Chapter 1

Introduction

Process is a phenomenon that takes place in time. In many practical situations, the result of a process at any time may not be certain. Such a process is called a stochastic process. As uncertainties lead to random variables, stochastic process requires a probabilistic setting. So many real world phenomena can be modelled and analysed by using the theory of stochastic processes where deterministic laws fail. This is called stochastic modelling. One of the most important part in stochastic modelling is the field of Queueing Theory.

1.1 Stochastic process

A stochastic process is a family of random variables and each random variable is a function of a parameter say time $t$. It is denoted by $\{X(t), t \in T\}$
where $T$ is an index set. The set of possible values of the random variable $X(t)$ is called its state space and a value of the random variable $X(t)$ is called a state. According to the nature of the state space and index set, a stochastic process is classified into four categories. (i) discrete time - discrete state space (ii) discrete time - continuous state space. (iii) continuous time - discrete state space (iv) continuous time - continuous state space. If the state space is discrete, then the stochastic process may be called as a chain, otherwise or in general we use the word process.

1.1.1 Poisson process

A continuous time stochastic process $\{X(t) : t \in T, T = [0, \infty)\}$ is called a Poisson process with parameter $\lambda$ if and only if it satisfies the following conditions. (i) $X(0) = 0$ (ii) the increments $X(s_i + t_i) - X(s_i)$, over an arbitrary finite set of disjoint intervals $(s_i, s_i + t_i)$ are independent random variables. (iii) for each $s \geq 0, t \geq 0, X(t) = X(s + t) - X(s) = n$ has the Poisson distribution $\frac{e^{-\lambda t}(\lambda t)^n}{n!}$ with mean $\lambda t$.

1.1.2 Markov process

A stochastic process is said to be Markov if $P(a < X_t \leq b | X_{t_1} = x_1, ..., X_{t_n} = x_n) = P(a < X_t \leq b | X_{t_n} = x_n)$ whenever $t_1 < t_2 < ... < t_n < t$. That is, it is a process with the property that the given the value of $X_t$, the values of $X_s, s > t$ do not depend on the values of $X_u, u < t$. That is the probability of any particular future behaviour of the process, when its present state is known exactly, is not altered by additional knowledge concerning
its past behaviour. A Markov process having a finite or denumerable state space is called a *Markov chain*. If the time is discrete, the MC is called discrete time Markov chain (DTMC) and if the time is continuous, it is called continuous time Markov chain (CTMC).

### 1.2 Queueing Theory

Queue is a waiting line. The imperfect matching between the customers and service facilities creates queues. Queueing theory originated as a very practical subject. It has many applications in telecommunications. The earliest works studied the telephone traffic congestion. The first work related to this is “The Theory of Probabilities and Telephone Conversations” which was published by A.K. Erlang in 1909. The theory of queues is applied in many other practical situations of traffic, internet, facility designs like banks, amusement parks, fastfood restaurants, hospitals and post offices.

#### 1.2.1 Queueing system

A system having arrivals, service facilities, and departures is called a queueing system. A diagrammatic representation of the queueing system is given in Figure 1.1. For a complete description of a queueing system, we consider the following characteristics.

1. **Arrival pattern of customers**
In a queueing system, time between two successive customer arrival is called \textit{inter arrival time}. Practically it is stochastic. So it is necessary to know its probability distribution. Also customers may arrive in batch or bulk. So the size of a batch is another random variable and it is necessary to identify its probability distribution. \textit{Customer behaviour} is another important one. A customer may decide not to enter the queue by seeing the queue too long. Then the customer is said to have \textit{balked}. Some customers after joining the queue, wait for some time, and leave the service system due to intolerable delay. In this case the customer is said to have \textit{reneged}. In case of parallel waiting lines, customers may move from one queue to another hoping to receive service more quickly. This is called \textit{jockeying}. An arrival pattern that does not change with time is called a \textit{stationary arrival pattern}. One that is not time independent is called
nonstationary.

2. Service pattern

Just like Arrival pattern, a probability distribution function is needed to describe service times. Service may be in single or in batch. Service process which depend on the number of customers waiting is called state dependent service. Like arrivals, service can be stationary or nonstationary with respect to time. If the system has no customers, then the server becomes idle. Then the server may leave the system for vacation with a random period of time. These vacations may be used by the server for doing other jobs. Server vacation period may be limited by some control policies like $N$, $D$, and $T$. An idle server starts service only when $N$ are present in the queue and once he starts serving, goes on serving till the system size become zero. This is called $N$-policy. Under $T$-policy, the service facility is turned off for a fixed period of time $T$, from the instant of each service completion leaving the system empty. According to $D$-policy, the service facility re-opens as soon as the total workload exceeds a critical level $D$.

