1. INTRODUCTION

1.1 THE BACKGROUND

A 'queue' is a waiting line of units or customers at a service station which provides certain service facilities. Queueing theory is concerned with mathematical models of queueing situations. The queueing system is described by the arrival of customers for service at a center with one or more service stations. The customers upon arrival, join a queue, if service is not immediately available and leave the system after being served. Examples of queueing phenomena are numerous and arise in such diverse areas as telephone traffic, machine shops, computer centres, supermarkets, restaurants, banks, car parks, hospital and airport operations, shipping docks, gas stations and so on. Examples such as reservoirs, inventory, traffic at intersections and social service systems can also be viewed as queueing systems by appropriate interpretation of the notions of arrival and service.

The mathematical study of queueing systems originated with the work of A.K. Erlang (cf. Brockmeyer et al. [1948]) between the years 1909 and 1920 while dealing with waiting line and trunking problems for telephone exchanges. Molina
[1927] and Fry [1928] extended Erlang's work and made some major contributions to the field. Their work stimulated the research of such early contributors as Pollaczek [1930], Khintchine [1932], Palm [1943] and Kendall [1951]. The literature on the theory of queues and on the diverse areas of its application has grown tremendously over the years. Cohen and Boxma [1985] give a survey of the evolution of queueing theory.

1.2 STRUCTURE OF QUEUEING MODELS

The basic process assumed by most queueing models is the following: The 'customers' requiring service are generated over time by an 'input source'. These customers enter the queueing system and join a queue. At certain times a member of the queue is selected by some rule known

Fig. 1 A Queueing System
as the 'queue discipline'. The required service is then provided to the customer by the 'service mechanism' after which the customer leaves the system. This process is depicted in Fig. 1.

From these descriptions it is clear that a queueing system is completely characterized by the following factors.

(i) **Input Process**: Customers demanding service emanate from a population known as 'input source' which may be either infinite or finite. Units arrive generally in a random fashion which may be described by the distribution of interarrival time. The capacity of the queueing system may be either unlimited or limited.

(ii) **Service Mechanism**: This includes the distribution of service time and the number of servers in the system. Some times service is provided in stages.

(iii) **Queue Discipline**: This refers to the manner in which the customers are selected for service. Examples are 'first-come first-served (FCFS)', 'service in random order (SIRO)' and 'priority rule'.

The following standard notations are used for various input/service distributions:

- $M$: exponential (Poisson) distribution
- $G$: general (arbitrary) service time distribution
- $GI$: general independent input distribution
- $E_k$: $k$ - Erlangian distribution
- $D$: deterministic
Traffic intensity is measured by the ratio of mean service time to mean interarrival time. This parameter usually denoted by $\rho$, has an important role in the theory.

1.3 BEHAVIOURAL ANALYSIS

In view of the random elements involved in queueing systems these are best described through, stochastic processes. Behavioural analysis of queueing systems provides important measures of performance and effectiveness of such systems through

(i) the queue length distribution,
(ii) waiting time distribution,
(iii) busy period analysis, and so on.

The analysis provides a description of the queueing characteristics such as the average queue length, average waiting time, busy period, server utilization and other features of interest to both users and operators of service providing facility. The literature in this area is very vast. Some of these aspects of queueing theory have been extensively discussed by Morse [1958], Takács [1962], Prabhu [1965], Cohen [1969], Gross and Harris [1974] and specifically on priority queues by Jaiswal [1968] and on bulk queues by Choudhry and Templeton [1983].

1.4 STATISTICAL ANALYSIS

The behavioural aspects of the queueing system can be
studied if the input and service time distributions are specified. However statistical techniques become necessary for the exact identification of probability models describing the input and service times. This involves collecting appropriate data and estimating the queueing parameters such as the mean arrival rate, mean service rate or the traffic intensity $\rho$ and testing hypotheses concerning these parameters. Bhat and Rao [1987] provide an excellent review of literature on statistical analysis of queueing systems.

Identification of probability model is a standard problem in statistical inference. This involves collection of appropriate data, estimation of the parameters of the model and applying tests for specific forms of the distribution. It is important to make use of the structural characteristics of a queueing process while employing estimation and hypothesis testing techniques to queues.

