CHAPTER I
GENERAL INTRODUCTION
[1- 51]
1.1 Introduction

A queue or waiting line is a familiar experience of our daily lives such as that are encountered directly or indirectly in banks, airports, gas stations, automobile traffic, hospitals, telephone traffic, computer systems, flexible manufacturing systems and communication systems etc. In many of these real life situations, it is desirable to reduce waiting by better scheduling, improved disciplines and allocation of service intensities etc. Since formation of a queue leads to loss of time and money to both the customers and administrators of the system, therefore it is beneficial to the society if it can be managed so that both the unit that waits and the one that serves get the most benefit. Many researchers have since been carried out to develop mathematical models and theories to study various irregularities and statistical fluctuation of such systems in order to modify in proper way so that better arrangement of the facility could be suggested to management. Queueing theory is the mathematical study of queue or waiting lines. It is a branch of stochastic process representing various physical and biological phenomenon of common occurrence.

It is formed when units (or customers) demanding service of some description arrive at a service facility where such service is available and when the service is not immediately available to the arriving units or because of incapability of the server(s) to cope with all the units at the same time and depart the system after being served or
sometimes without being served. Though a queue or waiting line is one of the most unpleasant experiences of life but unfortunately, this phenomenon happens to be common in congested, urbanized, "high-tech" societies and the reduction of the waiting time usually requires extra investments. Therefore knowing the effect of the investment on the waiting time we have to develop models and techniques to analyze whether to be invested or not.

Queueing theory is defined as a science of congestion, a natural phenomenon in real systems. A service facility gets congested if there are more customers than the server can possibly handle. Congestion being a natural phenomenon in real systems cannot be simply handled by increasing resources without limits. Therefore the queueing theory mainly faces the problem of balancing the tradeoff between the gains of less congestion by supplying more resources versus the expense of providing more service capability. This view defines queueing theory as a branch of optimization theory or more generally operations research, as the results are often used to make business decisions about the resources needed to provide service.

A queueing node consists of an input stream which indicates whether the arrivals occur singly or in groups or batches thereof, a waiting room or buffer (finite or infinite), a servicing facility consisting of one or more servers (channels) which can serve customers and a set of disciplines and policies specifying the rules for queueing and servicing customers or units. Queueing networks are formed by routing customers among multiple queueing nodes. A queueing network consisting of a single node is referred to as a queueing system. Here we identify the unit demanding service, whether it is human or otherwise, as the customers arrive at a facility where such service is available. A mechanism that provides the kind of service on customers or units feed
into either a *server* or a *service channel*. Mathematically, a basic queueing system consists of three components- *sources of service, the queue or waiting line* and *the server or several servers*. Customers or units may arrive in any fashion whatsoever. When the server is free, the incoming customer is provided service immediately. Customers arriving when the server is busy are put in a line and wait. When the server completes the service, any customer from the waiting line may be given access to the server, depending on the rules of *queue discipline*. The *queue discipline* specifies the disposition of blocked customer (customers who find all servers busy). The most common observed in everyday life is *first come first served*; though sometimes there are other possibilities have been considered too.

### 1.2 Historical aspects

The history of queueing theory dates back 100 years ago. The theory of queueing systems was developed to provide models for forecasting behaviors of systems subject to random demand, then; the earliest problems studied were those of largely in the field of telephone traffic congestion. The Danish mathematician, statistician and engineer **Agner Krarup Erlang (January 1, 1878–February 3, 1929)** was the pioneer investigator who developed telephone-traffic theory served the Copenhagen Telephone Exchange in the period 1909-1922. His works inspired engineers, mathematicians to deal with queueing problems using probabilistic method. In 1907, F. Johannsen, the managing director of that telephone company, published an article entitled *Waiting times and number of calls* (reprinted in *Post Office Electrical*
Journal, London, October 1910) as a prelude to queueing. But the method used in the paper was not mathematically rigorous.

In 1909, Erlang published his fundamental paper that has historic importance is “The theory of probabilities and telephone conversations” which is widely considered as marking the birth of queueing theory. He laid solid foundations for modern queueing theory by introducing Poisson distribution. Based on the previous works published in 1909, Erlang published his most significant notable paper in Danish and later in English, German and French entitled “Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges” in 1917 where he found that a telephone system can be modeled by Poisson customer arrivals and exponentially distributed service times. Erlang’s contributions in the field include some of the most important concepts and techniques of analysis; the notion of statistical equilibrium, the technique of writing the so-called balance-of-state equations. He was also responsible for tackling the optimization problem for the first time in the queueing theory. In a way, Erlang’s studies in the applications of analytical methods to operational problems appear to mark the beginning of the study of Operations Research.

Erlang’s impressive contribution on the application of the theory to telephony continued to stimulate an enormous volume of work in queueing. Besides Erlang, E.C. Molina published his paper “Application of the Theory of Probability to Telephone Trunking Problems” in 1927, which was followed by Thornton Fry’s “Probability and Its Engineering Uses” published in 1928, the motivation of the works has been the practical problem of congestion. In the following years the theory developed by several theoreticians and the diverse areas of its applications has grown and contributed
significantly to the technological progress. Some of the authors with important contributions to queueing are Pollaczek (1930, 1959), Palm (1938, 1943) and Khintchine (1932, 1960). One of the pioneering works of Pollaczek in the theory was the development of the well-known formula for a single channel with Poisson input and arbitrary output (Pollaczek-Khintchine). About in the 1950s, a large body of researchers contributes from diverse areas of research, particularly Operation research, management science and industrial engineering, has proved queueing theory to be a fertile field and active discipline of research area all over the world. From a theoretical standpoint, Kendall (1951, 1953) was the pioneer who viewed that imbedded Markov chains can be identified in the queue length process in systems $M/G/1$ and $GI/M/S$. The universally used symbolic notation to identify queueing systems was coined by Kendall (1953), stimulated considerable interest.

In tracking this growth, we may cite some of the significant contributions who laid the foundation for a vigorous growth in the application of queueing theory were Lindley (1952) and Lederman and Reuter (1956) on time dependent solutions, Takács (1962) on waiting time, Cox [1955(a)] on supplementary variables, Kendall (1951) on embedded Markov chains, Karlin and Mogregore [1957(a,b)] on birth and death processes etc. The apparent significance of queueing was also appealing to Kolmogorov (1931) and Fry (1928) in their historic works on queueing.

(1989), Sharma (1990), Takagi [1991, 1993], Dshalalow (1995) etc. that have made a profound impact on the direction of research in queueing theory.

Research on the theory of queues has developed many new general methods which are explicitly used for solving many particular problems. At the same time research has taken place in the development and application of new and old queueing models for particular application.

1.3 Basic terminology

The basic characteristics which provide an adequate description of a queueing system can be identified as follows:

(i) The input or arrival pattern of customers: The input or arrival pattern describes the manner in which the units demanding service called customers arrive in a service facility and join the system. It is specified either by the number arriving during a time interval or by the inter arrival time between successive arrivals. Let the successive occurrence of arrivals to the system be at times $t_1,t_2,\ldots$, then the inter arrival time defined by $u_n = t_{n+1} - t_n$ ($n = 1,2,\ldots$) is assumed to be independent and identically distributed random variables. The input process can be characterized by the distribution of the inter arrival times of the customers, denoted by $A(u)$, that is

$$A(u) = \Pr[u_n \leq u]$$

with the respective probability density function (pdf) $a(u) = \frac{dA(u)}{du}$
The number of arrivals to the system may be finite or infinite who may arrive individually or in batch of fixed or varying sizes. Again the source from which the arrivals occur may be finite or infinite. In this context, it may be mentioned that there are other factors of customer behavior such as balking, reneging and jockeying that require consideration as well.

(ii) The queue discipline: The queue discipline indicates the rule by which the units or customers are served followed by the server when a queue is formed. The commonly used rules are:

(1) FIFO – First In First Out: who comes early leaves earlier.

(2) LIFO – Last Come First Out: who comes later leaves earlier.

(3) RS – Random Service: the customer is served in random order.

(4) Priority.

Under priority queue discipline, customers are divided into two or more priority classes and services are rendered on the basis of their priorities, regardless of their time of arrival to the system. Again there are different kinds of priority queue discipline such as preemptive and non preemptive. In the former situation, the customer with the highest priority whenever entered in the system is served immediately by taking out the lower priority units from service, the service on the permitted units resumes only when there exists no higher priority units in the system. On the other hand, in the later situation, the highest priority customer join the head of the queue but cannot be served immediately until the customer presently in service is completed, even though this customer has a lower priority.
(iii) **The service pattern:** The service pattern, specified by the time taken to complete a service, indicates the manner in which the service is offered to customer at the service facility. The time required for servicing a unit (or a group, in case of batch arrival system) followed by server is called service time. Customers may be served either one by one or in a batch of fixed size or varying size by one or more servers having equal and different efficiencies. The service mechanism can be characterized by the number of servers $c \geq 1$ and the service times of the successive customers $s_1, s_2, \ldots \ldots$ which are supposed to be statistically independent and identically distributed random variables and also independent of the input process. The distribution function of the service times denoted by $B(s)$ is given by

$$B(s) = \Pr[\text{service time } \leq s]$$

The service process may depend on the number of customers demanding service. A server may provide service faster if the queue is building up or, on the contrary, may get flustered and become less efficient. The situation is referred to as state-dependent service in which service depends on the number of customers waiting. Although this term was not used in discussing arrival patterns, the problems of customer impatience can be looked upon as ones of state dependent arrivals, since the arrival behavior depends on the amount of congestion in the system.

