CHAPTER 1

INTRODUCTION

Queueing theory is a branch of applied probability theory. It is mostly a mathematically descriptive theory of practical importance and theoretical interest. A systematic study of the theory of queues provides a base of knowledge which can be applied to monitor and improve the efficiency of many queueing systems in the real world. A queue is a waiting line which is inevitably formed in front of some service facility. The study of queues is mainly applied in the fields of business (banks, booking offices, super markets), industries (servicing of automatic machines, production line storage), engineering (telephony, communication networks, electronic computers), transportation (airports, harbours, railways, buses, traffic operation in cities, postal services) and in every day life (elevators, restaurants, barber shops).

1.1 QUEUEING SYSTEM

A queueing system can be described by the arrival of customers for service, forming or joining the queue if service is not immediately available, and leaving the system after being served (or sometimes without being served).
By customers, we mean those demanding service, e.g. human customers at a bank counter or at a reservation counter, calls arriving at a telephone exchange, vehicular traffic at a traffic intersection, machines for repair before a repairman, airplanes waiting for take-off at a busy airport, merchandise waiting for shipment at a yard, computer programmes waiting to be run on a time-sharing basis etc.

The interarrival time between two consecutive arrivals of customers and the time needed for servicing a customer may be deterministic or stochastic. The case when both these characteristics are deterministic is trivial. We shall be generally concerned with stochastic interarrival and service times, and the theory will be essentially stochastic. The interarrival times and service times are assumed to be independent random variables. Further the two distributions given by the interarrival and service times are also taken to be independent.

1.2 CHARACTERISTICS OF A QUEUEING SYSTEM

The basic characteristics of a queueing system are as follows.

(i) The arrival process
(ii) The service process
(iii) The queue discipline
(iv) System capacity
(v) Service channels
1.2.1 THE ARRIVAL PROCESS

The arrival process or input describes the manner in which customers arrive and join the system. Arrivals may occur either singly or in a group. In case of group arrivals, the input is said to occur in bulk or batches.

It is important to know the reaction of a customer upon entering the system. A customer may decide to wait no matter how big the queue is or if the queue is too lengthy, he may decide not to enter it. If a customer decides not to enter the queue upon arrival, he is said to have balked. On the other hand, a customer may enter the queue, but after some time lose patience and decide to leave. In this case he is said to have reneged. In the event that there are two or more parallel queues, customers may switch over from one to another, that is, jockey for position. If an arrival process does not change with time, then it is called a stationary arrival process; otherwise it is called nonstationary.

1.2.2 THE SERVICE PROCESS

The service process describes the manner in which service is rendered. The uncertainties involved are the number of servers, the number of customers getting service at any time and the duration of service.
Customers may be served either singly or in batches. The service rate may depend on the number of customers waiting for service. A server may work faster if he sees that the queue is building up or, conversely, he may get flustered and become less efficient. The situation in which service depends on the number of customers waiting is referred to as state-dependent service. Service rate can be stationary or nonstationary with respect to time.

If service is done in batches, the queueing system is referred to as bulk service queueing system. The batches may be of fixed size or of variable size. Neuts [57] introduced the general bulk service rule. Under such a rule, a server starts service only if at least 'a' customers are present in the queue, the maximum service capacity being 'b' customers. In other-words, if there are 'x' customers waiting at the completion of a service, the following rule for service is followed.

(i) For $0 \leq x < a$, no service
(ii) For $a \leq x \leq b$, service is done for a batch of $x$ customers
(iii) For $x > b$ Service is done for a batch of $b$ customers and the remaining $(x-b)$ customers continue to wait in the queue.
In particular, if \( a = 1 \), the above rule will be called usual bulk service rule and if \( a = b = k \), the rule is then called fixed size bulk service rule.

1.2.3 THE QUEUE DISCIPLINE

The queue discipline indicates the way in which the customers form a queue and are served. The most common queue discipline is first in first out (FIFO), under which the customers are served in the strict order of their arrivals. Another queue discipline is last in first out (LIFO), according to which the last customer is served first. Yet another queue discipline is service in random order (SIRO), in which the customers are served randomly irrespective of their arrivals into the system.

