Chapter 1

INTRODUCTION

1.1 Queueing Theory

The study of the congestion in telephone network by A K Erlang lead to the origin of Queueing Theory. Over the years, the subject found its applications in diverse areas like Telecommunications, Traffic flow, Computer networking, Computing etc.. This is a branch of science that deals with the study of waiting lines. When customers/units requiring some kind of service gather at a service centre, a queue is formed. In Queueing Theory, we model such systems mathematically and predict some characteristics like average waiting time of a customer, average queue length etc..

The basic features that characterise a queueing system are the following:
Chapter 1. Introduction

a) **Arrival Pattern:** This describes the manner in which the units arrive and join the system. The customers may arrive in single or in batches. Time interval between any two consecutive arrivals is called the inter-arrival time. The arrival pattern is usually represented by the probability distribution of the inter-arrival time.

On arrival, if a customer sees a long queue, he may decide not to join the queue and may leave the station. This customer behaviour is called *balking*. Some customers join the queue, wait for a while but losing their patience, may leave the system without waiting further for service. This situation is referred to as *reneging*.

If there are more than one queue, customers have a tendency to switch from one queue to another. This is called *jockeying*.

b) **Service Pattern:** This indicates the manner in which the service is rendered. Like the arrivals, the service also is provided in single or in batches. The probability distribution of the service time describes the service pattern.

c) **Queue Discipline:** Queue Discipline tells us the rule by which the customers are taken for service. Some of the commonly used disciplines include *first in first out* (FIFO), *last in first out* (LIFO), *service in random order* (SIRO) and *server sharing*. In some systems customers may be given priorities so that the service is rendered in the order of their priorities.

d) **Number of service channels:** This refers to the number of servers providing service to the customers in the system.
1.2. Some basic concepts

1.2.1 Stochastic Process

A stochastic process or a random process is a collection

\[ \{X(t)/t \in T\} \]

of random variables where \( T \) is some index set. The index \( t \) is usually referred to as time. If the index set \( T \) is countable, then the process is called a discrete time process. Otherwise it is called a continuous time process. The set of all possible values of \( X(t) \) for each \( t \in T \) is the state space of the process.

1.2.2 Counting Process

A counting process \( \{N(t)/t \geq 0\} \) is a stochastic process if \( N(t) \) represents the total number of events occurred by time \( t \) such that

e) **Capacity of the system:** The capacity of the system is the maximum number of customers it can accommodate. It may be finite or infinite.

A queueing system is often analysed by modelling it as a Markov chain. Some basic concepts employed in this direction are given briefly in the following sections.
i) $N(t) \geq 0$.

ii) $N(t)$ is integral valued.

iii) If $s < t$, then $N(s) \leq N(t)$.

iv) For $s < t$, $N(t) - N(s)$ is the number of events occurred in the interval $(s, t]$.

A counting process $\{N(t)/t \geq 0\}$ is said to have independent increments if for all $t_1, t_2, ..., t_n$, $t_1 < t_2 < ... < t_n$, the random variables $N(t_2) - N(t_1), N(t_3) - N(t_2), ..., N(t_n) - N(t_{n-1})$ are independent. It has stationary increments if the distribution of $N(t) - N(s)$ depends only on $t - s$.

### 1.2.3 Markov Process

**Definition 1.2.1.** A stochastic process $\{X(t)/t \in T\}$ is called a Markov Process if

$$P[X(t_n) = x_n/X(t_{n-1}) = x_{n-1}, X(t_{n-2}) = x_{n-2}, ..., X(t_0) = x_0]$$

$$= P[X(t_n) = x_n/X(t_{n-1}) = x_{n-1}]$$

whenever $t_0 < t_1 < ... < t_{n-1} < t_n$ for every $n$.

A discrete time Markov Process is called a Markov chain. Thus a Markov chain is a stochastic process $\{X_n/n = 0, 1, 2, ...\}$.
for which

\[
P(X_n = x_n/X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) = P(X_n = x_n/X_{n-1} = x_{n-1})
\]

for \(x_0, x_1, \ldots, x_n\) in the state space.