The earliest paper on problem of estimation in queueing models seems to be that of Clarke [1957]. He obtains the maximum likelihood estimator (m.l.e) of the arrival rate $\lambda$ and the service rate $\mu$ for a $M|M|1$ queue. Here the system is observed for a fixed time and full information is used for estimation. It is assumed that the system is in equilibrium. This work was followed up by Beneš [1957] who obtained the m.l.e's of $\lambda$ and $\mu$ for a $M|M|\infty$ queue related to a simple telephone exchange model.
Wolff [1965] made use of the maximum likelihood estimation procedures as applicable to Markov processes developed by Billingsly [1961] for estimating the parameters of the birth-death type queueing model. The limiting distributions of these estimators were also obtained. Further the estimators and tests for the parameters of the $M|s,M|\infty$ and machine interference models are obtained as particular cases of the birth-death model. Cox [1965] also provides estimation and testing techniques for birth-death type queueing models on similar lines. Based on Billingsley's work, the technique of estimating the parameters of an $M|E_2|1$ queue is illustrated in Gross and Harris [1974, p.353].

Harris [1973] developed a new approach for estimating the arrival and service rates of a $M|G|1$ queue making use of waiting time data. Goyal and Harris [1972] consider the maximum likelihood estimation of the parameters assuming a state dependent service rate. Henningsen [1978] extended Clarke's work to the $M|G|1$ queue. Properties of the estimators in the case of $M|M|1$ queue have been studied by Jenkins [1972] and Saman and Tracy [1978].

Basawa and Prakasa Rao [1980] discuss the estimation problems for $M|E_k|1$ and $E_k|M|1$ systems. Basawa and Prabhu [1981] have considered estimation of queueing parameters and large sample tests for the $GI|G|1$ queue. Limiting distributions of these estimators also have been derived.
Quite a few other authors have also discussed the estimation problems in queues. For instance, Abou-El-Ata et al. [1979] give point estimators of $\lambda$, $\mu$, $\rho$ for a Markovian queue. Upendra and Shah [1980] have discussed the problem of $m.1$ estimation in $M|M|2$ queue with heterogeneous servers, while Baba [1982] discusses the case of birth-death type queueing model using Poisson sampling. Minimum variance unbiased estimation of the input/service parameters of $M|E_k|1$ system is discussed by Gupta [1982], a study of asymptotic properties of $m.1.e$'s in the case of birth-death queues is made by Allen [1983], efficient sequential estimation of the intensity rates in birth-death type queues is considered by Manjunath [1984], and estimation of the Poisson intensity in $M|G|=queue$ by considering an $M|G|=queue$ as a one dimensional mosaic process is dealt by Peter [1985].

Another approach to the estimation problem is through simulation techniques. A number of papers have been published in this area (cf. Bhat and Rao [1987]). Here the objective is to determine the $m.1$ or consistent estimators of the queueing parameters and the queueing characteristics such as mean queue length or mean waiting time.

The other aspect of statistical inference as applied to queues is that of tests concerning the arrival/service rates and the traffic intensity $\rho$. Wolff [1965] and Cox [1965] provide some test procedures concerning the queueing parameters such as the arrival rate and the service rate in the context of birth-death type queues. Lilliefors [1966] provides an
F-test for the traffic intensity $P$ in the case of a $M|M|1$ queue based on m.l.e's obtained by Clarke [1957]. Thiagarajan and Harris [1979] develop methods for testing the hypothesis that the service times of an $M|G|1$ queue are exponentially distributed based on waiting time data. Basawa and Prabhu [1981] have discussed some asymptotic tests for the arrival/service rates in the case of $G|G|1$ queue based on the m.l.e's and moment estimates.

One more problem is the distribution selection problem. Gross and Harris [1974] discuss these problems in great detail.

1.5 OPERATIONAL ANALYSIS

The literature in queueing theory has been principally devoted to modelling of queueing phenomena and the analysis of such systems. On the other hand operational aspects which are prescriptive in nature, deal with design and control of queueing systems.