'Priority' queue discipline allows priority in service to some customers in relation to other customers waiting in the queue. It is further subdivided into two categories viz. preemptive priority and nonpreemptive priority. In the preemptive case, the customer with higher priority is allowed to enter service immediately suspending even the service in progress to a customer with lower priority. In the nonpreemptive case, the higher priority customer 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.
It is assumed that, once a server who is able to provide service to a waiting customer becomes free, the customer immediately enters service without loss of time.

1.2.4 SYSTEM CAPACITY

The number of customers in the queue and in service put together is called system capacity. The system may have a limited or an unlimited capacity. A queue with limited waiting room can be viewed as one with forced balking where a customer is forced to balk if he arrives at a time when queue size is at its limit.

In many queueing problems such as assembling parts in a factory, undergoing medical checkup in a clinic etc., service is made up of several phases and is rendered by facilities arranged in series. Such queues in series are called tandem queues. Queues are allowed to build up in front of the servers. These intermediate queues, known as buffers, may again be finite or infinite.

1.2.5 SERVICE CHANNELS

A queueing system may have one or more service channels (servers) to provide service. The service channels may be arranged in parallel or in series or a combination of both depending on the nature of service required. It is generally assumed
that the service mechanisms of parallel channels operate independently of each other. A queueing system with only one server is called a single server model and a system with number of parallel servers is called a multiserver model. In case of multiserver models, the customers may form a single queue or parallel queues in front of each server.

A queueing system may have only a single stage of service such as the barber shop or it may have several stages. An example of a multistage queueing system would be a medical check up procedure, where each patient must proceed through several stages, such as medical history; ear, nose and throat examination; blood tests; electrocardiogram; eye examination; and so on. In some multistage queueing processes, 'feedback' (or recycling) may occur. It is common in manufacturing processes where quality control inspections are performed after certain stages, and parts that do not meet quality standards are sent back for reprocessing.

In a system with two units I and II of identical parallel servers in series separated by a finite intermediate buffer, the physical phenomenon 'blocking' occurs, i.e. when a service completion occurs in unit I and the buffer is full, the customer cannot leave his server in unit I and causes blocking of that server.
1.3 NOTATION

Kendall [35] has designed a very convenient and universally accepted notation to denote a queueing system. It consists of a five-part descriptor A/B/C/X/Y where A and B denote the interarrival time and service time distributions respectively, C is the number of parallel servers, X is the system capacity and Y the queue discipline. Thus, M/M/1/∞/FIFO means a Markovian queueing system with exponential interarrival and service times, single server, unlimited system capacity and first in first out queue discipline. In practice, this system is represented as M/M/1 with the understanding that the system capacity is infinite and FIFO queue discipline is followed, in case they are not specified. M/M(a,b)/C denotes a multiserver Markovian queueing system with the general bulk service rule.

1.4 SERVER VACATION

The non-availability of the server at the queueing system may be termed as server vacation. If the queue is empty or inadequate for service, the idle time of the server can be utilized in a useful way to perform additional jobs or for the preventive maintenance work in case the server is a machine. Sometimes maintenance work may have to be done even when the queue length is adequate for service. The idle time utilization should be aimed at minimizing the operational cost. The following
are some types of server vacations associated with general bulk service queueing systems.

1.4.1 SINGLE VACATION

The server on completion of service to a batch will start service again only if there are a minimum number of customers in the queue. Otherwise, the server will go for a vacation. On return from a vacation, he begins service if there are a minimum number of customers in the queue. If the queue size is less than the minimum, the server waits until the queue size reaches the minimum level to start his next service, i.e. the server takes only one vacation at a time.

1.4.2 REPEATED VACATION

All other assumptions being same as in the case of single vacation, in case of repeated vacation model if the server finds less than the minimum number of customers required for service at the end of a vacation, he takes another vacation. He will continue in this manner until he finds, upon returning from a vacation, the required minimum number of customers in the queue.

