Chapter 1

Introduction

Queueing models are effectively used in the design and analysis of telecommunication systems, traffic systems, service systems and many more. Queueing theory is a mathematical study of waiting lines or queues. The formation of a queue is a common phenomenon which occurs whenever the current demand for service exceeds the current capacity to provide that service. Queues may be seen at railway platforms, cinema ticket windows, billing counters of super markets etc. The main aim of queueing theory is to minimize the congestion and provide best level of service at minimum possible cost. The waiting lines arise from congestion which occurs from time to time as a result of irregularities in arrivals or in the length of time needed to service customers. Assuming that the irregularities mentioned above follow some probabilistic laws, the queueing theory attempts to study statistical fluctuations in these irregularities. It also helps to determine the probability of number of customers in system, the average system size, the average queue length, average waiting time of customers in the system, average server utilization etc.

Queueing theory grows considerably with the advent of Operational Research in late 1940’s and early 1950’s. It has its origin in the pioneer work done by Erlang (1909) in Copenhagen Telephone Exchange on the applications of probability theory to the congestion problem in telephony. The theory is then developed by Molina (1927), Fry (1928), Pollaczek (1923), Kolmogroff (1931), Khintchine (1932,1933), Crommelin (1932, 1934) and Palm (1937).

1.1 Queueing System and its Characteristics

Queueing System

A queueing system is generally defined as something having an input, an output and a transformation process in between. It consists of number of customers in the queue, customers in the service, the service facility (number of servers) and the
customers served. In a queueing system, customers requiring service form input, the serviced customers the output and service rendered is the transformation process.

**Characteristics of a queueing system**

The basic characteristics of a queueing system are as follows:

(i) **The input or the arrival pattern of customers**
   The customers arrive from outside into a queueing system either one by one or in batches of fixed or variable size. The arrivals may either occur at equidistant time intervals or occur at random time points. The mean arrival rate represents the flow of arrivals to the system. If the arrivals are random they may be better described by a probability distribution.

(ii) **The pattern of service or service mechanism**
    Service mechanism deals with how the customers are served. The customers may be served either one by one or in the batches of fixed or variable size. The service times may be constant or random. The mean service rate describes the intensity of service. The random service times can be modelled by a probability distribution.

(iii) **The number of servers or service channels**
    A system may have one or more than one server. Service channels are the number of parallel service centres in a queueing system that can serve customers simultaneously. In case of multi server queueing systems, there may be one single queue for servers or a single queue for each server.

(iv) **The capacity of the system**
    Capacity of the system refers to the number of customers a system can accommodate. When the capacity of the system is limited, the system is called finite capacity queueing system. The system in which there is no limit on the number of customers is known as infinite capacity queueing system.

(v) **The queue discipline**
    Queue discipline refers to the way how the customers are taken for service. The most commonly used queue discipline is first come, first served (FSFS). In this case the waiting customers are selected for service in the order of their arrivals. We can have random selection for service or service in random order (SIRO). The queue discipline, last come, first served (LSFS) is just opposite of first in first out discipline. A number of priority disciplines are also available where customers are given priorities upon entering the system.
1.2 Probability Distributions of Inter-arrival and Service Times

The commonly used probability distributions used for describing the inter-arrival and service times distributions are discussed in this section.

Kendall’s Notation of a queueing system

Kendall (1948) gives notation for describing the above first three characteristics of a queueing system viz. the inter-arrival time distribution, the service time distribution, and the number of service channels. Later, Lee (1966) added fourth and fifth characteristics to the notations i.e. the queue discipline and the maximum number in the system. The Kendall-Lee notation is A/B/C/D/E, where A represents the mean inter-arrival time distribution, B represents the mean service time distribution, C represents the number of servers, D represents the capacity of queueing system and E represents the queue discipline.

1. Markovian or the negative exponential distribution (M)
   This is the most commonly used distribution in queueing literature. Here the letter M represents that the inter-arrival time or service time distribution is exponential when the time is continuous.