The probability \(p_{i,j}^{n,n+1} = P(X_{n+1} = j/X_n = i)\) is called one step transition probability of the Markov chain. If the transition probabilities are independent of time, the chain is said to be homogeneous. If it depends on time, the chain is called non-homogeneous. For a homogeneous Markov chain, the one step transition probabilities are denoted by \(p_{i,j}\). The matrix \(P = [p_{i,j}]\) is called the one step transition matrix of the chain.

More generally the probability that starting from a state \(i\), the chain reaches the state \(j\) in exactly \(m\) transitions is called the \(m - \text{step transition probability}\). For a homogeneous chain this probability is denoted by \(p_{i,j}^{(m)}\). Thus

\[
p_{i,j}^{(m)} = P(X_{n+m} = j/X_n = i).
\]

A subset \(C\) of the state space of a chain is said to be closed if no state outside \(C\) can be reached from any state in \(C\). If the chain has no proper closed subset other than the state space itself, it is called an irreducible chain.

A state \(i\) is recurrent if and only if, starting from state \(i\),
the probability of returning to state $i$ after some finite time is certain. A non-recurrent state is said to be *transient*. For a recurrent state if the mean recurrence time is finite, it is called *positive recurrent*. The greatest common divisor of the recurrence times of a state is called its *period*.

If the period is one, the state is said to be *aperiodic*. A positive recurrent aperiodic state of a Markov chain is said to be *ergodic*. A Markov chain is ergodic if all its states are ergodic.

For a homogeneous Markov chain, the vector

$$\pi = (\pi_1, \pi_2, \pi_3, ...)$$

is called a *stationary probability vector* if for every $j$ in the state space,

$$\pi_j = \sum_i \pi_i p_{ij} \text{ such that } 0 \leq \pi_j \leq 1 \text{ and } \sum_j \pi_j = 1.$$ 

An irreducible chain has a stationary distribution if and only if all of its states are positive recurrent.

The probability vector $\varpi = (\varpi_1, \varpi_2, ...) \text{ is called the limiting distribution of the chain if } p_{ij}^{(n)} \to \varpi_j \text{ as } n \to \infty.$

If a positive recurrent chain is both irreducible and aperiodic, it has a limiting distribution.

**Theorem 1.2.1.** For an irreducible ergodic Markov chain, the limiting distribution exists and is same as its stationary distribution.
1.3 Modelling Tools

In this section we describe some tools we used in analysing the models introduced in this thesis.

1.3.1 Exponential Distribution

A random variable $X$ is said to have an exponential distribution with parameter $\lambda > 0$ if it has the probability density function

\[ f(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty \]

\[ = 0, \quad x < 0. \]

Its distribution function is given by $F(x) = 1 - e^{-\lambda x}, x \geq 0$.

The mean of this distribution is $\frac{1}{\lambda}$ and variance is $\frac{1}{\lambda^2}$. The moment generating function is $M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}$.

The important properties that make exponential distribution much useful in modelling queueing systems are the following:-

a) Memoryless property (Non-ageing property): This property implies that if $X$ denotes the duration of some activity, and if the the activity is still going on, the distribution of the duration of the remaining part of the activity is same as that of $X$, no matter when the activity has begun. In other words the remaining part of the activity can be treated
Chapter 1. Introduction

as a new activity.

\[ P(X \geq x + y | X \geq x) = P(X \geq y). \]

Exponential distribution is the only continuous distribution having this property.

b) **Minimum of two exponential variables is exponential:**

Let \( X_1 \) and \( X_2 \) be two exponential random variables with parameters \( \lambda_1 \) and \( \lambda_2 \) then \( \min(X_1, X_2) \) is exponential with parameter \( \lambda_1 + \lambda_2 \). Also \( P(X_i < X_j) = \frac{\lambda_i}{\lambda_i + \lambda_j} \).

