ON MULTI-SERVER QUEUES WITH CONSULTATION BY MAIN SERVER

Thesis submitted to the

Cochin University of Science and Technology

for the award of the degree of

DOCTOR OF PHILOSOPHY

under the **Faculty of Science**

by

RESMI T

Department of Mathematics

Cochin University of Science and Technology

Cochin - 682 022

AUGUST 2015

Certificate

This is to certify that the thesis entitled '**ON MULTI-SERVER QUEUES WITH CONSULTATION BY MAIN SERVER**' submitted to the Cochin University of Science and Technology by Ms.Resmi T for the award of the degree of Doctor of Philosophy under the Faculty of Science is a bonafide record of studies carried out by her under my supervision in the Department of Mathematics, Cochin University of Science and Technology. This report has not been submitted previously for considering the award of any degree, fellowship or similar titles elsewhere.

Dr. B. Lakshmy
(Research Guide)
Associate Professor
Dept. of Mathematics
Cochin University of Science and Technology
Kochi - 682 022, Kerala

Cochin-22
21.08.2015

Declaration

I, Resmi T, hereby declare that this thesis entitled '**ON MULTI-SERVER QUEUES WITH CONSULTATION BY MAIN SERVER**' contains no material which had been accepted for any other Degree, Diploma or similar titles in any University or institution and that to the best of my knowledge and belief, it contains no material previously published by any person except where due references are made in the text of the thesis.

Resmi T
Research Scholar
Registration No. 2952
Department of Mathematics
Cochin University of Science and Technology
Cochin-682 022, Kerala.

Cochin-22
21.08.2015

To

My Parents and Teachers

Acknowledgement

First of all I thank God Almighty for guiding me throughout my life and showering me with His blessings.

I would like to express my heart felt gratitude to Dr. A. Krishnamoorthy for his timely help and support to complete my research work. I am indebted to him to his support and affection.

I wish to express my sincere gratitude to Dr. B. Lakshmy, my guide and supervisor who encouraged me and gave me constant support. I thank her for her trust in me and allowing me to do research under her.

Dr. Romeo P. G, Head, Dept of Mathematics, CUSAT was a constant support for me. I thank him for his valuable suggestions.

I thank all other faculty members, Dr. Jathavedhan, Dr. M.N.N. Namboothiri, Dr.S.R. Chakravarthy, Dr.A. Vijayakumar, Ms. Meena, Dr. Kiran Kumar, Dr. Ambily for all the support and help I received. I also thank Mr. Syam Sunder for his timely help. I thank the office staff and library staff of the Dept. of Mathematics, CUSAT for their co-operation. I would like to thank the Principals and my colleagues at K.K.T.M. Govt. College, Pullut, Kodungallur for the care, concern and help they provided to me. I thankfully acknowledge the UGC for providing me Teacher Fellowship under FIP during 12^{th} plan.

I would like to thank my fellow research scholars Dr. Resmi Varghese, Dr. Pramod, Dr. Aparna, Dr. Manju, Dr. Raji George, Dr. Pamy, Ms.

Dhanya, Ms. Seethu, Mr. Pravas, Ms. Savitha, Ms. Jaya, Ms. Akhila, Mr. Manjunath, Mr. Gireesan, Ms. Binitha, Mr. Sathyan, and others for their love and support.

Words are not enough to express my feelings towards my family members for their love, support and care for me. I also thank all my relatives and friends.

Resmi T.

ON MULTI-SERVER QUEUES WITH CONSULTATION BY MAIN SERVER

Contents

Notations used

- $\boldsymbol{e}$ denotes column vector of 1's with appropriate dimension
- $\mathbf{0}$ is a column vector consisting of 0's with appropriate dimension
- $\underline{\mathbf{0}}$ is row vector consisting of 0's with appropriate dimension
- I is a matrix of appropriate order
- $\boldsymbol{e}_j(r)$ denotes column vector of dimension r with 1 in the j^{th} position and 0 elsewhere
- $\hat{I}_c = \begin{bmatrix} I_c & \mathbf{0} \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$
- $\hat{\mathbf{e}}_c = \begin{bmatrix} \mathbf{e}_c \\ 0 \end{bmatrix}$

Abbreviation used

PH	:	Phase Type;
MAP	:	Markovian Arrival Process;
$CTMC$	:	Continuous-time Markov Chain;
$FIFO$	:	First In First Out;
QBD	:	Quasi-Birth-Death;
$LIQBD$	:	Level Independent Quasi-Birth-Death;

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 $\boldsymbol{\pi} P = \boldsymbol{\pi}$, $\boldsymbol{\pi}\mathbf{e} = 1$, 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 $\boldsymbol{\pi}_i$ 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 $(D_0 + D_1)\mathbf{e}_m = 0$, $\mathbf{e}_m$ 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 $\boldsymbol{\delta}$ be the stationary probability vector of the generator $D = D_0 + D_1$. Then the constant $\lambda = \boldsymbol{\delta} D_1 \mathbf{e}_m$, 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 λ 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

$$Q = \begin{bmatrix} T & \mathbf{T}^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix},$$

where T is a matrix of order m and $\mathbf{T}^0$ is a column vector such that $T\mathbf{e} + \mathbf{T}^0 = \mathbf{0}$, $\mathbf{e}$ 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 $(\boldsymbol{\alpha}, \alpha_{m+1})$. Then the time until absorption, X is a continuous time random variable with probability distribution function $F(x) = 1 - \boldsymbol{\alpha} exp(Tx)\mathbf{e}$, 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 $f(x) = \boldsymbol{\alpha} exp(Tx)\mathbf{T}^0$, for $x \geq 0$. The Laplace Stieltjes transform $\tilde{f}(s)$ of $F(x)$ is given by $\tilde{f}(s) = \alpha_{m+1} + \boldsymbol{\alpha}(sI - T)^{-1}\mathbf{T}^0$, for $Re\ s \geq 0$. Hence the k^{th} non central moments of $F(x)$ is given by the formula $\mu_k^{'} = (-1)^k k!(\boldsymbol{\alpha} T^{-k}\mathbf{e})$, for $k \geq 1$. In particular, if $T = [-\boldsymbol{\mu}]$ and $\mathbf{T}^0 = [\boldsymbol{\mu}]$ with $\boldsymbol{\alpha} = (1)$, we get an exponential distribution with mean $\boldsymbol{\mu}$. 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 $R^2A_2 + RA_1 + A_0 = 0$ has spectral radius $sp(R)$ is less than 1. We can use the iterative formulas (see Neuts [44]) $R_n = -A_0(A_1 + R_{n-1}A_2)^{-1}$, 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.

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 μ_1 at the main server and μ_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 befell 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 tittle 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 interruptions 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 distributions. 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 important 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 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.

Chapter 2

Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

In this chapter we study three two-server queueing models with consultations given by the main server to the regular server. The service of the customer at the main server is interrupted when he is being served by the main server at the time of request of the regular server for consultation. It is not fair to interrupt a customer at the main server infinitely many times or to receive infinitely many consultations, if he is at the regular server, during his service. So we impose some upper bounds to control consultations and interruptions. In this aspect our model differs from that of Chakravarthy [6] in which a multi-server queueing system with consultations is discussed. There is no boundary on the number of interruptions to a customer at the main server and the regular server is free to get any number of consultations during the service of a customer. The main server gives immediate consultations to the regular servers. The request for consultation of the regular server is attended by the main server, even if there is a customer being served at the main server. Then that customer at the main server has to wait until the consultation is completed. At this stage the service of the customer at the main server is said to be interrupted. (So the word 'interruption' is associated with the customer at the main server when the main server is providing consultation to the regular server.) The service times at these servers follow independent phase type distributions.

We introduce upper bounds for interruptions, consultations; a super clock to get an 'approximate measure' of the total duration of interruption. We say it is an 'approximate measure' because the total interruption time can be greater than the duration of super clock if super clock expires during interruption and interruption continues to be completed. If super clock does not expire, duration of super clock is the total interruption time during the service of a customer at the main server. (But in model 2, duration of super clock is strictly equal to the total duration of interruption since the interruption will be removed and the service of the customer at the main server will be continued as soon as the super clock expires.) A maximum of M interruptions are allowed to a customer at the main server. No further interruptions are allowed to that customer after M interruptions. If the regular server needs further consultation at this time, he/she has to wait until the service at the main server is completed. After the service completion of the interrupted customer, the main server will immediately attend the consultation before taking a new customer from the queue for service. The maximum number of consultations possible to the regular server during the service of a particular customer is K.

If the super clock expires during consultation with one interrupted customer at the main server, then the present consultation is permitted to complete and no more interruption is allowed to befall to that particular customer at the main server. At this stage, if the regular server again needs a consultation, he has to wait until the completion of the service at the main server. After finishing the service, the main server will immediately attend the consultation. At this time, no customer is interrupted at the main server and so no super clock is present here.

So the main server offers consultation in the following manner:

(i) If the main server is idle, then the request for consultation will be attended immediately.

(ii) If the number of interruptions already befell to the customer at the main server is less than M and the super clock has not expired, then also the consultation will be provided immediately.

(iii) If either the customer at the main server has interrupted M times or the super clock has expired, then the regular server has to wait until the completion of the service of the present customer at the main server.

(iv) The regular server needs a further consultation only when the number of consultations already taken by him for the same customer is strictly less than K.

In model 3, we assume that interruption is not allowed to a customer at the main server if the service is in any one of the protected phases of service (which may be so costly to afford an interruption at these phases). So the consultation to the regular server will be denied if the main server is at the protected phases. All other assumptions are same as those in model 1.

A comparison of the three models is provided towards the end of this chapter.

2.1 Description of model 1

Here we consider a service system equipped with one main server and one regular server to which customers arrive according to a MAP with representation (L_0, L_1) , where L_0 and L_1 are matrices of order r. An arriving customer enters into service immediately if at least one server is free, else joins the queue of waiting customers. The service times at the main and regular servers follow independent phase type distributions with representations $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\beta}, U)$ with number of phases a and b, respectively. Write $\mathbf{T}^0 = -T\mathbf{e}$ and $\mathbf{U}^0 = -U\mathbf{e}$ where $\mathbf{e}$ is a column vector of 1's of appropriate order. The main server offers consultation to the regular server whenever it is needed. Requirement of consultation is a Poisson process with rate θ. The request for consultation by the regular server is attended by the main server. If there is a customer being served at the main server, that customer at the main server has to wait until the consultation is completed. At this stage the service of the customer at the main server is said to be interrupted. (So the word 'interruption' is associated with the customer at the main server when the main server is providing consultation to the regular server.) At most M interruptions are allowed to a customer at the main server. No further interruption is permitted to that customer after M interruptions. If the regular server needs consultation at this time, he/she has to wait until the service of the customer at the main server is completed. Once his service is completed, the main server will attend the consultation before taking a new customer from the queue for service. The maximum number of consultations possible to the regular server during the service of a particular customer is set as K. This is to ensure that customers in service at the regular server do not get too impatient to leave the system.

The duration of super clock, threshold clock and consultation clock follow independent phase type distributions with representations $(\boldsymbol{\gamma}, G)$, $(\boldsymbol{\eta}, E)$, $(\boldsymbol{\delta}, D)$ with number of phases c, d and f, respectively. We have $\mathbf{G}^0 = -G\mathbf{e}$, $\mathbf{E}^0 = -E\mathbf{e}$, and $\mathbf{D}^0 = -D\mathbf{e}$, respectively.

The threshold clock determines the restart or resumption of services at both the servers. Every time this clock starts anew when the regular server temporarily stops his service for consultation. If the regular server is waiting to get consultation, this clock starts ticking and continue during the time of consultation after the service at the main server. On the other hand, if regular server gets consultation immediately, the consultation process and threshold clock start together. If the threshold clock expires before the consultation process, then the services at both the servers are to be restarted. Otherwise the services will be resumed at the phases where they are interrupted.

A super clock is set to determine whether further interruption to a customer at the main server is to be allowed or not. This clock starts at the epoch of the first interruption of a particular customer at the main server and is freezed at the moment the consultation 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 during consultation with one interrupted customer at the main server, then the present consultation is permitted to continue until completion and no more interruption is allowed to befall to that particular customer at the main server. At this stage, if the regular server again needs a consultation, he has to wait until the completion of the service at the main server. After completing the service, the main server will immediately attend the consultation. Since there is no interrupted customer at the main server, super clock is in the 'off' mode (indicating that service is not interrupted at the main server.)

So the main server offers consultation to the regular server in the following manner:

(i) If the main server is idle, then the request for consultation will be attended immediately.

(ii) If the number of interruptions already befell to the customer at the main server is less than M and the super clock has not expired, then also the consultation will be provided immediately.

(iii) If either the customer at the main server has interrupted M times or the super clock has expired, then the regular server has to wait until the completion of the service of the present customer at the main server.

(iv) The regular server needs further consultation only when the number of consultations already taken by him for a particular customer is strictly less than K.

Notations :- We use the following notations in this model.

- $M_0 = M(c+1)$ and $M_1 = M_0 + 1$
- $\tilde{\boldsymbol{\alpha}} = \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha}$
- $\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, 0), \tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}, 0)$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$ and $\tilde{E} = \begin{bmatrix} E & E^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$
- $D^* = D \oplus \tilde{E}$ and $G^* = \tilde{G} \oplus D^*$
- $\dot{I} = \begin{bmatrix} \mathbf{0} & I_{M(c+1)} \end{bmatrix}$

Consider the queueing model

$X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t)\}$.

The variables are defined as follows:

$N(t)$	–	the number of customers in the system
$B_1(t)$	–	number of consultations already enjoyed by the regular server during the service of a particular customer
$B_2(t)$	–	number of interruptions already befell to a customer at the main server
$S_1(t)$	–	phase of the super clock
$S_2(t)$	–	phase of the consultation process
$S_3(t)$	–	phase of the threshold clock
$J_1(t)$	–	phase of the main server
$J_2(t)$	–	phase of the regular server
$U(t)$	–	phase of the arrival process

Here $S(t)$ denotes the status of the servers at time t such that

$$S(t) = \begin{cases} \tilde{0}, & \text{if only the regular server is busy} \\ 0, & \text{if the main together with or without} \\ & \text{the regular server is busy} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation} \\ & \text{with one interrupted customer at the main server} \\ 3, & \text{if the regular server is waiting for getting consultation} \\ & \text{after the present service at the main server} \end{cases}$$

Note that $B_2(t)$ is '0' means the customer at the main server has not interrupted yet and so super clock has not started. In this case the super clock has no role to play. So we do not consider the super clock variable $S_1(t)$ when $B_2(t) = 0$. Also, since super clock is associated with the interruption to a customer at the main server and no customer is present at the main server during the 'consultation only' mode, super clock is not 'present' at this mode.

$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \bigcup_{i=0}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as

$\psi(0) \quad = \{(0, u)\}$,

$\psi(1) = \psi(1, 0) \cup \psi(1, \tilde{0}) \cup \psi(1, 1)$ and

$\psi(i) = \psi(i, 0) \cup \psi(i, 1) \cup \psi(i, 2) \cup \psi(i, 3)$, for $i \geq 2$,

where

$\psi(1, 0) = \{(1, 0, 0, t_1, u)\} \cup \{(1, 0, k, l_1, t_1, u) : 1 \leq k \leq M\}$

$\psi(1, \tilde{0}) = \psi\{(1, \tilde{0}, j, t_2, u) : 0 \leq j \leq K\}$

$\psi(1, 1) = \{(1, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}$

$\psi(i, 0) = \{(i, 0, j, 0, t_1, t_2, u) \cup (i, 0, j, k, l_1, t_1, t_2, u) : 0 \leq j \leq K, 1 \leq$

$k \leq M\}$

$\psi(i,1) = \{(i,1,j,l_2,l_3,t_2,u) : 0 \leq j \leq K-1\}$

$\psi(i,2) = \{(i,2,j,k,l_1,l_2,l_3,t_1,t_2,u) : 0 \leq j \leq K-1, 0 \leq k \leq M-1\}$

$\psi(i,3) = \{(i,3,j,l_3,t_1,t_2,u) : 0 \leq j \leq K-1\}$

with $0 \leq l_1 \leq c, 1 \leq l_2 \leq d$, $0 \leq l_3 \leq f, 1 \leq t_1 \leq a$, $1 \leq t_2 \leq b$, $1 \leq u \leq r$ and for $i \geq 2$.

The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} L_0 & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{2.1}$$

where $B_1 = \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix} \otimes L_1$, $B_2 = \begin{bmatrix} \mathbf{e}_{M_1} \otimes T^0 \\ \mathbf{e}_{K+1} \otimes U^0 \\ \mathbf{0} \end{bmatrix} \otimes I_r$,

$$B_3 = \begin{bmatrix} I_{M_1} \otimes T & O & O \\ O & B_{31} & B_{32} \\ O & B_{33} & I_K \otimes D^* \otimes I_b \end{bmatrix} \oplus L_0,$$

$$B_4 = \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix} \otimes L_1 \text{ , } B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix} \otimes I_r,$$

$$A_0 = I \otimes L_1, \; A_1 = \begin{bmatrix} A_{11} & O & A_{12} & A_{13} \\ A_{14} & A_{15} & O & O \\ A_{16} & O & A_{17} & O \\ O & O & O & A_{18} \end{bmatrix} \oplus L_0,$$

$$A_2 = \begin{bmatrix} A_{21} & B_{53} & O \end{bmatrix} \otimes I_r.$$

Here A_0, A_1 and A_2 are square matrices of order C_0, B_3 is a square matrix of order C_1 and B_1, B_2, B_4, B_5 are matrices of orders $r \times C_1$, $C_1 \times r$, $C_1 \times C_0$ and $C_0 \times C_1$, respectively,

where

$C_0 = [M_1(K+1)ab + Kd(f+1)b + M_0Kd(f+1)ab + K(f+1)ab]r$, and

$C_1 = [M_1a + (K+1)b + Kbd(f+1)]r$.

