b) Neuts matrix-geometric algorithm
(c) Continued fraction method

Difference-differential equations play a key role in the solution of most queueing models. Some of the queueing systems are studied analytically
by deriving the corresponding difference - differential equations and solving them by applying Rouche’s theorem through suitable generating functions. The first method is discussed elaborately by Saaty (1961) and Gross & Harris (1998). Neuts (1981) developed the matrix-geometric algorithmic approach to study the steady state queueing models. This method involves real arithmetic and avoids the calculation of complex roots based on Rouche’s theorem.

1.2.2 Non-Markovian Queueing Models

Most of the real time situations are not Markovian and offer wide scope for investigation. As such, there is a practical need for models that do not depend on strict Markov assumptions. Queueing models having the inter arrival times and / or service times which are not exponentially distributed are known as non-Markovian queueing models. The analysis of such queues could be based on Markovian that can be extracted out of it. Following are some of the techniques or approaches that can be used to solve such models:

(a) Embedded Markov chain technique
(b) Supplementary variable technique
(c) Lindley’s integral equation method

In the 1950’s Kendall (1951, 1953) developed a procedure to convert the queue length processes in M/G/1 and G/M/s into Markov chains. Embedded Markov chain technique is the simplest technique that employs only the Markov chain apparatus and is used to find the stationary state probabilities at the chosen epochs. Supplementary variable method is known as a simple and convenient way of deriving the double transform of the queue length and the supplementary variables used in arbitrary epochs. For more details on supplementary variables used in non morkovian models one can

1.3 LITERATURE SURVEY

In this section, relevant literature survey is provided for queues with priorities, queues with server breakdowns, Multi server queues with Loss system, queues with two phase services, queues with vacation, bulk queues with vacation and closedown, queues with negative customers, queues with impatient customers and queues with feedback of customers. Evaluation of the performance of a variety of science systems in the fields of manufacturing, computer and communication systems, transportations etc., are dependent on these, they play a vital role.

1.3.1 Queueing Theory

Real time problems of queueing theory was first raised by calls in telephone exchange and was first investigated by Erlang, for more details refer Erlang (1909, 1918). The pioneer and much celebrated book on queueing theory were written by Takacs (1962).

1.3.2 Queues with Priorities

Queueing systems with priorities often occur in computer systems and other applications. Finite source single server queue with priorities is analysed by Jaiswal & Thiruvengadam (1967). Veran (1985) proposed a numeric method for the analysis of a single server queue with pre-emptive priorities. Multi server queues with general service times are considerably less in the literature because the corresponding models are analytically intractable. Federgruen & Groenevelt (1988) discuss M/G/c queueing system with finite number of distinct customers with various priority disciplines.
Bondi & Buzen (1984) give the mean response times of each priority level in M/G/m queue operating under pre-emptive resume scheduling.

1.3.3 Queues with Server Breakdown and Priorities

The problem of queueing systems with server failures and priorities is of continuing interest to many researchers. In this regard several models are developed and analysed. White & Christie (1958) introduced queueing systems with server interruptions. They pointed out the similarity of queueing with breakdowns to queueing with pre-emptive priority and worked out two models of breakdown effects. Single server queue with two classes of customers according to pre-emptive resume priority with server failures was analysed by Towsley & Tripathi (1991). They analysed the system with a Poisson bulk arrival, exponential service times, general repair times and exponential inter failure time. Single server queueing model with service interruption including priorities alone is considered by Gaver (1962), Madan (1972). Madan (1992) analysed a queueing system with two types of failures having pre-emptive priority to the repair of the major failures over the minor ones with constant arrival rate of customers.

1.3.4 Multi Server Loss System


1.3.5 Queueing Models with Two Phases of Service

A study on two phase services is made by Krishna & Lee (1990). Doshi (1991) gives the analysis of a two phase queueing system with general service times. A single server general service queue with two phases of heterogeneous service and unreliable server is considered by Choudhury & Deka (2012). Choudhury et al (2009) considers single server, bulk arrival with general service queue with an additional second phase of optional service and unreliable server. This system consists of a breakdown and a delay period under N-policy with the assumption that the server is working with any phase of service and may break down at any instant. Julia Rose Mary et al (2011) analysed a single vacation queueing system with second optional service channel under bi-level control policy.

