CHAPTER 1

INTRODUCTION

Queueing theory is a branch of applied mathematics utilizing concepts from the field of stochastic processes. It has been developed for the purpose of better understanding of queueing systems and for the sake of taking appropriate decisions to maintain the systems efficiently. The study of queueing systems finds applications in a variety of real life situations like business, industry, engineering, transportation, communication, computer and consumer activities.

1.1 QUEUEING SYSTEM

A queue or a waiting line is formed when units (customers) needing some kind of service, arrive at the facility which provides the service they seek. A queueing system can be described by the flow of units for service, waiting for service if it is not immediate and leaving the system after receiving service or sometimes even without receiving service. The units may be animate or inanimate, for example, persons waiting at a bank or at railway booking office, machines waiting to be repaired, airplanes at a busy airport waiting for take-off, computer programmes waiting to be run at a time sharing environment, etc.

1.2 CHARACTERISTICS OF QUEUEING SYSTEMS

The basic characteristics of a queueing system are arrival process, service mechanism, queue discipline, system capacity and service channels. These provide an adequate description of the system.
Arrival Process

Arrival or input process describes the manner in which units arrive and join the system. The source from which the units come may be finite or infinite. A unit may arrive either singly or in a group. In the case of group arrivals, the input is said to occur in bulk or batches. The interval between two consecutive arrivals is called the inter-arrival time.

Service mechanism

Service mechanism describes the manner in which service is rendered. Customer may be served either singly or in batches. The queueing system, where the service is done in batches, is called bulk service queueing system. The time required for servicing a unit (or a group) is called the service time.

Sometimes, the service rate may depend on the number of customers waiting for service. When the queue becomes longer, a server may work faster or, conversely, he may become less efficient. The situation in which service depends on the number of waiting customers is known as state dependent service.

Queue discipline

Queue discipline is the rule by which the customers are selected for service, when a queue is formed. The most common queue discipline is first-in first-out (FIFO). According to this rule the customers are served in the strict order of their arrival. Another queue discipline is last-in first-out (LIFO) under which the last customer is served first. Customers may also be served randomly irrespective of their arrivals into the system. This type of queue discipline is called service in random order (SIRO).
Another discipline is priority queue discipline, which allows service to be offered to customers depending on their priority in relation to other customers. There are two types in priority discipline viz., preemptive priority and non-preemptive priority. In the preemptive case, the customer with high priority is allowed to enter service immediately, suspending the service in progress to a customer with lower priority. In the non-preemptive case, the higher priority goes to the head of the queue but gets into service only after the completion of service in progress to the customer with lower priority.

System capacity

Some of the queueing processes admit the physical limitation to the amount of waiting room, so that when the waiting line reaches the maximum room capacity, no further customer is allowed to enter until space becomes available by a service completion.

Service channels

A queueing system may have one or more service channels to provide service. The service channels may be arranged in parallel or in series or a combination of both, depending on the design of the system's service mechanism. Queueing system with only one channel is called single server model. System with a number of parallel channels, providing identical service facilities, is called multichannel queueing system. In this model, customers may wait in a single queue until one of the service channels becomes ready to serve, or customers may form separate queues in front of each service channel. Queueing system may have a single stage of service or it may have several stages. In the multistage queueing system, a customer must pass successively through all the ordered stages for the completion of service.
1.3 NOTATION

Kendall [32] has proposed a convenient notation for queueing systems. A queueing process is described by a series of symbols and slashes such as A/B/X/Y/Z, where A indicates the interarrival time distribution of the customer, B the service time distribution, X the number of parallel service channels, Y the system capacity and Z the queue discipline. For example, M/M/c/$\infty$/FIFO represents a queueing system having Poisson input, exponential service, c service channels and there is no limit on the system capacity while the customers are served following first-in first-out queue discipline. In practice, this system is represented as M/M/c. That is, if the system capacity is infinite and the queue discipline is FIFO then the corresponding symbols can be omitted from the system representation.

1.4 BULK SERVICE

Bulk service queueing models are useful in many real-life transportation systems such as taxi stand, boat-house, bus stop and so on. There are many policies and rules according to which batches for bulk service may be formed. In literature the following types of bulk service are frequently considered.

(i) Usual bulk service: - Units are served in batches of not more than b. If the server, on completion of a service, finds more than b units waiting, he takes a batch of b for service while others wait. If he finds n units (1 ≤ n ≤ b), he takes all the n units in one batch for service.