Also interruption may take place while a service is going on, for reasons like server breakdown or the arrival of a high priority customer. On completion of interruption, the interrupted work may be resumed or repeated. Also working vacations and vacation interruptions are recent concepts in service pattern. In working vacation, a customer is served at a lower rate rather than completely stopping the service during a vacation. But in vacation interruption policy, the server will come back from the vacation without completing it.
3. Queue discipline

Queue discipline refers to the manner in which customers are selected for service when a queue has formed. The most common Queue discipline is first in-first out (FIFO) [or first come - first served(FCFS)]. Some others in common usage are last in-first out (LIFO) [or last come-first served(LCFS)], service in random order(SIRO), Service in priority(SIP). There are two general situations in priority disciplines: preemptive priority and non-preemptive priority. When the high priority customer enters the system, the low priority customer in service is preempted. This is called preemptive priority. But in non-preemptive priority, the high priority customer goes to the head of the queue but cannot get in to the service until the customer presently in service is served completely, even though this customer has low priority.

4. System capacity

Physical space of the waiting room is called system capacity. It may be finite or infinite. In the case of finite system capacity, customers may be forced to balk if the capacity is full. In such situations, work may be postponed. In this thesis, it is discussed in great detail.

5. Number of servers

There are queueing systems with number of parallel service stations, which can serve customers simultaneously. So the number of servers is essential to describe a queueing system.

6. Stages of service
There may be only one stage of service or may have several stages. In the case of several stages, a customer may not pass through all stages.

1.2.2 Notation of a queueing system

In the development of a queueing system, a notation has evolved to describe its essential characteristics, called Kendall-Lee notation. Here we notate a queueing system by $a/b/c/d/e$ where $a$ denotes the arrival pattern, $b$ the service pattern, $c$ the number of servers, $d$ the system capacity, $e$ the queue discipline. If the queueing system is represented by $a/b/c$, then it is understood that the system capacity is infinite and the queue discipline is FCFS. However this notation is not sufficient to describe the whole characteristics of modern queueing systems.

1.2.3 Analysis of queueing models

Queueing models can be classified in to *Markovian* and *non-Markovian* models. If the inter arrival time of customers and service times are exponentially distributed, then the queueing model is called Markovian queueing model. Queueing models with inter arrival times and/or service times which are not exponential distributions are called non-Markovian queueing models. Matrix geometric method developed by Neuts is useful for analysing complicated queueing models in steady state.

$M/M/1$ queue in continuous time is a simple Markovian birth-death queueing model. Let $1/\lambda$ be the mean inter arrival time and $1/\mu$ be the mean service time. Then to analyse its steady state behaviour, we first
form the steady state probability distribution $P_n$ for the system to have $n$ units, by using difference-differential equation method. Then $P_n = \rho^n(1 - \rho)$ if $\rho < 1$ where $\rho = \lambda/\mu$ called traffic intensity. It is a geometric distribution. If the system capacity is finite, we have the $M/M/1/K$ model, where $P_n = \frac{(1-\rho)\rho^n}{1-\rho K+1}$ if $\rho \neq 1$ and $P_n = \frac{1}{K+1}$ if $\rho = 1$ for $n > 0$ and $P_0 = \frac{1-\rho}{1-\rho K+1}$ if $\rho \neq 1$ and $P_0 = \frac{1}{K+1}$ if $\rho = 1$. This geometric nature of the solutions are the main motivating fact to the introduction of matrix geometric solutions for extended models.

### 1.3 Matrix analytic methods

Most of the modern queueing problems are difficult to analyse by making difference-differential equations and solving by the method of generating functions and Laplace transforms. This difficulty can be overcome by using Matrix analytic methods introduced by Neuts (see [45]). Here usual birth-death process can be extended to quasi-birth-death (QBD) process. If the tridiagonal elements in the intensity matrix of a birth-death process are matrices, then such a process is called a QBD process. Here the state space consists of states of the form $(i, j, \ldots)$. The first dimension is called the level of the process, while the other dimensions are called phases. The transitions are restricted to the same level or to the two adjacent levels. Thus it is possible only to move from $(n, j)$ to $(m, k)$ in one step if $m = n + 1$, $n$ or $n - 1$ for $n \geq 1$ and $m = 0, 1$ for $n = 0$. If the transitions rates are level independent, the resulting QBD process is called level independent quasi-birth-death process (LIQBD).
1.3. Matrix analytic methods

Let the infinitesimal generator of a LIQBD process be

\[
Q = \begin{bmatrix}
B_1 & B_0 \\
B_2 & A_1 & A_0 \\
& A_2 & A_1 & A_0 \\
& & A_2 & A_1 & A_0 \\
& & & \ddots & \ddots & \ddots \\
& & & & \ddots & \ddots & \ddots
\end{bmatrix}
\]

where the matrices \(B_0, B_2, A_0, A_2\) are non negative and the matrices \(B_1\) and \(A_1\) have non negative off diagonal elements but strictly negative diagonal elements. The row sums of \(Q\) are necessarily equal to zero.