2. Deterministic or constant inter-arrival / service times (D)
   The letter D in place of A or B or both in Kendall’s notation indicates that the inter-arrival time or service time is constant.

3. k-Erlang distribution \( E_k \)
   When inter-arrival times have Erlang distribution of order k that is each inter-arrival time (or service time) is made of k phases and time spent in each phase is negative exponential with a common mean \( \frac{1}{k\mu} \) (or \( \frac{1}{k\lambda} \)) where \( \lambda \) (or \( \mu \)) is the arrival rate (or service rate).

4. General distribution (G) and General Independent distribution (GI)
   The symbol G is used for a general or arbitrary distribution for inter-arrival time or service time depending on the position in Kendall’s notation. Also the symbol GI is used for general input distribution.

The notation A/B/C is used when the capacity of the system is infinite and queue discipline is first come, first served. The queueing system M/M/1/\( \infty \)/FCFS can be written as M/M/1. From the notation explained above, we see that

(i) M/M/1 stands for Poisson arrivals, exponential service times and one server.
(ii) M/D/c stands for Poisson arrivals, constant service times and c servers.
(iii) \( M/E_k/2 \) stands for Poisson arrivals, \( k \)-Erlang service times and two servers.

(iv) \( GI/G/c \) stands for the most general case of general input, general service times and \( c \) servers.

Queueing models can be classified as Markovian or non-Markovian queueing models.

**Markovian Queueing Models**

The queueing models in which both the inter-arrival time distribution and the service time distribution are Markovian in nature are known as Markovian queueing models. In continuous time queueing systems if the inter-arrival time distribution and the service time distribution is exponential then such systems are Markovian; while in discrete time queues if the inter-arrival time and service time distribution are geometric then such queueing systems are discrete-time Markovian queueing systems. For example, \( M/M/1, M/M/c \) and \( MX/M/1 \) etc.

**Non-Markovian Queueing Models**

If either the inter-arrival time distribution or service time distribution or both are not Markovian in nature then such queueing systems are known as non-Markovian queueing systems. For example, \( M/G/1, G/G/1 \) and \( G/M/1 \) etc.

### 1.3 Birth-Death Process

Consider, a continuous parameter stochastic process \( \{ X(t) : t > 0 \} \) with discrete state-space \( 0, 1, 2, 3, \cdots \). Suppose, this process describes a system that is in state \( E_n; n = 0, 1, 2, 3, \cdots \) at time \( t \), iff \( X(t) = n \) i.e. the system has a population of \( n \) elements or customers at time \( t \). Then the system is described by a birth-death process if there exists non-negative birth rates \( \lambda_n; n = 0, 1, 2, 3, \cdots \) and non-negative death rates \( \mu_n; n = 0, 1, 2, 3, \cdots \) such that following postulates are satisfied:

1. State changes are allowed from \( E_n \) to \( E_{n+1} \) or from \( E_n \) to \( E_{n-1} \) if \( n \geq 1 \), but from \( E_0 \) to state \( E_1 \) only that is, only one birth and no death can occur if the system is empty.

2. If at time \( t \) the system is in state \( E_n \), the probability that in the time interval \( (t, t+\delta t) \) the transition from \( E_n \) to \( E_{n+1} \) occurs is equals to \( \lambda_n \Delta t + o(\Delta t) \) and probability that the transition from \( E_n \) to \( E_{n-1} \) occurs is equal to \( \mu_n \Delta t + o(\Delta t) \). This postulate gives the probability of a birth and a death in a very small interval of length \( \Delta t \) when the system population is \( n \).