1.4.3 EXCEPTIONAL FIRST VACATION

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

1.4.4 GATED VACATION

In this case, the server, on return from a vacation, extends service only to those customers who were waiting at the time of his return to the system. The service to the subsequent arrivals is deferred until after the next vacation. One can imagine that when the server returns from a vacation, a 'gate' behind the last waiting customer closes and the server will serve only those customers in front of the gate before going for another vacation.

1.4.5 RANDOM VACATION

The random failure of a server irrespective of the queue size is regarded as server's random vacation. A machine used to produce an item may breakdown randomly and even the repair time may also be random. If the machine (the server) is idle, this idle time may be utilized for the preventive maintenance work.
1.4.6 LIMITED SERVICE VACATION

Sometimes, after producing a specific number of items, the machine (the server) may have to be sent for maintenance or stopped for sometime to make it fit for further production. The interval of time during which the server is not available in the system is called limited service vacation.

1.5 RELEVANT LITERATURE SURVEY

The theory of queues originated in 1909 with the publication of the paper "The theory of probabilities and telephone conversations" by A.K. Erlang. Interesting and fruitful interactions between theoretical structures and practical applications during the last eight decades have led to tremendous growth of the subject. In the early years of its development, the studies were confined to single arrival and single service systems. Bulk service queues have important applications in loading and unloading of cargoes at a sea-port, in traffic signal systems and in a mass transportation system. A review of relevant research papers on Markovian queueing models with general bulk service and suitable combinations of balking, jockeying, server breakdown, server vacation, feed back and blocking is given below.

Haight [29] seems to be the pioneer in introducing the concept of balking in queueing models. Singh [75,76] has
analyzed a two-heterogeneous-server single service queueing system with balking. The theory of bulk service queues originated with the work of Bailey [6]. Borthakur [9] and Medhi [49] have obtained the steady state probabilities for the number of customers in the queue and the waiting time distribution, respectively, for the M/M(a,b)/1 queue. Arora [2] and Ghare [26] have studied the multiserver Markovian queue with bulk service. Medhi and Borthakur [48] investigated M/M(a,b)/2 queue using Tauberian arguments. Later Medhi [50] derived the waiting time distribution for the M/M(a,b)/C queue by analytic methods. Following Neuts' [59] algorithmic approach, Neuts and Nadarajan [63] have studied the M/M(a,b)/C queue and obtained the stationary waiting time distribution. Sim and Templeton [74] used yet another recursive algorithm to obtain the steady state numerical results for the same model. Cosmetatos [17] and Sim and Templeton [73] have dealt with M/M(a,∞)/C queue by analytic methods. Recently Chaudhry et al. [14] have extended the model discussed in reference [48] to the two-heterogeneous-server case and obtained the steady state probabilities and busy period distributions.

Queueing systems (e.g., a computer facility), which are subject to breakdown, are not able to provide uninterrupted service to the customers. It is then important to see how the breakdowns affect the level of performance of the system. The significant contributors in this area are Avi-Itzhak and Naor [4].
Fond and Ross [22], Gaver [25], Nadarajan and Jayaraman [56], Neuts [58,61], Neuts and Lucantoni [60], Shanthikumar [69], Shogan [72] and Yechiali and Naor [82].

Jockeying among parallel queues is an interesting phenomenon in queueing theory. Such models have been investigated by Adan et al. [1], Disney and Mitchell [19], Elsayed and Bastani [21], Haight [30], Kao and Lin [32], Koenigsberg [37] and Rao and Posner [65]. Recently Yiqiang Zhao and Grassman [83] have studied the multiserver single service queues with jockeying and obtained explicit formulas for the equilibrium probabilities, the expected number of customers and the expected waiting time of a customer in the system.