We have

$$B_{31} = \begin{bmatrix} I_K \otimes (U - \theta I) & O \\ O & U \end{bmatrix}_{(K+1)b \times (K+1)b},$$

$$B_{32} = \theta \begin{bmatrix} I_K \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}} \\ O \end{bmatrix}_{(K+1) \times Kd(f+1)} \otimes I_b,$$

$$B_{33} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \tilde{\Delta}_b \end{bmatrix}_{Kd(f+1)b \times (K+1)b},$$

$$B_{41} = \begin{bmatrix} \mathbf{e}'_{K+1}(1) \otimes I_{M_1} \otimes I_a \otimes \boldsymbol{\beta} \\ I_{K+1} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ O \end{bmatrix}_{C_1 \times (K+1)M_1ab}, B_{42} = \begin{bmatrix} O \\ I \end{bmatrix}_{C_1 \times Kd(f+1)b},$$

$$B_{51} = \begin{bmatrix} \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_0 \times M_0 a},$$

$$B_{52} = \begin{bmatrix} I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_0 \times (K+1)b}, B_{53} = \begin{bmatrix} O \\ F \end{bmatrix}_{C_0 \times Kbd(f+1)},$$

$$A_{11} = \begin{bmatrix} I_K \otimes I_{M_1} \otimes (T \oplus U - \theta I) & O \\ O & I_{M_1} \otimes (T \oplus U) \end{bmatrix}_{M_1(K+1)ab \times M_1(K+1)ab},$$

$$A_{12} = \theta \begin{bmatrix} I_K \otimes P \\ O \end{bmatrix}_{M_1(K+1) \times M_0 d(f+1)} \otimes I_{ab},$$

$$A_{13} = \theta \begin{bmatrix} I_K \otimes P^* \\ O \end{bmatrix}_{M_1(K+1) \times K} \otimes \tilde{\boldsymbol{\eta}} \otimes I_{ab},$$

$$A_{14} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \Delta^0 \end{bmatrix}_{Kd(f+1)b \times M_1(K+1)ab},$$

$$A_{15} = I_K \otimes D^* \otimes I_b,$$

$$A_{16} = \begin{bmatrix} O & I_K \otimes \dot{I} \otimes D^0 \otimes \tilde{\Delta} \end{bmatrix}_{M_0 Kd(f+1)ab \times M_1(K+1)ab},$$

$$A_{17} = I_K \otimes I_M \otimes G^* \otimes I_{ab}, \; A_{18} = I_K \otimes (\tilde{E} \oplus T) \otimes I_b,$$

$$A_{21} = \begin{bmatrix} \tilde{F} \\ O \end{bmatrix}_{C_0 \times M_1(K+1)ab}.$$

Here

$$F = I_K \otimes \boldsymbol{\delta} \otimes I_{f+1} \otimes T^0 \otimes I_b,$$

$$P = \begin{bmatrix} diag(\tilde{\boldsymbol{\gamma}}, I_{M-1} \otimes \hat{I}_c) \\ O \end{bmatrix} \otimes (\boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}}),\ P^* = \begin{bmatrix} 0 \\ \mathbf{e}_{M-1} \otimes \hat{\mathbf{e}_c} \\ \mathbf{e}_{c+1} \end{bmatrix},$$

$$\tilde{F} = I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b + \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \boldsymbol{\beta},$$

$$\tilde{\Delta}_b = \begin{bmatrix} \mathbf{e}_f \otimes I_b \\ \mathbf{e}_b \otimes \boldsymbol{\beta} \end{bmatrix}, \Delta^0 = \begin{bmatrix} \mathbf{e}_f \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ \mathbf{e}_b \otimes \tilde{\boldsymbol{\alpha}} \otimes \boldsymbol{\beta} \end{bmatrix}, \tilde{\Delta} = \begin{bmatrix} \mathbf{e}_f \otimes I_{ab} \\ \mathbf{e}_{ab} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}.$$

P and P^* are matrices of orders $M_1 \times M_0 d(f+1)$ and $M_1 \times 1$ respectively.

2.2 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.

2.2.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. That is, $\boldsymbol{\pi}(A_0 + A_1 + A_2) = 0$; $\boldsymbol{\pi}\mathbf{e} = 1$.

~~The LIQBD description of the model indicates that the queueing system~~ is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{2.2}$$

That is, the rate of drift to the left has to be higher than that to the right. The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model, and hence the stability condition is known only implicitly. If we partition the vector $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)$$

and then using the structure of the matrices A_0 and A_2 , equation (2.2) is given by

$$\lambda < \boldsymbol{\pi}_0 \tilde{F} \mathbf{e} + \boldsymbol{\pi}_3 F \mathbf{e}. \tag{2.3}$$

For future reference, we define the traffic intensity ρ_1 as

$$\rho_1 = \frac{\boldsymbol{\pi} A_0 \mathbf{e}}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{2.4}$$

Note that the stability condition in (2.2) is equivalent to $\rho_1 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.3.

2.2.2 Steady state probability vector

Since the model studied as a QBD process, its steady-state distribution has a matrix-geometric solution under the stability condition. Assume that the stability condition holds. Let $\boldsymbol{x}$ denote the steady-state probability vector of the generator Q given in (2.1). That is,

$$\boldsymbol{x}Q = 0; \boldsymbol{x}\mathbf{e} = 1. \tag{2.5}$$

Partitioning $\boldsymbol{x}$ as

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...), \tag{2.6}$$

we see that, under the assumption that the stability condition (2.2) holds, the sub-vectors $\boldsymbol{x}_i$, $i \geq 3$ are obtained as (see, Neuts [44])

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3, \tag{2.7}$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{2.8}$$

$\boldsymbol{x}_0$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained using the boundary equations

$$\begin{aligned} \boldsymbol{x}_0 L_0 + \boldsymbol{x}_1 B_2 &= 0 \\ \boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_3 + \boldsymbol{x}_2 B_5 &= 0 \\ \boldsymbol{x}_1 B_4 + \boldsymbol{x}_2 (A_1 + RA_2) &= 0 \end{aligned} \tag{2.9}$$

The normalizing condition of (2.5) results in

$$\boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_1\mathbf{e} + \boldsymbol{x}_2(I - R)^{-1}\mathbf{e} = 1. \tag{2.10}$$

Once the rate matrix R is obtained, the vector $\boldsymbol{x}$ can be computed by exploiting the special structure of the coefficient matrices. We can use the iterative formulas (see Neuts [44]) $R_n = -A_0(A_1 + R_{n-1}A_2)^{-1}$, for $n \geq 1$, with an initial value R_0, which converges to R if $sp(R) < 1$.

2.2.3 Expected waiting time in queue

For computing expected waiting time in queue of a particular customer who joins as the m^{th} customer, where $m > 0$, in the queue, we consider the Markov process

$$Z(t) = \{(\tilde{N}(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t)) : t \geq 0\}$$

where

$\tilde{N}(t)$ is the rank of the customer and all other variables defined as earlier. The rank $\tilde{N}(t)$ of the customer is assumed to be i if he is the i^{th} customer in the queue at time t. His rank may decrease to 1 as the customers ahead of him leave the system either after completing their services (if $S(t) = 0$) or completing the consultation (if $S(t) = 1$). Since the customers who arrive after the tagged customer cannot change his rank, level-changing transitions in $Z(t)$ can only take place to one side of the diagonal. The absorbing state Δ_2 denote the tagged customer is selected for service. Thus the infinitesimal generator $\tilde{V}$ of the process $Z(t)$ takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ \underset{\sim}{\mathbf{0}} & 0 \end{bmatrix},$$

where

$$V = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 & & & \\ & \tilde{A}_1 & \tilde{A}_2 & & \\ & & \ddots & \ddots & \\ & & & \tilde{A}_1 & \tilde{A}_2 \\ & & & & \tilde{A}_1 \end{bmatrix}, V^0 = \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_{M_1} \otimes (T^0 \oplus U^0) \\ \mathbf{0} \\ T^0 \otimes \mathbf{e}_b \end{bmatrix},$$

with $\tilde{A}_1 = A_1^* - U_2$ and $\tilde{A}_2 = A_2^* + U_2$, where A_1^* and A_2^* are obtained from A_1 and A_2 if they are written as $A_1 = A_1^* \oplus L_0$ and $A_2 = A_2^* \otimes I_r$. Here $U_2 = \begin{bmatrix} O & O \\ A_{14} & O \\ O & O \end{bmatrix}$.

Now, the waiting time V of a customer, who joins the queue as the j^{th} customer is the time until absorption of the Markov chain $V(t)$. Thus the expected waiting time of this particular customer is given by the column vector,

$$E_V^{(j)} = [-\tilde{A}_1^{-1}(I + \sum_{i=1}^{j-1}(-\tilde{A}_2\tilde{A}_1^{-1})^i]\mathbf{e}.$$

The second moment of waiting time of the tagged customer is given by the column vector $E_{V^2}^j$ which is the first block of the matrix $2(-\tilde{V})^{-2}\mathbf{e}$. Hence the expected waiting time of a general customer in the queue is,

$$V_L = \sum_{j=1}^{\infty} x(j) E_V^{(j)}.$$

The second moment of V is

$$V_L^{(2)} = \sum_{j=1}^{\infty} x(j) E_{V^2}^{j}.$$

2.2.4 Performance measures

In this section we list a number of key system performance measures to bring out the qualitative aspects of the model under study. These are listed below along with their formulae for computation. Towards this end, we further partition the vectors $\boldsymbol{x}_i$, $i \geq 1$ as

$$\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$$

and

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}), i \geq 2.$$

Note that $\boldsymbol{x}_0$, $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{1\tilde{0}}$, $\boldsymbol{x}_{11}$, $\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$ and $\boldsymbol{x}_{i3}$ are vectors of dimensions r, $M_1 ar$, $(K+1)br$, $Kbd(f+1)r$, $M_1(K+1)abr$, $Kd(f+1)br$, $M_0 Kd(f+1)abr$, $K(f+1)abr$ respectively.

(1) Expected number of customers in the system

$$ES = \sum_{i=1}^{\infty} i \boldsymbol{x}_i \mathbf{e}. \tag{2.11}$$

(2) Expected number of customers in the queue

$$EQ = \sum_{i=2}^{\infty} (i-1) \boldsymbol{x}_{i1} \mathbf{e} + \sum_{i=3}^{\infty} (i-2)(\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i2}\mathbf{e} + \boldsymbol{x}_{i3}\mathbf{e}). \tag{2.12}$$

(3) Effective rate of consultation

$$ECo = \theta \sum_{j=0}^{K-1} \boldsymbol{x}_{1\tilde{0}j}\mathbf{e} + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \boldsymbol{x}_{i0j}\mathbf{e}. \tag{2.13}$$

(4) Effective rate of interruption

$$EI = \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \boldsymbol{x}_{i0j0}\mathbf{e} + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=1}^{M-1} \sum_{l_1=1}^{c} \boldsymbol{x}_{i0jkl_1}\mathbf{e} \tag{2.14}$$

(5) Fraction of time the main server is idle

$$F_{mi} = \boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_{1\tilde{0}}\mathbf{e}. \tag{2.15}$$

(6) Fraction of time the regular server is idle

$$F_{ri} = \boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_{10}\mathbf{e}. \tag{2.16}$$

(7) Fraction of time the main server is busy serving a customer

$$F_{mb} = \boldsymbol{x}_{10}\mathbf{e} + \sum_{i=2}^{\infty} \boldsymbol{x}_{i0}\mathbf{e} + \sum_{i=2}^{\infty} \boldsymbol{x}_{i3}\mathbf{e}. \tag{2.17}$$

(8) Fraction of time the regular server is busy serving a customer

$$F_{rb} = \boldsymbol{x}_{1\tilde{0}}\mathbf{e} + \sum_{i=2}^{\infty} \boldsymbol{x}_{i0}\mathbf{e}. \tag{2.18}$$

(9) Fraction of time regular server is getting consultation

$$F_{rc} = \sum_{i=1}^{\infty} \boldsymbol{x}_{i1}\mathbf{e} + \sum_{i=2}^{\infty} \boldsymbol{x}_{i2}\mathbf{e}. \tag{2.19}$$

(10) Fraction of time regular server is waiting to get consultation

$$F_{wc} = \sum_{i=2}^{\infty} \boldsymbol{x}_{i3}\mathbf{e}. \tag{2.20}$$

(11) Fraction of time main server remains interrupted

$$F_{min} = \sum_{i=2}^{\infty} \boldsymbol{x}_{i2}\mathbf{e}. \tag{2.21}$$

(12) Rate at which interruption completion takes place before threshold is realised

$$R_I^c b = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^0 \boldsymbol{x}_{i2jkl_1l_2l_3}\mathbf{e}. \tag{2.22}$$

(13) Rate at which interruption completion takes place after threshold is realised

$$R_I^c a = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_2=1}^{d} D_{l_2}^0 \boldsymbol{x}_{i2jkl_1l_20}\mathbf{e}. \tag{2.23}$$

(14) Rate at which consultation completion takes place before threshold

is realised

$$R_C^c b = \sum_{i=1}^{\infty}\sum_{j=0}^{K-1}\sum_{l_2=1}^{d}\sum_{l_3=1}^{f} D_{l_2}^0 \boldsymbol{x}_{i1jl_2l_3}\mathbf{e} + R_I^c b. \tag{2.24}$$

(15) Rate at which consultation completion takes place after the threshold is realised

$$R_C^c a = \sum_{i=1}^{\infty}\sum_{j=0}^{K-1}\sum_{l_2=1}^{d} D_{l_2}^0 \boldsymbol{x}_{i1jl_20}\mathbf{e} + R_I^c a. \tag{2.25}$$

(16) Rate at which service completion at the main server takes place without any interruption

$$R_S^c wi = \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{100t_1}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{j=0}^{K}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0j0t_1}\mathbf{e}. \tag{2.26}$$

(17) Rate at which service completion (with at least one interruption) at the main server takes place before super clock is realised

$$R_S^c b = \sum_{i=2}^{\infty}\sum_{j=0}^{K}\sum_{k=1}^{M}\sum_{l_1=1}^{c}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0jkl_1t_1}\mathbf{e} + \sum_{j=0}^{K}\sum_{k=1}^{M}\sum_{l_1=1}^{c}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{10jkl_1t_1}\mathbf{e}. \tag{2.27}$$

(18) Rate at which service completion (with at least one interruption) at the main server takes place after super clock is realised

$$R_S^c a = \sum_{i=2}^{\infty}\sum_{j=0}^{K}\sum_{k=1}^{M}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0jk0t_1}\mathbf{e} + \sum_{j=0}^{K}\sum_{k=1}^{M}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{10jk0t_1}\mathbf{e}. \tag{2.28}$$

(19) Rate at which service completion at the regular server takes place without any consultation

$$R_S^c wc = \sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{1\tilde{0}0t_2}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{t_1=1}^{a}\sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{i000t_1t_2}\mathbf{e}$$

$$+ \sum_{i=2}^{\infty}\sum_{k=1}^{M}\sum_{l_1=0}^{c}\sum_{t_1=1}^{a}\sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{i00kl_1t_1t_2}\mathbf{e}. \quad (2.29)$$

(20) Rate at which service completion (with at least one consultation) at the regular server takes place

$$R_S^c c = \sum_{j=1}^{K}\sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{1\tilde{0}jt_2}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{j=1}^{K}\sum_{t_1=1}^{a}\sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{i0j0t_1t_2}\mathbf{e}$$

$$+ \sum_{i=2}^{\infty}\sum_{j=1}^{K}\sum_{k=1}^{M}\sum_{l_1=0}^{c}\sum_{t_1=1}^{a}\sum_{t_2=1}^{b} U_{t_2}^0 \boldsymbol{x}_{i0jkl_1t_1t_2}\mathbf{e}. \quad (2.30)$$

2.3 Numerical results

For the arrival process we consider the following five sets of matrices for L_0 and L_1.

(i) Erlang (ERA) $L_0 = \begin{bmatrix} -5 & 5 & \\ & -5 & 5 \\ & & -5 \end{bmatrix}$, $L_1 = \begin{bmatrix} & & \\ & & \\ 5 & & \end{bmatrix}$.

(ii) Exponential (EXA)

$L_0 = [-1]$, $L_1 = [1]$

(iii) Hyper Exponential (HEA)

$$L_0 = \begin{bmatrix} -10 & 0 \\ 0 & -1 \end{bmatrix}, L_1 = \begin{bmatrix} 9 & 1 \\ 0.9 & 0.1 \end{bmatrix}.$$

(iv) MAP with negative correlation (MNA)

$$L_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{bmatrix}, L_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.02 & 0 & 1.98 \\ 445.995 & 0 & 4.505 \end{bmatrix}.$$

(v) MAP with positive correlation (MPA)

$$L_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{bmatrix}, L_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1.98 & 0 & 0.02 \\ 4.505 & 0 & 445.995 \end{bmatrix}.$$

All these five MAP processes are normalized so as to have an arrival rate of 4. However, these are qualitatively different in that they have different variances and correlation structures. The first three arrival processes, namely ERA, EXA, and HEA, correspond to renewal processes and so the correlation is 0. The arrival process labelled MNA has correlated arrivals with correlation between two successive inter-arrival times given by -0.4889 and the arrival process corresponding to the one labelled MPA has a positive correlation with value 0.4889. The ratio of the standard deviations of the inter-arrival times of these five arrival processes with respect to ERA are, respectively, 1, 2.2361, 5.0194, 3.1518, and 3.1518.

The purpose of this example to see how various performance measures behave under different scenario.

Let

$T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$, $U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$, $D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix}$,
$E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}$, $G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}$,
$\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$, $\boldsymbol{\delta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$, $\boldsymbol{\eta} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$,
$\boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$, $K = 3$, $M = 3$.

We choose the above matrices, vectors and values so that the stability condition $\rho_1 < 1$ is not violated.

We look at the effect of varying θ on the performance measures ρ_1, ES, EQ, EI and ECo. From the table 2.1 we can see that as θ increases the traffic intensity also increases. This results in a rapid accumulation of customers in system and in queue. Thus ES and EQ increase. The effective rates for interruption EI and for consultation ECo also increase as θ increases.

2.4 Description of model 2

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 moment the super clock realises and the service at the main server will be restarted or resumed according to the threshold clock. The regular server has to wait until the service completion at the main server to get the remaining consultation. After consultation, the regular server resumes or restarts the service in accordance with the threshold clock. Thus the total

Table 2.1: Effect of θ on various performance measures

	θ	ERA	EXA	HEA	MNA	MPA
ρ_1	1.0	0.3697	0.3808	0.3623	0.399	0.399
	1.5	0.3829	0.3988	0.3724	0.4257	0.4257
	2.0	0.3963	0.4169	0.3826	0.4526	0.4526
	2.5	0.4098	0.4349	0.3924	0.4796	0.4796
	3.0	0.4234	0.4529	0.4029	0.5068	0.5068
ES	1.0	2.4097	2.9833	4.6033	3.4646	0.9471
	1.5	4.2763	5.1493	7.2913	5.822	1.1145
	2.0	7.2931	8.2445	9.9817	9.0074	1.3083
	2.5	10.962	11.5465	12.013	12.2457	1.5302
	3.0	13.2435	13.4633	13.0224	14.0085	1.7821
EQ	1.0	1.4299	1.9467	3.5202	2.3434	0.1685
	1.5	3.1185	3.9301	6.0511	4.5111	0.2788
	2.0	5.9646	6.8711	8.6464	7.5533	0.4129
	2.5	9.5381	10.1048	10.6604	10.7442	0.5733
	3.0	11.8909	12.0959	11.7247	12.6007	0.7622
EI	1.0	0.1505	0.201	0.2245	0.2193	0.0489
	1.5	0.2389	0.3095	0.3128	0.3161	0.0724
	2.0	0.3269	0.4076	0.3729	0.3938	0.0954
	2.5	0.3906	0.4717	0.3999	0.4362	0.1180
	3.0	0.3964	0.4765	0.3964	0.4271	0.1403
ECo	1.0	0.2356	0.2755	0.3079	0.3456	0.0898
	1.5	0.3760	0.4221	0.4477	0.5002	0.1379
	2.0	0.5182	0.5571	0.5564	0.6282	0.1874
	2.5	0.6271	0.6502	0.6210	0.7058	0.2378
	3.0	0.6492	0.6660	0.6392	0.7059	0.2888

time duration of interruption to the main server is equal to the duration of the super clock whereas in model 1, the duration of interruption can be greater than the duration of super clock. The assumption in model 2 reduces the duration of interruption at the main server and so reduces the effective service time of a customer at the main server. (Effective service time at the main server includes the actual time necessary to get served and the sum of the durations of intermediate interruptions.)