3. The probability that in the time interval \( (t, t+\Delta t) \) more than one transition occurs is \( o(\Delta t) \). That is probability of more than one birth or death in a very small interval of length \( \Delta t \) is negligible.
Define, \( P_n(t) = P[X(t) = n] \) that the system is in state \( E_n \) at time \( t \), the differential difference equations are given by

\[
\frac{dP_n(t)}{dt} = -(\lambda_n + \mu_n)P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t), \quad n \geq 1 \quad (1.1)
\]

\[
\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (1.2)
\]

In steady-state, when time is very very large.

\[
\lim_{t \to \infty} \frac{dP_n(t)}{dt} = 0, \quad \text{and} \quad \lim_{t \to \infty} P_n(t) = P_n
\]

The above equations become

\[
0 = -(\lambda_n + \mu_n)P_n + \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}, \quad n \geq 1 \quad (1.3)
\]

\[
0 = -\lambda_0 P_0 + \mu_1 P_1 \quad (1.4)
\]

On solving these equations recursively, we have

\[
P_n = \frac{\lambda_{n-1}\lambda_{n-2}\lambda_{n-3}...\lambda_0}{\mu_n\mu_{n-1}\mu_{n-2}...\mu_1} P_0 \quad (1.5)
\]

Using \( \sum_{n=0}^{\infty} P_n = 1 \), we have

\[
P_0 = \left\{ 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} + ... + \frac{\lambda_0\lambda_1\lambda_2...\lambda_{n-1}}{\mu_1\mu_2\mu_3...\mu_n} + ... \right\}^{-1} \quad (1.6)
\]

The steady-state solution exits if

\[
\left\{ 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3} + ... + \frac{\lambda_0\lambda_1\lambda_2...\lambda_{n-1}}{\mu_1\mu_2\mu_3...\mu_n} + ...ight\} < \infty
\]

Thus, the equations (1.5) and (1.6) provide the steady-state solution to birth-death process described in (1.1)-(1.2).

### 1.4 Solutions of Queueing Models

The solution of a queueing system may either be time dependent (transient) or time independent (steady-state).

1. **Transient Solution**
   A queueing system is said to be in transient state when its operating characteristics (behaviour) are dependent on time. One can obtain steady state solution from the transient solution when the time is sufficiently large. Normally, it is difficult to obtain time dependent solution of a queueing system.

2. **Steady state solutions**
   Steady-state solutions provide the results which are independent of time. It describes the long run behaviour of queueing system.
1.5 Methods of Solving Queueing Models

Here, we describe some methods of solving queueing systems.

1. **Iterative (or Recursive) Method**
   In case of this method, we iteratively use the given difference equations to obtain a sequence of state probabilities $P_1, P_2, P_3, \ldots$, each in terms of $P_0$ and then use normalizing condition $\sum_{n=0}^{\infty} P_n = 1$ to find $P_0$ and all the state probabilities of the queueing system.

2. **Probability Generating Function Technique**
   The probability generating function $P(z) = \sum_{n=0}^{\infty} P_n z^n$, where $z$ a complex number with $|z| = 1$ can be used to find the steady state probabilities $P_1, P_2, P_3, \ldots$, of the queueing system. The procedure involves finding the power series expansion for state probabilities and the coefficient of $z^n$ gives the value of $P_n$ for $n = 1, 2, 3, \ldots$. In order to find $P_0$, one can use normalizing condition $\sum_{n=0}^{\infty} P_n = 1$ and the value of probability generating function $P(z)$ for $z = 1$. Moreover, when the series expansion cannot be found, one can use derivative of $P(z)$ at $z = 1$ to obtain the expected number in the system (i.e. $L_s = \sum_{n=0}^{\infty} nP_n$). This method is discussed in Gross and Harris (1985), Kleinrock (1975) and Saaty (1961).

3. **Matrix geometric Method**
   When the differential difference equations of a queueing system have repetitive structure, then this repetition allows one to determine a recursive solution for stationary state probabilities since it implies that if one knows the stationary state probability of any state $n$ then stationary probability for $n + 1$ can be determined. The stationary state probabilities for the repeating portion of the process thus have a geometric form. This form is found in all scalar state processes that have a similar repetitive structure. The matrix geometric method is developed by Neuts (1981).