### 1.3.2 Poisson Process

A counting process \( \{N(t)/t \geq 0\} \) is called a Poisson Process with rate \( \lambda > 0 \) if

i) \( N(0) = 0 \).

ii) It has stationary and independent increments.

iii) The distribution of \( N(t) \) is Poisson with mean \( \lambda t \).

\[ P(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n = 0, 1, 2, \ldots \]

A detailed description on Poisson process and related distributions is given in Medhi [53]. We state two important theorems.
Theorem 1.3.1. For a homogeneous Poisson process with mean \( \lambda t \), the inter-occurrence times are independently and identically distributed exponential random variables with mean \( \frac{1}{\lambda} \).

Theorem 1.3.2. If the interval between successive occurrences of an event \( E \) are independently and exponentially distributed with mean \( \frac{1}{\lambda} \), then the events \( E \) will form a Poisson process with mean \( \lambda t \).

1.3.3 Phase Type Distribution

Consider a Markov Process on the states \{1, 2, 3, ..., m, m+1\} with the infinitesimal generator

\[
Q = \begin{bmatrix}
T & T^0 \\
0 & 0
\end{bmatrix}
\]

where the \( m \times m \) matrix \( T \) satisfies \( T_{ii} < 0 \) for \( 1 \leq i \leq m \), and \( T_{ij} \geq 0 \), for \( i \neq j \). Also \( Te + T^0 = 0 \). Let initial probability vector of this process be \( (\alpha, \alpha_{m+1}) \) with \( \alpha e + \alpha_{m+1} = 1 \). Also assume that the states 1, 2, ..., \( m \) are transient so that absorption into the state \( m + 1 \) is certain.

Definition 1.3.1. A probability distribution \( F(.) \) on \([0, \infty)\) is said be a phase type distribution (PH-distribution) of order \( m \) with representation \( (\alpha, T) \) if and only if it is the distribution of the time until absorption of a finite Markov process defined in (1.1).

If \( F(.) \) is a phase type distribution described by the Markov
process defined in (1.1), then

\[ F(x) = 1 - \alpha.e^{Tx}, \quad x \geq 0. \]

For a PH distribution \( F(.) \) with representation \( (\alpha, T) \),

i) The distribution \( F(.) \) has a jump at \( x = 0 \) of magnitude \( \alpha_{m+1} \).

ii) The corresponding probability density function \( f(.) \) is given by

\[ f(x) = \alpha.e^{Tx}T^0, \quad x \geq 0. \]

iii) The Laplace-Stieltjes transform \( f(s) \) of \( F(.) \) is given by

\[ f(s) = \alpha_{m+1} + \alpha(sI - T)^{-1}T^0, \quad \text{for} \quad \text{Re}(s) \geq 0. \]

iv) The \( i^{th} \) raw moment \( \mu'_i \) is given by

\[ \mu'_i = (-1)^i i! \left( \alpha T^{-i} e \right), \quad i = 1, 2, 3, \ldots. \]

**Example 1.3.1. Erlang distribution**

A random variable \( X \) is said to follow an Erlang-\( k \) distribution, \( k = 1, 2, 3, \ldots \) if it has the probability density function

\[ f(x) = \frac{(\mu x)^{k-1}}{(k-1)!} \mu e^{-\mu x}. \]

The mean of this distribution is \( \frac{k}{\mu} \) and variance is \( \frac{k}{\mu^2} \). Its moment generating function is \( M_X(t) = (1 - \frac{t}{\mu})^{-k} \).
From the moment generating function it follows that the sum of $k$ mutually independent exponential random variables, each with common population mean $\frac{1}{\mu}$ is an Erlang-$k$ distribution with mean $\frac{k}{\mu}$.