Notations :- We use the following notations in this model.

- $M_1 = Mc + 1$
- $\tilde{\boldsymbol{\alpha}} = e^{'}_{M_1}(1) \otimes \boldsymbol{\alpha}$
- $\tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}, 0)$
- $\tilde{E} = \begin{bmatrix} E & E^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$
- $D^* = D \oplus \tilde{E}$

Consider the queueing model $X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t)\}$.
$N(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t)$ and $U(t)$ have the same meaning as described in model 1.
$S(t)$ takes one more value '4' in addition to the values taken by that in model 1 of this chapter.
$S(t) = 4$ if the regular server is waiting to get the remaining part of consultation after the completion of the present service at the main server. This happens when the super clock expires in the midst of a consultation. As soon as the super clock expires during the process of an interruption, the main server restarts or resumes the service of the customer with him according to the status of the threshold. At this stage, the regular server has to wait until the service completion at the main server to get his consultation completed. Since the interruption will be removed at the time

the super clock saturates, the super clock saturation point '0' is not in the interruption state; but it is present in the busy state.
$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Phi = \bigcup_{i=0}^{\infty} \phi(i).$$

The terms $\phi(i)$'s are defined as
$\phi(0) = \{(0, u)\}$,
$\phi(1) = \phi(1, 0) \cup \phi(1, \tilde{0}) \cup \phi(1, 1)$ and
$\phi(i) = \phi(i, 0) \cup \phi(i, 1) \cup \phi(i, 2) \cup \phi(i, 3) \cup \phi(i, 4)$, for $i \geq 2$, where
$\phi(1, 0) = \{(1, 0, 0, t_1, u)\} \cup \{(1, 0, k, l_1, t_1, u) : 1 \leq k \leq M\}$,
$\phi(1, \tilde{0}) = \{(1, \tilde{0}, j, t_2, u) : 0 \leq j \leq K\}$,
$\phi(1, 1) = \{(1, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}$
$\phi(i, 0) = \{(i, 0, j, 0, t_1, t_2, u) \cup (i, 0, j, k, l_1, t_1, t_2, u) : 1 \leq k \leq M\}$,
for $0 \leq j \leq K$; and
$\phi(i, 1) = \{(i, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}$,
$\phi(i, 2) = \{(i, 2, j, k, l_1, l_2, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1, 0 \leq k \leq M - 1\}$,
$\phi(i, 3) = \{(i, 3, j, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1\}$,
$\phi(i, 4) = \{(i, 4, j, l_2, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1\}$,
for $i \geq 2$ with $1 \leq l_1 \leq c$, $1 \leq l_2 \leq d, 0 \leq l_3 \leq f, 1 \leq t_1 \leq a$,
$1 \leq t_2 \leq b$, and $1 \leq u \leq r$.

The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} L_0 & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{2.31}$$

where

$$B_1 = \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix} \otimes L_1 \ , \ B_2 = \begin{bmatrix} \mathbf{e}_{M_1} \otimes T^0 \\ e_{K+1} \otimes U^0 \\ \mathbf{0} \end{bmatrix} \otimes I_r \ ,$$

$$B_3 = \begin{bmatrix} I_{M_1} \otimes T & & \\ & B_{31} & B_{32} \\ & B_{33} & I_K \otimes D^* \otimes I_b \end{bmatrix} \oplus L_0,$$

$$B_4 = \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix} \otimes L_1 \ , \ B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix} \otimes I_r,$$

$$A_0 = I \otimes L_1, \ A_1 = \begin{bmatrix} A_{11} & & A_{12} & A_{13} & \\ A_{14} & A_{15} & & & \\ A_{16} & & A_{17} & & A_{18} \\ & & & A_{19} & \\ & & & & A_{110} \end{bmatrix} \oplus L_0,$$

$$A_2 = \begin{bmatrix} A_{21} & A_{22} & O \end{bmatrix} \otimes I_r.$$

Here A_0, A_1 and A_2 are square matrices of order C_0, B_3 is a square matrix of order C_1 and B_1, B_2, B_4, B_5 are matrices of orders $r \times C_1$, $C_1 \times r$, $C_1 \otimes C_0$ and $C_0 \times C_1$ respectively, where

$C_0 = [M_1(K+1)ab + Kd(f+1)b + MKcd(f+1)ab + Kab + K(f+1)ab]r$,

and $C_1 = [M_1 a + (K+1)b + Kbd(f+1)]r$.

The blocks are defined as follows:

$$B_{31} = \begin{bmatrix} I_K \otimes (U - \theta I) & O \\ O & U \end{bmatrix}_{(K+1)b},$$

$$B_{32} = \theta \begin{bmatrix} I_K \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}} \\ O \end{bmatrix}_{(K+1)b \times Kd(f+1)} \otimes I_b,$$

$$B_{33} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \tilde{\Delta}_b \end{bmatrix}_{Kd(f+1)\times(K+1)b},$$

$$B_{41} = \begin{bmatrix} \mathbf{e}'_{K+1}(1) \otimes I_{M_1} \otimes I_a \otimes \boldsymbol{\beta} \\ I_{K+1} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ O \end{bmatrix}_{C_1 \times (K+1)M_1ab}, B_{42} = \begin{bmatrix} O \\ I \end{bmatrix}_{C_1 \times Kbd(f+1)},$$

$$B_{51} = \begin{bmatrix} \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_0 \times M_0 a},$$

$$B_{52} = \begin{bmatrix} I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_0 \times (K+1)b}, \ B_{53} = \begin{bmatrix} O \\ F \end{bmatrix}_{C_0 \times Kbd(f+1)},$$

$$A_{11} = \begin{bmatrix} I_K \otimes I_{M_1} \otimes (T \oplus U - \theta I) & O \\ O & I_{M_1} \otimes (T \oplus U) \end{bmatrix}_{M_1(K+1)ab},$$

$$A_{12} = \theta \begin{bmatrix} I_K \otimes P \\ O \end{bmatrix}_{M_1(K+1)\times McKd(f+1)} \otimes I_{ab},$$

$$A_{13} = \theta \begin{bmatrix} I_K \otimes P^* \\ O \end{bmatrix}_{M_1(K+1)\times K} \otimes I_{ab},$$

$$A_{14} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \Delta^0 \end{bmatrix}_{Kd(f+1)b \times M_1(K+1)ab},$$

$$A_{15} = I_K \otimes D^* \otimes I_b,$$

$$A_{16} = \begin{bmatrix} O & I_K \otimes \Delta^* \end{bmatrix}_{Mcd(f+1)ab \times M_1(K+1)ab},$$

$$A_{17} = I_K \otimes I_M \otimes (G \oplus D \oplus \tilde{E}) \otimes I_{ab},$$

$$A_{18} = I_K \otimes \mathbf{e}_M \otimes G^0 \otimes \mathbf{e}_d \otimes \bar{\Delta},$$

$$A_{19} = I_K \otimes T \otimes I_b, \ A_{110} = I_{K(f+1)} \otimes T \otimes I_b,$$

$$A_{21} = \begin{bmatrix} \tilde{F} \\ O \end{bmatrix}_{C_0 \times Kd(f+1)b}, \; A_{22} = \begin{bmatrix} O \\ F \\ \hat{F} \end{bmatrix}.$$

Here

$F = I_K \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}} \otimes T^0 \otimes I_b$,

$P = diag(\boldsymbol{\gamma}, I_{M-1} \otimes I_c) \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}}$, $P^* = \begin{bmatrix} 0 \\ \mathbf{e}_M \otimes \mathbf{e}_c \end{bmatrix}$,

$\Delta^* = \begin{bmatrix} O & I_M \otimes I_c \otimes D^0 \end{bmatrix} \otimes \tilde{\Delta}$,

$\bar{\Delta} = \begin{bmatrix} \mathbf{e}_f \otimes I_{ab} & 0 \\ O & \mathbf{e}_a \otimes \boldsymbol{\alpha} \otimes I_b \end{bmatrix}$,

$\tilde{F} = I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes \boldsymbol{\alpha} \otimes I_b + \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \boldsymbol{\beta}$,

$\hat{F} = I_K \otimes diag(I_f \otimes T^0 \otimes I_b, T^0 \otimes \mathbf{e}_b \otimes \boldsymbol{\beta})$,

$$\tilde{\Delta}_b = \begin{bmatrix} \mathbf{e}_f \otimes I_b \\ \mathbf{e}_b \otimes \boldsymbol{\beta} \end{bmatrix}, \; \Delta^0 = \begin{bmatrix} \mathbf{e}_f \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ \mathbf{e}_b \otimes \tilde{\boldsymbol{\alpha}} \otimes \boldsymbol{\beta} \end{bmatrix}, \; \tilde{\Delta} = \begin{bmatrix} \mathbf{e}_f \otimes I_{ab} \\ \mathbf{e}_{ab} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}.$$

P and P^* are matrices of orders $M_1 \times Mcd(f+1)$ and $M_1 \times 1$ respectively.

2.5 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.

2.5.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. That is, $\boldsymbol{\pi}(A_0 + A_1 + A_2) = 0$; $\boldsymbol{\pi}\mathbf{e} = 1$.
The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{2.32}$$

is satisfied. That is, the rate of drift to the left has to be higher than that to the right. The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model, and hence the stability condition is known only implicitly. If we partition the vector $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3, \boldsymbol{\pi}_4)$$

and then using the structure of the matrices A_0 and A_2, (2.32) is given by

$$\lambda < \boldsymbol{\pi}_0 \tilde{F}\mathbf{e} + \boldsymbol{\pi}_3 F\mathbf{e} + \boldsymbol{\pi}_4 \hat{F}\mathbf{e}. \tag{2.33}$$

For future reference, we define the traffic intensity ρ_2 as

$$\rho_2 = \frac{\boldsymbol{\pi} A_0 \mathbf{e}}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{2.34}$$

Note that the stability condition in (2.32) is equivalent to $\rho_2 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.6.

2.5.2 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$.
Note that $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$, for $i \geq 2$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3 \tag{2.35}$$

where R is the minimal non-negative solution of the matrix equation $R^2A_2 + RA_1 + A_0 = 0$. Knowing the matrix R, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained by solving the boundary equations

$$\begin{aligned} \boldsymbol{x}_0 L_0 + x_1 B_2 &= 0 \\ \boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_3 + \boldsymbol{x}_2 B_5 &= 0 \\ \boldsymbol{x}_1 B_4 + \boldsymbol{x}_2 (A_1 + RA_2) &= 0 \end{aligned}$$

subject to the normalizing condition
$\boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_1\mathbf{e} + \boldsymbol{x}_2(I - R)^{-1}\mathbf{e} = 1.$

2.5.3 Performance measures

In this section we list a number of key system performance measures to bring out the qualitative aspects of the model under study. These are listed below along with their formulae for computation. Towards this end, we further partition the vectors $\boldsymbol{x}_i$, $i \geq 1$ as

$$\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$$

and

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4}), i \geq 2.$$

Note that $\boldsymbol{x}_0$, $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{1\tilde{0}}$, $\boldsymbol{x}_{11}$, $\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$, $\boldsymbol{x}_{i3}$ and $\boldsymbol{x}_{i4}$ are vectors of dimensions r, M_1ar, $(K+1)br$, $Kbd(f+1)r$, $M_1(K+1)abr$, $Kd(f+1)br$, $MKcd(f+1)abr$, $K(f+1)abr$ and $Kabr$ respectively.
Even though the vectors $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, etc. in model 2 are different from those vectors in model 1, the expressions for ES, ECo, EI, F_{mi}, F_{ri}, F_{rb}, F_{rc}, F_{min}, R_C^cb, R_C^ca, R_S^cwi, R_S^cb, R_S^cwc and R_S^cc are similar to those in model 1. These values are obtained by using the equations (2.11), (2.13), (2.14), (2.15), (2.16), (2.18), (2.19), (2.21), (2.24), (2.25), (2.26), (2.27), (2.29) and (2.30). We get the following performance measures also.

(1) Expected number of customers in the queue

$$EQ = \sum_{i=2}^{\infty}(i-1)\boldsymbol{x}_{i1}\mathbf{e} + \sum_{i=3}^{\infty}(i-2)[\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i2}\mathbf{e} + \boldsymbol{x}_{i3}\mathbf{e} + \boldsymbol{x}_{i4}\mathbf{e}]. \quad (2.36)$$

(2) Fraction of time the main server is busy serving a customer

$$F_{mb} = \boldsymbol{x}_1\mathbf{e} + \sum_{i=2}^{\infty}\boldsymbol{x}_{i0}\mathbf{e} + \sum_{i=2}^{\infty}\boldsymbol{x}_{i3}\mathbf{e} + \sum_{i=2}^{\infty}\boldsymbol{x}_{i4}\mathbf{e}. \quad (2.37)$$

(3) Fraction of time regular server is waiting to get consultation

$$F_{wc} = \sum_{i=2}^{\infty} \boldsymbol{x}_{i3}\mathbf{e} + \sum_{i=2}^{\infty} \boldsymbol{x}_{i4}\mathbf{e}. \tag{2.38}$$

(4) Rate at which interruption completion takes place before threshold is realised

$$R_I^c b = \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{k=0}^{M-1}\sum_{l_1=1}^{c}\sum_{l_2=1}^{d}\sum_{l_3=1}^{f} D_{l_2}^0 \boldsymbol{x}_{i2jkl_1l_2l_3}\mathbf{e}. \tag{2.39}$$

(5) Rate at which interruption completion takes place after threshold is realised

$$R_I^c a = \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{k=0}^{M-1}\sum_{l_1=1}^{c}\sum_{l_2=1}^{d} D_{l_2}^0 \boldsymbol{x}_{i2jkl_1l_20}\mathbf{e}. \tag{2.40}$$

(6) Rate at which service completion (with at least one interruption) at the main server takes place after super clock is realised

$$R_S^c a = \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{l_2=1}^{d}\sum_{l_3=1}^{f}\sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i4jl_2l_3t_1}\mathbf{e}. \tag{2.41}$$

2.6 Numerical results

We consider the arrival processes ERA, EXA, HEA, MNA, MPA defined in the example of model 1. The purpose of this example to see how various performance measures behave under different scenario. Choose the matrices, vectors and values so that the stability condition $\rho_2 < 1$ is satisfied. We fix

$$T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}, U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}, D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix},$$

$$E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \boldsymbol{\delta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix},$$

$$\boldsymbol{\eta} = \begin{bmatrix} 0.5 & 00.5 \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}, K = 3, M = 3.$$

We look at the effect of varying θ on the performance measures ρ_2, ES, EQ, EI and ECo.

Table 2.2: Effect of θ on various performance measures

	θ	ERA	EXA	HEA	MNA	MPA
ρ_2	1.0	0.6043	0.6043	0.6043	0.6043	0.6043
	1.5	0.7180	0.7180	0.7180	0.7180	0.7180
	2.0	0.8227	0.8227	0.8227	0.8227	0.8227
	2.5	0.9186	0.9186	0.9186	0.9186	0.9186
ES	1.0	2.3627	2.9628	4.8839	3.4395	0.9215
	1.5	4.2166	5.2298	8.1508	5.9142	1.0628
	2.0	7.6748	8.9864	11.4539	7.5644	1.2185
	2.5	12.4492	13.0625	12.9123	13.692	1.3877
EQ	1.0	1.4552	2.0117	3.8925	2.4205	0.1676
	1.5	3.1734	4.1368	7.0363	4.7486	0.2670
	2.0	6.4915	7.7686	10.2969	6.5255	0.3805
	2.5	11.2291	11.8550	11.8539	12.4519	0.5075
EI	1.0	0.3491	0.4253	0.5211	0.5054	0.1103
	1.5	0.5842	0.6772	0.7718	0.7702	0.1692
	2.0	0.8502	0.9295	0.9447	0.8043	0.2290
	2.5	1.0411	1.0612	0.9605	1.1117	0.2889
ECo	1.0	0.2575	0.1783	0.0904	0.0700	0.5228
	1.5	0.3032	0.2025	0.0963	0.0757	0.7723
	2.0	0.2819	0.1821	0.0850	0.1336	1.0139
	2.5	0.1979	0.1273	0.06550	0.0433	1.2478

From the table 2.2 we can see that as θ increases the traffic intensity also increases. This results in rapid accumulation of customers in system and in queue. Thus ES and EQ increase. The effective rates for interruption EI and for consultation ECo also increase as θ increases.

2.7 Description of model 3

In a queueing system where the service process consists of certain number of phases, with service subject to interruptions, the concept of protecting a few phases of service (which may be so costly to afford an interruption) from interruption could be an important idea. Klimenok et. al. [29] studied a multi-server queueing system with finite buffer and negative customers where the arrival is BMAP and service is PH-type. 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. Klimenok and Dudin [28] extended the above paper by considering disciplines of complete admission and complete rejection. Further, Klimenok and Dudin [28] assumed an infinite buffer. Krishnamoorthy et. al. [32] introduced the idea of protection in a queueing system where the service process is subject to interruptions. They assume that the final $m - n$ phases of the Erlang service process with m phases are protected from interruption. Whereas if the service process belongs to the first n phases, it is subject to interruption and an interrupted service is resumed/repeated after some random time. There is no reduction (removal) in the number of customers due to interruption and no bound was assumed on the number of interruptions that can possibly occur in the course of a service.

Notations :- We use the following notations in this model.

- $M_0 = M(c+1), M_1 = M_0 + 1$
- $\tilde{\boldsymbol{\alpha}} = \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha}$

- $\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, 0)$ and $\tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}, 0)$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$ and $\tilde{E} = \begin{bmatrix} E & E^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$
- $\boldsymbol{\delta}^* = \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}}$ and $\boldsymbol{\gamma}^* = \tilde{\boldsymbol{\gamma}} \otimes (\boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}})$
- $D^* = D \oplus \tilde{E}$ and $G^* = \tilde{G} \oplus D^*$
- $\dot{I} = \begin{bmatrix} \mathbf{0} & I_{M_0} \end{bmatrix}_{M_0 \times M_1}$
- $\bar{I}_m = \tilde{\boldsymbol{\eta}} \otimes \begin{bmatrix} O & O \\ O & I_{a-m} \end{bmatrix}_{a \times a}$
- $\tilde{\mathbf{e}}_c = \begin{bmatrix} \mathbf{e}_c \otimes \bar{I}_m \\ \tilde{\boldsymbol{\eta}} \otimes I_a \end{bmatrix}$
- $I_m^* = \begin{bmatrix} I_m \\ O \end{bmatrix}_{a \times m}$

In this model, we assume that out of the $'a'$ phases at the main server, $m \leq a$ phases have the property that no interruptions are allowed to the main server (and therefore to the customer being served at the main server) if the service is at any one of these phases. If the regular server needs a consultation at this time, he/she has to wait until the service at the main server is completed. All other assumptions are same as those in model 1 of this chapter. Thus if either the customer at the main server has already interrupted M times or the super clock has expired or the service at the main server is at any one of the last $a-m$ protected phases, then the regular server has to wait until the completion of the service of the present customer at the main server.