4. **Continued Fractions Approach**
found in Jones and Thron (1980) and Bowman and Shenton (1989). A fraction of type
\[ \frac{a_1}{b_1 + \frac{a_2}{b_2 + \frac{a_3}{b_3 + \cdots}}} \]
Where \(a_n\) and \(b_n\) are real or complex numbers. This fraction can be terminated by \(a_1, b_1, a_2, b_2, \cdots, a_n, b_n\) and dropping all the remaining \(a_{n+1}, b_{n+1} \cdots\). The number obtained by this operation is called \(n^{th}\) convergent or \(n^{th}\) approximate and is denoted by \(\frac{A_n}{B_n}\). Both \(A_n\) and \(B_n\) satisfy the recurrence relation.
\[ U_n = a_n U_{n-2} + b_n U_{n-1} \]
with initial value \(A_0 = 0, A_1 = a_1\) and \(B_0 = 1, B_1 = b_1\).
There is a close connection between birth and death processes and continued fractions. This connection is exploited to find time-dependent solutions of certain birth and death processes and to determine birth and death rates.

5. **Matrix Method approach**
The computable matrix approach is used to derive state probabilities of finite capacity Markovian queueing systems. In this approach, a simple equation is presented to obtain the transient probabilities in terms of the eigenvalues of a symmetric tri-diagonal matrix. The closed form solution of finite source queueing model can be obtained. This method is developed by Sharma (1997) and is extensively applied to solve finite capacity, Markovian queueing models by Sharma and Nair (1991, 1996), Sharma and Dass (1988a, 1988b, 1988c, 1990) and Sharma and Maheshwar (1993, 1994).

### 1.6 Queueing with Customer Impatience

Queueing models have effectively been used in the design and analysis of telecommunication systems, traffic systems, service systems and many more. A number of extensions in the basic queueing models have been made and the concepts like vacations queueing, correlated queueing, retrial queueing, queueing with impatience and catastrophic queueing have come up. Of these, queueing with customer impatience has special significance for the business world as it has a very negative effect on the revenue generation of a firm. A customer is said to be impatient if he tends to join the queue only when a short wait is expected and tends to remain in the line if his wait has been sufficiently small. Customer impatience is of three types, Gross and Harris (1985).

1. **Balking**
When a customer sees a long queue upon his arrival and decides not to enter the queue is called balking.
2. Reneging

It is a type of customer behaviour in which a customer enters the queueing system for service, waits for some time, and if this wait is more than the expected wait he decides to leave the queue without getting service. This is known as reneging.

3. Jockeying

When there are number of parallel service channels having their own queues and the customers switch from one queue to another to receive service more quickly is known as jockeying.

The notion of customer impatience appears in the queueing theory in the work of Palm (1937), Udagawa and Nakamura (1957) and Haight (1957). Haight (1957) considers a model of balking for M/M/1 queue in which there is a greatest queue length at which an arrival would balk. This length is a random variable whose distribution is same for all customers. Haight (1959) studies a queue with reneging in which he study the problem like how to make rational decision while waiting in the queue and the probable effect of this decision. Finch (1959a) studies reneging in a GI/M/1 queue. Ancker and Gafarian (1963a) study M/M/1/N queueing system with balking and reneging and perform its steady state analysis. Ancker and Gafarian (1963b) obtain results for a pure balking system (no reneging) by setting the reneging parameter equal to zero. Rao (1968) studies queueing with balking and reneging in M/G/1 systems. Beccelli et al. (1989) study single server queues with impatient customers. Stanford (1979) studies general queueing systems in which the $n^{th}$ arrival may renege, i.e., withdraw from the system after entering, if he does not begin service within a specified (random) amount of time $Z_n$ after entering the system. A very nice review on queueing systems with impatient customers is presented by Wang and Jiang (2010). They survey various queueing systems according to various dimensions like customer impatience behaviors, solution methods of queueing models with impatient customers and associated optimization aspects.

