CHAPTER-1

INTRODUCTION
CHAPTER-I

INTRODUCTION

1.1 QUEUES IN GENERAL

Queueing theory is a vast field and its origin can be traced to the beginning of 20\textsuperscript{th} century. It was developed to provide models to predict the behaviour of systems that attempt to provide service for randomly arising demands. There are many valuable applications of the theory in the literature of probability, operation research, management science etc.

The pioneer work in the field of queueing theory was done by Erlang (1909), an engineer at the Copchagin telephone exchange, who published the paper "The theory of probabilities and telephone conversations". After Erlang work on the topic of queueing theory was continued by Molina(1927) who published his theory under the title of "Application of the theory of probability to telephone trunking problems". Fry (1928) in his book on "Probability and its engineering uses" throws further light on the above problem.

After 1950, with the emergence of operation research as a well recognized discipline, it was observed that the work done in connection with the telephone problem had a much wider scope. Queueing theory was studied extensively and has found applications in a wide variety of situations such as machine maintenance, road traffic, defence operations, inventory management etc.
Queueing theory involves study of behaviour of the system over time. A system is said to be in transient state if its operating characteristics are dependent on time and is said to be in steady state if its operating characteristics are independent of time.

In the work quoted below, emphasis is on indicating mathematical approach for analysis of queueing theory problems.


1.2 CHARACTERISTICS OF THE QUEUE

Though characterization of a queueing model differs from one situation to another, yet, the following characteristics are commonly used.

1) Arrival pattern of the customers:

The arrival pattern describes the method in which customers arrive at service facility for receiving the service. If the service facility is busy in serving other customers, then customer arriving for service will form a queue.

The statistical pattern of the arrival can be indicated through probability distribution of the number of arrivals at different points of time. This is a discrete random variable. The probability distribution of arrival pattern can be identified through analysis of the past data. The discrete random variable indicating the number of arrivals at different points of time and the continuous variable, the time between two successive arrivals are obviously interrelated for instance, if the distribution of number of arrivals in an interval is Poisson then the corresponding inter-arrival times are distributed as exponential. This property is frequently used to derive elegant results in waiting line problems.

Customers arriving for service mainly behave in following ways:

1) A customer is said to have balked if he does not join the queue upon arrival because the queue is too long.
2) A customer is said to have reneged if he leaves the queue after certain time due to loss of patience.
3) A customer who moves from one queue to another hoping to get service more quickly is said to be jockeying.
2) Service pattern of the customers:

Service pattern is the next component of waiting line system. Service time is a statistical variable and it is the time in which number of services completed in a given period of time. The service may be operated through a single-server or through several servers. The service pattern may depend on the number of customers waiting for service i.e. state of the system.

Service time may be constant or random variable. Service, like arrivals can be stationary or non-stationary with respect to time. Distribution of service time is commonly assumed to be negative exponential.

3) Queue discipline:

Queue discipline describes the manner in which customers are chosen for service when a queue is formed. The simplest queue discipline is (FCFS) “First Come First Served” according to which customers are served in order of their arrival. The next queue discipline is (LCFS) “Last Come First Served” i.e. customer coming at the last gets service first. Another queue discipline is (SIRO) “Service In Random Order”. There are priority based service disciplines also.

The discipline is said to be of preemptive nature if a customer is allowed to enter service immediately after entering into the system even if a customer with lower priority is already in service. The lower priority
customer’s service is interrupted. The discipline is said to be of non-preemptive nature if the customer of higher priority is served ahead of the queue but on completion of current service of lower priority customer.

4) System capacity:

The maximum number of customers allowed to enter the system to take service depending on the time or space constraints, is referred to as system capacity.

5) Number of service channels:

It refers to the number of parallel service channels available to give service to the customers simultaneously, to reduce the waiting time.

There are two service channels:
A) Single queue Multi-server
B) Multiple queue Multi-server

Generally, these channels operate independent to each other.