Early work in this area relates to design models of queueing systems. The design decision involves the selection of the input stream (e.g. the arrival rate or waiting room capacity), the queue discipline (e.g. the priority classes), or the service mechanism (e.g. the number of servers or the service rate). Using the theory of stochastic processes or simulation, one analyzes the long-run probabilistic behaviour of different system structures and selects an optimal one.
Given the queueing system design, the control problem involves the selection of an operating policy for dynamically adjusting the system parameters. The design models are referred to as 'static models' and the control models are called 'dynamic models'. Static models are dealt through standard optimization techniques while in dynamic models the methodology of semi-Markov decision processes is employed for establishing the existence of an optimal control strategy and characterizing its form. Some of the basic static and dynamic models are described in Gross and Harris [1974] and Hillier and Lieberman [1967]. Crabill et al., [1977] provide an extensive classified bibliography of research in this area. Teghem [1986] gives some recent developments in the field.

In the following sections we review some of the studies in design and control of queueing systems.

1.5.1 Static Models: Design models in queues were considered as early as in Erlang's time. His paper (cf. Brockmeyer et al. [1948]) 'on the rational determination of the number of circuits' deals with the optimum number of channels so as to reduce the probability of loss to the system. Bailey's [1952] simulation procedure to achieve a better appointment system in hospitals, Brigham's [1955] determination of the optimum number of clerks to be assigned to the tool crib counters, and Edie's [1956] study of traffic delays at toll booths are the other significant examples.

Morse [1958] considers some optimization problems with
simple cost functions in the case of $M|M|1$ queue with infinite and finite capacities, $M|M|s|s$ Erlang's loss system and machine interference problems. In all these models the optimal service rate $\mu$ is taken as the decision variable. Hillier [1963, 1964] considers the problem of determining the optimal service rate, optimum number of servers and service facilities in single and multiserver Markovian queues. Stidham [1970] treats similar problems in the case of $GI|G|1$ queue with nonlinear cost functions.

The literature on design models is quite meagre as compared to that on control models.

1.5.2 Dynamic Models: Here we consider three types of control processes: (i) arrival process control (ii) service process control and (iii) control of queue discipline.

(i) Arrival Process Control: The literature under this title can be broadly classified into the following four types:

a. Accept or reject a customer
b. Adjust the mean arrival rate
c. Arrival control by levying
d. Individual vs Social optimization

a. Accept or reject a customer: Here the arrival process is controlled by accepting or rejecting a customer. Naor [1969] considers this problem first for the $M|M|1$ queue where each arriving customer receives a fixed reward $r$. It is shown that the optimal joining policy is of the form: join iff $i < n^*_I$, where the critical number $n^*_I = \min \{i \mid (i+1)/\mu > r\}$.
Miller [1969] and Scott [1970] extend this to multiple server queues with several customer classes. Mine and Ohno [1970] obtain a policy for the M|G|1 queue of the type 'accept those customers who arrive during a prefixed time interval and reject those who come after this'. The optimal rejection time is also determined. Lippman [1975] develops an optimal policy of the type (s,S) policy for arrival control in a M|M|s system. Helm and Waldman [1984] have studied the problem of optimal customer admission to a GI|M|s type system when the service mechanism, arrival process and the cost structure are all dependent on a random environment. Customer admission problems have also been treated by Stidham [1984], Haia Wein-Shen and Scott [1983].

b. Adjust the Mean Arrival Rate: In this the customers are accepted by altering the mean arrival rate. Such control problems were first considered by Emmans [1972], Man [1973] and Lippman [1975]. Irwin and Kabak [1975] treat the problem of optimal allocation to a M|G|1 system. Johansen and Stidham [1980] have made some further investigations of a similar problem.

c. Arrival Control by Levying: The idea of reduction of queue through the use of price was first suggested by Leeman [1964]. Low [1974] examines the M|M|s finite capacity queue in which the decision maker dynamically selects the customer arrival rate so as to balance the linear holding costs against the higher entrance fee that leads to slower arrival rates. Anderson [1980] extends this result to a M|E_k|1 system.
Balachandran and Schaefer [1979] propose a model in which the customer makes use of information on his average waiting time to determine an optimal admission policy.

d. Individual vs Social Optimization: Acceptance or rejection of a customer to a system may be viewed from two different angles. First, from the point of view of the society by maximizing the benefit received by all the customers over a time horizon and second from the point of view of the arriving customer himself by taking a decision to join the queue or not so as to maximize his own benefit. These problems have been treated by many authors. For instance, see Yechiali [1971], Knudsen [1972], Lippman and Stidham [1977], Stidham [1978], Winston [1977,1978], Mandel [1983] and Whitt [1986]. In all these models the authors have found that individual and social optimization do not lead to the same optimal policy.