Balking is a common phenomenon of not joining a queue because the prospective arriving customer estimates the queue to be too long. The concept of balking finds its applications not only in daily life, but in computer communication, production systems and hospital management. The effect of balking on the state dependent M/M/1/N queue with reflecting barrier, reneging, additional server for longer queue etc. is discussed by many researchers. Abou-EI-Ata (1991) extends the model of Ancker and Gafarian (1963a) to study the state-dependent M/M/1/N queue with reneging and a general balk function. Al-Seedy and Kotb (1991) present transient solution of a state dependent queue: M/M/1 with balking. Abou-EI-Ata and Kotb (1992) study linearly dependent service rates for the queue: M/M/1/N with general balk function, reflecting barrier, reneging and additional server for longer queue. Abou-EI-Ata and Shawky (1992) study single server Markovian overflow

Wang and Chang (2002) study cost analysis finite M/M/R queueing system with balking, reneging and server breakdowns. A preemptive priority Markovian queue with state-dependent service and lower priority balking customers has been studied by Drelic and Woolford (2005). Yue et al. (2006) analyze M/M/c/N queue with balking, reneging and server vacations. Chakravarthy (2009) studies a disaster queue with Markovian arrivals and impatient customers. Adan et al. (2009) study queueing systems with vacations and synchronized reneging. Bae and Kim (2010) studies an G/M/1 queue with impatient customers. Ioannidis et al. (2010) present a queueing analysis for problems of inventory and admission control for make-to-stock production system with perishable inventory and impatient customers. Perel and Yechiali (2010) study M/M/c queue for $c = 1, 1 < c < \infty, c = \infty$ in a two phase (fast and slow) Markovian random environment, with impatient customers. They derive analytic results for queue length distributions using probability generating function technique. Mehandiratta (2011) discusses the application of queueing theory in health care, especially the health service planning. Recently, Senthil Kumar and Arumuganathan (2010) study single server batch arrival retrial queue with active breakdowns, two types of repair and second optional service (SOS). The server provides preliminary first essential service (FES) to the primary arriving customers or customers from retrial group. The customer under service decides probabilistically to remain in service or join the orbit during the breakdown. Liau (2011) develops a queueing model for estimating business loss. He uses Balking index and reneging rate to represent different configurations of balking behavior and reneging behavior respectively for different queueing systems. Pan (2011) considers M/M/1/N queueing model with variable input rates, variable service rates and impatient customers. He obtains the stationary distribution of the model.

Kapodistria (2011) gives a detailed analysis of a single server Markovian queue with impatient customers that perform synchronized abandonments. She considers standard assumption that customers perform independent abandonments and abandon the system simultaneously. Two abandonment scenarios are considered; in the first one all present customers become impatient and perform synchronized abandonments, while in the second scenario, the customer in service from the abandonment procedure. Choudhury and Medhi (2011) study balking and reneging in multi-server Markovian queueing systems. Kumar (2012) investigates a correlated
queueing problem with catastrophic and restorative effects with impatient customers which have special applications in agile broadband communication networks.

Queuing systems with customer impatience can be used for describing some types of perishable inventory systems as there is an analogy between queueing systems with reneging and perishable inventory systems. The on hand inventory can be considered as a queue, the demand completion as completion of service, the products coming in the form of replenishment as arrivals to the queueing system, and the life time of a product as the impatience (reneging) time. There is a considerable literature related to the modelling of perishable inventory systems using queueing systems with impatient customers. Customer reneging and product perishing are similar phenomena. A customer whose patience time expires leaves the queue and, likewise, a product made to stock whose lifetime expires is removed from the inventory.

In supply chains, the perishable items like vegetables, fruits etc. in the congestion situations become worthless if they are not supplied to the retailers (customers) at proper time as they may get spoiled i.e. the perishable items can be modelled by the reneged customers in the queueing modelling. The applications of queueing with customer impatience are mentioned in the work of Palm (1937), Gnedenko and Kovalenko (1968), Ancker and Gafarian (1963a, 1963b), Barrer (1957a, 1957b), Nahmias (1973), Graves (1978), and Kaspi and Perry (1983). Sasieni (1961) applies double ended queues and impatient customers to study inventory systems with perishable items.