Queueing models with server vacation have, in recent years, received much attention in the literature. The researchers of single server, single service queues with vacation include Chatterjee and Mukherjee [12], Cooper [16], Daniel and Krishnamoorthy [18], Fuhrmann [23], Fuhrmann and Cooper [24], Heyman [31], Lee and Lee [43], Levy and Yechiali [45], Lucantoni et al. [47], Scholl and Kleinrock [68], Shanthikumar [70] and Tian et al. [78]. Single server, bulk service queues with vacation have been analyzed by Chatterjee and Mukherjee [11], Lee et al. [44], Nadarajan and Subramanian [55] and Wortman and Disney [81]. Multiserver, single service queues with vacation have
been discussed by Kao and Narayanan [33], Levy and Yechiali [46], Mitrany and Avi-Itzhak [53] and Vinod [79]. Audsin Mohana Dhas et al. [3] have dealt with the two-heterogeneous-server general bulk service queue with server vacation. Neuts and Lucantoni [60] have analyzed a Markovian N server queue with breakdowns and repairs. Nadarajan and Audsin Mohana Dhas [54] have studied the multiserver general bulk service queue with vacation. Krishna Reddy et al. [39, 40] have analyzed the Markovian general bulk service queueing systems with vacation and additional server, and a tandem queue with general bulk service and servers' vacation.

Doshi [20] and Takagi [77] have presented excellent review papers discussing vacation models and provided details to illustrate how the seemingly diverse mix of problems are closely related in structure and can be understood in a common framework.

Tandem queues with feedback, general bulk service and blocking have important applications in quality controlled manufacturing processes. "Feedback" and "blocking" have been discussed by Konheim and Reiser [38] and Neuts [62]. Avi-Itzhak and Heyman [5], Krishna Reddy et al. [40], Latouche [41], Neuts [62] and Reiser and Konheim [66] are some notable researchers in this area.
1.6 SOLUTION METHODS

Queueing models are broadly classified into Markovian queueing models and non-Markovian queueing models. A queueing model is called Markovian if both the interarrival and service times follow the exponential distribution; if the interarrival and/or service time distributions are not exponential, then the queueing model under consideration is called non-Markovian. Several methods have been adopted for solving queueing problems. Some of them are outlined below.

1.6.1 THE DIFFERENTIAL-DIFFERENCE EQUATION METHOD

This analytic method is used to solve Markovian queueing models in transient (time dependent) and steady (time independent) states and discussed in detail by Gross and Harris [28], Kleinrock [36] and Saaty [67]. Based on the distributions of interarrival and service times, the differential-difference equations are derived. For solutions of these equations, a number of methods have been put forward. In the transient case, we have the method of generating function (using Rouche's theorem) of Bailey [7]; the combinatorial method of Champenowme [10]; the difference equation technique of Conolly [15]; the simple approach of Parthasarathy [64]; the method of Sharma [71]. In the steady-state case, one has the methods using the Rouche's theorem and iteration; the Rouche's theorem and generating functions; the Laplace transform.
1.6.2 THE MATRIX-GEOMETRIC ALGORITHMIC METHOD

Most of the queueing models require the application of Rouche's theorem. Neuts [59] broke completely new grounds and developed the matrix-geometric algorithmic method, not using Rouche's theorem, for the solution of Markovian and non-Markovian systems. This computational method uses matrix-method as an alternative to closed form analytic methods in solving steady state problems. The approach involves only real arithmetic and avoids the calculation of complex roots based on Rouche's theorem. A queueing problem is to be mathematically formulated. Then, by lexicographically or otherwise ordering the states, a rate matrix of infinite order is to be formed and partitioned in some convenient way into submatrices. Capitalizing the rich reservoir of results developed by Neuts and his associates, the stationary probability vector, the waiting time distribution and other performance measures of interest are computed. The results are verified by showing that the total probability is nearly equal to unity.

Other important methods include "imbedded Markov chain technique" of Kendall [35] and "supplementary variable technique" of Keilson and Kooharian [34]. Recently Grassman [27] has developed an algorithmic method known as "the state reduction method". This technique helps to solve even bulk
arrival and/or bulk service queueing problems including imbedded Markov chain models.