(ii) Fixed size bulk service: - A service batch may be of fixed size k. The server will wait till the number in queue reaches k units. If there are more than k in the queue, he takes a batch of k, while others wait.
(iii) General bulk service :- Neuts [53] has introduced the concept of
general bulk service rule. According to this rule, a service begins only when a units
are present in the queue. If the number of units in the queue is $n$, $a \leq n \leq b$, then all
the $n$ units are taken for service. If there are more than $b$ units waiting in the queue,
he takes a batch of $b$ units, while others wait in the queue. Multiserver general bulk
service Markovian queue is represented by $M/M(a,b)/c$.

(vi) Variable size bulk service :- The customers are served in batches
of variable capacity $Y_n$ where $Y_n$ is a random variable.

(v) Bulk service with accessible batches :- If a batch being served does
not utilize its full capacity for service, it may remain accessible for customers
arriving during the service time of the batch until its full capacity is attained. The
total service time is not altered by inclusion of such joining units in course of
ongoing service.

(vi) Bulk service with accessible and non-accessible batches :- Units
are served in batches with minimum $a$ and maximum $b$. If a batch being served is of
size less than some fixed integer $d$ ($a \leq d \leq b$), then it may remain accessible for the
customers arriving during the service time of the batch, as long as the number in the
batch is less than $d$. If a batch in service is of size greater than or equal to $d$, the
batch becomes non-accessible for the arriving units.

1.5 SERVER'S VACATION

Queueing system with server vacations arises as models of many diverse
fields such as computer, communication and production systems. The non-availability
of a server at the system is termed as server's vacation. The purpose of leaving the
system is manifold. The server may want to utilise his idle time for another task.
Or if the server is a machine, it may need some repair after completing a job. An exhausted bank teller may like to take a coffee break before getting ready for the next customer. Researchers have studied queueing models with different types of vacations.

**Repeated vacation**

As soon as the system becomes empty, the server leaves the system for a vacation. If, on return from a vacation, the server finds one or more customers waiting, he works until the system is emptied and then goes on another vacation. If the server returns from a vacation finds no customer waiting, one or more additional vacations are taken.

**Single vacation**

On return from a vacation, even if no customer is waiting, the server stays in the system and waits for at least one service completion. The server takes only one vacation at a time.

**Exceptional first vacation**

In the repeated vacation the duration of the first vacation and the subsequent vacations are assumed to have the same distribution. In exceptional first vacation, the duration of the first vacation is differently distributed from that of the subsequent vacations.

**Gated vacation**

On return from a vacation, the server serves only those who were waiting in the system at the time of his return to the system. The service of subsequent arrivals
is deferred until after the next vacation. In this type of model, one can imagine that when the server returns from a vacation, a gate behind the last waiting customer closes and the server serves all those waiting in the system, then go for another vacation after opening the gate.

**Random vacation**

The random failure of a server irrespective of the number the waiting line is considered as server's random vacation.

**1.6 CUSTOMER'S IMPATIENCE**

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 queue if his wait is expected to be sufficiently small. Generally, impatience may result in balking or reneging or jockeying.

**Balking**

Reluctance of a customer to join a queue upon arrival is called balking. An arriving customer may decide not to join the queue because of the length of the existing queue.

**Reneging**

Reluctance of a customer to remain in the queue after joining it and waiting for some time is called reneging. A customer, after joining the queue and waiting for some time, may decide to leave the system.
Jockeying

In case of parallel queues, customers in the longest queue have a tendency to shift to the shorter one. This type of customer's behaviour is known as jockeying.

1.7 MARKOVIAN AND NON-MARKOVIAN

Queueing models are classified into Markovian queueing models and non-Markovian queueing models. A queueing model is called Markovian, if both arrival process and service process follow Poisson distribution. Models in which arrivals and/or departures do not follow the Poisson process are called non-Markovian.

The differential - difference equation method and matrix - geometric method are used to solve Markovian queues. Phase technique, the technique of imbedded Markov-chain and the supplementary variable technique are usually employed in studying non-Markovian queues.

1.8 TRANSIENT AND STEADY STATE QUEUEING SYSTEMS

A queueing system is said to be in transient state when its operating characteristics (such as input, out put, mean queue length etc.) are dependent on time. Otherwise the system is said to be in steady state or equilibrium state. Solution of a queueing system, depending upon time is called a transient solution and independent of time, is called steady state solution.

Many applications of queueing theory involve queues which are emptied and restarted periodically (for example, banks, barber shops and traffic signals). These queues will never reach equilibrium state. Hence steady state solutions are not always adequate, and it is desirable to have time-dependent solutions.
Most of the analyses of queueing models are confined to steady state results. Very little seems to have been done to evaluate the corresponding transient results. This is because the transient solutions are, in general, mathematically more involved.