Consider the queueing model $X = \{X(t), t \geq 0\}$, where
$X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t)\}$.
$N(t)$, $S(t)$, $B_1(t)$, $B_2(t)$, $S_1(t)$, $S_2(t)$, $S_3(t)$, $J_1(t)$, $J_2(t)$ and $U(t)$ have the same meaning as those in model 1. Since there are no interruption from the $(m+1)^{th}$ phase onwards, these phases are not present when $S(t) = 2$ or 3.

$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \bigcup_{i=0}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as
$\psi(0) = \{(0, u)\}$,
$\psi(1) = \psi(1, 0) \cup \psi(1, \tilde{0}) \cup \psi(1, 1)$,
$\psi(i) = \psi(i, 0) \cup \psi(i, 1) \cup \psi(i, 2) \cup \psi(i, 3), i \geq 2$,

where

$$\begin{aligned}
\psi(1,0) &= \{(1,0,0,t_1,u)\} \cup \{(1,0,k,l_1,t_1,u) : 1 \le k \le M, 1 \le t_1 \le a\} \\
\psi(1,\tilde{0}) &= \{(1,\tilde{0},j,t_2,u) : 0 \le j \le K\} \\
\psi(1,1) &= \{(1,1,j,l_2,l_3,t_2,u) : 0 \le j \le K-1\} \\
\psi(i,0) &= \{(i,0,j,0,t_1,t_2,u) \cup (i,0,j,k,l_1,t_1,t_2,u) : 0 \le j \le K, \\
&\quad 1 \le k \le M, 1 \le t_1 \le a\}, \\
\psi(i,1) &= \{(i,1,j,l_2,l_3,t_2,u) : 0 \le j \le K-1\} \\
\psi(i,2) &= \{(i,2,j,k,l_1,l_2,l_3,t_1,t_2,u) : 0 \le j \le K-1, 0 \le k \le M-1, \\
&\quad 1 \le t_1 \le m\} \\
\psi(i,3) &= \{(i,3,j,l_3,t_1,t_2,u) : 0 \le j \le K-1, 1 \le t_1 \le a\}, \\
&\quad \text{for } 0 \le l_1 \le c, 1 \le l_2 \le d, 0 \le l_3 \le f, 1 \le t_2 \le b, 1 \le u \le r.
\end{aligned}$$

The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} L_0 & B_1 & & & \\ B_2 & B_3 & B_4 & & \\ & B_5 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{2.42}$$

Here A_0, A_1 and A_2 are square matrices of order C_0; B_3 is a square matrix of order C_1 and B_1, B_2, B_4, B_5 are matrices of orders $r \times C_1$, $C_1 \times r$, $C_1 \times C_0$ and $C_0 \times C_1$, respectively, where
$C_0 = [M_1(K+1)ab + Kd(f+1)b + M_0Kd(f+1)mb + K(f+1)mb]r$ and

$C_1 = [M_1a + (K+1)b + Kbd(f+1)]r$ and these matrices are defined as follows:

$$B_1 = \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix} \otimes L_1,\ B_2 = \begin{bmatrix} \mathbf{e}_{M_1} \otimes T^0 \\ \mathbf{e}_{K+1} \otimes U^0 \\ \mathbf{0} \end{bmatrix} \otimes I_r,$$

$$B_3 = \begin{bmatrix} I_{M_1} \otimes T & O & O \\ O & B_{31} & B_{32} \\ O & B_{33} & I_K \otimes D^* \otimes I_b \end{bmatrix} \oplus L_0,$$

$$B_4 = \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix} \otimes L_1,\ B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix} \otimes I_r,$$

$$A_0 = I \otimes L_1,\ A_1 = \begin{bmatrix} A_{11} & O & A_{12} & A_{13} \\ A_{14} & A_{15} & O & O \\ A_{16} & O & A_{17} & O \\ O & O & O & A_{18} \end{bmatrix} \oplus L_0,$$

$$A_2 = \begin{bmatrix} A_{21} & B_{53} & O \end{bmatrix} \otimes I_r.$$

Here the block matrices are

$$B_{31} = I_{K+1} \otimes U - \theta \begin{bmatrix} I_K & \mathbf{0} \\ \underline{\mathbf{0}} & 0 \end{bmatrix}_{(K+1)} \otimes I_b,$$

$$B_{32} = \theta \begin{bmatrix} I_K \\ \underline{\mathbf{0}} \end{bmatrix}_{(K+1)\times K} \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}} \otimes I_b,$$

$$B_{33} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \tilde{\Delta}_b \end{bmatrix}_{Kd(f+1)b\times(K+1)b},$$

$$B_{41} = \begin{bmatrix} \mathbf{e}'_{K+1}(1) \otimes I_{M_1} \otimes I_a \otimes \boldsymbol{\beta} \\ I_{K+1} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ O \end{bmatrix}_{C_1\times(K+1)M_1ab},$$

$$B_{42} = \begin{bmatrix} O \\ I_{Kd(f+1)b} \end{bmatrix}_{C_1 \times Kd(f+1)b},$$

$$B_{51} = \begin{bmatrix} \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_0 \times M_0 a},$$

$$B_{52} = \begin{bmatrix} I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_0 \times (K+1)b},$$

$$B_{53} = \begin{bmatrix} O \\ I_K \otimes \boldsymbol{\delta} \otimes I_{f+1} \otimes T^0 \otimes I_b \end{bmatrix}_{C_0 \times Kd(f+1)b},$$

$$A_{11} = I_{K+1} \otimes I_{M_1} \otimes (T \oplus U) - \theta \begin{bmatrix} I_K & \mathbf{0} \\ \underline{\mathbf{0}} & 0 \end{bmatrix} \otimes I_{M_1} \otimes I_{ab},$$

$$A_{12} = \theta \begin{bmatrix} I_K \\ \underline{\mathbf{0}} \end{bmatrix}_{(K+1) \times K} \otimes P \otimes I_m^* \otimes I_b,$$

$$A_{13} = \theta \begin{bmatrix} I_K \\ \underline{\mathbf{0}} \end{bmatrix}_{(K+1) \times K} \otimes P^* \otimes I_b,$$

$$A_{14} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \Delta^0 \end{bmatrix}_{Kd(f+1)b \times M_1(K+1)ab},$$

$$A_{15} = I_K \otimes D^* \otimes I_b,$$

$$A_{16} = \begin{bmatrix} O & I_K \otimes \dot{I} \otimes D^0 \otimes \tilde{\Delta} \end{bmatrix}_{M_0 Kd(f+1)rb \times M_1(K+1)ab},$$

$$A_{17} = I_K \otimes I_M \otimes G^* \otimes I_{mb},\ A_{18} = I_{Kd} \otimes (\tilde{E} \oplus T) \otimes I_b,$$

$$A_{21} = \begin{bmatrix} \tilde{T}^0 + \tilde{U}^0 \\ O \end{bmatrix}_{C_0 \times M_1(K+1)ab}.$$

Here

$$P = \begin{bmatrix} diag(\tilde{\boldsymbol{\gamma}}, I_{M-1} \otimes \hat{I}_c) \\ O \end{bmatrix}_{M_1 \times M_0} \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}},$$

$$P^* = \begin{bmatrix} \bar{I}_m \\ \mathbf{e}_{M-1} \otimes \tilde{\mathbf{e}_c} \\ \mathbf{e}_{c+1} \otimes \tilde{\boldsymbol{\eta}} \otimes I_a \end{bmatrix}_{M_1 a \times (f+1)a},$$

$$\tilde{T}^0 = I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes \boldsymbol{\alpha} \otimes I_b \ , \tilde{U}^0 = \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \boldsymbol{\beta} \ ,$$

$$\tilde{\Delta}_b = \begin{bmatrix} \mathbf{e}_f \otimes I_b \\ \mathbf{e}_b \otimes \boldsymbol{\beta} \end{bmatrix}, \Delta^0 = \begin{bmatrix} \mathbf{e}_f \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ \mathbf{e}_b \otimes \tilde{\boldsymbol{\alpha}} \otimes \boldsymbol{\beta} \end{bmatrix}, \tilde{\Delta} = \begin{bmatrix} \mathbf{e}_f \otimes (I_m^*)' \otimes I_b \\ \mathbf{e}_{mb} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}.$$

2.8 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.

2.8.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. That is, $\boldsymbol{\pi}(A_0 + A_1 + A_2) = 0$; $\boldsymbol{\pi}\mathbf{e} = 1$.
The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{2.43}$$

That is, the rate of drift to the left has to be higher than that to the right. The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model.

For future reference, we define the traffic intensity ρ_3 as

$$\rho_3 = \frac{\pi A_0 \mathbf{e}}{\pi A_2 \mathbf{e}}. \tag{2.44}$$

Note that the stability condition in (2.43) is equivalent to $\rho_3 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.9.

2.8.2 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$.
Note that $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3})$, for $i \geq 2$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3 \tag{2.45}$$

where R is the minimal non-negative solution of the matrix equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{2.46}$$

Knowing the matrix R, the vectors $\boldsymbol{x}_0$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained by solving the equation

$$\begin{bmatrix} \boldsymbol{x}_0 & \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} \begin{bmatrix} L_0 & B_1 & \\ B_2 & B_3 & B_4 \\ & B_5 & A_1 + RA_2 \end{bmatrix} = 0, \tag{2.47}$$

subject to the normalizing condition

$$\boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_1\mathbf{e} + \boldsymbol{x}_2(I - R)^{-1}\mathbf{e} = 1. \tag{2.48}$$

2.8.3 Expected waiting time in queue

For computing expected waiting time in queue of a particular customer who joins as the k^{th} customer, where $k > 0$, in the queue, we consider the Markov process

$$Z(t) = \{(\tilde{N}(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t)) : t \geq 0\},$$

where $\tilde{N}(t)$ is the rank of the customer and all other variables defined as earlier. The rank $\tilde{N}(t)$ of the customer is assumed to be i if he is the i^{th} customer in the queue at time t. His rank may decrease to 1 as the customers ahead of him leave the system either after completing their services (if $S(t) = 0$) or completing the consultation (if $S(t) = 1$). Since the customers who arrive after the tagged customer cannot change his rank, level-changing transitions in $Z(t)$ can only take place to one side of the diagonal. The absorbing state Δ_2 denote the tagged customer is selected for service. Thus the infinitesimal generator $\tilde{V}$ of the process $Z(t)$ takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ \underline{\mathbf{0}} & \mathbf{0} \end{bmatrix},$$

where $V = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 & & & & \\ & \tilde{A}_1 & \tilde{A}_2 & & & \\ & & \tilde{A}_1 & \tilde{A}_2 & & \\ & & & \ddots & \ddots & \\ & & & & \tilde{A}_1 & \tilde{A}_2 \\ & & & & & \tilde{A}_1 \end{bmatrix}$ and $V^0 = \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_{M_1} \otimes (T^0 \oplus U^0) \\ \mathbf{0} \\ T^0 \otimes \mathbf{e}_b \end{bmatrix}$,

with

$\tilde{A}_1 = A_1^* - U_2$ and $\tilde{A}_2 = A_2^* + U_2$, where A_1^* and A_2^* are obtained from A_1 and A_2 if they are written as $A_1 = A_1^* \oplus L_0$ and $A_2 = A_2^* \otimes I_r$. Here

$U_2 = \begin{bmatrix} O & O \\ A_{14} & O \\ O & O \end{bmatrix}$.

Now, the waiting time V of a customer, who joins the queue as the j^{th} customer is the time until absorption of the Markov chain $V(t)$. Thus the expected waiting time of this particular customer is given by the column vector,

$$E_V^{(j)} = \{-\tilde{A}_1^{-1}[I + \sum_{i=1}^{j-1}(-\tilde{A}_2\tilde{A}_1^{-1})^i]\}\mathbf{e}.$$

The second moment of waiting time of the tagged customer is given by the column vector $E_{V^2}^j$ which is the first block of the matrix $2(-\tilde{V})^{-2}\mathbf{e}$. Hence the expected waiting time of a general customer in the queue is,

$$V_L = \sum_{j=1}^{\infty} x(j) E_V^{(j)}.$$

The second moment of V is

$$V_L^{(2)} = \sum_{j=1}^{\infty} x(j) E_{V^2}^{j}.$$

2.8.4 Performance measures

The vectors $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, etc. in model 3 are different from those vectors in model 1. The expressions for ES, EQ, ECo, F_{mi}, F_{ri}, F_{mb}, F_{rb}, F_{rc}, F_{wc}, F_{min}, $R_C^c b$, $R_C^c a$, $R_S^c wi$, $R_S^c b$, $R_S^c a$, $R_S^c wc$ and $R_S^c c$ etc. are similar to those in model 1. These values are obtained by using the equations (2.11), (2.12),(2.13), (2.15), (2.16), (2.17),(2.18), (2.19), (2.20),(2.21), (2.24), (2.25), (2.26), (2.27), (2.28), (2.29) and (2.30).

Note that $\boldsymbol{x}_0$, $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11}, \boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}$, for $i \geq 2$ are vectors of dimensions r, M_1ar, $(K+1)br$, $Kbd(f+1)r$, $(K+1)M_1abr$, $Kd(f+1)br$, $M_0Kd(f+1)mbr$ and $K(f+1)mbr$, respectively.
We get the following performance measures also.

(1) Effective rate of interruption

$$EI = \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{t_1=1}^{m} \boldsymbol{x}_{i0j0t_1}\mathbf{e} + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=1}^{M-1} \sum_{l_1=1}^{c} \sum_{t_1=1}^{m} \boldsymbol{x}_{i0jkl_1t_1}\mathbf{e}. \tag{2.49}$$

(2) Rate at which interruption completion takes place before threshold

is realised

$$R_I^c b = \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{k=0}^{M-1}\sum_{l_1=0}^{c}\sum_{l_2=1}^{d}\sum_{l_3=1}^{f}\sum_{t_1=1}^{m} D_{l_2}^{0}\ \boldsymbol{x}_{i2jkl_1l_2l_3t_1}\mathbf{e}. \tag{2.50}$$

(3) Rate at which interruption completion takes place after threshold is realised

$$R_I^c a = \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{k=0}^{M-1}\sum_{l_1=0}^{c}\sum_{l_2=1}^{d}\sum_{t_1=1}^{m} D_{l_2}^{0}\ \boldsymbol{x}_{i2jkl_1l_20t_1}\mathbf{e}. \tag{2.51}$$

2.9 Numerical results

Let us assume

$$T = \begin{bmatrix} -12 & 3 & 1 & 2 \\ 3 & -15 & 1 & 2 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & 2 & -7 \end{bmatrix},\ U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix},\ D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix},$$

$$E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix},\ G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix},\ \boldsymbol{\alpha} = \begin{bmatrix} 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix},\ \boldsymbol{\delta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix},\ \boldsymbol{\eta} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix},\ \boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix},$$

$K = 3,\ M = 3.$

We choose these matrices, vectors and values such that ρ_3 is less than 1.

Referring to Table 2.3, as the rate of consultation θ increases, the traffic intensity ρ_3 increases and hence EI and ECo will increase. This results in an increase in F_{min} and F_{rc}. As θ increases, consultation is more frequent, so the main server will reach the upper bounds of number of

Table 2.3: Effect of θ on various performance measures

$\lambda = 4$

θ	1	1.5	2	2.5	3
ρ_3	0.5911	0.6910	0.7835	0.8694	0.9492
ES	2.4103	3.2431	3.9898	4.6096	5.1017
EQ	1.4926	2.2582	2.9576	3.5459	4.0178
EI	0.0759	0.1139	0.1472	0.1748	0.1973
Eco	0.2585	0.3685	0.4600	0.5339	0.5929
F_{mi}	0.4021	0.3503	0.3102	0.2800	0.2572
F_{ri}	0.5487	0.4798	0.4265	0.3861	0.3558
F_{mb}	0.4355	0.4193	0.4021	0.3853	0.3697
F_{rb}	0.2591	0.2471	0.2327	0.2175	0.2030
F_{min}	0.0320	0.0486	0.0636	0.0765	0.0872
F_{rc}	0.1613	0.2271	0.2806	0.3228	0.3559
F_{rw}	0.0298	0.0427	0.0533	0.0616	0.0682

Table 2.4: Effect of λ on various performance measures

$\theta = 3$

λ	3	3.5	4	4.5	5
ρ_3	0.5877	0.6856	0.7835	0.8815	0.9794
ES	2.0385	2.9399	3.9898	5.0653	6.0266
EQ	1.2498	2.0195	2.9576	3.9487	4.8572
EI	0.0927	0.1206	0.1472	0.1699	0.1871
Eco	0.3242	0.3964	0.4600	0.5105	0.5454
F_{mi}	0.4510	0.3757	0.3102	0.2553	0.2103
F_{ri}	0.5977	0.508	0.4265	0.3558	0.2965
F_{mb}	0.3462	0.3777	0.4021	0.4187	0.4268
F_{rb}	0.1641	0.2006	0.2327	0.2581	0.2756
F_{min}	0.0405	0.0525	0.0636	0.0727	0.0791
F_{rc}	0.2024	0.2448	0.2806	0.3070	0.3228
F_{rw}	0.0355	0.0448	0.0533	0.0602	0.0651

interruptions rapidly or super clock may realise frequently and thus the main server compels to complete the service of the customer at him before further consultations and this results in more waiting time of the regular server to get consultation. So F_{rw} also increases. Since F_{min}, F_{rc} and F_{rw} increase, the customers have to stay in the system and in queue for a longer time and this results in an increase in ES and EQ. This in turn make a decrease in F_{mi} and F_{ri}. Since main server has to spend more time in consultation, it gets less time to serve customers. So F_{mb} and F_{rb} decreases.

Referring to Table 2.4, as the arrival rate λ increases, the traffic intensity ρ_3 increases. The system is fed with more and more customers and therefore accumulation of customers increases. So ES and EQ increase. Thus EI and ECo will also increase. This results in a hike in F_{min} and F_{rc}. Thus F_{rw} also increases. As the arrival rate increases, there are more customers in the queue and therefore the servers have to spend longer time in service. Thus F_{mb} and F_{rb} increase. This in turn make a decrease in F_{mi} and F_{ri}.