Service, like arrivals, can be stationary or non stationary with respect to time. The dependence on time and dependence on state are not of similar type. The former is independent on the number of customers existing in the system, but rather depends on how long it has been in operation. The latter is independent on how long the system has been in operation, but depends only on the state of the system at a given time. It should be noted that a queueing system can be both non stationary and state dependent. Even if
the rate of service is high, it is very likely that some customers will be delayed by waiting in the queue. In general, since customers arrive and depart at irregular intervals, therefore the uncertainties involved are the number of servers, the number of customers being served at any time, the duration and mode of service; hence the queue length will assume no definite pattern unless arrivals and service are deterministic. Thus it follows that a probability distribution for queue lengths will be the outcome of two separate processes — arrivals and services, which are generally, though not universally, assumed mutually independent.

(iv) The system capacity: The number of customers that can wait at a time in a queueing system is an important factor to be considered. In some queueing processes there is a physical limitation to the amount of waiting room, so that when the space is filled to capacity, no further customers will be able to join the system until space becomes available as the result of a service completion. Since there is a finite limit to the maximum system size, therefore these are referred to as finite queueing situations. A queue under limited waiting room can be viewed as one with forced balking where a customer is forced to balk if it arrives when the queue size is at its limit. This is a simple case, since it is known exactly under what circumstances arriving customers must balk. Against this there may have an infinite capacity where the queue in front of the server(s) may grow to any length.

(v) The number of service channels: A system may have a single server where there is only one server serving all the existing units successfully or a number of parallel servers which can serve customers simultaneously. Fig. 1.3.1 reports an illustrative single server system, while Fig. 1.3.2 (a) and Fig. 1.3.2 (b) reveal two
variations of multichannel systems. The two multichannel systems differ in that the first has a single queue, while the later permits a queue for each channel. A hair styling salon having many chairs is an example of the first kind of multichannel system under the condition that no customer is waiting for any particular stylist, while a supermarket or fast-food restaurant might fit the second kind. It is generally assumed that the service mechanisms of parallel channels operate independently of each other.

![Single server queueing system](image1)

**Fig. 1.3.1: Single server queueing system**

![Multi channel queueing system](image2)

**Fig. 1.3.2: Multi channel queueing system.**

**(vi) Stages of service:** A queueing system may have only a single stage or several stages of service. An example of a multistage queueing system would be a
physical examination procedure, where each patient must proceed through a number of stages, such as medical history; ear, nose and throat examination; blood tests; electrocardiogram; eye examination; and so on. In case of some multistage queueing processes recycling, which is common in manufacturing process, or feedback may occur. Under recycling quality control inspections are performed after certain stages and parts that fail to meet quality standards are sent back for reprocessing. Similarly, a telecommunications network may process messages through a randomly selected sequence of nodes, with the possibility that some messages will require rerouting on occasion through the same stage. A multistage queueing system with some feedback is shown in Fig.1.3.3.

![Fig.1.3.3: A multistage queueing system with feedback.](image)

The six characteristics of queueing systems mentioned in this section are sufficient to completely describe a process under study. The major characteristics that are to be studied to understand the behavior of a queueing system are the queue length i.e. the number of customers (or units) waiting in the queue or present in the system, the length of the busy period i.e. the length of time interval when the server will be continuously busy and the length of idle period which is the length of the interval from the instant the server becomes free for the first time to the next arrival when it becomes engaged.
1.4 The anatomy of queueing system

From the perspective of applications of queueing theory results to realistic problems, the growth identifying major developments and directions has occurred at a phenomenal rate. Over the last century, a large variety of queueing systems have been successfully modeled and investigated extensively by researchers which could be treated as reasonable models of real-world phenomena. Their significant amount of research gave rise to a large body of literature and the creation of a huge collection of analytical queueing models.

The analysis of queueing theory includes the following points [e.g., see Medhi (2003)]:

1. Stochastic behavior of various random variables or stochastic processes that arises and evaluated related performance measures.
2. Various method for solving various queueing models like- exact, transform, algorithmic, asymptotic, numerical approximations etc.
3. Nature of solution as time dependent or limiting etc.
4. Behavior of the system- transient state behavior or steady state behavior.
5. Control and design of queues for comparing the behavior and performances under various situations as well as queue disciplines, service rules, strategies etc.
6. Optimization of specific objective functions involving performance measures associate with cost functions etc.

Usually, queueing problems are classified according to their model, processes and methods. Research of various queueing models generally consists of three phases:
(a) Formation of model, (b) Identification of some processes of interest and (c) Devising a solution method for the requisite statistics.

1.4.1 Description of various models and their notations

To describe a queueing model one uses the following terminologies:

(i) Input or arrival process
(ii) Queueing discipline
(iii) Buffer size or waiting room capacity
(iv) Number of servers and their capacities
(v) Service discipline and service process
(vi) Vacation discipline and vacation process
(vii) Network configuration or routing.

A very convenient notation to describe a queueing model was originated by Kendall (1953) which is being universally used throughout the queueing literature. The basic representation identifying the above elements has the form

\[ A/B/m/C_w/C_s/Q/(q,N) \]

The complete description of each part of 7-tuple is as follows:

The first part A of the notation stands for arrival process i.e. inter-arrival time distribution. On the other hand \( A^X \) stands for batch arrivals.

The common arrival processes are – (i) Renewal process, (ii) Semi-Markov process (SM), (iii) Markov arrival process (MAP), (iv) Batch Markov arrival process (BMAP).
The second term B identifies service process while $B^x$ denotes batch service and the common service time distributions are- (i) M-Exponential, (ii) D-deterministic or constant, (iii) $E_k$ - k Erlang, (iii) $H_n$ - Hyper-exponential of type n, (iv) PH- Phase type distribution, (v) G- Arbitrary or general distribution and (vi) SM- Semi-Markov.

The third symbol $m$ represents the capacity of the servicing facility i.e. the numbers of servers or service channels that may either be 1 or more than 1(one) or may be finite or infinite.

$C_w$ specifies the limitation on waiting room capacity of the system according to the prescribed model which may be finite or infinite.

$C_s$ is the restriction on system capacity ($s \geq 1$), which is the maximum number of customers in a servicing batch.

$Q$ indicates the queue or service discipline used at the service facility in which the customers are served. If $Q$ is omitted then the service discipline is always FIFO.

Finally, the last part stands for the size of the finite or infinite source from which the arrivals occur and also denotes the busy period discipline which specifies the rules by which the system starts and ends a busy period. Under these rules, the server stops processing new customers when the queue length drops below a certain level (say) $q \ (\geq 1)$ and it resumes service when the queue length reaches or exceeds level $N \ (\geq 1)$. Busy periods alternate with periods during which the server suspends service. In the latter case, the server either idles or takes vacation or takes vacation and idles. If the arrival process follows the bulk arrival process or if the server leaves for vacations, the queue length at the beginning of a new busy period need not be exactly $N$. The presence of the notation $q$ as mentioned above is associated with the term $q$-
quorum or just quorum. The rule of entering a busy period with a specified value $N$ is referred to as $N$-policy.

In most practical situations only the first three symbols are necessary to denote any kind of model and the rests are demanded according to the importance of the existing model. Also, the various characteristic descriptions are suitably modified to cover more complicated models. Thus by an $M/G/1$ we mean a single server queueing system having exponential inter arrival time distribution and general (arbitrary) service time distribution.

1.4.2 Various queueing processes under study

In all the analysis of queueing systems the ultimate objective is to understand the behavior of their underlying processes with the hope that informed and intelligent decisions can be made in their management. The basic processes of most interest in the research of queueing systems are –

(a) The queueing process: The queue size or queue length is the distribution of the number of customers waiting in the queue, in front of the service facility for receiving service, excluding the one being served, if any. In the context of a queueing system the number of customers with time as the parameter is a stochastic process. Let $Q(t)$ be the distribution of the number of customers in the system (including those being in service) at time $t \geq 0$ then the queueing process is denoted by $\{Q(t)\}$. In order to manage the system efficiently we have to observe the behavior of the process $Q(t)$ over time. The determination of the probability distribution of this discrete random
variable is very important for the design of the system as it enables assessment of the need to provide an extra server or a larger waiting space to accommodate large queue.

(b) The waiting time process: Waiting time is the amount of time a new arrival has to wait at the service facility until its service starts. The actual waiting time distribution is given by \( \{W_n\} \) where \( W_n \) denotes the waiting time of the \( n^{th} \) arrival in the queue (excluding his service time). The virtual waiting time distribution (Takács process) is \( \{W(t)\} \) where \( W(t) \) is the length of time an arrival has to wait had he arrived at time \( t \). The probability distribution of this continuous random variable is very important from the customer point of view.