In the call centres usually a calling customer hangs up before service agent and thus gets reneged. In packet switched communication networks with time critical traffic, a packet loses its value if it is not transmitted within a given time interval. The patients (customers) who leave the emergency rooms in hospitals without been seen are also considered as reneged customers. Kidney transplant waiting system can be considered as a queue with reneging, where reneging occurs because a customer that is waiting for a kidney may die.

1.7 Motivation

This study is motivated by the fact that customer impatience leads to loss of potential customers. It has become a highly challenging problem in the current era of cut-throat competition. Customers are the backbone of any business, because without customers there will be no reason for a business to operate. Therefore, the concept of customer retention assumes a tremendous importance for the business management. Firms are employing a number of customer retention strategies to
sustain their businesses. Studies across a number of industries have revealed that the cost of retaining an existing customer is only about 10% of the cost of acquiring a new customer, so customer retention is essential for a company (Wang and Lei, 2010). It is envisaged that a reneging customer can be retained with certain probability if some customer retention mechanism is employed. That is, there is a probability that a reneging customer can be retained. Therefore, we incorporate the probability of retaining a reneging customer into various Markovian queueing models with customer impatience.

1.8 Queueing with Retention of Reneged Customers

Queueing with customer impatience has been a hot topic of research since last six decades. It finds its applications in various areas like call centers, packet-switched communication networks, hospitals, perishable inventory systems etc. Customer impatience is harmful for any business as without customers there is no business. But, if the firms employ certain customer retention strategies then there are chances that some or all of the reneging customers may be retained in the queueing system. That is, there is a probability that a reneging customer may be retained.

Kumar and Sharma (2012a) takes this idea into consideration and incorporate it into an \( M/M/1/N \) queueing system with reneging. According to them, each customer upon joining the queue waits for a certain length of time \( T \) (say) for his service to begin. If it does not begin by then he may renge with probability \( p \) or may remain in the queue for his service with probability \( q (= 1 - p) \) if certain customer retention strategy is applied. The probability \( q \) is termed as the probability of retaining a reneging customer. They derive steady-state solution of the model and obtain some important performance measures like average system size, average reneging rate, average retention rate etc. They also perform the sensitivity analysis of the model. They found that the introduction of the probability of retaining a reneging customer has positive impact on the expected system size.

Kumar and Sharma (2012b) incorporate the concept of retention of reneged customers into an \( M/M/1/N \) queueing system with reneging and balking. They derive the steady-state solution of the model. The cost-profit analysis of the model is performed by Kumar and Sharma (2012c) wherein they show the application of the model in supply-chain management. Kumar and Sharma (2013c) propose the application of \( M/M/1/N \) queueing system with reneging and retention of reneged customers to study the spoilage prevention strategies in supply chains. They perform the cost-profit analysis of the model in order to study the long run impact of different spoilage prevention strategies on the revenue generation of a supply chain firm.
Kumar and Sharma (2012e) propose a multi-server finite capacity Markovian queueing system with retention of reneged customers for the study of product replacement strategies in perishable inventory systems. The related cost-profit analysis is performed. The multi-server finite capacity Markovian queueing systems with reneging and retention of reneged customers are further analyzed by Kumar and Sharma (2013a, 2013b). Kumar and Sharma (2014c) obtain the steady-state solution of multi-server finite capacity Markovian queueing system with balking, reneging and retention of reneged customers. Kumar (2013) perform economic analysis of an $M/M/c/N$ queueing system with balking, reneging and retention of reneged customers. The time dependent solution of the model is studied. Sharma and Kumar (2012a, 2012b, 2014c) study feedback queueing system with customer impatience and retention of reneged customers. Kumar and Sharma (2014a, 2012d) study single and multi-server Markovian queueing systems with discouraged arrivals, reneging and retention of reneged customers. They derive the steady-state solution of the model and provide some important performance measures.