Let \(x\) be a stationary vector. Then \(xQ = 0\) and \(xe = 1\) where \(e\) is a column vector of ones of appropriate order. Let \(x\) be partitioned by the levels into subvectors \(x_i\) for \(i \geq 0\). Then \(x_i\) has the matrix geometric form \(x_i = x_1R^{i-1}\) for \(i \geq 2\) where \(R\) is the minimal non negative solution to the matrix equation \(A_0 + RA_1 + R^2A_2 = 0\) and the vectors \(x_0, x_1\) are obtained by solving the equations \(x_0B_1 + x_1B_2 = 0\) and \(x_0B_0 + x_1(A_1 + RA_2) = 0\) subject to the normalising condition \(x_0e + x_1(I - R)^{-1}e = 1\).

For the existence of stationary solution, spectral radius of \(R\); \(sp(R) < 1\), which is analogous to the condition \(\rho < 1\) in familiar \(M/M/1\) queueing model. \(R\) is called rate matrix. Once \(R\) is determined, the geometric nature of the solution is established. If the matrix \(A = A_0 + A_1 + A_2\) is irreducible, then \(sp(R) < 1\) if \(\pi A_0 e < \pi A_2 e\) where \(\pi\) is the stationary probability vector of the generator matrix \(A\). That is \(\pi\) is the solution of \(\pi A = 0\) and \(\pi e = 1\). One can use the iterative formula \(R = -A_0(A_1 + R_{n-1}A_2)^{-1}\) for \(n \geq 1\) with an initial value \(R_0\) which converges to \(R\) if \(sp(R) < 1\).
The following are some distributions frequently used in queueing theory.

1. Exponential and Geometric distribution

Consider a Poisson process \( \{N(t), t \geq 0\} \) with parameter \( \lambda \) where \( N(t) \) represents total number of arrivals in an interval of duration \( t \). Then the time between two successive arrivals will follow exponential distribution with probability density function \( f(x) = \lambda e^{-\lambda x}, \ x \geq 0 \). The distribution function is \( F(x) = 1 - e^{-\lambda x} \). The exponential distribution is the only one continuous distribution which exhibits Markovian property. This property states that the probability that a customer currently in service has \( t \) units of remaining service is independent of how long it has already been in service. That is, \( P[T \leq t_1 | T \geq t_0] = P[0 \leq T \leq t_1 - t_0] \). The only other distribution to exhibit this property is the geometric distribution which is the discrete analogue of the exponential distribution. Probability density function of geometric distribution is \( f(x) = pq^x \) where \( 0 < p < 1 \) and \( q = 1 - p \).

2. Continuous time phase type distribution

Let \( \{X_t, t \geq 0\} \) be a finite MC with statespace \( \{1, 2, \ldots, m+1\} \) and generator \( Q = \begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix} \). Here 1, 2, ..., \( m \) are transient states and \( m + 1 \) or 0 is absorbing state. \( T \) is a square matrix of order \( m \) satisfying \( T_{ii} < 0 \) for \( 1 \leq i \leq m \) and \( T_{ij} > 0 \) for \( i \neq j \). Also \( Te + T^0 = 0 \) where \( e \) is a column vector of ones of order \( m \). Let the initial probability vector be \( (\alpha, \alpha_{m+1}) \) with \( \alpha \) a row vector of dimension \( m \), so that \( \alpha e + \alpha_{m+1} = 1 \). Let \( Z = \inf[t \geq 0, X_t = m + 1] \) be the random variable of the time until absorption in state \( m + 1 \). The distribution of \( Z \) is called phase
1.3. Matrix analytic methods

Type distribution. We denote it by $PH(\alpha, T)$. The dimension $m$ of $T$ is called the order of the phase type distribution. The states $1, 2, \ldots m$ are called phases. If $Z$ follows $PH(\alpha, T)$, the distribution function of $Z$ is given by $F(t) = P(Z \leq t) = 1 - \alpha . \exp(Tt).e$, $\forall t \geq 0$ and the density function is $f(t) = \alpha . \exp(Tt).T^0$, $\forall t \geq 0$. It is possible to approximate any distribution on the non-negative real numbers by a PH-distribution. Moments of $Z$ are given by $E(Z^n) = (-1)^n n! \alpha T^{-n} e$, $\forall n \in N$. So $E(Z) = -\alpha T^{-1} e$ is the mean time to absorption.