(c) The busy period process: The busy period is the duration of the interval from the moment the service starts with arrival of a unit at an empty counter to the moment the server becomes free for the first time. The distribution of the busy period is the length of time during which the server remains busy. The busy period is a random variable and is a continuous process which alternates with idle or vacation periods. In the former case, the server remains idle as long as the system has fewer than \( N \geq 1 \) customers. In the latter case, the server leaves the system for vacation, which may be related to maintenance or service of secondary customers.

(d) The Output process: The output process refers to the flow of completely processed customers. Each of these processes can be studied in 'transient' or 'stationary' mode where analysis conventionally aimed at their one dimensional distributions. A time dependent distribution is known as the transient distribution as opposed to the steady state or stationary distribution in which case the underlying
process settles down to what is commonly termed as a state of equilibrium under certain conditions. Stochastic processes usually approach steady state quite rapidly. Other special characteristics may be of interest in conjunction with above processes or independently, such as various statistical data and optimization.

1.4.3. Basic Queueing Model:

A lot of development contributed by several researchers has grown on computational aspects of queueing model and their works have led to extensive applications of various methods. To develop the operations of queueing systems A. K. Erlang applied Markov process methodology successfully to simple queueing models and inspired many researchers to study more general Markov process. Some classical birth and death queues of Erlang are as follows:

(a) Erlang's system $M/M/1/\infty$: Under this model we consider a queueing system in which arrivals occur one at a time in a Poisson process having parameter $\lambda$ and customers are served in the order of their arrival by a single server at the service facility with service time being independent and exponentially distributed with common parameter $\mu$. If $Q(t)$ be the number of customers at time $t$ waiting in the queue including the one being served, if any, then the queueing process $\{Q(t)\}$ will be a particular case of the general birth and death process. The transition probabilities $p_m(t) = P\{Q(t) = n / Q(0) = i\}$ for $t > 0$ and $p_m(0) = \delta_y$ satisfies the forward Kolmogorov differential difference equations of the birth and death processes

\[
\begin{align*}
    p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\
    p_n'(t) &= - (\lambda + \mu) p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t), \quad n > 0
\end{align*}
\]

(1.4.3.1)
Let $\rho = \frac{\lambda}{\mu}$ = Traffic intensity of the system (also known as utilization factor). Now under the equilibrium condition imposed on $1 + \rho + \rho^2 + \ldots = \infty$, i.e. $\rho < 1$, the steady state distribution $\lim_{t \to \infty} p_n(t) = p_n$ exists and is independent of the initial state. In equilibrium, the equations in (1.4.3.1) reduce to a system of linear equations solving which we have $p_n = (1 - \rho)p^n$; $n \geq 0$

(b) Erlang's loss system $M/M/s/0$: We consider a model which envisages that a customer, who finds, on arrival, that all the $s$ servers are busy, leaves the system without being served and is lost to the system. This type of model known as a ($s$-channel) loss system, which was investigated initially by Erlang, assumes that customers arrive at a $s$ server system in accordance with a Poisson process with rate $\lambda$ and the service times of the servers are exponentially distributed with distribution function $G$. Assume $G$ to be continuous having density $g$ and hazard rate function $\lambda(t)$. That is, $\lambda(t) = g(t)/G(t)$ is the instantaneous probability intensity that a $t$-unit-old service will end. This system models telephone traffic where incoming calls are lost or blocked if all lines are busy. If $N(t)$ denotes the number of channels busy in a telephone exchange with a finite number $s$ of channels, in which customers do not wait for service, then the queueing process $\{N(t), t \geq 0\}$ is again Markov and in particular, is a birth and death process. The equilibrium of such a system exists unconditionally and the steady state probabilities $p_n = \lim_{t \to \infty} P[N(t) = n]$, $n = 0,1,2,\ldots,s$ satisfies the following famous Erlang loss formula which is denoted by $B(s, \rho)$
This formula is also true for the case of generally distributed service times i.e. for the system $M/G/s/0$. In 1932, F. Pollaczek gave the first intuitive proof of this formula for this general case. A numerically convenient formula to compute $B(s, \rho)$, suggested by Jagerman (1974), is via

$$[B(s, \rho)]^{-1} = \sum_{j=0}^{\infty} s^{(j)} \rho^{-j}$$

where $s^{(j)} = \{1, j = 0
\begin{align*}
& s(s-1)....(s-j+1), \quad j \geq 1
\end{align*}$

and it is useful when $\rho > s$.

Thus, (1.4.3.2) admits of Fortet’s integral representation,

$$[B(s, \rho)]^{-1} = \rho \int_{0}^{\infty} e^{-\rho y} (1 + y)^s dy = \int_{0}^{\infty} e^{-y} \left( 1 + \frac{y}{\rho} \right)^s dy.$$  

which permits evaluation of loss function for non-integral number $s$ of servers in practical traffic-related problems.

1.4.4 Some useful techniques:

To solve the standard queueing models, we briefly describe some of the generally applied mathematical techniques.

(1) Birth and Death processes: Markovian queueing models, where arrivals occur in accordance with a Poisson process and the service times are independently and
exponentially distributed, have been widely used to help decision making in the early stages of queueing theory in the telephone industry. The underlying Markov process that represents the number of customers is known as birth and death process used in population models. In 1924, G. Yule first introduced birth and death processes as pure birth processes (i.e. if death is impossible) which is referred to as a Yule process. Later, the theory was developed by Feller (1968) while the more comprehensive treatment in this context was found in Cohen (1982). These processes are characterized by the property that a birth and death process is a continuous parameter Markov chain which changes only through transitions from a state to its immediate neighbors.

A continuous-time Markov chain \( \{X(t), t \in T\} \) with state space \( S = \{0, 1, 2, \ldots\} \) and with rates
\[
q_{i,i+1} = \lambda_i \quad \text{(say),} \quad i \geq 0 \\
q_{i,i-1} = \mu_i \quad \text{(say),} \quad i \geq 1 \\
q_{i,j} = 0, \quad j \neq i \pm 1, j \neq i, i \geq 0 \\
\]
and
\[
q_i = -(\lambda_i + \mu_i), \quad i \geq 0, \quad \mu_0 = 0
\]
is called

(i) a pure birth process, if \( \mu_i = 0 \quad \text{for} \quad i \geq 1 \)

(ii) a pure death process, if \( \lambda_i = 0, \quad \text{for} \quad i \geq 0, \)

and (iii) a birth and death process if some of the \( \lambda_i \)'s and some of the \( \mu_i \)'s are positive [e.g. see Medhi (2003)]

Now using the following Chapman-Kolmogorov forward equation
\[
P'_y(t) = \sum_{k \neq i} p_{ik}(t)q_{ky} + q_j p_y(t),
\]
we get the Chapman-Kolmogorov forward equations for the birth and death process.
For \( i, j = 1, 2, \ldots \),

\[
p'_j(t) = -\left( \lambda_j + \mu_j \right) p_j(t) + \lambda_{j-1} p_{j-1}(t) + \mu_{j+1} p_{j+1}(t)
\]

and \( p'_{0j}(t) = -\lambda_0 p_{0j}(t) + \mu_1 p_{1j}(t) \).

With the boundary conditions \( p_{ij}(0+) = \delta_{ij}, \quad i, j = 0, 1, \ldots, \)

where \( P_j(t) = \Pr\{X(t) = j\}, \quad j = 0, 1, \ldots, t > 0 \) and assuming the time at \( t = 0 \), the system starts at state \( i \), so that \( P_j(0) = \Pr\{X(0) = j\} = \delta_{ij}. \)

In matrix notation,

\[
P'(t) = P(t)Q
\]

Solving above equation, we get

\[
P(t) = P(0)e^{Q(t)}, \quad \text{with } P(0) = I; \quad \text{we have}
\]

\[
P(t) = e^{Qt} = I + \sum_{n=1}^{\infty} \frac{Q^nt^n}{n!}
\]

... (1.4.3.3)

Assume that the eigenvalues \( d_i \) of \( Q \) are all distinct, \( d_i \neq d_j; \quad i, j = 0, 1, \ldots, m \). Let \( D \) be the diagonal matrix having \( d_0, d_1, \ldots, d_m \) as its diagonal elements. Then there exists a non singular matrix \( H \) (whose column vectors are right eigen vectors of \( Q \)) such that \( Q \) can be written in the canonical form

\[
Q = HDH^{-1}
\]

Then \( Q^n = HD^nH^{-1} \)

and substituting in (1.4.3.3), we get \( P(t) = H\Delta(t)H^{-1} \);

where \( \Delta(t) \) is the diagonal matrix with diagonal elements \( e^{d_i}, i = 0, 1, \ldots, m \).

This process is of particular interest in queueing theory as several queueing systems can be modeled as birth and death processes. They also have interesting applications in other diverse fields such as economics, biology, ecology, reliability
theory etc. Karlin and Moregore [1957(a,b)] further developed the theory from application point of view in queueing theory.

It should be noted that an analytical solution can be obtained, especially when \( m \) is small, it becomes difficult when \( m \) is large. For such cases, numerical methods have been put forward. The randomization technique of construction in numerical analysis gives very useful formula for computation of \( P_y(t) \) or \( \Pi_j(t) \) i.e., transition probabilities of an uniformizable Markov process \( \{X(t), t \geq 0\} \).