(ii) Service Process Control: Following are a few main models which comes under this title:

a. Server On-Off Models
b. Models dealing with Removable number of Servers
c. Rate Control Models

a. Server On-Off Models: Several authors have dealt with the problem of service process control by turning the server on or off during the operation of the system. Obviously the question is to determine when, it is optimal to turn off the server and when to turn on. Heyman [1968] considers an
M|G|1 system based on a cost function of the type used in inventory models and obtains a policy of the form: 'turn the server on when n customers are present, and turn off when the system is empty.' This policy is known as N-policy. Tijms [1976] develops a policy called D- policy of the type 'turn the server off only when the system is empty and turn the server on only when the work load exceeds a level D.' Heyman [1977] develops a T - policy for the M|G|1 system in which the server is switched on only T time units after the end of the last busy period. Toshikazu [1981] uses diffusion approximation for determining an optimal operating policy for the M|G|1 system. Some related work has been done by Yadin and Naor [1963], Sobel [1969], Bell [1971], Baker [1973], Lippman [1975], Shivalzian [1979] and Michael [1985].

b. Removable number of Servers: The control of queues through removable servers was first proposed by Romani [1957]. In his model additional servers are added if the queue length exceeds a critical value so as to prevent the queue from building up too large. Moder and Phillips [1962] modify the Romani model by imposing a limit to the number of additional servers and including the case of deletion of servers if the queue length drops below a critical value. McGill [1969] and Magazine [1971] have also discussed these problems. Bell [1975] treats the control problem in a M|M|s system and shows that an optimal control policy might require turning one or more servers off even when there are
customers waiting to be served. Huang et al. [1977] extend Bell's work and develop optimal policies when the system is observed periodically. Winston [1978] takes up the same problem when the arrival rate depends on the number of customers present in the system. Szarkowicz and Knowles [1985] extend the work by Huang et al. with less restrictive assumptions. Infinite horizon models have also been considered by these authors. Teghem [1975] discusses this problem for the M|G|1 queue. Teghem [1985, 1986] makes a comparative analysis of N,D and T - policies and provides a survey of control of queues through removable servers.


c. Rate Control Models: Control of queues by varying the service rates has been considered by many authors. The earliest work was by Yadin and Naor [1967]. They propose the following model: The service intensities are changed in a discrete manner and chosen from a set \( \{ \mu_1, \mu_2, \ldots, \} \) \( \mu_i < \mu_{i+1} \) \( (i=1,2,\ldots) \). Two sets of integers \( [R_1, R_2, \ldots] \) and \( [S_1, S_2, \ldots] \) are specified, the integers being in the
increasing order of magnitude. Then the policy is to increase the service intensity from $\mu_i$ to $\mu_{i+1}$ if the queue size exceeds $R_{i+1}$. On the other hand, the intensity is decreased to $\mu_{i-1}$ if the queue size drops below $S_{i-1}$. The sequences $\{R\}$ and $\{S\}$ are optimally determined.

Gebhard [1967] treats a very similar problem of service rate switching policy with only two possible service rates. Crabill [1972], Sabeti [1973] discuss the service rate control problem in $M|M|1$ system where again the server can operate at only one of a finite number of service rates. The modification here is that a reward is attached for each service completions. Cramér [1971] and Miller [1969] deal with this problem for the $M|M|s$ system with a finite capacity in which the customers are distinguished by the reward associated with their acceptance into the queue. Cohen [1976] considered the optimal switching policy with only two levels of service rate for the $M|G|1$ queue.


Singh [1970] considers the allocation of service rates in the case of two servers. He compares the efficiency of a nonhomogeneous server system over the homogeneous server system. Lin and Kumar [1984] also discuss a similar problem. Such a problem for multiple server system has been considered by Tahara and Nishida [1975], Akihiko and Tahara [1975] and
Nath and Enns [1981].