3. Discrete time phase type distribution

Let $Z = \text{min}[n \in N_0, X_n = m + 1]$ denote the time until absorption in the state $m + 1$. The transition probability matrix (TPM) has the form $P = \begin{bmatrix} T & T^0 \\ 0 & 1 \end{bmatrix}$ where $T$ is a square matrix of order $m$ such that $I - T$ is non-singular and $Te + T^0 = e$. The distribution of $Z$ is called a discrete PH-distribution. $P(Z = n) = \alpha T^{n-1} T^0$ and $P(Z \leq n) = 1 - \alpha T^n e$, $\forall n \in N$. The mean time to absorption is given by $E(Z) = \alpha(I - T)^{-1} e$.

4. Erlang distribution

An Erlang distribution with $n$ degrees of freedom (or stages) and parameter $\lambda$ is the distribution of the sum of $n$ exponential random variables with parameter $\lambda$. It has the density function $f(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}$, $\forall t \geq 0$. It can be represented as the holding time in the transient state set $\{1, 2, \ldots n\}$ of a MC with absorbing state $n + 1$ where the only possible transitions occur from a state $k$ to the next state $k + 1$ ($k = 1, 2, \ldots n$) with rate $\lambda$ each. This can be approximated to PH distribution with
Chapter 1. Introduction

\[ \alpha = (1, 0, 0, \ldots, 0), \]

\[
T = \begin{bmatrix}
-\lambda & \lambda \\
-\lambda & \lambda \\
& \ddots & \ddots \\
& & -\lambda & \lambda \\\n& & & -\lambda
\end{bmatrix},
T^0 = \begin{bmatrix}
0 \\
0 \\
\vdots \\
\lambda
\end{bmatrix}.
\]

In the discrete case, general Erlang distribution with \( m \) stages is approximated to PH distribution with \( \alpha = (1, 0, 0, \ldots, 0) \),

\[
T = \begin{bmatrix}
s_{11} & s_{12} \\
s_{22} & s_{23} \\
& \ddots & \ddots \\
& & s_{mm}
\end{bmatrix},
T^0 = \begin{bmatrix}
0 \\
0 \\
\vdots \\
s_m0
\end{bmatrix}.
\]

where \( Te + T^0 = e \).

1.4 Summary of the thesis

In many real life situations a work may be postponed for several reasons. It may be to attend a more important job or to go on vacation. To bring the postponed work in the usual service track, we introduce \( N \)-policy. Considering the physical limitation of a system, finite capacity queues are more realistic than infinite capacity queues. But this will result in overflow of jobs and make considerable loss to the system. Models with
postponement are an alternative to finite capacity queues to minimise such a loss.

A paper to deal with postponed work was introduced by Deepak et.al. (see [16]). They analysed such a system in the stationary case and provided a number of system performance measures. No further development in this is reported so far. Nevertheless this notion of postponement of work has been introduced into inventory by a few researchers (see Krishnamoorthy and Islam [36], Arivarignan et.al. [2], Paul Manuel et.al. [46], Sivakumar and Arivarignan [48]).

The thesis entitled “Queues with postponed work under N-policy” is divided in to 6 chapters.

Chapter 1 is an introductory chapter containing basic definitions and terminologies of stochastic process, queueing theory and matrix analytic methods. We provide a brief of the work done so far in queues with postponed work; some associated work is reviewed in this chapter.

Chapter 2 describes an $M/PH/1$ queue with postponed work under $N$-policy. Here we extend the model described in [16] by introducing $N$-policy for transfer of customers from the pool. When a buffer having capacity $K$ is full, newly arriving jobs are not necessarily lost. They can accept the offer of joining a pool of postponed work having infinite capacity with probability $\gamma$. With probability $1 - \gamma$, such customers do not join the system. When at the end of a service, if there are postponed customers, the system operates as follows. If the buffer is empty, the one ahead of all waiting in the pool gets transferred to the buffer for immediate service. If the buffer contains $y$ jobs, where $1 \leq y \leq L - 1$; $2 \leq L \leq K - 1$ at a
service completion epoch, then again the job at the head of the buffer starts service and with probability $p$, the head of the queue in pool is transferred to the finite buffer and positioned as the last among the waiting customers in the buffer. With probability $q = 1 - p$, no such transfer takes place.

