Chapter 1

Introduction

1.1 Preliminaries

Queueing Theory is the mathematical study of queues or waiting lines. Queues abound in every day life - in computer networks, in traffic islands, in communication of electro-magnetic signals, in telephone exchange, in bank counters, in super market checkouts, in doctor’s clinics, in petrol pumps, in offices where paper works to be processed and many other places.

Originated with the published work of A. K. Erlang in 1909 [16] on congestion in telephone traffic, Queueing Theory has grown tremendously in a century. Its wide range applications includes Operations Research, Computer Science, Telecommunications, Traffic Engineering, Reliability Theory, etc.
The congestion in a service system adversely affects the profit and good will of the system. To control this congestion effectively, a thorough knowledge about the relationships between congestion and delay is inevitable. Queueing Theory provides all the tools for this analysis.

We explain some fundamental concepts in waiting line analysis.

1.1.1 Markov process

A Markov Process is a stochastic process with the property that, given the value of $X_t$, the values of $X_s$, $s > t$, do not depend on the values of $X_u$, $u < t$. If the time is discrete, the Markov process is called discrete time Markov chain; otherwise continuous time Markov chain. If a Markov chain is irreducible and positive recurrent, there exists a unique solution to the linear system $\pi P = \pi$, $\pi e = 1$, where $P$ is the one step transition probability of the Markov chain. If, moreover, the chain is aperiodic, the probabilities $P[X_t = i]$ will converge to $\pi_i$ as $i \to \infty$.

1.1.2 Markovian arrival process

A Markovian arrival process (MAP) is a Markov process $(N(t), J(t))$ with state space $\{(i, j) : i \geq 0; 1 \leq j \leq m\}$ with infinitesimal generator $Q^*$ having the structure

$$Q^* = \begin{bmatrix} D_0 & D_1 \\ D_0 & D_1 \\ & \ddots & \ddots \end{bmatrix}.$$
Here $D_0$ and $D_1$ are square matrices of order $m$; $D_0$ has negative diagonal elements and nonnegative off-diagonal elements, $D_1$ has nonnegative elements and $(D_0 + D_1)e_m = 0$, $e_m$ being a column vector of 1’s of dimension $m$. We define an arrival process associated with this Markov process as follows. An arrival occurs whenever a level state transition occurs into a state in the $D_1$ block, and there is no arrival otherwise. Here $N(t)$ represents the number of arrivals in $(0, t]$ and $J(t)$ the phase of the Markov process at time $t$. Let $\delta$ be the stationary probability vector of the generator $D = D_0 + D_1$. Then the constant $\lambda = \delta D_1 e_m$, referred to as the fundamental rate, gives the expected number of arrivals per unit time in the stationary version of the MAP. It should be noted that in general MAP is a non-renewal process. However, by appropriately choosing the parameters of the MAP the underlying arrival process can be made as a renewal process. To sum up, MAP is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, Markov-Modulated Poisson process and superpositions of these. One of the most significant features of MAP is the underlying Markovian structure and fits ideally in the context of matrix analytic solutions to stochastic models. Often, in model comparisons, it is convenient to select the time scale of the MAP so that the stationary arrival rate $\lambda$ has a certain value. That is accomplished, in the continuous MAP case, by multiplying the coefficient matrices $D_0$ and $D_1$, by the appropriate common constant. For further details on MAP and their usefulness in stochastic modelling, we refer to [36], [46] and for a review and recent work on MAP we refer the reader to [7]. Chakravarthy [9] and Krishnamoorthy et al. [33] provide an account of more recent works in this area.
1.1.3 Phase type distributions

Consider a finite state space Markov chain with \( m \) transient states and one absorbing state. The infinitesimal generator \( Q \) of this Markov chain be partitioned as