1.7 PROFILE OF PRESENT WORK

This thesis contains the study of some Markovian queueing models with general bulk service rule and suitable combinations of balking, server breakdown, jockeying, server vacation, feedback and blocking.

Chapter two deals with the problem of finding the waiting time distribution of a two-heterogeneous-server Markovian queue with general bulk service rule and balking. After formulating the problem mathematically, the corresponding differential-difference equations are obtained analytically for the transient case and solved for the steady state, using Rouche's theorem and iteration. Explicit expressions for the equilibrium probabilities of the number of customers in the queue and the waiting time distribution are derived. The expected waiting time of a customer in the queue is found to satisfy the Little's formula in the non-balking case.

In chapter three, a single server queueing system with general bulk service that alternates stochastically between operational state and repair state is considered. The system, when
operational, functions as a single server Poisson queue with general bulk service and when it fails, no service takes place but customers continue to arrive according to a Poisson process with different arrival rate. The durations of the operating and repair periods follow exponential and phase-type distributions, respectively. The state space is identified and the steady state probability vector of the number of customers in the queue and the stability condition are obtained using the matrix-geometric algorithmic approach. Numerical results are also presented.

The fourth chapter analyzes a jockeying problem with two queues in parallel before the two heterogeneous servers. The arrivals are Poisson, the service times of the servers with general bulk service are exponentially distributed and bulk jockeying takes place instantaneously among the queues. The model is mathematically formulated and the matrix-geometric method is used to find the steady state queue length and the stability condition. Numerical results are also presented for particular values of the parameters of the model.

Chapter five contains two queueing models with Poisson arrivals, n heterogeneous exponential servers, each of them serving customers in groups according to general bulk service rule and server vacations. In both of the models, at least one server remains in the system at a time. Model I deals with single
vacation of servers, in which if the server finds \((a-1)\) customers or less at the end of a vacation, stays idle in the system waiting for the queue size to become 'a' and then starts servicing. Model II discusses the case of repeated vacations of servers, in which upon terminating the first vacation period, if the server finds \((a-1)\) customers or less, he immediately takes another vacation. The server continues in this manner till the queue size becomes 'a' or more on his return from a vacation. The state space is identified and the steady state probability vector of the number of customers in the queue, the stability condition and the mean queue length are derived using matrix-geometric approach. Numerical results are also presented.

The sixth chapter is devoted to the study of exponential tandem queues with feedback, general bulk service and blocking. The model consists of an input queue of infinite capacity, a subsidiary system consisting of units I and II each having a queue of finite capacity and a buffer of finite capacity in unit III to which customers arrive after service completion in the subsidiary system. Customers arrive into the input queue according to a Poisson process with rate \(\lambda\). In the subsidiary system, unit I has \(J\) parallel exponential servers each with service rate \(\alpha\) and unit II has \(C\) parallel exponential servers each with service rate \(\beta\). Unit III contains a single exponential server. The service in this unit is done according to general bulk service.
rule with rate $\mu$. The maximum number of customers in the subsidiary system waiting or receiving service should not exceed $M$, a given positive integer satisfying $J+C \leq M$. Customers in the input queue are allowed to enter the subsidiary system subject to the capacity constraint $M$ and join the queue in unit I. Upon completion of service in unit I, the customer leaves the subsidiary system instantaneously with probability $(1-\theta)$, $0 \leq \theta \leq 1$, and enters into the buffer or joins the queue in unit II with probability $\theta$. In the former case, if the buffer is full, the served customer can not leave his server in unit I and so causes blocking of the server. In the latter case, the customer will return to unit I queue upon completion of service in unit II and in this way, such customers make a number of passes through unit II before their eventual departure from the subsidiary system. The inter-arrival times, the service times in units I and II of the subsidiary system and the branching events at departures from unit I are assumed to be mutually independent. The problem is mathematically modelled and the stationary probability vector of the number of customers in the queues and the stability condition are obtained explicitly using the matrix-geometric method. Numerical results are also presented for particular values of the parameters of the model.

The final chapter contains some conclusions and suggestions for further work on related problems.

* * * * *