2.10 Comparison of the three models

Now we present a comparison of the three models analysed in this chapter. Recall that in model 1, the interruption is allowed to continue until its completion even when the super clock realises, whereas in model 2, the interruptions instantly terminated when the super clock realises. In model 3, some phases of service at the main server are protected from interruption. We compare the performance measures, namely, the traffic intensities ρ_1, ρ_2, ρ_3, expected number of customers in the system ES and expected number of customers in the queue EQ. Let expected number of customers in the system ES and expected number of customers in the queue EQ be denoted by ES_i and EQ_i for the respective models $i = 1, 2$ and 3. Here a model with least number of customers waiting in the queue EQ is considered as the most efficient one. We check the traffic intensities ρ_i's in each case. Let G_r denote the rate of the super clock. For convenience, we denote the models 1,2 and 3 as M_1, M_2 and M_3, respectively. Let us assume the following matrices and parameters:

$$T = \begin{bmatrix} -12 & 3 & 1 & 2 \\ 3 & -15 & 1 & 2 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & 2 & -7 \end{bmatrix}, U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}, D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix},$$

$$E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \boldsymbol{\delta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix},$$

$$\boldsymbol{\eta} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \lambda = 4, \theta = 1, M = 3 \text{ and } K = 3.$$

Table 2.5: Effect of G_r on various performance measures

G_r	ρ_1	ρ_2	ρ_3	ES_1	ES_2	ES_3	EQ_1	EQ_2	EQ_3
0.2	0.6207	0.5891	0.6338	3.2826	2.8073	3.1793	2.1836	1.7059	2.1690
0.5	0.6210	0.5560	0.6338	3.2850	2.3273	3.1792	2.1935	1.2732	2.1693
1	0.6213	0.5209	0.6338	3.2869	1.9438	3.1789	2.1935	0.9380	2.1695
2	0.6217	0.4838	0.6338	3.2883	1.6469	3.1785	2.1981	0.6900	2.1696
4	0.6220	0.4528	0.6338	3.2891	1.4657	3.1780	2.2015	0.5487	2.1697

Referring to table 2.5 we can see that as G_r increases, there is a slight increase in ρ_1, a considerable decrease in ρ_2 whereas ρ_3 remains a constant. We see that if G_r is high, the super clock realises at a faster rate and this will result in a faster completion of interruption and hence the service completion in case of model 2. Remember that service completion happens only at the main server and the regular server waits to get the remaining consultation after the main server's service completion. Thus in M_2, even though the regular server waits to get the remaining consultation, by this time the service at main server will be completed. This decreases ρ_2 of the system.

In M_2, the service at the main server will be restarted or resumed as soon as the super clock expires during an interruption, while in M_1 and M_3 the interruption will be continued even after the expiry of the super clock. In M_3 there are no interruption at all from some protected phases of service at the main server. Thus we get a comparison of expected number of customers in the system and in the queue as $ES_2 < ES_3 < ES_1$ and $EQ_2 < EQ_3 < EQ_1$.

As G_r has a rapid increase, ES_2 and EQ_2 decrease rapidly. But ES_1 and EQ_1 increase slightly as G_r increases. ES_3 has a slight decrease and EQ_3 has a negligible increase. This shows that the rate of the super clock G_r has a considerable effect in model M_2 when compared with the other two models. This is exactly what we are expected because in M_2, the interruption will be stopped as soon as the super clock expires, but the interruption will be continued until its completion in the other two models. So the rate of super clock has no direct effect on the values of ES and EQ in models M_1 and M_3.

Thus we can see that M_2 is the most efficient model and M_1 is the least one for the data in hand.

Chapter 3

Two-server queue with consultations in random and Markovian environments

In chapter 2, the consultation is due to a single factor and the duration of consultation follows phase type distribution. In this chapter, we consider consultations are due to L factors in random environment in model 1 and those in Markovian environment in model 2.

3.1 Description of model 1

Here we consider a service system with a main server and a regular server to which customers arrive according to a Poisson process with rate λ. The service times of customer at the main and regular servers have phase type distributions with representations ($\boldsymbol{\alpha}$,T) and ($\boldsymbol{\beta}$,U) respectively. Write $T^0 = -T\mathbf{e}$ and $U^0 = -U\mathbf{e}$ where $\mathbf{e}$ is a column vector of $1's$ of appropriate order. The main server offers consultation to the regular server whenever it is needed. Let $f_1, f_2, ..., f_L$ be L random environmental factors due to which consultations occur . Requirement of consultation in random environment is a Poisson process with rate θ where the i^{th} factor occurring with probability δ_i, i=1,2,...,L. The duration of consultation for the i^{th} factor is exponentially distributed with parameter ξ_i. Even though combinations of these factors are possible, in this chapter, we consider single factors only. The threshold clock determines the restart or resumption of services at both the servers. The duration of threshold clock has exponential distribution with parameter ω. The assumptions regarding status of servers, number of interruptions and consultations and super clock are same as those in model 1 of chapter 2.

Notations :- We use the following notations in this model.

- $M_0 = M(c+1)$ and $M_1 = M_0 + 1$
- $C_1 = M_1(K+1)ab + 2KLb + 2M_0KLab + 2KLab$
- $C_0 = M_1a + (K+1)b + 2KLb$
- $\dot{I}_1 = \begin{bmatrix} O & I_{M_0} \end{bmatrix}_{M_0 \times M_1}$
- $\tilde{\boldsymbol{\alpha}} = e'_{M_1}(1) \otimes \boldsymbol{\alpha}$
- $\boldsymbol{\eta}^* = (1, 0)$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$, $\tilde{E} = \omega \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$

- $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_L)'$

- $\tilde{\xi} = diag(\xi_1, \xi_2, ..., \xi_L)$

Consider the queueing model $X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), C(t), J_1(t), J_2(t)\}$.
Here $C(t)$ is the environmental factor due to which consultation is in progress/ waiting to get consultation and all other variables are as those in model 1 in chapter 2.
$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as
$\psi(1) = \psi(1, 0) \cup \psi(1, \tilde{0}) \cup \psi(1, 1)$ and
$\psi(i) = \psi(i, 0) \cup \psi(i, 1) \cup \psi(i, 2) \cup \psi(i, 3)$, for $i \geq 2$,
where
$\psi(1, 0) = \{(1, 0, 0, t_1)\} \cup \{(1, 0, k, l_1, t_1) : 1 \leq k \leq M\}$,
$\psi(1, \tilde{0}) = \{(1, \tilde{0}, j, t_2) : 0 \leq j \leq K\}$,
$\psi(1, 1) = \{(1, 1, j, l_2, l_3, t_2) : 0 \leq j \leq K - 1\}$,
and for $i \geq 2$,
$\psi(i, 0) = \{(i, 0, j, 0, t_1, t_2) \cup (i, 0, j, k, l_1, t_1, t_2) : 0 \leq j \leq K, 1 \leq k \leq M\}$,
$\psi(i, 1) = \{(i, 1, j, l_2, l_3, t_2) : 0 \leq j \leq K - 1\}$,
$\psi(i, 2) = \{(i, 2, j, k, l_1, l_2, l_3, t_1, t_2) : 0 \leq j \leq K - 1, 0 \leq k \leq M - 1\}$,
$\psi(i, 3) = \{(i, 3, j, l_2, l_3, t_1, t_2) : 0 \leq j \leq K - 1\}$,
with $0 \leq l_1 \leq c, l_2 = \{1, 0\}, 1 \leq l_3 \leq L, 1 \leq t_1 \leq a$ and $1 \leq t_2 \leq b$.

The infinitesimal generator matrix Q is given by

$$Q = \begin{bmatrix} -\lambda & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{3.1}$$

where the block matrices appearing in Q are as follows:

$B_1 = \lambda \begin{bmatrix} \tilde{\boldsymbol{\alpha}} & \underline{\mathbf{0}} \end{bmatrix}$, $B_2 = \begin{bmatrix} \mathbf{e}_{M_1} \otimes T^0 \\ \mathbf{e}_{K+1} \otimes U^0 \\ \mathbf{0} \end{bmatrix}$,

$B_3 = \begin{bmatrix} I_{M_1} \otimes T & O & O \\ O & B_{31} & B_{32} \\ O & B_{33} & B_{34} \end{bmatrix} - \lambda I$,

$B_4 = \lambda \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix}$, $B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix}$,

$A_0 = \lambda I$, $A_1 = \begin{bmatrix} A_{11} & O & A_{12} & A_{13} \\ A_{14} & B_{34} & O & O \\ A_{16} & O & A_{17} & O \\ O & O & O & A_{18} \end{bmatrix} - \lambda I$,

$A_2 = \begin{bmatrix} A_{21} & B_{53} & O \end{bmatrix}$.

Here A_0, A_1 and A_2 are square matrices of order C_1;
B_3 is a square matrix of order C_0 and B_1, B_2, B_4, B_5 are matrices of orders $1 \times C_0$, $C_0 \times 1$, $C_0 \times C_1$ and $C_1 \times C_0$, respectively.

Here

$$B_{31} = \begin{bmatrix} I_K \otimes (U - \theta I) & O \\ O & U \end{bmatrix}_{(K+1)b},$$

$$B_{32} = \theta \begin{bmatrix} I_K \\ O \end{bmatrix}_{(K+1)\times K} \otimes \boldsymbol{\eta}^* \otimes \boldsymbol{\delta} \otimes I_b,$$

$$B_{33} = \begin{bmatrix} O & I_K \otimes \Delta_b^{'} \end{bmatrix}_{2KLb\times(K+1)b}, \; B_{34} = I_K \otimes \tilde{\nabla} \otimes I_b,$$

$$B_{41} = \begin{bmatrix} \mathbf{e}_{K+1}^{'}(1) \otimes I_{M_1} \otimes I_a \boldsymbol{\beta} \\ I_{K+1} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ O \end{bmatrix}_{C_0\times(K+1)M_1ab},$$

$$B_{42} = \begin{bmatrix} O \\ I_{2KLb} \end{bmatrix}_{C_0\times 2KLb},$$

$$B_{51} = \begin{bmatrix} \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_1\times M_0a},$$

$$B_{52} = \begin{bmatrix} I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_1\times(K+1)b},$$

$$B_{53} = \begin{bmatrix} O \\ I_{KL} \otimes I_2 \otimes T^0 \otimes I_b \end{bmatrix}_{C_1\times 2KLb},$$

$$A_{11} = \begin{bmatrix} I_K \otimes I_{M_1} \otimes (T \oplus U - \theta I) & O \\ O & I_{M_1} \otimes (T \oplus U) \end{bmatrix}_{M_1(K+1)ab},$$

$$A_{12} = \theta \begin{bmatrix} I_K \otimes P \\ O \end{bmatrix}_{M_1(K+1)\times M_0K} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\eta}^* \otimes I_{ab},$$

$$A_{13} = \theta \begin{bmatrix} I_K \otimes P^* \\ O \end{bmatrix}_{M_1(K+1)\times K} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\eta}^* \otimes I_{ab},$$

$$A_{14} = \begin{bmatrix} O & I_K \otimes \hat{\Delta} \end{bmatrix}_{2KLb\times M_1(K+1)ab},$$

$$A_{16} = \begin{bmatrix} O & I_K \otimes \dot{I}_1 \otimes \Delta^* \end{bmatrix}_{2M_0KLab\times M_1(K+1)ab},$$

$A_{17} = I_K \otimes I_M \otimes (\tilde{G} \oplus \tilde{\nabla}) \otimes I_{ab}$, $A_{18} = I_{KL} \otimes (\tilde{E} \oplus T) \otimes I_b$,

$$A_{21} = \begin{bmatrix} \tilde{T}^0 + \tilde{U}^0 \\ O \end{bmatrix}_{C_1 \times M_1(K+1)ab}.$$

Here

$$P = \begin{bmatrix} diag(\tilde{\boldsymbol{\gamma}}, I_{M-1} \otimes \hat{I}_c) \\ O \end{bmatrix}_{M_1 \times M_0}, \; P^* = \begin{bmatrix} 0 \\ \mathbf{e}_{M-1} \otimes \hat{\mathbf{e}}_c \\ \mathbf{e}_{c+1} \end{bmatrix}_{M_1 \times 1},$$

$\tilde{T}^0 = I_{K+1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes \boldsymbol{\alpha} \otimes I_b$, $\tilde{U}^0 = \mathbf{e}_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \boldsymbol{\beta}$,

$$\Delta_b' = \begin{bmatrix} \boldsymbol{\xi} \otimes I_b \\ \boldsymbol{\xi} \otimes \mathbf{e}_b \otimes \boldsymbol{\beta} \end{bmatrix}, \; \tilde{\nabla} = -I_2 \otimes \tilde{\xi} + \tilde{E} \otimes I_L,$$

$$\hat{\Delta} = \begin{bmatrix} \boldsymbol{\xi} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_b \\ \boldsymbol{\xi} \otimes \mathbf{e}_b \otimes \tilde{\boldsymbol{\alpha}} \otimes \boldsymbol{\beta} \end{bmatrix}, \; \Delta^* = \begin{bmatrix} \boldsymbol{\xi} \otimes I_{ab} \\ \boldsymbol{\xi} \otimes \mathbf{e}_{ab} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}.$$

3.2 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study. We first establish the stability condition of the queueing system.

3.2.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. That is, $\boldsymbol{\pi}(A_0 + A_1 + A_2) = 0$; $\boldsymbol{\pi}\mathbf{e} = 1$. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\lambda < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{3.2}$$

That is, the rate of drift to the left has to be higher than that to the right. The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model.

For future reference, we define the traffic intensity ρ_1 as

$$\rho_1 = \frac{\lambda}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{3.3}$$

Note that the stability condition in equation (3.2) is equivalent to $\rho_1 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 3.3.

3.2.2 Steady state probability vector

Since the model studied as a QBD process, its steady-state distribution has a matrix-geometric solution under the stability condition. Assume that the stability condition (3.2) holds. Let $\boldsymbol{x}$ denote the steady-state probability vector of the generator Q given in equation (3.1). That is,

$$\boldsymbol{x}Q = 0; \boldsymbol{x}\mathbf{e} = 1. \tag{3.4}$$

Partitioning $\boldsymbol{x}$ as

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...) \tag{3.5}$$

we see that the sub-vectors of $\boldsymbol{x}$, under the assumption that the stability condition (3.2) holds, are obtained as (see, Neuts [44])

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3 \tag{3.6}$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{3.7}$$

$\boldsymbol{x}_0$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained using the boundary equations

$$\begin{aligned} -\lambda \boldsymbol{x}_0 + \boldsymbol{x}_1 B_2 &= 0 \\ \boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_3 + \boldsymbol{x}_2 B_5 &= 0 \\ \boldsymbol{x}_1 B_4 + \boldsymbol{x}_2 (A_1 + RA_2) &= 0 \end{aligned} \tag{3.8}$$

The normalizing condition of (3.4) results in

$$\boldsymbol{x}_0 + \boldsymbol{x}_1 \mathbf{e} + \boldsymbol{x}_2 (I - R)^{-1} \mathbf{e} = 1. \tag{3.9}$$

Once the rate matrix R is obtained, the vector $\boldsymbol{x}$ can be computed by exploiting the special structure of the coefficient matrices.

3.2.3 Performance measures

The vectors $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, etc. in this model are different from those vectors in model 1 of chapter 2. The expressions for ES, EQ, ECo, EI, F_{mi}, F_{ri}, F_{mb}, F_{rb}, F_{rc}, F_{wc}, F_{min}, $R_S^c wi$, $R_S^c b$, $R_S^c a$, $R_S^c wc$ and $R_S^c c$ etc. are similar to those in model 1 of chapter 2. These values are obtained by using the equations (2.11), (2.12), (2.13), (2.14), (2.15), (2.16), (2.17), (2.18), (2.19), (2.20), (2.21), (2.26), (2.27), (2.28), (2.29) and (2.30).

Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3})$, for $i \geq 2$. Here $\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11}, \boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}$, for $i \geq 2$ are vectors of dimensions M_1a, $(K+1)b$, $2KLb$, $(K+1)M_1ab$, $2KLb$, $2M_0KLab$ and $2KLab$, respectively. Now we compute some more performance measures.

(1) Rate at which interruption completion takes place before threshold is realised

$$R_I^c b = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_3=1}^{L} \xi_{l_3} \boldsymbol{x}_{i2jkl_1 1 l_3} \mathbf{e}. \tag{3.10}$$

(2) Rate at which interruption completion takes place after threshold is realised

$$R_I^c a = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_3=1}^{L} \xi_{l_3} \boldsymbol{x}_{i2jkl_1 0 l_3} \mathbf{e}. \tag{3.11}$$

(3) Rate at which consultation completion takes place before threshold is realised

$$R_C^c b = \sum_{i=1}^{\infty} \sum_{j=0}^{K-1} \sum_{l_3=1}^{L} \xi_{l_3} \boldsymbol{x}_{i1j1l_3} \mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_3=1}^{L} \xi_{l_3} \boldsymbol{x}_{i2jkl_1 1 l_3} \mathbf{e}. \tag{3.12}$$

(4) Rate at which consultation completion takes place after the thresh-

old is realised

$$R_C^c a = \sum_{i=1}^{\infty}\sum_{j=0}^{K-1}\sum_{l_3=1}^{L}\xi_{l_3}\boldsymbol{x}_{i1j0l_3}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{j=0}^{K-1}\sum_{k=0}^{M-1}\sum_{l_1=0}^{c}\sum_{l_3=0}^{L}\xi_{l_3}\boldsymbol{x}_{i2jkl_1 0l_3}\mathbf{e}. \quad (3.13)$$

3.3 Numerical results

We analyse the effect of the parameters λ and θ on the key performance measures.

Let us choose the following data so that the system is stable.

$T = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$; $U = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$; $G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}$;

$\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$; $\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$; $\boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$;

$\boldsymbol{\delta} = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$; $\boldsymbol{\xi} = \begin{bmatrix} 1 & 1.5 & 2 \end{bmatrix}^T$; $M = 3$; $K = 3$.

Referring to Table 3.1, as the rate of consultation θ increases, the traffic intensity ρ_1 will increase and hence EI and ECo will increase. This results in an increase in F_{min} and F_{rc}. As θ increases, consultation is more frequent, so the main server will reach the upper bounds of number of interruptions or super clock may realise frequently, and main server compels to complete the service of the customer at him before further consultations and this results in more waiting time of the regular server to get consultation. Thus F_{rw} also increases. Since F_{min}, F_{rc} and F_{rw} increase, the customers have to stay in the system and in queue long time and this results in an increase in ES and EQ. This in turn make a decrease in F_{mi} and F_{ri}. Since main server has to spend more time in consultation,

Table 3.1: Effect of θ on various performance measures

$\lambda = 2$

θ	3	3.5	4	4.5	5	5.5	6
ρ_1	0.5309	0.5803	0.6266	0.6699	0.7103	0.7481	0.7835
ES	1.6367	1.9958	2.4136	2.8976	3.4560	4.0969	4.8290
EQ	0.9242	1.2177	1.5694	1.9869	2.4778	3.0482	3.7014
EI	0.1537	0.1784	0.2026	0.2261	0.2487	0.2702	0.2903
ECo	0.2906	0.3432	0.3960	0.4482	0.4993	0.5489	0.5963
F_{mi}	0.5321	0.4974	0.4626	0.4281	0.3944	0.3618	0.3306
F_{ri}	0.6797	0.6358	0.5918	0.5482	0.5054	0.4639	0.4241
F_{mb}	0.2574	0.2537	0.2501	0.2463	0.2426	0.2388	0.2351
F_{rb}	0.1017	0.1046	0.1073	0.1099	0.1121	0.1141	0.1158
F_{min}	0.1348	0.1600	0.1855	0.2108	0.2358	0.2602	0.2835
F_{rc}	0.2105	0.2489	0.2874	0.3255	0.3629	0.3991	0.4338
F_{rw}	0.0081	0.0107	0.0134	0.0164	0.0195	0.0226	0.0258

it gets lesser time to serve customers. So F_{mb} decreases. As θ increases, possibility for restart of the service is high and F_{rb} increases.

Referring to Table 3.2, as the arrival rate λ increases, the traffic intensity ρ_1 increases. The system is fed with more and more customers and therefore accumulation of customers increases. So ES and EQ increase. Thus EI and ECo will also increase. This results in a hike in F_{min} and F_{rc}. Thus F_{rw} also increases. As the arrival rate increases, there are more customers in the queue and therefore the servers have to spend longer time in service. Thus F_{mb} and F_{rb} increase. This in turn make a decrease in F_{mi} and F_{ri}.