(2) The **Supplementary variable technique**: The use of supplementary variable technique, which is an important technique to obtain a transient solution of a non-Markovian system, was first introduced by Kosten (1973). Later in 1951 Kendall indicated an alternative approach to determine the distribution of the original processes in order to solve the various queueing processes. Kendall’s concept was fully explored by Cox in 1955 [e.g. see Cox (1955(b))] who introduced a supplementary variable to construct the forward Kolmogorov equations for the \( M/G/1 \) queueing system. Earlier Kendall (1953) considered this technique, which he labeled as ‘augmentation’, but later preferred the imbedded Markov chain technique as leading to simpler calculations. In spite of this, the supplementary variable technique has been used by many authors to solve a large amount of queueing problems, and their solutions seem to be pretty good, and the mathematics involved becomes less cumbersome.

In the supplementary variable technique a non Markovian process \( \{N(t), t \geq 0\} \), where \( N(t) \) gives the state of the system or the system size in continuous time, is made Markovian by the inclusion of one or more supplementary variables. Therefore the basic idea under this technique is to augment the state \( N(t) \) with additional information i.e. supplement it with a variable to a Markov process which will enable
one to write down the differential equations, as in the case of a Markovian system. There are two kinds of supplementary variable, in general, one of which is known as the elapsed service time i.e. time already spent in service by time $t$ of a customer receiving service and the other is the remaining or residual service time of the customer in service, and for both cases the approaches of deriving the double transforms are different. The hazard rate function and the boundary conditions will be needed when elapsed service time is considered as supplementary variable. On the other hand, when we use the remaining service time as supplementary variable, none of these are needed and leads to much simpler calculations. Thus if the elapsed service time by time $t$ be denoted by $Q(t)$ then the couplet $\{N(t), Q(t)\}$ becomes a two dimensional homogeneous Markov process. By using Laplace and z-transforms, the corresponding system of the Kolmogorov partial differential equations are replaced by functional equations. For example, if we consider an $M/G/1$ queueing model then we have required only one supplementary variable that is the time since the last departure. Suppose that a general distribution of a random variable has the hazard function $\mu(x)$ which can be expressed as,

$$\mu(x) = \eta(x) \exp\left\{-\int_0^x \eta(y) \, dy\right\},$$

where $\eta(x)$ is the conditional completion rate for service at time $x$ and then $\mu(x)dx = Pr\{\text{service will be completed in } (x, x+dx) \text{ given that service time exceeds } x\}$.

Let $Q_n(x,t)$ be the joint probability density that there are $n$ customers in the system at epoch $t$ and a customer is present in service who has been there for time $x$. Then we can write the following equations which governs the supplementary variable technique.
\[ Q_n(x + \Delta, t + \Delta) = Q_n(x, t)(1 - \lambda \Delta)(1 - \eta(x) \Delta) + Q_{n-1}(x, t) \lambda \Delta, n \geq 1 \]

and \[ Q_0(x + \Delta, t + \Delta) = Q_0(x, t)(1 - \lambda \Delta)(1 - \eta(x) \Delta). \]

Therefore, the generating function, \( G(z, x, t) = \sum_{n=0}^{\infty} Q_n(x, t) z^n = G \) (say) of \( Q_n(x, t) \), which satisfies the following relation

\[ \frac{\partial G}{\partial t} + \frac{\partial G}{\partial x} + [\lambda + \eta(x)]G = \lambda z G. \]

Solving the above equation which is a partial differential equation of the Lagrangian type, the following result can be obtained

\[ G(z, x, t) = G(z, 0, t - x) \exp\{-\lambda x(1 + z) - N(x)\}, \]

where \( N(x) = \int_0^x \eta(y) dy = -\log[1 - A(x)] \); \( A(x) \) is the common probability distribution function of the independent customers at service time \( x \).

When the system is assumed to be ergodic with \( G_\infty(z, x) = \lim_{t \to \infty} G(z, x, t) \), the above relation simplifies to \( G_\infty(z, x) = G_\infty(z, 0) \exp\{-\lambda x(1 - z) - N(x)\} \).

Later, Conolly (1960) has obtained the transient state solution of the queue length distribution of \( G1/M/1 \) queueing system. To obtain the busy period distribution of \( G1/M/1 \) and \( G1/E_k/1 \) queueing system, Keilson and Kooharian (1960) used this technique by considering the couplet \( \{N(t), X(t)\} \), where the supplementary variable \( X(t) \) is the service time already received by the customer in service, if any. Jaiswal (1962) made heavy use of this method for \( M/G/1 \) queueing system to study the busy period distribution of the 'head of the line priority queue'. Borthakur and Medhi (1974) applied the technique to derive the queue length distribution for \( M^x/G(a,b)/1 \).
model. Chaudhry and Templeton (1983), Medhi and Templeton (1992), Borthakur et. al. (1987) and Takagi (1991) and Choudhury (2000) have also introduced this technique to study the various aspects of different queueing models.

(3) **Embedded Markov process technique:** Kendall [1951, 1953] developed the technique of embedded Markov process to convert the queue length processes in $M/G/1$ and $G/M/s$ into Markov chains by using the concept of regeneration points. The concept of regeneration points is credited to Palm (1943). For a Markov process, the whole parameter space is a set of regeneration points. A process possessing a set of regeneration points is called regenerative.

In the $M/G/1$ queueing system, the inter arrival time has the memory less property and we consider departure epochs as points of regeneration while in the $G/M/s$ system, the service time has the memory less property and arrival epochs are considered to be points of regeneration. Thus if in the $M/G/1$ system there exists an increasing sequence $\{t_n\}$ of regeneration points i.e. if $t_0 = 0, t_1, t_2, \ldots \ldots$ be the points of departure of customers such that transition probabilities $p_y$ associated with $\{Q_{n+1}|Q_n\}$, where we define $Q_n = Q(t_n)$, can be calculated easily, then analysis of the embedded Markov chain $\{Q_n, n = 0,1,2,\ldots\}$ will supply valuable information about the process $Q(t)$. The departure points of the process constitute the parameter space and the number of customers in the system will be the state space $S = \{0,1,2,\ldots\}$ whose transition probability matrix (TPM), $P = (p_y)$ can be expressed in explicit form. Under the equilibrium condition $\rho < 1$, the chain is ergodic and its limiting probability vector
\( \Pi = (\pi_0, \pi_1, \pi_2, \ldots) > 0 \) (vector), where \( \pi_i = \lim_{t \to \infty} p_{ij}^{(r)} \); \( i, j \in S \), is the unique left (stochastic) fixed point of the matrix TPM \( P \) i.e. \( \Pi = \Pi P \) with \( \sum_{j=0}^{\infty} \pi_j = 1 \)

The probabilities of transition \( p_{ij} = P\{Q_{n+1} = i / Q_n = j\} \) can be expressed as

\[
p_{ij} = \begin{cases} k_{j-i+1}; & \text{if } i > 0 \\ k_j; & \text{if } i = 0 \end{cases}
\]

Where \( k_j = \Pr \{j \text{ units arrive during a service time } 'B' \} \)

\[
p_{ij} = P(X_n = j) = \int_0^\infty \frac{e^{-\lambda t} (\lambda t)^j}{j!} dB(t); \ j \geq 0
\]

Such that its \( z \) transform of the service time is given by \( K(z) = \beta^* (\lambda - \lambda z) \)

We then have the Kolmogorov equation given by

\[
\pi_j = \pi_0 k_j + \sum_{i=1}^{j+1} \pi_i k_{j-i+1}; \ j \geq 0
\]

Let \( \Pi(z) \) be the probability generating function of \( \{\pi_j; j \geq 0\} \) then we define

\[
\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j; \ |z| \leq 1
\]

Evaluating \( \pi_j \) and solving the above equation gives the following Kendall’s formula

\[
\Pi(z) = \frac{\pi_0 (1-z) \beta^*(\lambda - \lambda z)}{[\beta^*(\lambda - \lambda z) - z]}
\]

where \( \pi_0 = 1 - \rho \) is obtained under the normalizing condition \( \Pi(1) = \sum_{j=0}^{\infty} \pi_j = 1 \)

Following Kendall’s embedded Markov chain theory, Bailey (1954) and Downton (1955) made an equilibrium study of bulk-service queue and obtained
expressions for the expected numbers waiting and the expected waiting time for the chi-square service-time distribution. Lindley (1952) gave integral equations for waiting time distributions defined at embedded Markov points in the general queue GI/G/1. Wishart (1956) developed waiting times for the systems GI/E_k/1 and D/E_k/1. Recently a considerable effort has been devoted to the study of the queue with server vacation introducing this embedded Markov process technique and an excellent survey of such systems is found in Miller (1964), Scholl and Kleinrock (1983), Fuhrmann (1984), Fuhrmann and Cooper (1985), Doshi (1986), Takagi (1991), Choudhury [2002 (a)] etc.

1.5 Some important criteria

In this section we discuss below some of the important criteria and theorems that are extensively applied in various queueing models.