In real life situations the service rates of servers are not always homogeneous rather they are heterogeneous. Heterogeneous queueing systems are used as modeling tool for various computer communication systems. Recently Kumar and Sharma (2014b, 2014c) incorporate the concept of retention of reneged customers in the queueing systems with two-heterogeneous servers facing customer impatience. Further, Kumar and Sharma (2014d) propose a finite capacity Markovian queueing model with reneging, retention of reneged customers and two heterogeneous servers to study an insurance business with surrenders. In insurance business, due to dissatisfaction of service and better alternative elsewhere customers may decide to surrender their policies before maturity. This cause monitory loss to the insurance firm. If the insurance firm employ certain customer retention strategies then there are chances that customers may decide to continue their policies upto maturity. The long run behavior of the system is studied and the performance measures like average number of policies in the system, average rate of surrender, average rate of retention are obtained. Finally, the cost-profit analysis of the model is presented.

1.9 Queues with Discouragement

In case of discouragement, the customers arrive in a Poisson fashion with rate that depends on the number of customers present in the system at that time e.g. in case of single server, the arrival rate is $\frac{\lambda}{n+1}$. In case of multi-server Markovian queueing model, there are $c$ servers and the customers arrive to a multi-server system in a Poisson fashion with rate $\lambda$, and a customer finding every server busy, arrives with arrival rate that depends on the number of customers present in the system at that time i.e. if there are $n(n > c)$ customers in the system, the new customer enters the
system with rate $\frac{\lambda}{(n-c)+1}$.

Queueing models where potential customers are discouraged by queue length are studied by many researchers in their research work. Natvig (1975) studies the single server birth-death queueing process with state dependent parameters $\lambda_n = \frac{1}{(n+1)} \lambda$, $n \geq 0$ and $\mu_n = n$, $n \geq 0$. He reviews state dependent queueing models of different kind and compare his results with $M/M/1$, $M/D/1$ and $D/M/1$ and the single server birth-and-death queueing model with parameters $\lambda_n = \lambda$, $n \geq 0$ and $\mu_n = \mu$, $n \geq 1$ numerically. Cuortois and Georges (1971) study finite capacity $M/G/1$ queueing model where the arrival and the service rates are arbitrary functions of the current number of customers in the system. They obtain results for expected value of time needed to complete a service including waiting time distribution and limited probability distribution of the congestion. Hadidi (1974) carries out analyzes of busy period processes for $M/Mn/1$ and $Mn/M/1$ queueing models with state dependent service and arrival rates. He also obtains results for busy period and transient state probabilities. Von Doorn (1981) obtains exact expressions for transient state probabilities of the birth death process with parameters $\lambda_n = \lambda$, $n \geq 0$ and $\mu_n = \mu$, $n \geq 1$.

Sharma and Maheswar (1993) study finite capacity single server Markovian queue with state dependent parameters and obtain closed form computable transient probabilities using matrix method. Knossel et al. (1994) study state-dependent GI/G/1 queueing system characterized by the unfinished work $U(t)$ in the system at time $t$. They consider the limit of short inter-arrival times and small service requests and compute asymptotic approximations to the stationary density of the unfinished work, including the stationary probability of finding the system empty. Ammar et al. (2012) study single server finite capacity Markovian queue with discouraged arrivals and reneging using matrix method. They compare their numerical results for expected queue length with Monte Carlo Simulation to show the efficiency and accuracy of the method used.