Table 3.2: Effect of λ on various performance measures

$\theta = 2$

λ	2	2.5	3	3.5	4	4.5
ρ_1	0.4224	0.5280	0.6336	0.7392	0.8448	0.9504
ES	1.0672	1.8760	3.2097	5.3483	8.6509	13.5100
EQ	0.4818	1.0897	2.2077	4.0546	6.6089	9.3050
EI	0.1035	0.1611	0.2281	0.2983	0.3646	0.4212
ECo	0.1871	0.2755	0.3701	0.4617	0.5380	0.5840
F_{mi}	0.6003	0.4952	0.3929	0.2992	0.2196	0.1566
F_{ri}	0.7652	0.6536	0.5337	0.4162	0.3115	0.2258
F_{mb}	0.2644	0.3057	0.3394	0.3654	0.3820	0.3843
F_{rb}	0.0957	0.1410	0.1895	0.2365	0.2756	0.2991
F_{min}	0.0861	0.1340	0.1890	0.2453	0.2940	0.3242
F_{rc}	0.1352	0.1992	0.2676	0.3337	0.3874	0.4172
F_{rw}	0.0039	0.0062	0.0090	0.0119	0.0144	0.01600

3.4 Description of model 2

In this model the consultation is in Markovian environment. Let $f_1, f_2, .., f_L$ be L Markovian environmental factors due to which there occurs consultations. Requirement of consultation is a Poisson process with rate θ where the i^{th} factor occurring with probability δ_i, i=1,2,...,L. Let the environmental factors are related to each other by the transition probability matrix E. The duration of consultation and threshold clock are same as those in model 1 of this chapter. All other assumptions regarding the number of interruptions, consultations and super clock are same as those in the model 1 of chapter 2.

Notations :- We use the following notations in this model.

- $M_0 = M(c+1)$ and $M_1 = M_0 + 1$
- $K_1 = KL + 1,\ K_2 = 2KL$
- $C_1 = K_1 M_1 ab + K_2 b + K_2 M_0 ab + K_2 ab$
- $C_0 = M_1 a + K_1 b + K_2 b$
- $\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, 0),\ \tilde{\boldsymbol{\eta}} = (1, 0)$
- $\dot{I}_1 = \begin{bmatrix} \mathbf{0} & I_{M_0} \end{bmatrix}_{M_0 \times M_1},\ \dot{I}_2 = \begin{bmatrix} \mathbf{0} & I_{KL} \end{bmatrix}_{KL \times K_1}$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$ and $E^* = \begin{bmatrix} \boldsymbol{\delta} & \underline{\mathbf{0}} \\ O & I_{K-1} \otimes E \\ O & O \end{bmatrix}$

Consider the queueing model $X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), B_1(t), F(t), B_2(t), S_1(t), S_2(t), J_1(t), J_2(t)\}$.
The variable $F(t)$ is described as follows:
If the regular server is busy serving a customer after a consultation, then $F(t)$ represents the environmental factor due which that consultation has occurred; if the regular server is getting a consultation or waiting to get a consultation, then $F(t)$ represents the environmental factor for which that consultation is going on.
Thus $F(t) = i$, where $1 \leq i \leq L$ for the following cases:
(1)$N(t) = 1$ and $S(t) = \tilde{0}$ or $N(t) \geq 2$ and $S(t) = 0$ with $1 \leq B_1(t) \leq K$
(2)$N(t) = 1$ and $S(t) = 1$
(3)$N(t) \geq 2$ and $S(t) = \{1, 2, 3\}$.

The variables $N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), J_1(t), J_2(t)$ are the same as those in model 1.
Note that $B_1(t)$ is '0' means the the regular server has not obtained a consultation yet and so the phases of the environmental factors $F(t)$ do not present when $B_1(t) = 0$.
$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as
$\psi(1) = \psi(1,0) \cup \psi(1,\tilde{0}) \cup \psi(1,1)$ and
$\psi(i) = \psi(i,0) \cup \psi(i,1) \cup \psi(i,2) \cup \psi(i,3)$, for $i \geq 2$,
where
$\psi(1,0) = \{(1,0,0,t_1)\} \cup \{(1,0,k,l_1,t_1) : 1 \leq k \leq M\}$,
$\psi(1,\tilde{0}) = \{(1,\tilde{0},0,t_2) \cup (1,\tilde{0},j,l_3,t_2) : 1 \leq j \leq K\}$,
$\psi(1,1) = \{(1,1,j,l_2,l_3,t_2) : 0 \leq j \leq K-1\}$,
$\psi(i,0) = \{(i,0,0,0,t_1,t_2) \cup (i,0,j,l_3,0,t_1,t_2) \cup (i,0,0,k,l_1,t_1,t_2)$
$\cup\, (i,0,j,l_3,k,l_1,t_1,t_2) : 1 \leq j \leq K, 1 \leq k \leq M\}$,
$\psi(i,1) = \{(i,1,j,l_2,l_3,t_2) : 0 \leq j \leq K-1\}$,
$\psi(i,2) = \{(i,2,j,l_2,l_3,k,l_1,t_1,t_2) : 0 \leq j \leq K-1, 0 \leq k \leq M-1\}$,
$\psi(i,3) = \{(i,3,j,l_2,l_3,t_1,t_2) : 0 \leq j \leq K-1\}$,
for $i \geq 2$ with $0 \leq l_1 \leq c$, $l_2 = \{1,0\}$, $1 \leq l_3 \leq L$,
$1 \leq t_1 \leq a$ and $1 \leq t_2 \leq b$.

The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} -\lambda & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{3.14}$$

where

$B_1 = \lambda \begin{bmatrix} \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix}$, $B_2 = \begin{bmatrix} \mathbf{e}_{M_1} \otimes T^0 \\ \mathbf{e}_{KL+1} \otimes U^0 \\ \mathbf{0} \end{bmatrix}$,

$B_3 = \begin{bmatrix} I_{M_1} \otimes T & O & O \\ O & B_{31} & \theta E^* \otimes \tilde{\boldsymbol{\eta}} \otimes I_b \\ O & \dot{I}_2 \otimes \xi \otimes \Delta & I_K \otimes \nabla \otimes I_b \end{bmatrix} - \lambda I,$

$B_4 = \lambda \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix}$, $B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix}$,

$A_0 = \lambda I,\ A_1 = \begin{bmatrix} A_{11} & O & A_{12} & A_{13} \\ A_{14} & A_{15} & O & O \\ A_{16} & O & A_{17} & O \\ O & O & O & A_{18} \end{bmatrix} - \lambda I,$

$A_2 = \begin{bmatrix} A_{21} & B_{53} & O \end{bmatrix}$.

Here A_0, A_1 and A_2 are square matrices of order C_1;

B_3 is a square matrix of order C_0 and B_1, B_2, B_4, B_5 are

matrices of orders $1 \times C_0$, $C_0 \times 1$, $C_0 \times C_1$ and $C_1 \times C_0$, respectively.

The block matrices are given by:

$B_{31} = I_{K_1} \otimes U - J \otimes I_b,$

$$B_{41} = \begin{bmatrix} \mathbf{e}'_{K_1}(1) \otimes I_{M_1} \otimes I_a \otimes \boldsymbol{\beta} \\ I_{K_1} \otimes \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha} \otimes I_b \\ O \end{bmatrix}_{C_0 \otimes K_1 M_1 ab}, \ B_{42} = \begin{bmatrix} O \\ I_{K_2 b} \end{bmatrix}_{C_0 \otimes K_2 b},$$

$$B_{51} = \begin{bmatrix} \mathbf{e}_{K_1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_1 \otimes M_1 a},$$

$$B_{52} = \begin{bmatrix} I_{K_1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_1 \otimes K_1 b},$$

$$B_{53} = \begin{bmatrix} O \\ I_{K_2} \otimes T^0 \otimes I_b \end{bmatrix}_{C_1 \otimes K_2 b},$$

$$A_{11} = I_{K_1} \otimes I_{M_1} \otimes (T \oplus U) - J \otimes I_{M_1} \otimes I_{ab},$$

$$A_{12} = \theta E^* \otimes P \otimes \tilde{\boldsymbol{\eta}} \otimes I_{ab}, \ A_{13} = \theta E^* \otimes P^* \otimes \tilde{\boldsymbol{\eta}} \otimes I_{ab},$$

$$A_{14} = \dot{I}_2 \otimes \xi \otimes \tilde{\Delta}, \ A_{15} = I_K \otimes \nabla \otimes I_b, \ A_{16} = \dot{I}_2 \otimes \xi \otimes \dot{I}_1 \otimes \Delta^*,$$

$$A_{17} = I_K \otimes \tilde{\nabla} \otimes I_{ab}, \ A_{18} = I_{KL} \otimes (H \oplus T) \otimes I_b,$$

$$A_{21} = \begin{bmatrix} \tilde{T}^0 + \tilde{U}^0 \\ O \end{bmatrix}_{C_1 \otimes K_1 M_1 ab}.$$

Here

$$J = \theta \begin{bmatrix} I_{(K-1)L+1} & O \\ O & O \end{bmatrix}_{K_1 \otimes K_1},$$

$$P = \begin{bmatrix} diag(\tilde{\boldsymbol{\gamma}}, I_{M-1} \otimes \hat{I}_c) \\ O \end{bmatrix}, \; P^* = \begin{bmatrix} 0 \\ \mathbf{e}_{M-1} \otimes \hat{\mathbf{e}_c} \\ \mathbf{e}_{c+1} \end{bmatrix},$$

$$\tilde{T}^0 = I_{K_1} \otimes \mathbf{e}_{M_1} \otimes T^0 \otimes \boldsymbol{\alpha} \otimes I_b, \; \tilde{U}^0 = \mathbf{e}_{KL+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \boldsymbol{\beta},$$

$$\Delta = \begin{bmatrix} I_b \\ \mathbf{e}_b \otimes \boldsymbol{\beta} \end{bmatrix}, \; H = \omega \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}, \; \nabla = I_K \otimes (\boldsymbol{\xi} \oplus H) \otimes I_b,$$

$$\tilde{\nabla} = I_K \otimes [\boldsymbol{\xi} \otimes I_{2M_0} + I_{LM} \otimes (\tilde{G} \oplus H)] \otimes I_{ab},$$

$$\tilde{\Delta} = \begin{bmatrix} \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha} \otimes I_b \\ \mathbf{e}_b \otimes \mathbf{e}'_{M_1}(1) \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}, \; \Delta^* = \begin{bmatrix} I_{ab} \\ \mathbf{e}_{ab} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{bmatrix}.$$

3.5 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study. We first establish the stability condition of the queueing system.

3.5.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{3.15}$$

The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model.

Define the traffic intensity ρ_2 as

$$\rho_2 = \frac{\boldsymbol{\pi} A_0 \mathbf{e}}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{3.16}$$

Note that the stability condition in equation (3.15) is equivalent to $\rho_2 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 3.6.

3.5.2 Steady state probability vector

Since the model studied as a QBD process, its steady-state distribution has a matrix-geometric solution under the stability condition. Assume that the stability condition (3.15) holds. Let $\boldsymbol{x}$ denote the steady-state probability vector of the generator Q given in equation (3.14).
Partitioning $\boldsymbol{x}$ as

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots\ldots\ldots\ldots) \tag{3.17}$$

we see that the sub-vectors of $\boldsymbol{x}$, under the assumption that the stability condition (3.15) holds, are obtained as (see, Neuts [44])

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3 \tag{3.18}$$

where R is the minimal non-negative solution to the matrix quadratic equation:

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{3.19}$$

$\boldsymbol{x}_0$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained using the boundary equations

$$-\lambda \boldsymbol{x}_0 + \boldsymbol{x}_1 B_2 = 0$$

$$\boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_3 + \boldsymbol{x}_2 B_5 = 0 \qquad (3.20)$$

$$\boldsymbol{x}_1 B_4 + \boldsymbol{x}_2 (A_1 + RA_2) = 0$$

The normalizing condition results in

$$\boldsymbol{x}_0 + \boldsymbol{x}_1 \mathbf{e} + \boldsymbol{x}_2 (I - R)^{-1} \mathbf{e} = 1. \qquad (3.21)$$

Once the rate matrix R is obtained, the vector $\boldsymbol{x}$ can be computed by exploiting the special structure of the coefficient matrices.

3.5.3 Performance measures

The vectors $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, etc. in this model are different from those vectors in model 1 of chapter 2. The expressions for ES, EQ, ECo, EI, F_{mi}, F_{ri}, F_{mb}, F_{rb}, F_{rc}, F_{wc}, F_{min}, $R^c_S a$, $R^c_S wc$ and $R^c_S c$ are similar to those in model 1 of chapter 2. These values are obtained by using the equations (2.11), (2.12), (2.13), (2.14), (2.15), (2.16), (2.17), (2.18), (2.19), (2.20), (2.21), (2.28), (2.29) and (2.30).
Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3})$, for $i \geq 2$.
Here $\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11}, \boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}$, for $i \geq 2$ are vectors of dimensions $M_1 a$, $K_1 b$, $K_2 b$, $K_1 M_1 ab$, $K_2 b$, $M_0 K_2 ab$ and $K_2 ab$, respectively.

Now we compute some more performance measures.

(1) Rate at which interruption completion takes place before threshold is realised

$$R_I^c b = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \xi_{j_1} \boldsymbol{x}_{i2jj_1kl_11}\mathbf{e}. \tag{3.22}$$

(2) Rate at which interruption completion takes place after threshold is realised

$$R_I^c a = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \xi_{j_1} \boldsymbol{x}_{i2jj_1kl_10}\mathbf{e}. \tag{3.23}$$

(3) Rate at which consultation completion takes place before threshold is realised

$$R_C^c b = \sum_{i=1}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \xi_{j_1} \boldsymbol{x}_{i1jj_11}\mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \xi_{j_1} \boldsymbol{x}_{i2jj_1kl_11}\mathbf{e}. \tag{3.24}$$

(4) Rate at which consultation completion takes place after the threshold is realised

$$R_C^c a = \sum_{i=1}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \xi_{j_1} \boldsymbol{x}_{i1jj_10}\mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{j_1=1}^{L} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \xi_{j_1} \boldsymbol{x}_{i2jj_1kl_10}\mathbf{e}. \tag{3.25}$$

(5) Rate at which service completion at the main server takes place

without any interruption

$$R_S^c wi = \sum_{t_1=1}^{a} \boldsymbol{x}_{100t_1} T_{t_1}^0 + \sum_{i=2}^{\infty} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i000t_1} \mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=1}^{K} \sum_{j_1=1}^{L} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0jj_10t_1} \mathbf{e}. \tag{3.26}$$

(6) Rate at which service completion at the main server (with at least one interruption) takes place before super clock is realised

$$R_S^c b = \sum_{i=2}^{\infty} \sum_{k=1}^{M} \sum_{l_1=1}^{c} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i00kl_1t_1} \mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=1}^{K} \sum_{j_1=1}^{L} \sum_{k=1}^{M} \sum_{l_1=1}^{c} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0jj_1kl_1t_1} \mathbf{e}. \tag{3.27}$$

(7) Rate at which service completion at the main server (with at least one interruption) takes place after super clock is realised

$$R_S^c a = \sum_{i=2}^{\infty} \sum_{k=1}^{M} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i00k0t_1} \mathbf{e} + \sum_{i=2}^{\infty} \sum_{j=1}^{K} \sum_{j_1=1}^{L} \sum_{k=1}^{M} \sum_{t_1=1}^{a} T_{t_1}^0 \boldsymbol{x}_{i0jj_1k0t_1} \mathbf{e}. \tag{3.28}$$

3.6 Numerical examples

Let $U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$; $T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$; $G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}$;

$\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$; $\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$; $\boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$; $\boldsymbol{\delta} = \begin{bmatrix} 0.3 & 0.5 & 0.2 \end{bmatrix}$;

$$E = \begin{bmatrix} 0.3 & 0.3 & 0.4 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}; \boldsymbol{\xi} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T; M = 3; K = 3.$$

The stability condition $\rho_2 < 1$ is satisfied for the above matrices, vectors and values.

Table 3.3: Effect of θ on various performance measures

$\lambda = 3$

θ	3	3.5	4	4.5	5
ρ_2	0.7211	0.7849	0.8445	0.9001	0.9519
ES	4.4827	5.9070	7.6951	9.9176	12.6401
EQ	3.3316	4.5717	6.0350	7.6359	9.2266
EI	0.3886	0.4544	0.5156	0.5706	0.6186
ECo	0.5579	0.6477	0.7288	0.7971	0.8482
F_{mi}	0.3388	0.2896	0.2446	0.2045	0.1696
F_{ri}	0.4435	0.3799	0.3214	0.2692	0.2236
F_{mb}	0.3135	0.3053	0.2962	0.2859	0.2737
F_{rb}	0.1947	0.1969	0.1972	0.1981	0.1993
F_{min}	0.2461	0.2882	0.3263	0.3580	0.3811
F_{rc}	0.3471	0.4030	0.4532	0.4950	0.5256
F_{rw}	0.0140	0.0181	0.0222	0.0261	0.0294

Referring to Table 3.3, as the rate of consultation θ increases, traffic intensity ρ_2 will increase and so EI and ECo will increase. This results in an increase in F_{min} and F_{rc}. As θ increases, consultation is more frequent, so the main server will reach the upper bounds of number of interruptions or super clock may realise frequently and main server compels to complete the service of the customer at him before further consultations and this results in more waiting time of the regular server to get consultation. Thus F_{rw} also increases. The possibility for restart of the service at the regular server will increase and so F_{rb} also increases. Since F_{min}, F_{rc} and F_{rw} increase, the customers have to stay in the system and in queue for longer time and this results in an increase in ES and EQ. Thus the idle time of the servers F_{mi} and F_{ri} decrease. Since main server has to spend more time in consultation, it gets lesser time to serve customers. So F_{mb} decreases.

Table 3.4: Effect of λ on various performance measures

$\theta = 2$

λ	3	3.5	4	4.5	5
ρ_2	0.5808	0.6775	0.7743	0.8711	0.9679
ES	2.4768	3.9794	6.2527	9.6005	14.3406
EQ	1.5690	2.8554	4.7760	7.2039	9.6201
EI	0.2504	0.3313	0.4116	0.4848	0.5465
ECo	0.3657	0.4607	0.5479	0.6166	0.6544
F_{mi}	0.4446	0.3535	0.2716	0.2024	0.1471
F_{ri}	0.5796	0.4695	0.3662	0.2763	0.2030
F_{mb}	0.3282	0.3599	0.3853	0.4015	0.4036
F_{rb}	0.1865	0.2349	0.2794	0.3143	0.3334
F_{min}	0.1573	0.2079	0.2569	0.2969	0.3199
F_{rc}	0.2272	0.2861	0.3397	0.3805	0.4004
F_{rw}	0.0067	0.0090	0.0113	0.0131	0.0142

Let us analyse the results of table 3.4. As the arrival rate λ increases, the traffic intensity ρ_2 increases. The system is fed with more and more customers and therefore accumulation of customers increases. So ES and EQ increase. Thus EI and ECo will also increase. This results in a hike in F_{min} and F_{rc}. Thus F_{rw} also increases. As the arrival rate increases, there are more customers in the queue and therefore the servers have to spend longer time in service. Thus F_{mb} and F_{rb} increase. This in turn make a decrease in F_{mi} and F_{ri}.