(1) The Pollaczek-Khinchine formula: One of the most elegant and widely used formulas in queueing theory is the Pollaczek-Khinchine formula that was first incorporated by Felix Pollaczek (1892-1981) in 1930 and recast two years later in probabilistic terms by Aleksandr Khinchine (1894-1959). The formula can be applied when inter arrival time distribution of units to the workstation are exponential, and the service times follow a general distribution. Assume that the input process is Poisson with intensity \( \lambda \) and the service times are IID random variables having a general distribution with DF \( B(t) \) and mean \( \frac{1}{\mu} \). Assume that the steady-state condition \( \rho < 1 \) exists. Let \( W \) and \( W_q \) be the waiting time in the system and in the queue, respectively with \( W(t) \) and \( W_q(t) \) as their corresponding distribution functions. In particular, under
the steady state condition $\rho < 1$, $p_n = \lim p_m(t)$ as $t \to \infty$, exists and $\pi_n = \lim p_m^{(m)}$ as $m \to \infty$, exists. Then we have, for a Poisson input process

$$p_n = \pi_n = \Pr \{\text{departing customer leaves } n \text{ in the system}\} \quad (1.5.1)$$

for all $n$. We have the stationary distribution of the process $\{W\}$ given by

$$V(s) = \sum_{n=0}^{\infty} \pi_n s^n = W^*(\lambda - \lambda s)$$

where $W^*(s)$ is the Laplace Stieltjes transform (LST) of $W$ which satisfies the formula

$$W^*(s) = \frac{s(1-\rho)B^*(s)}{s - \lambda[1 - B^*(s)]} \quad (1.5.2)$$

where $B^*(s)$ is the LST of the service time distribution. Noting that $W = W_q + V(W_q$ and $V$ are independent), we obtain the expression for LST of $W_q$ i.e. for $W^*_q(s)$ as

$$W^*_q(s) = \frac{s(1-\rho)}{s - \lambda[1 - B^*(s)]} \quad (1.5.3)$$

The relations (1.5.2) and (1.5.3) are known as Pollaczek-Khinchine formula.

(2) **Little's formula:** One of the most important and useful relationships in queueing theory that hold under fairly general conditions is Little’s law, developed by J. D. C. Little in 1961. Little related the long-term mean number of customers in the system to the long-term average customer waiting times as follows:

$$L = \lambda W$$

where $\lambda$ is the mean arrival rate, $L$ is the average number of units in the system, and $W$ is the average waiting time of a unit in the system in steady state. If $L_Q$ and $W_Q$ be the average number of units in the queue and average waiting time of a unit in the
queue under the steady state condition then they are connected by a similar formula as given by

\[ L_q = \lambda W_q \]

The Little’s formula is of great universality and holds for all queueing systems irrespective of the form of inter arrival and service time distributions and discrepancy that may be within the system. It is valid for any system in steady-state under some very general conditions.

(3) Distributional form of Little’s law: As Little’s law was first appeared in 1961, its simplicity and importance has established it as a basic tool of queueing theory. Little’s law relates the average number in a system or queue, \( N \), with the average time in the system, \( T \) under very broad conditions. This was observed by Kleinrock (1975), Fuhrmann and Cooper (1985) and Svoronos and Zipkin (1986). However, Keilson and Servi (1988) demonstrated that for many systems, the relationship between the queue length and the time in the system can be characterized beyond just average value. For a given class of customers if

(C-1) the arrival process is a Poisson with rate \( \lambda \).

(C-2) all the arriving customers enter the system and remain in the system until they are served.

(C-3) the customer leaves the system at a time in order of the arrival.

(C-4) for any time \( t \), the arrival process after time \( t \) and the time in the system of any customer arriving statistically independent.

Then the relationship that holds between the probability distribution of the number in the system and the time in the system follows the simple formula
\[ \pi_N(z) = \alpha_T(\lambda - \lambda z) ; \]

where \( \pi_N(z) = E[z^N] = \sum P_z[N = n]z^n \) is the probability generating function of \( N \) and \( \alpha_T(\theta) = E[e^{-\theta T}] = \int e^{-\theta t} f_T(t)dt \) is the Laplace Stieltjes transform of \( T \).

The above relationship is known as distributional form of Little's law. In 1990, Keilson and Servi demonstrated that for many systems, the relationship between the queue length and the time in the system can be characterized beyond just their average value.

(4) Poisson Arrival See Time Average (PASTA): An interesting property of Poisson arrivals in a system is that such arrivals behave like random arrivals i.e. an observer from a Poisson input stream sees or finds the same system state distribution as a random observer having nothing to do with the system. For such systems under the steady-state condition \( \rho < 1 \), we have \( a_n = p_n \) for all \( n \geq 0 \) subject to the condition that there is neither bulk arrival nor bulk service. The result is also valid in the transient state i.e.

\[ a_n(t) = p_n(t), \text{ for } n \geq 0, t \geq 0. \]

Thus it can also be expressed as the requirement is that the future arrivals are independent of the current number of customers in the system, which is satisfied by the Poisson arrival process. This property can also be stated as Poisson arrivals see time average (PASTA). In queues with Poisson arrivals, the fraction of arrivals that find \( n \) customers in the system is equal to the fraction of time with \( n \) customers.

A simple proof of PASTA is given by Wolf (1982) under the lack of anticipation assumption (LAA).
Let $N(t)$ be the number of customers in the system in a queue at time $t$ where $A(t)$ denote the number of customers that arrive during $(0,t)$. Assume that the arrival process $A(t); t > 0$ is Poisson with finite rate $\lambda$. We further impose that lack of anticipation (LAA); for every $t > 0$, $N(t-0)$ on $A(t); t > 0$ through almost $A(u); 0 < u < t$. On the basis of these assumptions, for every $t \geq 0$ and $u > 0$, we have

$$\Pr[N(t-0) \in B] = \Pr[N(t-0) \in B|A(t) - A(t-u) > 1]$$

for every Boreal set $B$.

Furthermore, as an analogue of the PASTA property in the continuous time system, there is a BASTA (Bernoulli Arrival See Time Average) [Boxma and Groendijk (1988)] or GASTA (Geometric Arrival See Time Average) [Halpin (1983)], property in the discrete time system.

(5) Chapman Kolmogorov equation: In mathematics, specifically in probability theory, in particular the theory of Markovian stochastic processes and yet more specifically in the queueing theory where the systems are studied with the help of birth and death process, the Chapman Kolmogorov equation is an identity relating the joint probability distributions of different sets of coordinates on a stochastic process. The equation was developed independently by both the British mathematician Sydney Chapman (1888-1970) and the Russian mathematician Andrey Kolmogorov (1903-1987). The Chapman Kolmogorov equation, which is satisfied by the transition probabilities of a Markov chain, is given by

$$P_{jk}^{(m+n)} = \sum_{i \in S} P_{ji}^{(m)} P_{ik}^{(n)}$$
where $p_{jk}^{(m)} = \Pr\{X_{m+n} = k | X_n = j\}$ gives the probability that from the state $j$ at $n$th trial, the state $k$ is reached at $(m+n)$th trial in $m$ steps.

(6) **Foster's criterion:** In 1953, F. G. Foster developed the Foster's theorem to draw conclusions regarding the positive recurrence of Markov chains with countable state spaces. Let us consider an aperiodic, irreducible discrete-time Markov chains; $n = 1, 2, ....$ on a countable state space $S$ with a transition probability matrix $P = (p_{ij}) \forall i, j \in S$. Then $L_n$ is recurrent non null if there exists a non negative solution to the system. Let us consider the function $V$ such that $V : S \to [0, \infty)$ and define $d(i)$, known as a generalized drift in state $i$, as follows

$$d(i) = E[V(X_{n+1}) - V(X_n)]$$

$$d(i) = E[V(X_{n+1}) - V(X_n) | X_n = i]$$

$$= \sum p_{ij} (V(j) - V(i)).$$

provided the expectations exist. The quantity $d(i)$ is called drift if $V(i) = i \forall i \in S$. The theorem, called Foster's criterion, is given below:

Let $\{X_n, n \geq 0\}$ be an irreducible discrete time Markov chain (DTMC) defined on a countable state space $S$. If there exists a function $V : S \to [0, \infty)$ and $\epsilon > 0$ with a finite set $A \subset S$ such that

$$|d(i)| < \infty \text{ for } i \in A$$

$$|d(i)| < -\epsilon \text{ for } i \not\in A$$

the DTMC is positive recurrent.
**Pakes's Lemma:** In 1969, under the influence of Foster, A. G. Pakes derived the so called Pakes's Lemma, a result which is probably the most often used in establishing stability for a one dimensional Markov chain. Let \( \{X_n, n \geq 0\} \) be an irreducible discrete time Markov chain (DTMC) on state space \( S \) and define \( D(i) = E(X_{n+1} - X_n / X_n = i), i \in S \) then DTMC is positive recurrent if \( D(i) < \infty \forall i \) and \( \limsup_{n \to \infty} D(i) < 0 \).