6) Stages of service:

A queueing system may have a single stage or multiple stages of service. Single stage of service is defined as where a customer gets service from a single counter.

Multistage queueing system is the one where a customer gets complete service through several counters one after another in series.
.3 BRIEF REVIEW OF THE WORK ON QUEUEING THEORY

Most of the research studies in the field of queueing theory arise out of different combinations of the above characteristics. We outline here some such studies with no attempt to exhaust the field.

Cobham (1954) studied priority problems in queues. Haight (1957) studied arrival pattern of such customers who do not join the queue on arrival since the queue is too long and the customer is said to have balked. Barrer (1957a, b) studied queues with impatient customers having indifferent clerks and ordered service simultaneously. Haight (1959) studied queues with reneging. Neuts (1965) studied busy period of a queue with batch service.


Blackburn (1972) gave solutions on optimal control over queues with balking and reneging. Crabill (1972) investigated a service facility with variable exponential service time and constant arrival rate.

Fa.kinos(1980) considered the M/G/K group arrival loss, where the system has k servers whose customers arrive in accordance with a compound Poisson process and operates in a prescribed way. Kao and Chiusin(1990) obtained the matrix geometric solution of jockeying problem. Simple procedures were developed for computing the stationary probability vector of the continuous time Markov chain.

Hlynka et.al.(1994) considered a situation with two parallel queues and two possibly heterogeneous servers. He assumed that inter-arrival times are distributed with mean $1/\lambda$ and service times with rates $\mu_1$ and $\mu_2$. Rubinstein and Melamed (1997) in their book obtained modern methods of simulation.

1.4 REVIEW OF THE RELEVANT WORK

Queueing theory originated as a very practical subject, but regrettably most of the recent literature has been of little direct practical value. It is clear that emphasis in the recent literature on the exact solution of queueing problems with clever mathematical tricks became secondary to model building. Most real problems do not correspond exactly to a mathematical model but very little of the literature deals with approximate solutions, sensitivity analysis and the like.

Queueing models are helpful to satisfy customer's demand. There are many applications of the theory such as traffic flow (vehicles, aircraft) and scheduling (patients in hospitals, jobs on machines, programs on a computer) etc.
In our present work we have investigated queues with blocking and
ring applicable in the field of medical science and computer
ce, we review only related studies.

The pioneer work on the field of medical science was done by
y (1952) who studied queues in out-patient department and gave
interesting results. Hunt (1956) treated a modified series model to
e a two station sequential series queue in which no waiting is allowed
ven stations, but where a queue with no limit is permitted in front of
ion I. He obtained steady state equations for this model, the expected
em size \( L \), and the maximum allowable \( \rho \) for steady state to be
ved.

Reich (1957) obtained waiting times of tandem queues.
ke(1964) studied loss and delay in tandem queues. Burke (1969) also
ved a three station series queue with first and third having a single
ver but the middle station having multiple servers. Schwartz (1974)
uded systems belonging to a class of models called lane selection
models with one server of each type and each with its own queue. Weber
1978) discussed a situation with several identical parallel servers. Roque
1980) also studied queueing models with lane selections.
Approximations for departure processes and queues in series have been

Green (1985) considered a LS type model with two types of servers
amely general-use and limited-use servers and two types of customers.
General-use servers can provide service to either customer type while
limited-use servers can be used only for one of the two. This type of
queueing situations arises in a bank that has tellers, who can perform any
ordinary banking service and machines that can be used only for certain kinds of transactions such as check-cashing.

Worthington (1987) studied queueing models for hospital waiting lists and derived some useful results.

Perros (1994) gave a good reference on queueing networks with blocking where blocking means if the customer got his service completed at this stage but unable to get service immediately from next stage, since server in next stage is busy. Arrival at stage I when the system is blocked is turned away. The probability of such event is called rate of loss call, which is considered to be an important measure of efficiency for the system without queues.