Now consider a random variable $X$ with phase type probability distribution $F(.)$ represented by $(\alpha, T)$ where

\[
\alpha = (1, 0, 0, ..., 0)^{1\times m} \quad \text{and} \\
T = \begin{bmatrix}
-\mu & \mu & 0 & \ldots & 0 \\
0 & -\mu & \mu & 0 & \ldots & 0 \\
0 & 0 & -\mu & \mu & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & -\mu & \mu \\
0 & 0 & \ldots & 0 & -\mu & \mu \\
\end{bmatrix}^m
\]

Since the corresponding Markov process always start from the first phase, the time until absorption, $X$ is the sum of time spent in each of the $m$ phases. Hence $X$ is the sum of $m$ exponentially distributed random variables with mean $\frac{1}{\mu}$. That is the distribution of $X$ is Erlang-$m$. Thus Erlang distribution is a phase type distribution.

**Example 1.3.2. Exponential distribution**

When $k = 1$ the Erlang distribution reduces to Exponential distribution. Hence Exponential distribution can be considered as an Erlang-1 distribution. Therefore exponential distribution is a phase type distribution with a single phase.
1.4 Matrix Analytic Methods

When Queueing theory found its applications in several new areas like computer networking, mobile phone communications etc., the usual methods like Method of generating functions, Methods using Transforms etc. failed to provide much tractability in the analysis of many queueing models especially when the distribution of inter-arrival time or service time is not exponential. But the introduction of Matrix analytic methods gave us the ability to analyse much complicated Stochastic models in an algorithmic way and to numerically explore the problems more deeply. In this thesis, the Matrix analytic methods are used to analyse quasi-birth-and-death processes.

1.4.1 Level independent quasi-birth-and-death processes

Consider a Markov process with state space

\[ E = \{(i, j), i \geq 0, 1 \leq j \leq m\}. \]

We partition the state space as

\[ E = \bigcup_i E_i \text{ where } E_i = \{(i, j), 1 \leq j \leq m\}. \]

The states in \( E_i \) are said to be in level \( i \). Such a Markov process is called a level independent quasi-birth-and-death process.
(LIQBD) if its infinitesimal generator is the irreducible tridiagonal matrix $Q$ given by

$$Q = \begin{bmatrix}
B_0 & A_0 & & \\
B_1 & A_1 & A_0 & & \\
& A_2 & A_1 & A_0 & \\
& & A_2 & A_1 & A_0 \\
& & & \cdots & \cdots & \cdots
\end{bmatrix}.$$ 

Then we have the following theorem (Neuts [49]).

**Theorem 1.4.1.** The process $Q$ is positive recurrent if and only if, the minimal non-negative solution $R$ to the matrix quadratic equation

$$R^2 A_2 + RA_1 + A_0 = 0$$

has spectral radius less than one and the finite system of equations

$$x_0 (B_0 + RB_1) = 0$$
$$x_0 (I - R)^{-1} e = 1$$

has a unique positive solution for $x_0$. If the matrix $A = A_0 + A_1 + A_2$ is irreducible, then $\text{sp}(R) < 1$ if and only if, $\pi A_0 e < \pi A_2 e$, where $\pi$ is the stationary probability vector of the matrix $A$. The stationary probability vector $x = (x_0, x_1, x_2, \ldots)$ of $Q$ is given by

$$x_i = x_0 R^i, \ i \geq 0.$$
To find the solution $R$ of equation (1.1), we use the iterative formula

$$R_n = -A_0 (A_1 + R_{n-1}A_2)^{-1}, n = 1, 2, 3, .... \quad (1.4)$$

with an initial approximation $R_0$. If $sp(R) < 1$ then $R_n$ converges to $R$. More powerful iterative methods can be found in Latouche and Ramaswami [50].