Some of the methods used to study the transient behaviour of queues are given below:

(i) Spectral method of Ledermann and Reuter [36]
(ii) Combinatorial method of Champernowne [10]
(iii) Difference equation technique of Conolly [15]
(iv) Method of Parthasarathy [56] by defining generating function in a special way and using properties of Bessel functions.
(v) Method of Sharma [64] by using the properties of real, symmetric and tridiagonal matrix.

1.9 RELEVANT LITERATURE SURVEY

Queueing theory had its origin in 1909, when Danish telephone engineer, Agner Krarup Erlang published his paper ‘The theory of probabilities and telephone conversations’. It is interesting and also important to note that this field is continuously presenting challenging problems to the many capable investigators working in the field throughout the world.

Queues with balking and reneging

Haight has introduced the concept of balking in queueing models. Haight [24] and Homma [27] have studied M/M/1 queue with balking in equilibrium state. Singh [71, 72] has analysed a two-server Markovian queue with balking. Homma [28] has also studied balking problem for a general input distribution and an
exponential service time distribution in the multichannel case and derived equilibrium results. Barrer [6] has analysed M/M/1 system with reneging and obtained queue length distribution. Ancker and Gafarian [1] have investigated M/M/s/N system with reneging and obtained various steady state results. Ancker and Gafarian [2,3] have also combined balking and reneging and studied a single channel queue in equilibrium with Poisson input and exponential service time.

Queues with additional servers

In many situations, when there are too many people waiting to be served in front of a service facility, the system opens another service facility to reduce congestion. This happens normally in banks, booking offices and supermarkets.

Singh [73] has discussed a Markovian queue with number of servers depending upon queue length. Garg and Khanna [21] have considered the steady state behaviour of a queueing system with queue dependent additional server facility, wherein arrivals occur in batches of variable size. Romani [59] and Phillips [57] have obtained the steady state probabilities of queueing problems with variable number of service channels assuming that when the waiting size increases to some preassigned fixed number N, then with each arriving unit, a new channel is made available and is cancelled at the termination of service, if there is no unit waiting, with the exception of one channel which remains open at all times. Bidhi Singh [7] and Murari [48,49] have discussed queues with additional servers.

Non-Markovian time-dependent queuing systems

In many real life situations, we come across systems, where departures depend upon departures or no departures at the preceding time marks. One such situation is a musical entertainment service station wherein musicians,
singers, etc., play musical instruments and sing songs to entertain people. People come and join the audience and enjoy the programme. If the quality of the programme declines, departures from the audience will begin to take place and the departures at subsequent time mark will be influenced by those at the preceding time. Murari [50] has studied a queueing system with correlated arrivals and correlated phase type service. Sharda and Indu Garg [63] have analysed a time-dependent queueing problem with departures having random memory. Sharma [68] has considered a queueing system with arrivals in batches of variable size and correlated departures. Sharma [69] has also investigated a non-Markovian time-dependent queueing system and derived mean queue length.

Queues with vacation

In recent years some researchers have concentrated on the analysis of queueing system with vacation. Courtosis [16], Scholl and Kleinrock [62], Fuhrman and Cooper [20], Levy and Yechiali [37], Heymann [25], Ramachandran Nair [58], and Manoharan and Krishnamoorthy [39] have investigated single server queueing systems with vacation. Levy and Yechiali [38] have studied the system M/M/s with server's vacation. Nadarajan and Subramanian [52] have analysed a general bulk service queueing system with server's vacation. Nadarajan and Audsin Mohana Dhas [51] have considered multiserver general bulk service queue with vacation. Krishna Reddy, Nadarajan and Kandasamy [35] have discussed Markovian general bulk service queueing system with vacation and additional server. They have derived steady state solutions using Matrix-geometric method.
Queues with bulk service

bulk service queueing situations are very common in every day life. Runnenberg [60], Bloemen [8], Jaiswal [29], Neuts [53], Chaudhry and Templeton [12]. Arora [4] and Ghare [22] have considered queues with usual bulk service rule. Neuts [53] has studied the transient state distribution of the number of customers in the system for the M/G(a,b)/1 model. Borthakur [9] and Medhi [41] have obtained the steady state probabilities for the number in the queue and the waiting time distribution, respectively, for the M/M (a,b)/1 queue. Medhi and Borthakur [43], Neuts and Nadarajan [54] and Sim and Templeton [70] have extended the results of Borthakur [9] and Medhi [41] to the M/M (a,b)/c queue.