**Burke's theorem:** In 1956, this theorem was established along with a proof by Paul J. Burke who studied the output from a queue and observed that the output from a Poissonian queue is also Poissonian. Let \( L(t) \) be a stochastic process whose sample paths are step functions with unit jumps. Let the points of increase be labeled as \( t_n + 0; n = 1, 2, \ldots \) and let the points of decrease be labeled as \( t_n - 0; n = 1, 2, \ldots \). Let \( L_n = L(t_n - 0) \) and \( L_n^+ = L(t_n + 0) \) where ' \( t_n - 0 \)' means the points immediately before ' \( t_n \)' and ' \( t_n + 0 \)' means the points immediately after \( t_n \). Then if either

\[
\Pi_k^+ = \lim_{n \to \infty} \Pr \{L_n^+ = k\} \quad \text{or} \quad \Pi_k^- = \lim_{n \to \infty} \Pr \{L_n^- = k\}
\]

exist, so does the other, and they are equal. In the application of this theorem to the queue size \( L(t) \) and points \( t_n \) correspond to the arrival and departure points respectively.

**Ergodicity property:** A stochastic process is ergodic when its time average converges to its ensemble average as time \( \to \infty \). When a stochastic process is ergodic,
estimates obtained using one long sample path have been found to be equally accurate as estimates obtained from a large number of shorter sample paths.

Ergodicity is associated with the problem of determination of measures of a stochastic process $X(t)$ from a single realization instead of a large number of realizations, as is often done in the analysis of simulation output. That is to say, the process $X(t)$ is ergodic in the most general sense if, with probability one, all its measures can be determined or well approximated from a single realization, $X_0(t)$ of the process. Since statistical measures of the process are usually expressed as time
averages, the Figure 1.5.1 reveals that \( X(t) \) is ergodic if time averages equal ensemble averages, that is, expected values. One may not always be interested in all, but sometimes only in certain measures (or moments) of a process. Then we can define ergodicity with respect to these moments, and a process might thus be ergodic for certain moments, but not for others. However, in queueing theory, we are typically interested in fully ergodic processes, that is, processes which are ergodic with respect to all moments.

If we are considering stationary processes then statistical averages will be time-independent; hence concern with respect to ergodicity centers only on convergence of time averages, e.g. see Karlin and Taylor (1975) and Heyman and Sobel (1982). In queueing theory, the processes of interest are in general not stationary, and thus our interest in ergodicity involves the convergence of both time and ensemble averages.

(10) **Stochastic decomposition property:** One of the most remarkable results that deals with the queueing systems with server vacation is “Stochastic Decomposition” result which allows the system behavior to be analyzed by considering separately the distribution of system size with no vacation and additional system size due to vacation. For an \( M/G/1 \) queue with server vacation, the queue length at a departure epoch could be decomposed into two random variables; one of which is the queue length at a departure epoch of the standard \( M/G/1 \) queue without vacation (a vacation-independent random variable) and the other one is the queue length at a random epoch when the server is on vacation (a vacation-related random variable).

For queues with Poisson input, the decomposition property has been investigated, among others, by Gaver (1962), Cooper (1970) and Fuhrmann (1984).
However, for generalized vacation this important result was first established by Fuhrmann and Cooper (1985) for $M/G/1$ type queues by using the embedded Markov chain technique where they showed that the queue size distribution can be decomposed into distributions of two or more than two random variables while Choudhury (2008) extended the result for two phases queueing system. Harris and Marchal (1988) extended Fuhrmann and Cooper's result to the state dependent case. Doshi [1990(b)] obtained this decomposition property for virtual workload process. Keilson and Servi (1990) discussed that the decomposition result holds for the distributional form of Little's law. Yang and Templeton (1987) have found that this property also takes place in retrial queues. Further, Yang et. al (1994) have developed an approximation method to determine the stationary distribution for $M/G/1$ retrial queues with general retrial times based on this property. Moreover, Artalejo and Falin (1994) studied this property for various retrial models and they cited that retrial queues can also be considered as a special type of vacation model in which the vacation begins at the end of each service time.

1.6 Types of queuing models

Queueing systems can be broadly classified into various categories depending upon their following aspects viz.

1.6.1 Depending upon the number of servers in the system

(a) Single server queue: When there is only one server to serve customers in the system, then it is known as single server queue. Under this queue, customers arrive at a service facility (single or in a group) to receive the desired service and attended by
only one server. Students’ arriving at a library counter is an example of a single server facility where customer arrivals and service times are random.

The single server queue is stable if mean service time is less than the mean inter arrival time. Giving stress on the problem of busy period Kendall (1951) had investigated the $M/G/1$ queueing model, where a single server serves jobs that arrive according to a Poisson process and have an arbitrary (general) service time distribution. Prabhu (1960) investigated the $M/G/1$ model with the help of combinatorial method. The single server $M/M/1$ queue having Poisson input and exponential service time is the simplest of the queueing models used in practice and various aspects of it has been studied by many authors, among others, by Bailey (1954), Conolly (1958), Takacs (1962), Cohen (1982), Neuts (1989), Sharma (1990) etc. Lindley (1952) studied the general single server queueing system $GI/G/1$ and obtained the waiting time distribution and derived its limiting form by using integral equation technique while Conolly (1958) established the time dependent solution of $GI/G/1$ queueing model by applying the supplementary variable technique. Takacs (1965) also studied $M/G/1$ model. Besides these, a large number of scientists had been working on various queueing models for single server. Some of the notable scholars are Cohen [1976, 1982], Rosenlund (1983), Pegden and Rosenshine (1982), Hubbard et. al. (1986), Abate and Whitt (1987, 1988), Takagi (1991) etc.

(b) Multi server queue: In a multi server queueing system there is more than one server to serve customers. Erlang (1917) was the originator who first introduced the multi server Markovian queue $M/D/s$ queue, with Poisson arrivals, deterministic service times and $s$ servers, along with the $M/M/s$ queue (Erlang delay model) and the $M/M/s/s$ queue (Erlang loss model) under the steady state. The multi server queue
M/M/s is the model, with Poisson arrivals, exponential service times and \( s \) servers providing service independently of each other, used most in analyzing service stations with more than one server such as banks, checkout counters in stores, check-in counters in airports, and the like. The transient state solution for multi server Markovian queues were investigated by Satty (1961), Jackson and Henderson (1967). The busy period distribution of multi server queue was initially observed by Chaudhry and Templeton (1973). Due to its numerous applications in different fields, many researchers like Sharfal and Parthasarathy (1989), Everitt and Downs (1984), Grassman (1983) and Kelton and Law (1985) etc. studied various aspects of multi server Markovian queue. Notable researchers who observed the non-Markovian multi server queues are Kendall (1953) and Carter and Cooper (1972), Kingman (1964), Loris-Teghem (1973), Rosenlund (1983) and De-Smit (1973).

(c) **Tandem queue:** A queueing system consisting of sequentially interconnected queues in which service facilities are located in series through which each customer must progress prior to leaving the system is known as the queues in tandem or series. Jackson (1954) was the pioneer who considered queues in series as a model specifying a queueing system for the overhaul of aircraft engines, where stages of overhaul involve sequential operations such as stripping, inspecting, repairing, assembling, and testing. Later some extension of Jackson's model had been developed by Reich (1957), Cohen (1956), Sacks (1960) and Burke (1976). Boxma (1986) outlined a review of results relating to (i) two queues in series and (ii) two parallel queues with a single server.
(d) **Network of queues:** Most real life queueing systems are usually networks that have been found to be useful in formulation and modeling of computer, communication and other such systems. In such a system we may have a group of nodes $k$ (say) where each node represents a service facility of some kind with, say, with $r_i$ servers at node $i$, $i = 1, 2, \ldots, k$. Customers can arrive from outside the system at any of the $k$ nodes for processing and may depart from the system from any node. In a queueing network, a group of servers operating from the same facility is identified as a node. There is no restriction as to which node customers may proceed to after completing service at a particular node, unlike in the case of the tandem queueing systems that has restrictions in its definition. Jackson (1957) considered a more general network of queues which is known as a Jackson network. Several aspects of network queues have been extensively investigated by Lemoine (1978), Kobyashi [1978, 1983], Gelenbe and Mitrani (1980), Geist and Trivedi (1982) etc.

1.6.2 **Depending upon various phases of service**

As the growth of queueing theory identifying major developments and directions in practical field expands gradually, various phases of queueing systems, where customers are served in various stages as required, have come into consideration. The method depending upon various phases of service was initially coined by Erlang himself while by using complex valued probabilities a generalization of the method was proposed by Cox [1955(b)]. Gaver (1954) has obtained an extended method of phases. Neuts (1975) has greatly extended these ideas by introducing the notion of
probability distributions of phase type, of which the Erlangian and hyper-exponential distributions are very special cases.

(a) **Single stage service:** In this system customers arrive at service facility for the desired service and attended by one or more servers but the server may provide a single stage of service to all arriving customers and their service will complete in that single stage of service.