### 1.4.2 Level dependent quasi-birth-and-death processes

A Level dependent quasi-birth-and-death process (LDQBD) is a Markov process with state space

$$E = \{ (i, j) / i \geq 0, 1 \leq j \leq n_i \}$$

whose infinitesimal generator $Q$ is given by

$$Q = \begin{bmatrix}
A_{1,0} & A_{0,0} \\
A_{2,1} & A_{1,1} & A_{0,1} \\
& A_{2,2} & A_{1,2} & A_{0,2} \\
& & A_{2,3} & A_{1,3} & A_{0,3} \\
& & & \cdots & \cdots & \cdots \\
& & & & \cdots & \cdots & \cdots \\
& & & & & \cdots & \cdots & \cdots 
\end{bmatrix}.$$  

The state space is partitioned into different levels where level $i$ is given by $E_i = \{ (i, j) / 1 \leq j \leq n_i \}, i = 0, 1, 2, ...$. The tran-
sitions are to the adjacent levels alone. But the transition rate will depend on the level in which the process is then. Assuming that the process is irreducible, we have the following theorems (Latouche and Ramaswami [50]).

**Theorem 1.4.2.** If an LDQBD is aperiodic and positive recurrent, its limiting probability vector \( \pi = \{\pi_1, \pi_2, \pi_3, \ldots\} \) satisfies the relation

\[
\pi_n = \pi_{n-1} R_n, \quad n \geq 1
\]

where the matrices \( R_n \) are the minimal non-negative solutions of the system of equations

\[
R_n R_{n+1} A_{2,n+1} + R_n A_{1,n} + A_{0,n} = 0.
\]

**Theorem 1.4.3.** The LDQBD is positive recurrent if and only if there exists a strictly positive solution of the system

\[
\pi_0 = \pi_0 \left( A_{1,0} + R_1 A_{2,1} \right)
\]

normalized by

\[
\pi_0 \sum_{n \geq 0} \left( \prod_{1 \leq k \leq n} R_k \right) e = 1.
\]

To calculate \( R_n \) we use Neuts - Rao Truncation method (Neuts and Rao [51]) for retrial queues. In this method an upper level \( N \) is selected such that the transitions between the levels higher than \( N \) are independent of the level using the approxi-
A_{2,i} = A_{2,N}, A_{1,i} = A_{1,N} \text{ and } A_{0,i} = A_{0,N} \text{ for } i > N.

For retrial queues this makes sense since if the number of retrying customers are very large, most of the retrials fail. So the retrials exceeding a large number $N$ will have no effect on the system. Then $R_N$ is the the minimal non-negative solution of the equation

$$R_N^2 A_{2,N} + R_N A_{1,N} + A_{0,N} = 0.$$ 

Using this $R_N$ we can find the steady state vector $\pi_N$ which converges to $\pi$ as $N \to \infty$.

In more general cases we choose the method proposed by Bright and Taylor [52].

### 1.5 Summary of the thesis

In our day to day life, in processes like banking, internet, business, agriculture, scientific experiments etc., we are faced with different kinds of interruptions like a power failure affecting a banking procedure or the working of certain machinery. Though the facilities are improving/increasing each day, which reduce the severity of interruptions, increasing needs bring them back.
For an example consider the following situations. Suppose we have a computer with a backup of one hour. This backup is sufficient for many purposes like saving a work that got interrupted due to power failure. However, suppose that we have to run certain program, which requires more than one hour for execution on this computer. While running such a program, we are worried about a power failure. This example applies to many other situations, as the world is run by computers.

Another example from day to day life: Consider a work person who uses a power tool that runs on electricity, who has more than one work sites to attend. If a power failure lasts for long or repeats randomly on one work site, he may choose to shift to another site with constant power supply.

An emergency call for an ambulance service getting interrupted due to network issues is yet another example for interruptions in day to day life.

An ideal world, as one would expect, be that which is free from all types of interruptions. However, as we discussed earlier, we do not belong there yet. Facing the reality of interruption, we hope to minimize its severity by introducing some protection mechanism. Analysis of queueing models with service interruption and protection is therefore important.

Naturally, researchers got involved in modelling these scenarios in a queueing theory perspective. A brief review on the related works is added in the beginning of each chapter.
abundance of works in this direction tells the practical importance. This is the reason for the selection of a few systems in which the service process is susceptible for interruption and having some measures for immunizing the server from it in this thesis. The thesis is arranged in 6 chapters including the present introductory chapter.