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

where \( T \) is a matrix of order \( m \) and \( T^0 \) is a column vector such that \( Te + T^0 = 0 \), \( e \) being a column vector consisting of 1’s of appropriate dimension. For the eventual absorption into the absorbing state it is necessary and sufficient that \( T \) be nonsingular. The initial state of the Markov chain is chosen according to a probability vector \((\alpha, \alpha_{m+1})\). Then the time until absorption, \( X \) is a continuous time random variable with probability distribution function \( F(x) = 1 - \alpha \exp(Tx)e \), for \( x \geq 0 \). The density function \( f(x) \) of \( F(x) \) is either identically zero or strictly positive for all \( x \geq 0 \). In the latter case \( f(x) \) is given by \( f(x) = \alpha \exp(Tx)T^0 \), for \( x \geq 0 \). The Laplace Stieltjes transform \( \tilde{f}(s) \) of \( F(x) \) is given by \( \tilde{f}(s) = \alpha_{m+1} + \alpha(sI - T)^{-1}T^0 \), for \( Re \ s \geq 0 \). Hence the \( k^{th} \) non central moments of \( F(x) \) is given by the formula \( \mu_k' = (-1)^k k!(\alpha T^{-k}e) \), for \( k \geq 1 \). In particular, if \( T = [-\mu] \) and \( T^0 = [\mu] \) with \( \alpha = (1) \), we get an exponential distribution with mean \( \mu \). The class of PH distributions include the distributions such as hyper exponential, Erlang and generalized Erlang also as its special cases. Most importantly any continuous time distribution on non negative real line can be approximated by phase type distributions. Phase type distributions are well suited for applying matrix analytic methods. For further details of PH distribution see [35], [5], [44].
1.1.4 Quasi-birth-and-death process

A level independent quasi-birth-and-death (LIQBD) process is a Markov process on the state space \( E = \{(i, j) : i \geq 0; 1 \leq j \leq m\} \) with infinitesimal generator \( \tilde{Q} \), given by

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

The one step transitions are allowed only between the states belonging to the same level or adjacent levels. Hence the name quasi-birth-and-death process. The number of boundary level states may vary and the complexity increases with the number of boundary levels. However, with suitable modifications we can handle more complicated boundary behavior. The generator \( \tilde{Q} \) is assumed to be irreducible. The matrix \( A = A_0 + A_1 + A_2 \) is the generator matrix of a finite state Markov process. The process \( \tilde{Q} \) is positive recurrent if and only if the minimal nonnegative solution \( R \) of the matrix quadratic equation \( R^2 A_2 + RA_1 + A_0 = 0 \) has spectral radius \( sp(R) \) is less than 1. We can use the iterative formulas (see Neuts [44])

\[
R_n = -A_0(A_1 + R_{n-1}A_2)^{-1}, \quad \text{for } n \geq 1,
\]

with an initial value \( R_0 \), which converges to \( R \) if \( sp(R) < 1 \). Although level dependent quasi-birth-and-death process arises in a natural way, it does not appear in this thesis.
1.1.5 Kronecker product and Kronecker sum

Let $A$ and $B$ be matrices of orders $m \times n$ and $p \times q$ respectively, then the Kronecker product of $A$ and $B$, denoted by $A \otimes B$ is a matrix of order $mp \times nq$ whose $(i, j)^{th}$ block matrix is given by $a_{ij}B$. If $A$ and $B$ are square matrices of order $m$ and $n$ respectively then the Kronecker sum of $A$ and $B$, denoted by $A \oplus B$ is defined as $A \otimes I_n + I_m \otimes B$. For further details on Kronecker products and sums, we refer the reader to [21] and [37].

1.2 Motivation of the present work

In this modern world, demand for almost all types of services is very high. In order to keep up the good will, the service providers have to appoint more counters. Thus arises the case of multi-server queueing systems. The services provided by these channels can be of the same type or of entirely different types or they may contain some common elements. In the first case, only one queue of customers is formed and each server is fed up by this queue. But in the other types, different queues are to be maintained.