1.3.6 Queues with Vacation

The queueing model with server vacations (server absences) has been well studied in the past three decades and successfully applied in many areas such as manufacturing/service and computer/communication network systems. The aim of studying the queueing model with vacation is to utilize the idle time (vacation period) of the server, by which the total average cost involved may be minimized. Excellent surveys on the earlier works of vacation models have been found in Doshi (1986), Takagi (1991), Tian & Zhang (2006). There have been significant contributions to vacation queues by considering MAP and BMAP as input processes. For more details, Ferrandiz (1993), Matendo (1994), Kasahara et al (1996), Chang et al (2002) etc.


1.3.7 Queues with Vacation and Closedown

Very little works are in the literature by considering queues with single service with vacation and closedown. Using supplementary variable technique, Sakai et al (1998) analysed a finite-capacity M/G/1 queueing system with a mixture of the single vacation, close-down and set-up times which arises in Switched Virtual Channel Connection (SVCC) mechanism for an IP over ATM networks. Using supplementary variable technique, Moreno (2009) analysed a infinite capacity Geo/G/1 queueing system with multiple

1.3.8 Bulk Queueing Models with Vacations

The literature endorses that there have been significant contributions to bulk queueing models with vacations. Lee et al (1992) have analysed M/G(1,b)/1 queue with single vacation and obtained the probability generating function of the queue length distributions at arbitrary and departure epochs. Further in (1996) they discussed a fixed batch service queue with single and multiple vacations. The operating characteristics of an M$^X$/G/1 queueing system with unreliable server and single vacation are analysed by Haridass & Arumuganathan (2008).

The steady state analysis of the M$^X$/G/1/N queue with vacation time is analysed by Baba (1988). Banik et al (2006) considers a finite buffer single server queue with Batch Markovian Arrival Process (BMAP) with the assumption that the server serves a limited number of customers before leaving for vacation(s). They analysed single and multiple vacation policies along with two possible rejection strategies, namely, partial batch rejection and total batch rejection.

The Geo$^X$/G/1 queue subjected to multiple vacations governed by a geometrically distributed timer is analysed by Fiems & Bruneel (2002). Chang & Choi (2005) have analysed a discrete time finite-buffer queueing system with bulk arrival, bulk service in which the customers are served by a single server in batches of random size with multiple vacations.
1.3.9 **Bulk Queueing Models with Vacations and Closedown**

Few research work of bulk queues with vacations and closedowns are available in the literature. Arumuganathan & Jeyakumar (2004) analyse a single server infinite capacity bulk service queue with multiple vacations and closedown times. The detail study of BMAP/G/1/N queue with vacations under exhaustive service discipline was performed by Niu et al (2003), where they have included setup time, close-down time, single/multiple vacations. Applying supplementary variable technique, Frey & Takahashi (1999) considers an $M^X/GI/1$ finite capacity queue with close-down time, vacation time and exhaustive service discipline under the partial batch acceptance strategy as well as whole batch acceptance strategy.

1.3.10 **Queues with Impatient Customers**

Queueing models with impatient customers have been studied by various authors in the past and extensive literature is available in these kinds of models. A long wait already experienced in the queue or a long wait anticipated by a customer upon arrival makes the customer impatient. Altman & Yechiali (2006, 2008) analysed the models in a different way. They assumed that the customers become impatient only when the server(s) is (are) on ‘vacation’ and unavailable for service. That is, if upon arrival, no servers are ready to serve then the customer becomes impatient. Wang et al (2010) gives an extensive review on queueing systems with impatient customers. Single-server queues with impatient customers are considered by Baccelli et al (1984). General customer impatience in the queue GI/G/1 is analysed by Daley (1965). Adan et al (2009) gives a detailed analysis of queueing models with vacations and impatient customers, where the source of impatience is due to absence of the server. Van Houdt et al (2003) gives the algorithmic procedure to calculate the Delay distribution of (im)patient customers in a discrete time DMAP/PH/1 queue with age-dependent service times.
Stanford (1990) provides a bibliographical update on the subject of impatience in general queueing systems and pointed out some of the similarities and differences in the results from the existing papers on the subject. Barrer (1957) treated a queueing problem characterized by the impatience of the customers and the indifference of the clerks. Ancker & Gafarian (1963) studied $M/M/1/N$ queuing system with balking and reneging, and its steady state analysis. They also obtained the results for a pure balking system (no reneging) by setting the reneging parameter equal to zero. A single server queueing system with Poisson input and general service times under the assumption that the customers are lost, when his service has not begun within a fixed time after his arrival, is studied by De Kok & Tijms (1985).