In Chapter 2, we consider a single server queueing system where the service time distribution is phase type. The service process may face some interruptions during the service. The interruption occurs according to a Poisson process. Interruptions are assumed to occur only when a service is in progress and not when the server is idle. The interrupted service is either resumed or repeated based on which of the two renewal processes, started simultaneously with the interruption, renews first. Customers arrive according to a Poisson process with different means depending on whether the server is interrupted or not. The customers waiting in the queue for service may leave the system without waiting further for service while the server is interrupted. Stability of the above system is analysed and steady state vector is calculated using Neuts-Rao truncation. A thorough numerical study of various performance measures such as mean and variance of waiting time of a customer are carried out.

Chapter 3 analyses a single server retrial queueing model with service interruptions, resumption/repeat of interrupted service. On arrival if a customer finds an idle server, he is immediately taken for service. If the server is busy when a customer
arrives, this customer goes to an orbit of infinite capacity from where he makes repeated attempts for service according to a Poisson process. After an unsuccessful retrial he rejoins the orbit with probability $p$ or leaves the system without waiting for service with probability $q = 1 - p$. The service time durations follow PH distribution when there is no interruption. The service process is subject to interruptions, which occur according to a Poisson Process. The interrupted service is either resumed or repeated as in the model described in chapter 2. The system is found to be always stable if $q > 0$. The case $q = 0$ is also analysed. Using Matrix analytic method, expressions for important system characteristics such as expected service time, expected number of interruptions etc. are obtained. System performance measures are numerically explored and the effect of service interruptions in a retrial set up is studied.

Chapter 4 is devoted to a model in which service time distribution is Erlang of order $m$. The server is subjected to interruption. The arrival of customers as well as the occurrence of interruptions is according to a Poisson Process. The interrupted server is taken for repair immediately. The repair time follows exponential distribution. The interrupted service is either resumed or restarted after repair according to the time taken for repair to be done. As a means to reduce the impact of interruption, a protective mechanism is employed. To reduce the chance of a service reaching completion being restarted all over again, the final $n$ phases of service are immunized from interruption. Thus a service that completed the first $m - n$ phases
will no longer face any interruption. The condition for stability is determined and the service process is thoroughly analysed. The steady state probabilities are evaluated using Matrix Analytic method. Many system performance measures like expected number of customers in the system, expected waiting time, expected number of interruptions during a service, expected interruption duration etc. are also investigated. A cost analysis is done numerically to find the number of phases to be protected at an optimum cost.

In Chapter 5, we consider a single server queueing system where customers arrive according to a Poisson process. Service time distribution is exponential. The service process is subject to interruptions, which occurs according to a Poisson process. We assume that during interruption, the customer being served waits there until his service is completed. The interrupted service is restarted after repair. Repair time is exponentially distributed. To minimize the loss due to the interruptions, some protection is given to the server. There will be no interruption if the server is in protected mode. But the way in which the server is protected differs from the method adopted in the previous chapter. Here the server is brought to the protected mode after a random time from the start of the service. Stability of the above system is analysed and steady state vector is calculated. Explicit formulas for system performance measures such as expected number of customers in the system, expected interruption rate, waiting time of a customer in the system etc. are also obtained. A cost analysis is also done numerically to
find the time after which the service has to be protected at an optimum cost.

In Chapter 6, a system similar to one discussed in the previous chapter is analysed, but in this model the service time has Erlang-$m$ distribution. The strategy used to protect the service is the same as the one used in chapter 5. The condition for stability of the system is obtained. The service process is well studied and important system performance measures are evaluated. A cost analysis is made to determine the optimum time at which the protection is to be started in a cost effective manner. A comparison between the two strategies of protection is also done.

It may be noted that the protection mechanism introduced in chapter 4 looks similar to the N-policy in queueing system. In contrast those introduced in chapters 5 and 6 are similar to the T-Policy.
Chapter 1. Introduction