Mazumdar and Liu (1999) have investigated transition (Markov) models for the analysis of survival times in clinical research and described a method to estimate survival time distributions. Singh (2001) considered two stages of service with blocking and derived solution using D-operator and obtained expression for waiting time in the system.

Gray et. al.(2002) considered two models in which more than one server is available to service a queue., but only one server at a time is actually used. All servers are subject to breakdown and are repaired one at a time. In one model there are two ranked servers, a primary server and a back up server; when operable, the primary server always serves the queue. In the second model, there are finite member of unranked servers.
1.5 Problems under investigation:

The problems under investigation essentially deals with single and multi-server queues in series with two or more stages of service. The purpose of the present work is to exploit the concepts of blocking and branching to certain practical situations occurring in the field of medical science and computer sciences.

In chapter I we have described queues in general, its characteristics and mentioned some important studies. Review of work related to our studies and symbols and notations used in the thesis are also mentioned in this chapter.

In chapter II we have considered a queueing model with blocking having three stages of service in series. Probability equations have been derived for the system using birth and death models and solutions obtained by using D-operator and matrix algebra. Expression for waiting time has been obtained and it is seen through illustration that waiting time becomes less when blocking arises. The probability equations have also been obtained through stochastic balance procedure for verification.

Chapter III deals with two stages of service with blocking at first stage and branching at the second stage assuming same types of specialists at second stage. Probability equations have been derived by two methods as in chapter I and solutions have been obtained by using D-operator and matrix algebra. Expression for waiting time has been obtained and it is noticed that waiting time in two stages of service with branching is little less as compared to three stages of service without branching.
In chapter IV we have considered two stages of service with blocking at first stage and branching at second stage with two different types of specialists. Probability equations have been derived by two methods and solutions have been derived by using D-operator and matrix algebra. After comparing the waiting time required with that of chapter II, it is observed that waiting time is little more. This is due to increase in blocked state. This is the cost, the patient has to pay for getting appropriate treatment.

In chapter V we deal with two stages of service having two different types of customers. The first type of customers are those who stand in a queue on finding the server of stage I busy. They are known as elective customers. The second type of customers are those who return on finding the server of stage I busy. They are known as emergency customers. Steady state equations have been derived and expression for waiting time has been obtained.

In chapter VI we consider a queueing model developed for a system where there are two service counter at first stage and one service counter at second stage. This has been done with special reference to a computer system in which there are two terminals and one printer at two stages of service respectively. Steady state equations have been derived and expression for waiting time has been obtained.

Chapter VII is an extension work of chapter VI where in the first case it is assumed that no queue is allowed in front of stage I(terminals) but queue is allowed in front of stage II(printer) and in the second case we consider queue both in front of stage I(terminals) and stage II(printer). Probability equations have been derived and on comparing waiting time of this chapter with the previous one it is seen that waiting time increases in the latter case.
1.6 SYMBOLS AND NOTATIONS

In this thesis we shall make use of the following symbols and notations:

\( n \) = number of customers in the system, both waiting and in service
\( \lambda \) = average number of customers arriving per unit time
\( \lambda_n \) = mean arrival rate of customers of type \( n \)
\( \mu \) = average number of customers being served per unit time
\( \mu_n \) = mean service rate of server \( n \)
\( \lambda/\mu = \rho \) traffic intensity
\( s \) = number of service channels served
\( P_n(t) \) = probability that there are \( n \) customers in the system at any time \( t \), both waiting and in service.
\( P_{ij} \) = transition probability when a customer makes transition from state \( i \) to state \( j \)
\( P_{n_1,n_2,\ldots,n_k}(t) \) = Probability of \( n_1 \) customers at stage I, \( n_2 \) customers at stage II, \ldots, \( n_k \) customers at stage k in a series queue at time \( t \)
\( L_s \) = expected number of customers in the system
\( W_s \) = expected waiting time of a customer in the system
\( W_q \) = expected waiting time of a customer in the queue
\( L_q \) = expected number of customers in the queue
\( L/L>0 \) = expected length of non-empty queue