Gross and Harris [23], Kleinrock [34], Chiamsiri and Leonard [14] and Sivasamy [74] have investigated bulk service queueing systems with accessible batches. Sivasamy [75] has also analysed a bulk service queue with accessible and non-accessible batches and derived steady state distributions of queue length, waiting time and occupation time.

1.10 PROFILE OF PRESENT WORK

Transient analysis of some interesting queueing models are presented in this dissertation.

Chapter two is devoted for the analysis of finite capacity Markovian model with balking and reneging. There are c homogeneous servers and the maximum system capacity is N. A customer, who is in the queue, reneges the system after waiting for a certain time, which is an exponential random variable. Expressing the Laplace transform of system of equations in a matrix form and using the properties...
of a symmetric tridiagonal matrix, the transient and steady state probabilities are obtained for two models of this category. Numerical results are also presented. The two models differ from each other only in the balking behaviour. In model 1, when all the c servers are busy, an arriving customer joins the system with probability \(1 - \frac{n}{N-c}\), where \(n\) is the number of customers in the system. In model 2, an arriving customer, who finds the system size as \(n\), joins the system with probability 
\[
e_n = \begin{cases} 
1, & \text{if } n < c \\
\beta/(n+1-c), & \text{if } c \leq n < N \\
0, & \text{if } n = N
\end{cases}
\]
where \(\beta\) is the measure of a customer's willingness to join the queue, when all the servers are busy.

Chapter three deals with the study of infinite Markovian queueing system with two heterogeneous servers and balking. If the system is empty, an arriving customer joins the faster server with probability \(\theta\) and the slow server with probability \((1 - \theta)\). If only one server is free, he chooses the free server. If both are busy, an arriving customer joins the system with probability \(\beta\). The transient probabilities are derived and the steady state probabilities are deduced. Numerical results for steady state probabilities, expected system length and expected queue length are presented.

The queueing system M/M/c with balking and reneging is analysed in chapter four. When all the c servers are busy and if the number of customers in the system is less than a fixed number \(N\) (\(> c\)), an arriving customer joins the queue and waits for service a certain time, which is an exponential random variable. If the service does not begin by then, he departs from the system. When the system size increases to \(N\) then the service rate of each channel is increased and the
customers are not allowed to renege. If an arriving customer finds the number of customers in the system as greater than N, he joins the system with probability $\beta$. The transient probabilities and the probability density function of a busy period are obtained in terms of Bessel functions. Numerical results of performance measures are tabulated for selected parameters.

By defining the probability generating function in a special way, a differential equation is obtained. Then, by using the properties of Bessel functions in the solution of this differential equation, the probabilities of queueing models in chapters three and four are extracted.

In chapter five, Markovian queueing system with $c$ additional heterogeneous service channels is considered. The system starts with one regular channel, which is always opened irrespective of queue length. If the number of customers in the system is $m(1)$, then one additional service channel is started, which will be dropped at the termination of service, if the system size becomes less than $m(1)$. When both the channels are operating, if the system size increases to $m(2) > m(1)$, the second additional service channel is opened and will be dropped at the termination of service, if the system size becomes less than $m(2)$ and so on. The number of additional channels is limited to a maximum of $c$. For the queueing system with infinite room capacity, the transient solutions and the probability density function of a busy period for any intermediate number (k) of channels busy are derived in terms of Bessel functions. The transient probabilities are obtained for the system with finite room capacity. Numerical results are also presented.

Sixth chapter is devoted for the analysis of non-Markovian time dependent queueing system with server's vacation. The arrivals, departures and the return of server from vacation can take place only at the transition marks, where
the intertransition times are governed by a Poisson process. The departure of a unit at a time mark $t_r$ depends upon departure or no departure at $t_{r-1}$ and $t_{r-2}$. The arrivals are statistically independent. Two models of this category are considered. An arriving unit cannot depart at the same time mark in model 1, whereas in model 2 it can. For each model, the Laplace transform of the probability generating function of the queue length distribution in transient state and the expected queue length are obtained. The numerical results are included for particular values of parameters.

The queuing system $M/M_i(a,b)/2$ is investigated in the seventh chapter. In channel 1, there is a provision for an arriving customer to join the batch in service, if the size of the batch is less than some fixed integer $d$ ($a \leq d \leq b$). The server in channel 2, serves the customers according to general bulk service rule. The Laplace transform of the transient probabilities are expressed in terms of the unique root lying inside the unit circle $|z| = 1$, obtained by using Rouche's theorem. The busy period distribution and waiting time distribution are derived. Analytical expressions are derived for expected busy period, expected queue length and the expected waiting time of a customer. Little's formula is verified for this model. Numerical results are presented for particular values of parameters.