Chapter 4

Three-server queues with consultation by main server controlled by an upper bound on number of interruptions

4.1 Introduction

In the previous chapters, we considered two-server queueing models with one main server and one regular server. The consultations were controlled by number of interruptions to the main server, consultations to the regular server during the services of particular customers at the servers and duration of super clock. In this chapter, we study two three-server queueing models with one main server and two identical regular servers where consultations are given by the main server to the regular servers. The important assumptions of this chapter are

(i) the service times at the main and regular servers are independent phase type distributions with representations $(\boldsymbol{\alpha}, T)$ and $(\boldsymbol{\beta}, U)$ with number of phases a and b, respectively. Note that $\mathbf{T}^0 = -T\mathbf{e}$ and $\mathbf{U}^0 = -U\mathbf{e}$

(ii) M denotes the upper bound of number of interruptions to the customer at the main server

(iii) duration of consultation follows exponential distribution with parameter ξ

The main server offers consultation to the regular server whenever it is necessary. Requirement of consultation arises according to a Poisson process with rate θ_i, if there are i busy regular servers, where $i = 1, 2$. When both the regular severs need consultation, a queue is formed for consultation and it is provided in FIFO basis. In order to distinguish the regular servers, we denote them $\Re_1$ and $\Re_2$. After getting consultation, they resume the services at the phases where they were suspended.

In model 1, the arrival process is a Poisson process with rate λ. In model 2, we consider independent arrival processes to the main server and regular servers. There is a finite buffer of size K at the main server. The arriving customer to the main server will be lost when the buffer is full.

4.1.1 Notations

We use the following notations in the sequel.

- $\tilde{\boldsymbol{\alpha}} = \mathbf{e}'_{M+1}(1) \otimes \boldsymbol{\alpha}$
- $\nabla = diag(I_b \otimes \boldsymbol{\beta}, \boldsymbol{\beta} \otimes I_b)$, $\nabla_1 = \begin{bmatrix} \nabla & O \end{bmatrix}$
- $\Delta = diag(I_b \otimes U^0, U^0 \otimes I_b)$, $\Delta_1 = \begin{bmatrix} \Delta \\ O \end{bmatrix}$, $\Delta_2 = \begin{bmatrix} \Delta \otimes I_n & O \end{bmatrix}'$, $\Delta_3 = diag(\Delta \otimes I_M, O)$ and $\Delta_4 = diag(\Delta, O)$
- $\hat{\Delta} = diag(I_b \otimes U^0 \otimes \boldsymbol{\beta}, U^0 \otimes \boldsymbol{\beta} \otimes I_b)$, $\hat{\Delta}_1 = diag(\hat{\Delta}, O)$, $\Delta^* = diag(I_b \otimes U, U \otimes I_b)$
- $\ddot{I} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

4.2 Description of model 1

Here customers arrive according to a Poisson process with rate λ. An arriving customer enters into service immediately if at least one server is free, else joins a queue. The customer will be served by the main server whenever the main server and at least one of the regular servers is free. If both the regular servers are idle and the main server is busy, then the customer will approach any one of the regular server with probability 1/2.

Consider the queueing model

$$X = \{X(t), t \geq 0\},$$

where

$$X(t) = \{N(t), S(t), B_1(t), B_2(t), J_1(t), J_2(t), S_1(t), J_3(t)\}.$$

Here

$N(t)$ – the number of customers in the system

$J_i(t)$ – phase of the regular server $\Re_i$, $i = 1, 2$

$S_1(t)$ – number of interruptions already befell to a customer at the main server

$J_3(t)$ – phase of the main server

Here $S(t)$ denotes the status of the servers at time t such that

$$S(t) = \begin{cases} \tilde{0}, & \text{if the regular server(s) is busy and main server is idle} \\ 0, & \text{if the main server is busy together with} \\ & \text{regular server(s) is (are) busy or idle} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation} \\ & \text{with one interrupted customer at the main server} \\ 3, & \text{if the regular server is waiting for getting consultation} \\ & \text{after the present service at the main server} \\ 4, & \text{if the regular server is waiting for getting consultation} \\ & \text{after the service at the main server followed by} \\ & \text{the present interruption} \end{cases}$$

The variable $B_1(t)$ appears only when $N(t) = 1$ and $S(t) = \tilde{0}$ or $N(t) = 2$ and $S(t) = 0$.
$B_1(t) = \{1, 2\}$ according to $\Re_1$ or $\Re_2$ is busy.
Now consider the variable $B_2(t)$.
If $N(t) \geq 1$ and $S(t) = \{1, 2\}$, then

$$B_2(t) = \begin{cases} 1 \text{ (or 2)}, & \text{if } \Re_1 \text{ is getting consultation and } \Re_2 \text{ is busy or idle} \\ & \text{(or vice versa)} \\ 3 \text{ (or 4)}, & \text{if both regular servers are in a queue for consultation} \\ & \text{with } \Re_1 \text{ is getting consultation in the first place} \\ & \text{and } \Re_2 \text{ in the second place (or vice versa)} \end{cases}$$

If $N(t) \geq 1$ and $S(t) = 3$, then $B_2(t)$ takes the same values $\{1, 2, 3, 4\}$ according to the above definition with 'getting consultation' is replaced by 'waiting to get consultation.'
If $N(t) \geq 3$ and $S(t) = 4$, then

$$B_2(t) = \begin{cases} 1 \text{ (or 2)}, & \text{if } \Re_1 \text{ (or } \Re_2\text{) is waiting to get consultation} \\ & \text{after the present interruption followed by the} \\ & \text{service completion at the main server} \end{cases}$$

$\{X(t), t \geq 0\}$ is a CTMC with state space

$$\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as

$$\begin{aligned}
\psi(1) &= \{(1,0,k,j_3) : 0 \le k \le M\} \cup \{(1,\tilde{0},m,j_m) \\
&\quad \cup (1,1,m,j_m) : 1 \le m \le 2\} \\
\psi(2) &= \{(2,0,m,j_m,k,j_3) : 0 \le k \le M, 1 \le m \le 2\} \cup \{(2,\tilde{0},j_1,j_2)\} \\
&\quad \cup \{(2,1,m,j_1,j_2) : 1 \le m \le 4\} \\
&\quad \cup \{(2,2,m,j_m,k,j_3) : 1 \le m \le 2, 0 \le k \le M-1\} \\
&\quad \cup \{(2,3,m,j_m,j_3) : 1 \le m \le 2\} \\
\psi(i) &= \{(i,0,j_1,j_2,k,j_3) : 0 \le k \le M\} \cup \{(i,1,m,j_1,j_2) : 1 \le m \le 4\} \\
&\quad \cup \{(i,2,m,j_1,j_2,k,j_3) : 1 \le m \le 2, 0 \le k \le M-1\} \\
&\quad \cup \{(i,2,m,j_1,j_2,k,j_3) : 3 \le m \le 4, 0 \le k \le M-2\} \\
&\quad \cup \{(i,3,m,j_1,j_2,j_3) : 1 \le m \le 4\} \\
&\quad \cup \{(i,4,m,j_1,j_2,j_3) : 1 \le m \le 2\} \\
&\quad \text{ where } 1 \le j_1, j_2 \le b \text{ and } 1 \le j_3 \le a.
\end{aligned}$$

The infinitesimal generator matrix Q is given by

$$Q = \begin{bmatrix}
-\lambda & D_0 & & & & & \\
B_1 & C_1 & D_1 & & & & \\
 & B_2 & C_2 & D_2 & & & \\
 & & B_3 & A_1 & A_0 & & \\
 & & & A_2 & A_1 & A_0 & \\
 & & & & \ddots & \ddots & \ddots
\end{bmatrix} \tag{4.1}$$

Here A_0, A_1 and A_2 are square matrices of order K_3; C_1 and C_2 are square

matrices of orders K_1 and K_2, respectively. D_0, D_1, D_2, B_1, B_2 and B_3 are matrices of orders $1 \times K_1$, $K_1 \times K_2$, $K_2 \times K_3$, $K_1 \times 1$, $K_2 \times K_1$ and $K_3 \times K_2$, respectively, where
$K_1 = (M+1)a + 4b$, $K_2 = 4(M+1)ab + 5b^2$ and
$K_3 = 5(M+1)ab^2 + 4b^2$.
The block matrices are defined as follows:

$$D_0 = \lambda \begin{bmatrix} \tilde{\boldsymbol{\alpha}} & \mathbf{0} \end{bmatrix}, \; B_1 = \begin{bmatrix} \mathbf{e}_{M+1} \otimes T^0 \\ \mathbf{e}_2 \otimes U^0 \\ O \end{bmatrix},$$

$$C_1 = \begin{bmatrix} I_{M+1} \otimes T & O \\ O & C_{11} \end{bmatrix} - \lambda I, \; D_1 = \lambda \begin{bmatrix} D_{11} & D_{12} & D_{13} \end{bmatrix}',$$

$$B_2 = \begin{bmatrix} B_{21} & B_{22} & B_{23} \end{bmatrix}, \; D_2 = \lambda \begin{bmatrix} D_{21} & D_{22} \end{bmatrix}',$$

$$B_3 = \begin{bmatrix} B_{31} & B_{32} & B_{33} \end{bmatrix},$$

$$A_0 = \lambda I, \; A_1 = \begin{bmatrix} A_{11} & & A_{12} \\ A_{13} & A_{14} & \\ A_{15} & & A_{16} \end{bmatrix} - \lambda I,$$

$$A_2 = \begin{bmatrix} A_{21} & B_{32} & A_{22} \end{bmatrix}, \text{ where}$$

$$C_{11} = diag(I_2 \otimes U, O) + \begin{bmatrix} -\theta_1 & \theta_1 \\ \xi & -\xi \end{bmatrix} \otimes I_{2b},$$

$$D_{11} = \begin{bmatrix} \frac{1}{2}\mathbf{e}_2' \otimes \boldsymbol{\beta} \otimes I_{(M+1)a} & O \end{bmatrix}_{(M+1)a \times K_2},$$

$$D_{12} = \begin{bmatrix} I_{2b} \otimes \tilde{\boldsymbol{\alpha}} & O \end{bmatrix}_{2b \times K_2},$$

$$D_{13} = \begin{bmatrix} O & \nabla & O \end{bmatrix}_{2b \times K_2},$$

$$B_{21} = \begin{bmatrix} \mathbf{e}_2 \otimes U^0 \otimes I_{a(M+1)} \\ O \end{bmatrix}_{K_2 \times (M+1)a},$$

$$B_{22} = \begin{bmatrix} I_{2b} \otimes \mathbf{e}_{M+1} \otimes T^0 & \mathbf{e}_2^{'} \otimes \Delta & O \end{bmatrix}^{'}_{K_2 \times 2b},$$

$$B_{23} = \begin{bmatrix} O & \Delta_1 & O & I_{2b} \otimes T^0 \end{bmatrix}^{'}_{K_2 \times 2b},$$

$$D_{21} = \begin{bmatrix} \nabla \otimes \mathbf{e}_2 \otimes I_{(M+1)a} & O \end{bmatrix}_{2(M+1)ab \times K_3},$$

$$D_{22} = \begin{bmatrix} diag(I_{b^2} \otimes \tilde{\boldsymbol{\alpha}}, I_{4b^2}, \nabla_1 \otimes I_{aM}, \nabla_1 \otimes I_a) & O \end{bmatrix}_{5b^2+2(M+1)ab \times K_3},$$

$$B_{31} = \begin{bmatrix} \mathbf{e}_2^{'} \otimes \Delta \otimes I_{(M+1)a} & I_{b^2} \otimes \mathbf{e}_{M+1} \otimes T^0 \\ O & O \end{bmatrix}_{K_3 \times 2(M+1)ab+b^2},$$

$$B_{32} = \begin{bmatrix} O & \hat{\Delta}_1 & O & I_{4b^2} \otimes T^0 & O \end{bmatrix}^{'}_{K_3 \times 4b^2},$$

$$B_{33} = \begin{bmatrix} O \\ diag(\Delta_2 \otimes I_a, \Delta_1 \otimes I_a) \\ O \end{bmatrix}_{K_3 \times 2(M+1)ab},$$

$$A_{11} = U \oplus U \oplus I_{M+1} \otimes T - 2\theta_2 I_{(M+1)ab^2},$$

$$A_{12} = \theta_2 \begin{bmatrix} \mathbf{e}_2^{'} \otimes I_{b^2} \otimes \hat{I}_M & O & \mathbf{e}_2^{'} \otimes I_{b^2} \otimes \hat{\mathbf{e}}_M & O \end{bmatrix} \otimes I_a,$$

$$A_{13} = \xi \begin{bmatrix} \mathbf{e}_2 \otimes I_{b^2} \otimes \tilde{\boldsymbol{\alpha}} \\ O \end{bmatrix},$$

$$A_{14} = diag(\Delta^*, O) - \xi I_{4b^2} + \begin{bmatrix} -\theta_1 I_2 & \theta_1 I_2 \\ \xi \ddot{I} & O \end{bmatrix} \otimes I_{b^2},$$

$$A_{15} = \xi \begin{bmatrix} \mathbf{e}_2 \otimes I_{b^2} \otimes \dot{I}_M \\ O \end{bmatrix} \otimes I_a, \; A_{16} = \begin{bmatrix} F_2 & & F_3 \\ & F_4 & \\ & F_5 & -\xi I_{2ab^2} \end{bmatrix},$$

$$A_{21} = \begin{bmatrix} U^0 \otimes \boldsymbol{\beta} \oplus U^0 \otimes \boldsymbol{\beta} \oplus \mathbf{e}_{M+1} \otimes T^0 \otimes \boldsymbol{\alpha} \\ O \end{bmatrix}_{K_3 \times (M+1)ab^2},$$

$$A_{22} = \begin{bmatrix} O \\ diag(\Delta_3 \otimes I_a, \Delta_4 \otimes I_a, O) \end{bmatrix}_{K_3 \times 4(M+1)ab^2}.$$

Here

$$F_2 = diag(\Delta^* \otimes I_{(M+1)a}, O) - \xi I_{4b^2(2M-1)}$$

$$+ \begin{bmatrix} -\theta_1 I_{2Mb^2} & \theta_1 I_{2b^2} \otimes \hat{I}_{M_1} \\ \xi \ddot{I} \otimes I_{M_1 b^2 a} & O \end{bmatrix} \otimes I_a,$$

$$F_3 = \theta_1 \begin{bmatrix} I_{2b^2} \otimes \hat{\mathbf{e}}_{M-1} \otimes I_a \\ O \end{bmatrix},$$

$$F_4 = diag(\Delta^*, O) \otimes I_a - I_{4b^2} \otimes T + \begin{bmatrix} -\theta_1 I_{2ab^2} & \theta_1 I_{2ab^2} \\ O & O \end{bmatrix},$$

$$F_5 = \xi \begin{bmatrix} \ddot{I} & O \end{bmatrix} \otimes I_{ab^2}.$$

4.3 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study. We first establish the stability condition of the queueing system.

4.3.1 Stability condition

Let $\boldsymbol{\pi}$ denotes the steady-state probability vector of the generator $A_0 + A_1 + A_2$. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{4.2}$$

The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model. Define the traffic intensity ρ_1 as

$$\rho_1 = \frac{\boldsymbol{\pi} A_0 \mathbf{e}}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{4.3}$$

Note that the stability condition in equation (4.2) is equivalent to $\rho_1 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 4.4.

4.3.2 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$, $\boldsymbol{x}_2 = (\boldsymbol{x}_{20}, \boldsymbol{x}_{2\tilde{0}}, \boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \boldsymbol{x}_{23})$ and
$\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$, where $i \geq 3$. The vector $\boldsymbol{x}$ satisfies the condition

$$\boldsymbol{x}Q = 0; \boldsymbol{x}\mathbf{e} = 1, \tag{4.4}$$

where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_3 R^{j-3}, j \geq 4 \tag{4.5}$$

where R is the minimal non-negative solution of the matrix equation

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{4.6}$$

Knowing the matrix R, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are obtained by solving the equations

$$\begin{aligned} -\lambda \boldsymbol{x}_0 + \boldsymbol{x}_1 B_1 &= 0 \\ \boldsymbol{x}_0 D_0 + \boldsymbol{x}_1 C_1 + \boldsymbol{x}_2 B_2 &= 0 \\ \boldsymbol{x}_1 D_1 + \boldsymbol{x}_2 C_2 + \boldsymbol{x}_3 B_3 &= 0 \end{aligned}$$

$$\boldsymbol{x}_2 D_2 + \boldsymbol{x}_3 (A_1 + RA_2) = 0 \tag{4.7}$$

subject to the normalizing condition

$$\boldsymbol{x}_0 + \boldsymbol{x}_1\mathbf{e} + \boldsymbol{x}_2\mathbf{e} + \boldsymbol{x}_3 (I - R)^{-1}\mathbf{e} = 1. \tag{4.8}$$

Once the rate matrix R is obtained, the vector $\boldsymbol{x}$ can be computed by exploiting the special structure of the coefficient matrices.

4.3.3 Performance measures

Note that $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{1\tilde{0}}$, $\boldsymbol{x}_{11}$, $\boldsymbol{x}_{20}$, $\boldsymbol{x}_{2\tilde{0}}$, $\boldsymbol{x}_{21}$, $\boldsymbol{x}_{22}$, $\boldsymbol{x}_{23}$, $\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$, $\boldsymbol{x}_{i3}$ and $\boldsymbol{x}_{i4}$ are vectors of orders $(M+1)a$, $2b$, $2b$, $2(M+1)ab$, b^2, $4b^2$, $2Mab$, $2ab$, $(M+1)ab^2$, $4b^2$, $2(2M-1)ab^2$, $4ab^2$ and $2ab^2$.
Now we compute some performance measures.