1.3.11 Queues with Negative Customers

During the last decade considerable attention has been paid to studying queueing systems with negative arrivals. After the introduction of negative customers by Gelenbe (1991), research on queueing systems with negative arrivals has been greatly motivated by researchers along with some practical applications such as computer, neural networks, manufacturing systems and communication networks etc. Extensive studies on these themes can be found in Gelenbe & Schassberger (1992), Henderson (1993), Gelenbe (1993), Chao (1995), Artalejo & Gomez-Corral (1998). A different kind of negative arrivals, called disasters, was introduced by Jain & Sigman (1996). A disaster kills simultaneously all the customers in the system and has no impact on the empty system. It is also called mass exodus by Chen & Renshaw (1997) or queue flushing by Towsley & Tripathi (1991).
1.3.12 Queues with Feedback of Customers

Many queueing situations have the property that the customers may be served repeatedly, if the customer is not satisfied with the service provided by the server. These queueing models arise in many real-life situations. There have been wide literatures on feedback of customers with single and multiple types of feedback. Epema (1991) considers a general single server time sharing model with multiple queues and job classes, priorities and feedback. Geom$^X$/G/1 queue with Bernoulli feedback with server vacations and setup/closedown times are discussed by (Jia & Chen 2010, Jia et al 2011). Parthasarathy & Vasudevan (2009) considers a single server queue with multi-class Bernoulli feedback queue with gate mechanism. Disney et al (1984) gives the stationary queue-length and waiting-time distributions in single-server feedback queues.

1.4 OBJECTIVES OF THE WORK

The main objective of this study is to develop analytical treatment of some discrete and continuous time queueing models and to obtain performance measures. This research proposes different queueing models arising in various real time situations. The analytical treatment of all these discrete and continuous time models is obtained using supplementary variable technique and embedded Markov chain technique. The proposed models are theoretically developed and numerically illustrated with some of the parameters. Main objectives of this research are:

- to develop the theoretical framework for various discrete and continuous time queueing models.
- to motivate each model through real time situation.
- to obtain some important performance measures.
to analyze the performance measures with numerical illustration.

➢ to obtain some interesting particular cases and special cases.

1.5 THESIS ORGANISATION

The present study is devoted to the analysis of some discrete and continuous queueing models with server breakdown/repair, closedown and vacation. Priority queues of loss system with different types of customers, feedback, impatient and negative customers are also developed. The thesis is organised as follows and is schematically represented in Figure 1.2:

Chapter 2 concentrates on the loss system with pre-emptive priorities and server breakdown. It is assumed that two types of customers arrive, namely, high priority customers (type 2) and low priority customers (type 1). The type 1 and type 2 customers are assumed to arrive according to Poisson process with rates $\lambda_1$ and $\lambda_2$ respectively. The customers are served immediately if the servers are available. However, if all ‘C’ servers are busy, then on the arrival of a high priority customer, the service of a low priority customer is pre-empted. On the other hand if all the servers are busy with high priority customers, arrival of low and high priority customers are a loss to the system. The service times of each is considered to follow general distribution. The time after which a server breaks down and the time taken to repair the same, by a single repair facility, is assumed to follow exponential distribution with parameter $\alpha$ and $\beta$ respectively. The steady state distribution of such underlying system is obtained together with the performance measures. A numerical illustration of the obtained results is also presented. The results are validated using simulation.
Chapter 3 analyses the MAP/G/1/N queue with an additional second phase of optional service in which customers arrive according to Markovian arrival process. There is a single server which provides both the phases of service, namely, First Phase of Essential Service (FPES) and Second Phase of Optional Service (SPOS). At the completion of FPES
(denoted by $S_1$), the customer may leave the system with probability $q (=1 - p)$ or may be provided with the SPOS (denoted by $S_2$) with probability $p (0 \leq p \leq 1)$. On completion of FPES or SPOS, if the number of waiting customers is at least one, the server starts FPES according to FIFO rule. On the other hand, if the server finds an empty system at any service completion epoch (either FPES or SPOS), the server avails single or multiple vacation according to the indicator function $\delta_s$. If $\delta_s = 1$, the server avails single vacation and becomes idle(dormant) when the system becomes empty, otherwise starts FPES. If $\delta_s = 0$, the server avails multiple vacation until the number of waiting customers is at least one, then starts FPES. On the other hand, if a customer on arrival finds $N$ customers in the waiting room, then the customer leaves the system without being served and this is assumed to be a loss to the system. The queue length distributions at various epochs, such as pre-arrival, arbitrary and departure have been obtained using embedded Markov chain technique and supplementary variable technique. Waiting time of a customer in the queue is also obtained. Numerical illustrations are also provided.