In a multi-server queueing system providing same type of services, some of the servers (trainees or less experienced ones) need clarifications or help frequently. So an experienced server provides timely clearances together with serving customers. Such queueing systems with consultations given by a server (namely, main server) to the fellow servers are common in banks, super market check outs, hospitals, etc.
1.2. Motivation of the present work

Chakravarthy [6] introduced a multi-server queueing system with consultations. There are $c$ servers. One of these $c$ servers are referred to as the main server and the others as the regular servers. The main server provides preemptive priority to the regular servers on FIFO basis for consultation. Thus the service of the customer at the main server will be interrupted when a consultation occurs. The service of the interrupted customer at the main server will be resumed after all consultations are completed. The regular servers receive any number of consultations during the service of a customer. The service times are exponentially distributed with mean $\mu_1$ at the main server and $\mu_2$ at the identical regular servers. Queueing system with consultation has many applications in daily life. One such example is given in the above mentioned work.

Krishnamoorthy et.al [33] discussed a single server queueing model with interruptions to the server controlled by a finite number of interruptions and a super clock. When the number of interruptions already befall to the server reaches the upper bound, no further interruptions are allowed to the customer being served. A super clock is started at the epoch of the first interruption to a customer’s service and is freezed at the moment the interruption is over. When the next interruption to the same customer strikes, the super clock starts from the earlier position where it stopped ticking and so on. If the super clock expires, no further interruptions are permitted to the present customer. A threshold clock starts at the epoch of each interruption and it ends with the completion of that interruption. After each interruption, the service will be resumed or restarted according to the realisation of the threshold clock. The arrival process is MAP, the interruption occurs according to a Poisson Process and the service time,
durations of interruption, threshold clock and super clock follow mutually independent phase type distributions.

Queues with service interruptions was first studied by White and Christie [53] with exponentially distributed interruption duration. At the end of an interruption the service will be resumed. Some of the earlier papers which analyse queueing models with service interruptions, assuming general distributions for the service and interruption durations, are by Gaver [18], Keilson [26], Avi-Izhak and Naor [1] and Fiems et. al [17].

Klimenok et. al. [29] discussed a multi-server queueing system with finite buffer and negative customers. They assumed that a negative customer can delete an ordinary customer in service if the service of a customer goes on in any of the unprotected phases; whereas if the service of the customer is protected from the effect of the negative customers, the interruption has no effect on the service process. Klimenok and Dudin [28] extended the above paper by considering disciplines of complete admission and complete rejection. They assumed the system to have an infinite capacity waiting room.

Krishnamoorthy et. al. [32] introduced the idea of protection in a queueing system where the service process is subjected to interruptions. They assumed that the final $m - n$ phases of the Erlang service process with $m$ phases are protected from interruptions.
Bhaskar Senguptha [3] dealt with a queueing system in an alternating random environment. Here the server is subject to random breakdown and cannot serve until it is repaired. During the break down period, some arriving customers are diverted to another service facility. Thus the arrival rate and service rate of the customers who arrive during the break down period are different from those arrive at the busy period of the server.

1.3 Summary of the thesis

The title of the thesis is “On Multi-Server Queues with Consultation by Main Server.” Here ‘consultation by main server’ means the consultation is provided by the main server to the regular server(s). This thesis consists of six chapters including the introductory chapter. Chapters 2, 3 and 6 analyse two-server queues; chapter 4 analyses three server queues and chapter 5 analyses a multi-server queue. In all these models (except in chapter 6) one of the servers is referred to as ‘main server’ and the other(s) as ‘regular server(s)’. The main server provides consultation to the regular servers with a preemptive priority over customers. The arrival processes in chapter 2 are MAP and those in other chapters are Poisson Processes. The service times at the servers follow mutually independent phase type distributions except in chapter 5, where service times at regular servers follow exponential distributions.