(b) **Two phases of service:** In this model the service discipline involves two phases of heterogeneous service where the server provides a first phase of service (FPS) which is essential to all the arriving units. At the end of the first essential service a customer may either get the second phase of service or depart from the system. Recently, there have been several contributions, in which the server may provide a second phase service. Doshi (1991) studied a two phase queuing with general service times. Besides Doshi, Madan [2000(a,b)] studied an $M/G/1$ queue where a first phase of regular service (FRS) is provided to all the arriving units by the server, whereas only some of them received a second phase of service (SPS). Medhi (2002) generalized the model by incorporating a general distribution for SPS. Choudhury [2003(a)] investigated this model in depth and generalized for batch arrival case in [2003(b)].

(c) **Multistage service:** In this model the system provides more than two stages of service to a customer one after another. In general a multistage queue may be a complex network with feedback.

1.6.3 Depending upon service mechanism of the system:

Very often we come across the situations in queueing theory where a customer
may demand more than one service or different kinds of service from different servers and may be required to pass through different service channels. From these points of view, the following types of queueing models can be formulated.

1.6.3.1 Priority queues: Queue disciplines that assign priority service as one of the alternatives to FIFO service discipline are common in service systems. The priority may depend on factors such as customer class, the type of service, and even the length of service. In priority schemes customers with the highest priorities, regardless of their position in the line, are selected for service ahead of those with lower priorities. In priority situations, the priority accorded to a class of customers can be preemptive or non preemptive. In preemptive priority queue the arrival of a higher priority customer will preempt a lower priority customer for service while the lower priority customer is in service, the preempted lower customer being attended to only after serving all the units of higher priority classes. In non preemptive priority queue also known as head of the line queue, there is no interruption and the higher priority customer will enter service on the completion of the ongoing service at the time of its arrival.

1.6.3.2 Queues with vacation: Queueing system with server Vacation makes the queueing models more realistic and flexible in studying real-world queueing situations in which the server works on primary and secondary (vacation) customers and can be used to model production, communication and computer systems. As far as the primary customers are concerned, the server engaged in doing other work such as maintenance activities or servicing secondary customers is equivalent to the server taking a vacation and is unavailable to the primary customers over occasional periods

The service discipline considered in a vacation model can be broadly classified into two categories viz. (i) Exhaustive and (ii) Non exhaustive.

(i) **Exhaustive service discipline:** In this service we have assumed the situations where at the end of each busy period the system becomes empty and the server starts a vacation of random length of time. When the server returns from vacation and finds one or more customers waiting, he goes on serving continuously until there is no customer left in the system is called exhaustive service discipline. Therefore, the server in an exhaustive service system cannot take a vacation until the system becomes empty. We discuss briefly some exhaustive service discipline models.

(a) **Single vacation model:** Under the single vacation policy, the server takes only one vacation at the end of each busy period. After returning from the single vacation, the server either serves the waiting customer, if any, or waits for the arrival of a customer to start the next service. This type of model can be successfully applied in preventive maintenance of any production system. Consider a machine that can produce certain items undergoes preventive maintenance whenever required. The period of random lengths of preventive maintenance may be considered as period of server vacation. While the server is on vacation i.e. preventive maintenance of the machine is going on, any item arriving at the system will have to wait.

(b) **Multiple vacation models:** In the multiple vacation models the server keeps on serving the waiting customers until the system becomes empty and whenever no units left to be served in the system then he may take a vacation of random length of
time. If on return from a vacation, he finds at least one customer waiting, the server begins to serve the customer in the queue. Otherwise, the server takes a sequence of vacations for as long as the system is empty. Such type of model usually can be encountered again in production system in which a sequence of preventive maintenance is required to be performed. Whenever the system is left with no primary jobs, the server has to undertake a segment of the maintenance work. If on completion, the server finds some primary jobs present in the system then it will serve these jobs until the system is idle again. Otherwise a second maintenance work is done and so on. Levy and Yechiali [1975, 1976] were first to study such type of model for single unit arrival case while a generalization of the model for batch arrival case was given by Baba (1986), also see Borthakur and Choudhury (1997) for detailed analysis.

(c) Randomized vacation policy: As a generalization of multiple and single vacation policy model Takagi (1991) and Leung (1992) studied a model in which they first incorporated the notion of a randomized vacation policy (RVP), where the server takes at most \( M \) (say) numbers of random vacations. Whenever the system becomes empty, the server immediately goes for a vacation. Upon returning from a vacation, the server begins to serve the customer if the server finds at least one customer waiting in the queue to be served. Otherwise, if no customers are found waiting at the end of a vacation, the server either remains idle or leaves for another vacation. This pattern continues until the number of repeated vacations taken reaches \( M \). If no units are found by the end of the \( M \) -th vacation the server remains idle and waits for the arrival of a customer to start the next service. Choudhury [2002(b)] generalized this model for a batch arrival queueing system as a modification of multiple and single vacation model. Such a model could be appeared in control of queues. Later, Ke and Chu (2006)
successfully modeled this model and obtained optimal randomized policy under a suitable cost structure. More recently, Ke. et al [2010(b,c)] have made an extensive investigation on RVP queueing systems.

(d) **Unreliable server:** In most of the previous studies of queueing theory, it is assumed though seemed to be practically unrealistic that the server is available in the service station on a permanent basis and the service station never fails. But in real life situations, a remarkable and unavoidable situation in the service facility that a queueing system may experience is its break down at random time and hence the server will not be able to continue providing service unless the system is repaired. Queueing models with service breakdowns or some other kind of interruptions have been studied extensively by numerous researchers, among others, White and Christie (1958), Heathcote (1959), Keilson (1962), Gaver (1962), Avi-Itzhak and Naor (1963), Thirurengadan (1963), Mitrany and Avi-Itzhak (1968), Sengupta (1990), Ibe and Trivedi (1990), Li et al. [1997(b)], Tang (1997), Takin and Sengupta (1998), Madan [2003(a)], Li and Lin (2006), Fiems et al. (2008), Krishnamoorthy et al. (2009). Since, in many real-life situations, servers may well be subject to lengthy and unpredictable breakdowns while serving a customer therefore it may not be feasible to start the repairs immediately due to the non-availability of the server, in which case the system may also be turned off. Since the performance of systems like computer communication networks, flexible manufacturing systems, etc. may be heavily affected by the service station breakdown and delay in repair; these systems with a repairable service station are well worth investigating from both the queueing theory viewpoint and the reliability point of view.
(ii) **Non exhaustive Service**: Under non exhaustive service discipline the server may take a vacation even when there are some customers still left in the system without being served. Some specific types of non exhaustive service are discussed below briefly

(a) **Limited service vacation model**: In this model, there is a limitation on the amount of work performed to serve the customers waiting in the system during a given service period. Under this policy, the maximum number of customers that can be served before the server starts a vacation is limited to a fixed value of ‘m’ (say). This discipline allows the server to go for a vacation, regardless of the number of customers, as soon as the service limit is attained. Thus in this model, the server may go for a vacation if the system becomes empty or after serving ‘m’ consecutive customers from the beginning of a busy period. This type of service discipline was extensively investigated by Skinner (1967), Gelenbe and Mitrani (1980) and Lavenberg (1983) etc.

(b) **Gated service vacation system**: According to this principle when the server returns from a vacation, he accepts and starts servicing only those arbitrary numbers of customers $N$ (say) that are already present in the system at the beginning of the service period, other customers that arrive during a service period are served at the beginning of the next vacation period. This discipline looks like a gate that is closed when the $N$ units, that the server finds waiting upon return from a vacation, are kept inside the gate and others who arrive after the server begins service during the current service period will wait outside the gate to be served after the server vacation. More information on this class of model can be found in Takagi (1991) and in recent monograph of Tian and Zhang (2006).

(c) **Bernoulli vacation schedule**: Queues with server vacations under Bernoulli vacation schedule have emerged as an important area of queueing theory where after
completion of each service, the server may go for a vacation of random length \( V \). In this schedule the vacations follow a Bernoulli distribution, that is, the server may decide to go for a vacation with probability \( p \) \( (0 \leq p \leq 1) \) or may stay in the system and continues to provide service to the next customer, if any, with probability \( q = 1 - p \) after each customer service. If \( p = 1 \), this policy corresponds to the limited service system with multiple vacation model and \( p = 0 \) reduces to the exhaustive service system multiple vacation model. Keilson and Servi (1986) were initiators who incorporated the concept of Bernoulli scheduled server vacation.

1.6.3.3 Retrial queueing models: Retrial queues (or queues with repeated attempts) characterized by the feature that a customer who finds the server busy upon arrival is obliged to abandon the service area and repeat his demand in order to seek service again after a random amount of time called retrial time. Between trials, the blocked customer joins a group of unsatisfied customers called "orbit" or 'retrial group'. The general structure of a retrial queue is shown in Figure 1.6.3.

![Figure 1.6.3: General structure of a retrial queue](image)

It is obvious from the Figure 1.6.2 that retrial queues can also be considered as a special type of queueing networks. In the most general form these networks contain two nodes: the main node where blocking is possible and a delay node for repeated trials.
The early studies of retrial queues by Kosten (1947), Wilkinson (1956) and Cohen (1957) shows that these type of queues are suitable mathematical models for the modeling of subscribers’ behavior in telephone networks. Since the pioneering works published in 1950s, retrial queues have been widely used as powerful tool to provide stochastic modeling of many problems arising in telecommunication, computer networks and in daily life. Yang and Templeton, Falin, Kulkarni and Liang have given explicit survey on retrial queueing systems. A bibliographical information was provided by Arttalejo [1999(a), 1999(b)].Two specific monographs were recently appeared by Falin and Templeton (1997) and Arttalejo and Gomez-Corral [2008] where they proposed the fundamental methods and results of this topic.