Chapter 4 gives the cost analysis of a finite capacity queue with batch service, multiple vacations and closedown times in which the arriving customers are served by a single server in batches of maximum size ‘$b$’ with the minimum threshold value ‘$a$’. Customers arrive according to a Markovian Arrival Process (MAP). On completion of a service, if the queue length is less than ‘$a$’, then the server performs a closedown work and then leaves for a vacation of random length. When the server returns from vacation and if the queue length is still less than ‘$a$’ the server avails another vacation and continuous to do so until the server finds at least ‘$a$’ customers waiting for service in the queue. After the completion of a service, if the number of customers in the queue is at least a specified value ‘$a$’ then the server will continue the batch service with general bulk service rule. If the server finds at
least ‘a’ customers at the completion of closedown job, it immediately starts service according to general bulk service rule. Probability generating function of the queue length distribution at arbitrary epoch is obtained. Some key performance measures are also obtained. Cost model is discussed with Numerical illustration.

Chapter 5 investigates the discrete time infinite capacity queueing system with correlated arrival and negative customers served by two state Markovian server. Positive customers are generated according to the first order Markovian arrival process with geometrically distributed lengths of On periods and Off periods. The arrivals of positive customers are considered in the stochastic manner and are stored in the infinite waiting room on a first-come, first-serve basis. In the waiting room, the positive customers wait for some time, until finally; they receive service from the server. Further, the geometrically distributed arrival of negative customer removes the positive customers if any, and has no effect when the system is empty. The server state is a two state Markov chain which alternates between Good and Bad states with geometrically distributed service times. Closed form expressions for mean queue length, mean waiting time, sojourn time are obtained. Numerical illustrations are also presented.

Chapter 6 investigates the discrete time single server general service infinite capacity queueing system, On-Off source arrivals with negative customers and Bernoulli feedback. Positive customers are generated according to the first order Markovian arrival process with geometrically distributed lengths of On periods and Off periods. Service times of these positive customers are generally distributed. After service completion, an unsatisfied customer immediately joins the queue with probability \( p \) and a satisfied customer leaves the system permanently with probability \( q \) (\( p + q = 1 \)). Further, the geometrically distributed arrival of negative customer
removes the positive customers if any, and has no effect when the system is empty. Using Z–transform technique, closed form expressions for mean queue length, mean waiting time, sojourn time are obtained. Numerical illustrations are also presented.

Chapter 7 studies the discrete time finite buffer queue with server subject to starting failures and impatient customers. Whenever the starting failure occurs, the server is immediately sent to repair station and customers who are patient can join the queue with probability $\beta$ and impatient customers may leave the system with compliment probability $\overline{\beta}$. The inter arrival time of customers are assumed to be discrete time renewal process arrivals and service/repair times are mutually independent and are geometrically distributed. Using supplementary variable and embedded Markov chain techniques, probability generating function of actual waiting time distribution in the system and some performance measures have been carried out. Finally, numerical examples are presented to illustrate the effect of the parameters on several performance characteristics.

Chapter 8 concludes the thesis by presenting the overview of all the proposed continuous and discrete time queueing models and their scope for future enhancements.