$N$-policy ensures an early service for pooled customers. If the pool contains at least one postponed work, continuously served customers from the buffer since the last transfer under $N$-policy, is counted at each service completion epoch. When it reaches a pre-assigned number $N$, then the one ahead of all waiting in the pool gets transferred to the buffer for immediate service. The $N$-policy introduced here differs from the classical $N$-policy as explained below. In the classical case, $N$ customers are to queue up to start the new service cycle once the system becomes empty. However in the present case $N$-policy is applied to determine a priority service to be given to a customer from the pool.

A stability condition based on first passage time probability and stationary distribution has been obtained. We derived the expected waiting time of a tagged customer (i) in the buffer and (ii) in the pool, the expected duration (i) between two consecutive transfers under $N$-policy (ii) for the first $N$-policy transfer in a busy cycle and the expectation of FIFO violation. Several system performance measures and an optimization problem involving $N$ are discussed. Numerical illustrations are also provided.

In Chapter 3, we modify the model discussed in chapter 2. The entry to the buffer is restricted by the system with the increasing number of work in the buffer. If the buffer is empty, an arriving customer can enter in to it and his service starts immediately. Otherwise there is a probability depending on the number of work in the buffer. But at every time, when a
customer is not allowed to enter the buffer, he may join a pool of postponed work having infinite capacity with probability $\delta$. If the buffer is full, a customer may select the pool with probability $\gamma_1$. But at that time, system may reject him with probability $\gamma_2$. Usual transfer from the buffer to the pool with some probability $p$ and $N$-policy is considered for the service of postponed work. We studied its long run behaviour. Several system performance measures, and numerical illustrations are provided. By treating server and customer as players, we give a game theoretic approach to the model and found the mixed strategies of the players and the value of the game.

Chapter 4 discusses an $M/M/1$ Queue with Postponed work and service interruption under $N$-policy. At a service completion epoch, if the buffer size drops to a pre-assigned level or below, a postponed work is transferred to the buffer for immediate service with some probability. During the service of such a pooled customer, if the buffer size rises to a pre-assigned higher level, then the postponed work at server will be interrupted, again postponed and wait at the head of the queue in pool. Just after the interruption, we start to count the number of continuously served customers from the buffer. When it reaches a pre-assigned number $N$ at a service completion epoch, the interrupted pooled customer gets transferred to the buffer for immediate service, and further interruption is not allowed for such a work. We studied its long run behaviour and obtained several system performance measures. Several numerical illustrations are also provided.

Chapter 5 analyses a discrete time Geo/$PH_d$/1 queue with postponed work under $N$-policy. It is the discrete time counter part of the continuous time model discussed in chapter 2. Continuous time models describe the
event in a very short interval of time. But in this discrete time queueing system, time axis is divided into intervals of equal length called slots, and where all queueing activities take place at the slot boundaries. Both arrival and departure may happen in a slot. We consider late arrival system. That is departures occur at the moment immediately before the slot boundaries and arrivals occur at the moment immediately after the slot boundaries. The time between two successive arrivals is governed by a geometrical law with parameter $\alpha$ and service time of each customer by a discrete phase-type distribution. The model is studied as a quasi birth-death(QBD) process and a solution of the classical matrix geometric type is obtained.

In Chapter 6 we consider two models of discrete time $Geo/E_d/1$ queues with postponed work and protected stages. If a buffer having finite capacity is not full, a higher priority customer can enter it and a lower priority customer is directed to a pool of postponed work having infinite capacity. When the buffer is full, new arrivals of higher priority customers cannot join the system and will leave the system permanently. At that time, a new arrival of lower priority customer will join the pool with probability $\gamma$ or it is lost to the system for ever with probability $1 - \gamma$. At a service completion epoch, if buffer size drops to a pre-assigned lower level or below a postponed work is transferred to the buffer for immediate service with some probability. During the service of such a pooled customer, if the buffer size raises to a pre-assigned higher level, the postponed work at server, serving in unprotected stages will be lost for ever in the model-1 and will be interrupted in the model-2. Interrupted work is again postponed and wait at the head of the queue in the pool. After the interruption, when the continuously served customers from the buffer
1.4. Summary of the thesis

reaches a pre-assigned number $N$, at a service completion epoch, the service of interrupted customer will suddenly repeat. We study its long run behaviour and obtained certain system performance measures. Several numerical illustrations are also provided.

In chapter 7, we compare the performance of all the models discussed through chapters 2 to 6. Concluding remarks and some further possible investigations are also included.