Langen [1982] and Jo and Stidham [1983] have used the phase method (approximating the service distribution by a mixture of Erlang distributions) for the service-rate control problems. Lam Yeh and Thomas [1983] extend the treatment of rate-control problems to the case where the arrival parameter is assumed unknown. Verma [1986] discusses the multiobjective optimization problem for an $M|M|1$ system.

1.5.3 Control of Queues through Queue discipline:

In all of the above models it is assumed that customers are serviced according to 'first-come first-serve' basis. There are many situations, however, where the order of service is a controllable variable. Priority models, scheduling models and allocation of customers deal with such control techniques. Phipps [1956] shows that by serving the customers with shortest processing time, queue length can be reduced. There are studies where the concept of 'bribing' or purchase of priorities have been incorporated. See for instance, Balachandran [1970, 1972], Klienrock [1967], Tijms [1974]. Heyman [1985] uses a queue discipline for a multiserver system, where the priority class of a customer is determined only when he starts receiving the service. Nawijn [1985] uses a somewhat different approach. Here the server chooses between two actions: either he immediately starts service or he admits the newly arrived customer to the system, but delays service pending the next arrival. This process is
continued. The author obtains an optimal policy. Crabill et al. [1977] provide some additional references.

1.5.4 Control of Bulk Queues: Some researchers have examined the control of bulk queues. These problems include the optimal batch size for service, optimal time for batch service, optimal service rate etc.. This was first considered by Deb [1971]. Deb and Serfozo [1973] examine a batch service queue where each batch size and its time of service are subject to control. They obtain a policy of the type:

at a review point when \( x \) customers are waiting, serve \( \min \{x, Q\} \) customers (\( Q \) being the service capacity) iff \( x \) exceeds a certain optimal level \( M \). This is referred to as a control limit policy. Such policies have also been discussed by Deb [1974], Ignall and Kolesar [1974], Kosten [1973]. Weiss and Pliska [1982] develop a somewhat different kind of policy based on 'cost-rate process'. Unlike the control limit concept, here the policy is to serve all the customers as soon as the cost rate first exceeds the minimum average cost. Deb [1984] extends the control problem of bulk queues when the arrival process is compound Poisson. Maki [1985a, 1985b] discusses the control problems with a constraint on the waiting time. Choudhry and Templeton [1983] give an extensive bibliography of literature on bulk queues including control problems.

1.5.5 Control problems in Machine Interference: This problem has been analyzed as a particular case of birth-death type queueing control and design problems. A few noted references are Hillier [1964], Winston [1977], Smith [1978] and Derman
1.5.6 Control of Tandem Queues: A queue is called a tandem queue if the service is made up of several phases and is rendered by facilities arranged in series. These are commonly applied to production processes, computer systems and information processing. In the literature on control of such tandem queues most of the authors have considered only two-station tandem queues. For instance, see Rosberg et al. [1982], Hajek [1984], Ghoneim and Stidham [1985] and Nishimura [1986]. Dang [1979] discusses the optimal dynamic operating policies for a general tandem queueing system.

1.5.7 Control of Network Queues: Another class of queueing systems are network of queues in which customers move from one queueing facility to another in some random fashion until they depart from the system at various points. Here each node of the network may be treated as a queueing system by itself. Network of queues have very important applications in communication systems, computer systems, production processes and traffic flow. The theoretical developments can be seen in Klienrock [1976]. The control of network of queues have been considered by Brandwajn [1976], Dong-Wan and Pliska [1977], Ephremides et al. [1980], Larsen [1981], Verma [1984], Pan and Morimura [1986] and Weber and Stidham [1987]. Ralph Disney and Konig [1985] have given a survey of literature in queueing network theory and provide about 300 references.
1.6 SCOPE OF THE THESIS

In the present thesis we discuss some applications of optimization and statistical techniques in the design and control of queueing systems. As compared to the literature on the behavioural aspects of queueing theory, the literature on statistical inference as applied to queues is very meagre. Most of the published articles on inference in queues relate to estimation and tests of hypotheses concerning the arrival/service rates of Poisson/exponential distributions and birth-death type of queues.