In chapter 2 we analyse three distinct queueing models equipped with two servers, namely a main server and a regular server. The main server not only serves customers but also provides consultation to the regular
server with a preemptive priority over customers. Thus the customers at
the main server undergo interruptions during their service. The upper
bound of interruptions to a customer at the main server and the upper
bound of consultations for the regular server are respectively denoted by
$M$ and $K$. A super clock also determines whether to attend further inter-
ruptions during the service of a customer at the main server. A threshold
clock is set to determine whether the services at both the servers are to
be restarted or resumed after consultation. The arrivals of customers to
the system follow MAP and requirement of consultation follows a Poisson
process; the durations of consultation, threshold clock and super clock
follow mutually independent phase type distributions. The service times
at the servers are assumed to follow mutually independent phase type dis-
tributions. In model 1, the interruption is allowed to continue even when
the super clock is saturated. In model 2, the interruption will be stopped
at the time the super clock realises and the service at the main server will
be restarted or resumed according to the status of the threshold clock.
The regular server will wait to get the remaining consultation after the
present service completion at the main server. In model 3, in addition
to the assumptions in model 1, we assume that some phases of the main
server are protected from interruption of service. When the main server is
at any one of these protected phases, the regular server has to wait until
the service completion at the main server to get consultation. Implicit
expressions for stability of the systems are derived in all the three models.
We compute expected waiting time of a customer in queue. Some impor-
tant performance measures are studied numerically. Finally a comparison
of the three models is presented.

In chapter 3 we consider two two-server queueing models with con-
1.3. Summary of the thesis

consultations. In model 1, consultation for the regular server is in random environment and in model 2, consultation is in Markovian environment and the environmental factors are related to each other by a transition probability matrix. In both models the arrival of customers and requirement of consultation follow independent Poisson processes, the duration of consultations caused by distinct factors follow independent exponential distributions and the duration of the threshold clock follows an exponential distribution. All other assumptions regarding number of interruptions, consultations and super clock are same as those in model 1 of chapter 2. We establish stability conditions in both the models. Some important performance measures are studied numerically.

Two queueing models equipped with three servers, namely a main server and two i.i.d regular servers are dealt in chapter 4. The upper bound for interruptions possible to a customer at the main server is $M$. No bound is imposed on the number of consultations to the regular servers. The requirement of consultations follow independent Poisson processes. Duration of services provided at the main server and the regular servers are assumed to follow mutually independent phase type distributions. In model 1, arrival of the customers to the system is assumed to follow a Poisson Process. Whereas in model 2, arrivals to the main server and regular servers follow independent Poisson processes and there is a finite buffer at the main server such that an arriving customer to the main server will be lost when the buffer is full. The stability condition is established in each model. Expected number of interruptions to the main server during the service of a particular customer is evaluated and a cost function is analysed in model 2. Some performance measures are studied numerically.

In chapter 5 we analyse a multi-server queueing model with $c + 1$
servers, namely one main server and $c$ regular servers. The main server provides consultation to the regular servers in a FIFO basis with a preemptive priority over customers. The arrivals to the system and requirement of consultation follow independent Poisson processes; the service time at the main server follows phase type distribution and the service times at the regular servers follow independent and identically distributed exponential distribution. The duration of consultation follows an exponential distribution. An explicit expression for stability of the system is obtained. The expected number of interruptions to a customer at the main server is evaluated. A cost function is also analysed. Some important performance measures are studied numerically.

We consider a two-server queueing model in chapter 6. In this model the servers provide consultations to each other with a preemptive priority over the customers being served. Thus customers at both the servers undergo interruptions during their services. There are no upper bounds on the number of interruptions to the customers at the servers. The customer arrival to the system and requirement of consultations of the servers follow independent Poisson processes. Duration of consultation follow independent exponential distributions. Each server is free to have any number of consultations with the other server during the service of a customer. The service times of customers at these servers are assumed to follow mutually independent phase type distributions. An explicit expression for system stability is derived and some performance measures are studied numerically. Two particular cases of this model are considered and a comparison of the respective performance measures of the three models is presented.

The thesis ends with a conclusion of the work done and the scope of further study.