The effect of discouraged arrivals in a multi-server Markovian queueing model with retention of reneged customers is studied by Kumar and Sharma (2012d). They obtain the steady-state solution. Some useful measures of effectiveness are derived and discussed. The effect of discouraged arrivals on the expected system size is studied. The cost-profit analysis of the model is performed. The model is compared with the model by Kumar and Sharma (2013e) to study the effect of discouraged arrivals on the total expected profit of the system. Some important queueing models are also derived as the special cases of the model.
1.10 Feedback Queues

In case of feedback, after the completion of service, each customer may rejoin the system as a feedback customer for receiving another regular service with probability $p_1$ or he can leave the system with probability $q_1$ where $p_1 + q_1 = 1$. The customers both newly arrived and those that are fed back are served in order in which they join the tail of original queue. We do not distinguish between the regular arrival and feedback arrival. The customers are served according to first come, first served rule.

Finch (1959b) introduces the concept of feedback in cyclic queues. Takacs (1963) studies queue with feedback to determine the stationary process for the queue size and the first two moments of the distribution function of the total time spent in the system by a customer. Davignon and Disney (1973) study single server queues with state dependent feedback. Santhakumaran and Thangaraj (2000) consider a single server queue with impatient and feedback customers. They study $M/M/1$ queueing model for queue length at arrival epochs and obtain result for stationary distribution, mean and variance of queue length.

Sharma and Kumar (2012a) study single-server finite capacity, Markovian queue with feedback and retention of reneged customers. They perform steady-state analysis of the model. Sharma and Kumar (2012b) further study $M/M/1/N$ feedback queueing model with balking and retention of reneged customers in the same year. They obtain steady-state solution of the model. They derive important performance measures of the model. Some queueing models are derived as special cases of the model. The infinite capacity case of models in (2012a, 2012b) is solved in Kumar and Sharma (2013d).

1.11 Queues with Heterogeneous-Servers

Heterogeneous servers serve their customers with different service rates. We consider queueing system with two heterogeneous servers in chapter 8. The service times of servers are exponentially distributed with parameters $\mu_1$ and $\mu_2$ such that $\mu_1 > \mu_2$. A unit arrives to find both the server engaged, it waits in a line in order of arrival and the unit in front of the queue occupies the server that falls vacant first. When only one of the server is free, the arriving unit occupies the free server. When both the channels are free, the arriving unit chooses fast server with probability $\pi_1$ and slow server with probability $\pi_2$, such that $\pi_1 + \pi_2 = 1$.

Service stations which are not mechanically controlled like checkout counters, grocery stores, banks and departments etc. have heterogeneous service because one can
not expect human servers to work at constant rate. A two-heterogeneous-server
Markovian queueing system has been extremely studied by a number of researchers
following pioneering work of Morse (1958) who considers the situation of some hyper-
exponential distributions for service times with parallel channels. Saaty (1960) ob-
tains time dependent solution of multi-server Markovian queue with heterogeneous
servers. Krishanamoorthy (1963) considers a Poisson queue with two-heterogeneous
servers with modified queue disciplines. The queue disciplines when more than one
server is free, in the first discipline, the incoming unit prefers the server with fast
service rate than the one with the slow service rate and in the second discipline,
the incoming unit wait for service in front of fastest server than to go into the slow-
est server if the estimate of total time spent in the system is any less by doing so.
Further, the steady-state solution, transient solution and busy period distribution
for the first discipline and the steady-state solution for the second discipline are
obtained.

Singh (1970) extends the work of Krishanamoorthy (1963) on heterogeneous
servers by incorporating balking to compares results with homogeneous server queue
to show the conditions under which the heterogeneous system is better than the
corresponding homogeneous system. Sharma and Dass (1988c) analyze M/M/2/N
queueing system with heterogeneous servers to derive the probability density func-
tion of the busy period. The mean and variance of the queueing system is also found.
Moreover, the time-dependent solution of a limited space double channel Markovian
queueing model with heterogeneous servers using matrix method is found by Sharma
(1963) by considering three heterogeneous server instead of two server. El-Paoumy
(2008) consider batch arrivals and balking while El-Paoumy and Nabwey (2011),
and El-Sherbiny (2012) include balking, and general balking function in a Markovian
queueing system with reneging and two heterogeneous server to derive study-state
solution using iterative method.