In many queueing applications, a performance characteristic of great importance and interest is the traffic intensity $\rho$ (ratio of mean service time to mean interarrival time). Since $\rho$ is the ratio of two population parameters it does not possess an unique inverse. That is, an infinite combinations of these two parameters leads to a single value of $\rho$. Therefore, it is desirable to develop test procedures for $\rho$ rather than for the arrival or service rates individually. This is particularly of interest from the point of view of design and control of queues. For example, a queueing system may be designed to operate at a desired optimum level $\rho_0$ of traffic intensity. Then the purpose of a control technique is to signal in time any change in $\rho$ from the design level $\rho_0$ and take corrective actions such as increase or decrease of the service rate, arrival rate, number of servers or some combinations of these to bring back $\rho$ to the design level $\rho_0$. To this end, control of
Queues can be seen as problems in statistical inference for the system parameters.

In chapter 2, we discuss a method to detect changes in the traffic intensity based on the sequential probability ratio test (SPRT). The method is applicable to M|G|1 and GI|M|s type queues, machine interference problems and bulk queues. The procedure has been developed using the SPRT as applicable to a Markov dependent sequence of observations. The procedure of implementing control through the SPRT is explained. Extensive numerical methods have been applied in order to obtain the operating characteristics and average sample number of the test procedure. A numerical illustration in the case of M|M|1 queue has been provided.

In chapter 3 we discuss some statistical inference problems for $\rho$ in M|E_k|1 and E_k|M|1 queueing systems. It may be recalled that in the case of M|G|1 queueing system the number of arrivals occurring during the successive service times form a sequence of i.i.d random variables. In particular for the M|E_k|1 system the distribution of this random variable involves only the parameter $\rho$. Similarly for the GI|M|1 system the number of service completions during the successive interarrival times forms a sequence of i.i.d random variables, and in particular for E_k|M|1 queue the distribution of these random variables involve only the single parameter $\rho$. Based on these facts, a detailed study has been made regarding the maximum likelihood
estimation of $\rho$, study of the properties of these estimators, large sample tests and confidence estimation, and small sample tests for simple and composite hypotheses concerning $\rho$. Numerical illustrations have been included.

A large sample test for $\rho$ is discussed in Chapter 4. The test statistic is developed through a logarithmic transformation of the observations on service times and interarrival times. The test is based on normal approximation and is applicable to single and multiserver queues with general input and service time distributions. These test procedures are simple to use and no steady state conditions are required.

In queueing problems there are situations where the exact forms of the input/service time distributions are not known. Accordingly, Chapter 5 discusses some applications of nonparametric statistical inference such as estimation, tests of hypothesis and confidence estimation of the queueing parameters. Certain distribution-free test procedures for testing the exponentiality of queues are also developed. A few standard nonparametric tests are applied for testing the stability of queues, through developing a test procedure for testing $H_0: \rho \geq 1$ against $H_1: \rho < 1$. These tests are compared with the parametric tests through the vehicle of asymptotic relative efficiency.

Static models deal with design of queueing systems. This involves optimal determination of service rate, number
of servers, number of service facilities or some combinations of these. The models available in the literature so far are unconstrained optimization models. However, for the better system utilization it is desirable to have constraints on the mean queue length, mean waiting time, busy period, server utilization etc. In Chapter 6, we discuss a variety of constrained optimization problems related to single and multiple server queues. Applications of linear programming, nonlinear techniques, nonlinear algorithms and goal programming are discussed. Numerical illustrations are provided.

A large number of communication systems and time shared computer systems or other types of computer systems have been modelled as queueing systems in the literature. In Chapter 7, we discuss some unconstrained as well as constrained optimization problems related to such queueing systems. Problem of determining the optimal CPU rate, and problem of maximizing the throughput in simple feedback queueing systems and cyclic queueing systems are discussed.

Chapter 8 discusses some optimization problems related to tandem queues. The problem of optimal allocation of service rates, number of servers etc. , with constraints have been considered. The problems have been formulated as simple multistage dynamic programming problems and provided with solutions. Numerical illustrations have been given in some cases.