1.6.3.3.1 Types of retrial models: Depending upon the response of the customer’s behavior in the orbit or retrial group various retrial policies have been studied. Here we define some retrial policy very briefly.

(i) Classical retrial policy: The pioneering studies of retrial queues present the concept of retrial time as an alternative to the classical models of telephone systems. In this context each blocked customer in the orbit generates a stream of repeated requests independently of each other, so that the intervals between successive repeated attempts are exponentially distributed with rate \( n\mu \) (say), depending on the orbit size \( n \). This type of retrial policy is known as classical retrial policy and was investigated by Yang and Templeton (1987) and Falin (1990).

(ii) Constant retrial policy: Recent applications to communication protocols and local area networks show that there are queueing situations with a constant retrial policy in which the retrial rate is independent of the number of customers (if any) in the


. orbit, i.e., the retrial rate is \((1 - \delta_{on})\alpha\), where \(\delta_y\) denotes Kronecker's delta. This constant retrial policy was initiated by Fayolle (1986), who investigated a telephone exchange model as an \(M/M/1\) retrial queue where the customers in the retrial group form a queue and only the customer at the head of the orbit queue can request service after exponentially distributed retrial time with rate \(\alpha\). Farahmand (1990) viewed this discipline a retrial queue with FCFS orbit.

(iii) Linear retrial policy: Artalejo and Gomez-Corral (1997) introduced the linear retrial policy by incorporating both possibilities by assuming that when there are \(n\) customers in the system the time intervals between successive repeated attempts are exponentially distributed random variable with parameter 
\[ \mu_n = \alpha(1 - \delta_{no}) + \eta \mu, \]
where \(\mu\) can be considered as the retrial rate per customer and \(\alpha\) the rate at which the server demands service for customer whenever he is idle and \(\delta_y\) denotes Kronecker's delta function. Such type of retrial policy is known as linear retrial policy.

(iv) Models with generalized retrial time: Retrial queueing systems with general retrial time have received considerable attention during last decade. Kapyrin (1977) introduced the concept of general retrial time for the first time by assuming that each customer in the orbit generates a stream of repeated attempts that are independent of the customer in the orbit and state of the server. However, this methodology was found to be incorrect by G. Falin in 1986. Subsequently, Yang et al. (1994) have developed an approximation method to obtain the steady state performance for the model of Kapyrin.
(v) **Batch arrival retrial queue**: In batch arrival retrial queues, it is assumed that at every arrival epoch a batch of $k$ primary units arrive with probability $p_k$. If the server is busy in the arrival epoch, then whole batch joins the orbit. Whereas if the server is free, then one of the arriving customer starts its service immediately and the rest joins the orbit. This type of retrial model was first incorporated by Falin (1976).

(vi) **Other advanced retrial queues**: The retrial literature is vast and rich so it is possible to find a great number of variants and generalizations including systems with priorities e.g. see Choi and Chang (1999), models with negative arrivals and disasters e.g. see Artalejo and Gomez Correl (1999), polling systems e.g. see Langaris (1999), overloading e.g. see Anisimov (1999), Bernoulli vacation with repeated attempts e.g. see Choudhury (2007), two phase service with unreliable server e.g. Choudhury and Deka (2008) etc.

1.6.3.3.2 **Comparison between Classical $M/G/1$ and $M/G/1$ retrial queueing systems:**

We now consider a single server queueing system having poison input process with rate $\lambda$ and general service time distribution with distribution function $B(x)$. If a primary customer finds the server free upon arrival, he immediately starts its service and leaves the service area described after being served. On the other hand, any customer who finds server busy is to leave the system, but retries his luck after an exponential time with parameter $\theta \geq 0$.

Here the system state at time $t$ can be by means of a bivariate process $X(t) = \{(C(t), N(t)) ; t \geq 0\}$, where $C(t)$ is the number of busy server and $N(t)$ is the
number of customers in the orbit (queue, in case of classical queueing system). Under the above assumptions we assume that $\rho = \lambda \beta, < 1$ (where $\beta$ is the first moment service time distribution), so that our queueing models are stable and the limiting probabilities

$$P_j = \lim_{t \to \infty} \Pr\{N(t) = j\} \quad \text{and} \quad Q_{i,j} = \lim_{t \to \infty} \Pr\{X(t) = (i,j)\}; \quad \text{for} \ i = 0,1 \ \text{and} \ j \geq 0$$

(where $i = 0$ denotes server is free and $i = 1$ denotes server is busy) exist and positive.

Then both sequences $\{P_j; j \geq 0\}$ and $\{Q_{i,j}; j \geq 0, i \in (0,1)\}$ can be computed recursively with the help of the following equations e.g. see Artalejo and Falin (2002)

$$P_0 = 1 - \rho, \quad P_i = \frac{1 - a_0}{a_0} P_0, \quad P_2 = \frac{1 - a_0 - a_1}{a_0} (P_0 + P_1)$$

and

$$P_{j+1} = \frac{1 - \sum_{i=0}^{j} a_i^*}{a_0} \sum_{i=0}^{j} P_i + \sum_{i=0}^{j} \sum_{k=j-i+2}^{j} \frac{a_k}{a_0}; \quad j \geq 2.$$

Similarly, we can compute for $j \geq 0$,

$$Q_{0,j} = \frac{\lambda}{\lambda + j \theta} \pi_j; \quad Q_{i,j} = \frac{(j+1)\theta}{\lambda} Q_{0,j};$$

$$\pi_0 = (1 - \rho) \exp \left\{ -\frac{\lambda}{\theta} \int_{0}^{\theta} \frac{1 - \beta^* (\lambda - \lambda u)}{\beta^* (\lambda - \lambda u) - u} \, du \right\},$$

and

$$\pi_j = \sum_{i=0}^{j} \pi_i \frac{\lambda}{\lambda + i \theta} a_{j-i} + \sum_{i=1}^{m} \frac{i \theta}{\lambda + i \theta} \pi_{j-i+1} \quad j \geq 0,$$

where $\beta^*(s)$ is the Laplace Stieltjes transform of the service time distribution

and $a_j = \int_{0}^{\theta} \frac{(\lambda x)^j}{j!} dB(x); \ j \geq 0$, is the probability that exactly $j$ customers arrive during the service time. In fact, the sequence $\{\pi_j; j \geq 0\}$ corresponds to the distribution.
of the embedded Markov chain at service completion epochs. The recursions provide an efficient method of computing $P_j, Q_{i,j}$ and $\pi_j$.

An alternative solution in terms of generating functions

$$P(z) = \sum_{j=0}^{\infty} z^j P_j, \quad Q_i(z) = \sum_{j=0}^{\infty} z^j Q_{i,j}; \quad i \in \{0,1\},$$

is given by

$$P(z) = \frac{(1-\rho)(1-z)\beta^*(\lambda-\lambda z)}{\beta^*(\lambda-\lambda z) - z} \quad (1.6.1)$$

$$Q_0(z) = (1-\rho) \exp \left\{ -\frac{\lambda}{\theta} \int_0^z \frac{1-\beta^*(\lambda-\lambda u) \, du}{\beta^*(\lambda-\lambda u) - u} \right\} \quad (1.6.2)$$

$$Q_1(z) = \frac{1-\beta^*(\lambda-\lambda z)}{\beta^*(\lambda-\lambda z) - z} Q_0(z) \quad (1.6.3)$$

Note that the solution for both standard and retrial queues are given in terms of the Laplace Stieltjes transform of the service times but the retrial model exhibits a more complex expression mainly due to integral arising in the right hand side of (1.6.2).

In particular, the corresponding expectations are given by

$$E[X] = \rho + \frac{\lambda \beta_z}{2(1-\rho)}, \quad E[C] = \rho \quad \text{and} \quad E[N] = \frac{\lambda^2}{(1-\rho)} \left\{ \frac{\beta_1 + \beta_z}{\theta} \right\}$$

Here $E[X]$ denotes the mean value of $X(t)$ as $t \to \infty$ and $\beta_z$ represents second moment of the service time distribution.
PART-I

SOME VACATION QUEUEING MODELS WITH TWO PHASES OF SERVICE SUBJECT TO SERVICE INTERRUPTION

Chapter II  A Poisson input queue with two phases of heterogeneous service under Bernoulli vacation schedule [52-82]

Chapter III  A batch arrival queueing system with two phases of Heterogeneous service and delayed repair under Bernoulli vacation schedule [83-117]

Chapter IV  A batch arrival unreliable server delaying repair queue with two phases of service and Bernoulli vacation under Multiple vacation policy [118-155]

Chapter V  A Poisson input queue with two phases of heterogeneous service and Bernoulli vacation schedule under Randomized vacation policy [156-191]

Chapter VI  A batch arrival queueing system with two phases of service and Bernoulli vacation schedule under Randomized vacation policy [192-233]