(1) Expected number of customers in the system

$$ES = \sum_{i=1}^{\infty} i\boldsymbol{x}_i\mathbf{e}.$$

(2) Expected number of customers in the queue

$$EQ = \sum_{i=3}^{\infty}(i-2)\boldsymbol{x}_{i1}\mathbf{e} + \sum_{i=4}^{\infty}(i-3)[\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i2}\mathbf{e} + \boldsymbol{x}_{i3}\mathbf{e} + \boldsymbol{x}_{i4}\mathbf{e}].$$

(3) Effective rate of interruption

$$EI = \theta_1 \sum_{j=0}^{M-1} \boldsymbol{x}_{20j}\mathbf{e} + \theta_2 \sum_{i=3}^{\infty}\sum_{j=0}^{M-1} \boldsymbol{x}_{i0j}\mathbf{e} + \theta_1 \sum_{i=3}^{\infty}(\boldsymbol{x}_{i21} + \boldsymbol{x}_{i22})\mathbf{e}.$$

(4) Fraction of time the main server is idle

$$F_{mi} = \boldsymbol{x}_0 + \boldsymbol{x}_{1\tilde{0}}\mathbf{e} + \boldsymbol{x}_{2\tilde{0}}\mathbf{e}.$$

(5) Fraction of time the main server is busy serving a customer

$$F_{mb} = \boldsymbol{x}_{10}\mathbf{e} + \sum_{i=2}^{\infty}(\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i3}\mathbf{e}).$$

(6) Fraction of time main server is interrupted

$$F_{min} = \sum_{i=2}^{\infty}(\boldsymbol{x}_{i2}\mathbf{e} + \boldsymbol{x}_{i4}\mathbf{e}).$$

(7) Fraction of time the first regular server is idle

$$F_{ri} = \boldsymbol{x}_0 + \boldsymbol{x}_{10}\mathbf{e} + \boldsymbol{x}_{1\tilde{0}2}\mathbf{e}.$$

(8) Fraction of time the first regular server is busy serving a customer

$$F_{rb} = \boldsymbol{x}_{1\tilde{0}1}\mathbf{e} + \boldsymbol{x}_{2\tilde{0}}\mathbf{e} + \boldsymbol{x}_{212}\mathbf{e} + \sum_{i=3}^{\infty}(\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i12}\mathbf{e} + \boldsymbol{x}_{i22}\mathbf{e} + \boldsymbol{x}_{i32}\mathbf{e}).$$

(9) Fraction of time the first regular server is under consultation

$$F_{rc} = \sum_{i=1}^{\infty}\boldsymbol{x}_{i11}\mathbf{e} + \sum_{i=2}^{\infty}\boldsymbol{x}_{i21}\mathbf{e}.$$

(10) Fraction of time first regular server is waiting to get consultation after the present service at the main server

$$F_{wcs} = \sum_{i=2}^{\infty}\boldsymbol{x}_{i31}\mathbf{e}.$$

(11) Fraction of time first regular server is waiting to get consultation after the present consultation

$$F_{wcc} = \sum_{i=2}^{\infty}(\boldsymbol{x}_{i13}\mathbf{e} + \boldsymbol{x}_{123}\mathbf{e}).$$

(12) Fraction of time first regular server is waiting to get consultation after the present service at the main server and consultation to the regular server

$$F_{wcsc} = \sum_{i=3}^{\infty} \boldsymbol{x}_{i41}\mathbf{e}.$$

4.4 Numerical examples

In this section we examine the effect of λ, θ_1 and θ_2 on various performance measures.
Let us choose the following data so that the stability condition $\rho_1 < 1$ is satisfied. Let
$T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$; $U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$;
$\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$; $\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$; $\xi = 2$; $M = 3$.

From table 4.1 we see that as λ increases the traffic intensity ρ_1 increases as is to be expected. The system is fed with more customers and so more customers are accumulated in the system and in queue. So ES and EQ increase. The main server has to serve more customers which results in an increase in F_{mb}. As number of customers increases, effective rate of interruption to the main server EI and thus the fraction of time

Table 4.1: Effect of λ on various performance measures

$\theta_1 = 3, \theta_2 = 2$

λ	1	1.5	2	2.5	3
ρ_1	0.2161	0.3242	0.4322	0.5403	0.6483
ES	1.1568	1.3054	1.639	2.3727	3.9522
EQ	0.0157	0.0892	0.3201	0.9073	2.2726
EI	0.0432	0.0956	0.1665	0.2542	0.3568
F_{mi}	0.0121	0.0224	0.0317	0.0376	0.0386
F_{rc}	0.0254	0.0566	0.0969	0.1421	0.1871
F_{min}	0.0227	0.0519	0.0936	0.1486	0.2170

Table 4.2: Effect of θ_1 on various performance measures

$\lambda = 3, \theta_2 = 2$

θ_1	2	2.5	3	3.5	4
ρ_1	0.5407	0.5946	0.6483	0.7018	0.7549
ES	2.2804	2.9499	3.9522	5.4581	7.7758
EQ	0.8050	1.3847	2.2726	3.5974	5.4917
EI	0.2476	0.3029	0.3568	0.4078	0.4538
F_{mi}	0.0488	0.0439	0.0386	0.0329	0.0270
F_{mb}	0.2622	0.2453	0.2280	0.2107	0.1935
F_{min}	0.1514	0.1841	0.2170	0.2492	0.2790

the main server stay in interrupted state F_{min} also increase.

We see from table 4.2 that as θ_1 increases, EI also grows faster since it depends directly on θ_1. This results in a hike in F_{min}. So the main server gets lesser time to serve customers which results in a decrease in F_{mb}. As F_{min} increases the main server gets lesser time to be idle and so F_{mi} decreases. As a whole there is a rapid accumulation of customers

Table 4.3: Effect of θ_2 on various performance measures

$$\lambda = 3, \theta_1 = 2$$

θ_2	2	3	4	5	6	7	8
ρ_1	0.5407	0.6258	0.6844	0.7271	0.7594	0.7848	0.8052
ES	2.2804	2.7377	3.2558	3.8234	4.4300	5.0669	5.7268
EQ	0.8050	1.1914	1.6399	2.1378	2.6717	3.2282	3.7953
EI	0.2476	0.2881	0.3245	0.3568	0.3852	0.4100	0.4315
F_{mi}	0.0488	0.0446	0.0408	0.0375	0.0346	0.0321	0.0298
F_{mb}	0.2622	0.2479	0.2346	0.2224	0.2113	0.2013	0.1923
F_{min}	0.1514	0.1844	0.2142	0.2406	0.2639	0.2843	0.3021

in the system and in the queue, and hence ES and EQ increase. Since the effective service time (the sum of the time taken for actual service completion and intermediate consultations) increases, the traffic intensity ρ_1 will increase.

From table 4.3 we see that as θ_2 increases EI increases since EI depends directly on θ_2 and so F_{min} also increases. So the idle time F_{mi} of the main server decreases. Since the fraction of interrupted time of main server increases, the main server gets lesser time to serve customers and so F_{mb} decreases. The accumulation of customers increases, since the time for the service completion of customers increases. Thus both ES and EQ increases. Here also ρ_1 increases.

4.5 Description of model 2

In this model, we consider a 3-server queueing system with different arrival processes. The arrivals to the main server and regular servers are independent Poisson processes with rates λ_1 and λ_2 respectively. The customers to the main and regular servers are called Type 1 and Type 2 customers, respectively. There is a finite buffer of size K at the main server. An arriving Type 1 customer will be lost when the buffer is full. An arriving Type 2 customer enters into service immediately if at least one regular server is free, else joins a queue. If both the regular servers are free, that customer is free to choose any one of the regular servers with probability 1/2.

4.5.1 Notations

In addition to the notations defined in (4.1.1) we use the following notations also in model 2:

- $\mathtt{P}_1 = \begin{bmatrix} \mathbf{0} & I_{K-1} \\ 0 & \underline{\mathbf{0}} \end{bmatrix}_{K \times K}$, $\mathtt{P}_2 = \begin{bmatrix} \underline{\mathbf{0}} & 0 \\ \mathtt{I}_{K-1} & \mathbf{0} \end{bmatrix}_{K \times K}$,
  $\mathtt{P}_3 = \begin{bmatrix} \underline{\mathbf{0}} & 0 \\ \mathtt{I}_K & \mathbf{0} \end{bmatrix}_{K+1 \times K+1}$
- $\lambda = \lambda_1 + \lambda_2$
- $L_1 = diag(\lambda I_{K-1}, \lambda_2)$, $L_2 = diag(\lambda I_K, \lambda_2)$

- $J_1 = I_{K(M+1)a+1}$
- $Z_1 = diag(1, 0)$
- $K_1 = (K+1)(M+1)a + 1$
- $K_2 = 2K_1b + 2(K+1)b + 2KMab + 2Kab$
- $K_3 = K_1b^2 + 4(K+1)b^2 + 2KMab^2 + 2K(M-1)ab^2 + 4Kab^2 + 2Kab^2$

Consider the queueing model $X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), B_1(t), B_2(t), J_1(t), J_2(t), S_2(t), S_1(t), J_3(t)\}$.
Here $N(t)$ is the number of type 2 customers in the system, $S_2(t)$ is the number of type 1 customers in the system and $S(t)$ denotes the status of the servers at time t such that

$$S(t) = \begin{cases} 0, & \text{if the main together with or without regular server(s) are busy} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation with one interrupted customer at the main server} \\ 3, & \text{if the regular server is waiting for getting consultation after the present service at the main server} \\ 4, & \text{if the regular server is waiting for getting consultation after the service at the main server followed by the present interruption} \end{cases}$$

All other variables are as those defined in model 1 of this chapter.
$\{X(t), t \geq 0\}$ is a CTMC with state space

$$\Psi = \bigcup_{i=0}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as
$\psi(0) = \{(0,0) \cup (0,j,k,t_3) : 1 \leq j \leq K+1, 0 \leq k \leq M\}$;
$\psi(1) = \psi(1,0) \cup \psi(1,1) \cup \psi(1,2) \cup \psi(1,3)$ and
$\psi(i) = \psi(i,0) \cup \psi(i,1) \cup \psi(i,2) \cup \psi(i,3) \cup \psi(i,4)$, for $i \geq 2$,

where
$\psi(1,0) = \{(1,0,l,t_l,0) \cup (1,0,l,t_l,j,k,t_3) : 0 \leq j \leq K+1\}$,
$\psi(1,1) = \{(1,1,l,t_l,j) : 0 \leq j \leq K\}$,
$\psi(1,2) = \{(1,2,l,t_l,j,k,t_2) : 1 \leq j \leq K, 0 \leq k \leq M-1\}$ and
$\psi(1,3) = \{(1,3,l,t_l,j,t_2) : 1 \leq j \leq K\}$, for $1 \leq l \leq 2$;
and for $i \geq 2$,
$\psi(i,0) = \{(i,0,t_1,t_2,j,k,t_3) : 0 \leq j \leq K+1, 0 \leq k \leq M\}$,
$\psi(i,1) = \{(i,1,l,t_1,t_2,j) : 0 \leq j \leq K\}$,
$\psi(i,2) = \{(i,2,l,t_1,t_2,j,k,t_3) : 1 \leq j \leq K, 0 \leq k \leq M-1\}$,
$\psi(i,3) = \{(i,3,l,t_1,t_2,j,k,t_3) : 1 \leq j \leq K\}$ and
$\psi(i,4) = \{(i,4,l_1,t_1,t_2,j,t_3) : 1 \leq j \leq K\}$;
with $1 \leq t_1, t_2 \leq b$ and $1 \leq t_3 \leq a$, $1 \leq l \leq 4$.
The infinitesimal generator matrix Q is given by

$$Q = \begin{bmatrix} B_0 & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & & \end{bmatrix} \tag{4.9}$$

Here A_0, A_1 and A_2 are square matrices of orders K_3, B_0, B_3 are square matrices of orders K_1 and K_2 respectively. B_1, B_2, B_4 and B_5 are matrices of orders $K_1 \times K_2$, $K_2 \times K_1$, $K_2 \times K_3$ and $K_3 \times K_2$ respectively.

$$B_0 = \begin{bmatrix} -\lambda & \mathtt{B}_{01} \\ \mathtt{B}_{02} & \mathtt{B}_{03} \end{bmatrix},\ B_1 = \tfrac{\lambda_2}{2}\begin{bmatrix} \mathbf{e}_2' \otimes \boldsymbol{\beta} \otimes J_1 & O \end{bmatrix},$$

$$B_2 = \begin{bmatrix} \mathbf{e}_2 \otimes U^0 \otimes J_1 \\ O \end{bmatrix},\ B_3 = \begin{bmatrix} \mathtt{B}_{31} & \mathtt{B}_{32} & \mathtt{B}_{33} \\ \mathtt{B}_{34} & \mathtt{B}_{35} & \\ \mathtt{B}_{36} & \mathtt{B}_{37} & \mathtt{B}_{38} \end{bmatrix},$$

$$B_4 = \lambda_2 \begin{bmatrix} diag(\nabla \otimes \mathbf{e}_2 \otimes J_1, \mathtt{B}_{41}, \mathtt{B}_{42}, \mathtt{B}_{43}) & O \end{bmatrix},$$

$$B_5 = \begin{bmatrix} diag(\mathbf{e}_2' \otimes \Delta \otimes J_1, \mathtt{B}_{51}, \mathtt{B}_{52}, \mathtt{B}_{53}) \\ O \end{bmatrix},$$

$$A_0 = \lambda_2 I, \quad A_1 = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{14} & A_{15} & \\ A_{16} & A_{17} & A_{18} \end{bmatrix},$$

$$A_2 = diag(\mathbf{e}_2' \otimes \hat{\Delta} \otimes \mathbf{e}_2 \otimes J_1, A_{21}, A_{22}, A_{23}, O).$$

Here

$$\mathtt{B}_{01} = \lambda_1 \begin{bmatrix} \boldsymbol{\alpha} & \mathbf{0} \end{bmatrix}_{1 \times K(M+1)a},\ \mathtt{B}_{02} = \begin{bmatrix} \mathbf{e}_{M+1} \otimes \mathtt{T}^0 \\ \mathbf{0} \end{bmatrix}_{K(M+1)a \times 1},$$

$$\mathtt{B}_{03} = I_{(M+1)K} \otimes \mathtt{T} + \lambda_1 \mathtt{P}_1 \otimes \mathtt{I}_{(M+1)a} + \mathtt{P}_2 \otimes \mathbf{e}_{M+1} \otimes \mathtt{T}^0 \otimes \boldsymbol{\alpha} - \mathtt{L}_1 \otimes \mathtt{I}_{a(M+1)},$$

$$\mathtt{B}_{31} = I_2 \otimes (I_b \otimes B_0 + (U - \theta_1 I_b) \otimes J_1),$$

$$\mathtt{B}_{32} = \theta_1 I_{2b} \otimes Z_1,\ \mathtt{B}_{33} = \theta_1 \begin{bmatrix} I_{2b} \otimes E_1 & I_{2b} \otimes E_2 \end{bmatrix}_{2K_1 b \times 2K(M+1)ab},$$

$$\mathtt{B}_{34} = \xi I_{2b} \otimes \Omega,\ \mathtt{B}_{35} = I_{2b} \otimes (-\xi I_{K+1} - L_2 + \lambda_1 P_3),$$

$$\mathtt{B}_{36} = \begin{bmatrix} \xi I_{2b} \otimes G_1 \\ O \end{bmatrix}_{2K(M+1)ab \times 2K_1 b},\ \mathtt{B}_{37} = \begin{bmatrix} O \\ I_{2b} \otimes T^* \end{bmatrix}_{2K(M+1)ab \times 2(K+1)b},$$

$$\mathtt{B}_{38} = diag(I_{2b} \otimes F_1, I_{2b} \otimes F_2)_{2Kab \times 2Kab},$$

$$\mathtt{B}_{41} = \begin{bmatrix} \nabla & O \end{bmatrix}_{2b \times 4b^2} \otimes I_{K+1},$$

$$\mathtt{B}_{42} = \begin{bmatrix} \nabla \otimes I_{KM} & O \end{bmatrix}_{2KMb \times 2K(2M-1)b^2} \otimes I_a,$$

$$\mathtt{B}_{43} = \begin{bmatrix} \nabla & O \end{bmatrix}_{2b \times 2b^2} \otimes I_{Ka},$$

$$\mathtt{B}_{51} = \begin{bmatrix} \Delta \\ O \end{bmatrix}_{4b^2 \times 2b} \otimes I_{K+1},\ \mathtt{B}_{52} = \begin{bmatrix} \Delta \otimes I_{KM} \\ O \end{bmatrix}_{2K(2M-1)b^2 \times 2KMb} \otimes I_a,$$

$$\mathtt{B}_{53} = \begin{bmatrix} \Delta \\ O \end{bmatrix}_{4b^2 \times 2b} \otimes I_{Ka},$$

$$A_{11} = I_{b^2} \otimes B_0 + (U \oplus U - 2\theta_2 I_{b^2}) \otimes J_1,$$

$$A_{12} = \theta_2 \begin{bmatrix} \mathbf{e}_2' \otimes Z_1 & O \end{bmatrix}_{K_1 \times 4(K+1)} \otimes I_{b^2},$$

$$A_{13} = \theta_2 \begin{bmatrix} M_1 & M_2 & O \end{bmatrix}_{K_1 b^2 \times 4K(M+1)b^2} \otimes I_a,$$

$$A_{14} = \xi \begin{bmatrix} \mathbf{e}_2 \otimes I_{b^2} \otimes \Omega \\ O \end{bmatrix}_{4(K+1)b^2 \times K_1 b^2},$$

$$A_{15} = \begin{bmatrix} H_1 & \theta_1 I_{2b^2(K+1)} \\ \xi \ddot{I} \otimes I_{b^2(K+1)} & H_2 \end{bmatrix}_{4(K+1)b^2 \times 4(K+1)b^2},$$

$$A_{16} = \xi \begin{bmatrix} \mathbf{e}_2 \otimes I_{b^2} \otimes G_1 \\ O \end{bmatrix}_{2K(2M-1)ab^2 \times K_1 b^2},$$

$$A_{17} = \begin{bmatrix} O \\ I_{4b^2} \otimes T^* \\ O \end{bmatrix}_{2K(2M-1)ab^2 \times 4(K+1)b^2},$$

$$A_{18} = \begin{bmatrix} V_1 & O & V_2 \\ & V_3 & \\ O & V_4 & V_5 \end{bmatrix}_{2K(2M-1)ab^2 \times 2K(2M-1)ab^2} \otimes I_a,$$

$$A_{21} = diag(\hat{\Delta}, O)_{4b^2 \times 4b^2} \otimes I_{K+1},$$

$$A_{22} = diag(\hat{\Delta} \otimes I_{KM}, O)_{2K(2M-1)b^2 \times 2K(2M-1)b^2} \otimes I_a,$$

$$A_{23} = diag(\hat{\Delta}, O)_{4b^2 \times 4b^2} \otimes I_{Ka}.$$

We describe the following terms:

$$E_1 = \begin{bmatrix} \mathbf{0} \\ I_K \otimes \hat{I}_M \otimes I_a \end{bmatrix}_{K_1 \times KMa},$$

$$E_2 = \begin{bmatrix} \mathbf{0} \\ I_K \otimes \hat{\mathbf{e}}_M \otimes I_a \end{bmatrix}_{K_1 \times Ka},$$

$$\Omega = \begin{bmatrix} 1 & 0 \\ O & I_K \otimes \tilde{\boldsymbol{\alpha}} \end{bmatrix},$$

$$G_1 = \begin{bmatrix} O & I_K \otimes \dot{I}_M \otimes I_a \end{bmatrix},$$

$$T^* = \begin{bmatrix} I_K \otimes T^0 & \mathbf{0} \end{bmatrix}_{Ka \times K+1},$$

$$F_1 = (\xi I_K - L_1 + \lambda_1 P_1) \otimes I_{Ma},$$

$$F_2 = I_K \otimes T + (-L_1 + \lambda_1 P_1) \otimes I_a,$$

$$M_1 = \begin{bmatrix} \mathbf{e}_2' \otimes I_{b^2} \otimes E_3 & O \end{bmatrix}_{K_1 b^2 \times 2K(2M-1)b^2},$$

$$M_2 = \begin{bmatrix} \mathbf{e}_2' \otimes I_{b^2} \otimes E_4 & O \end{bmatrix}_{K_1 b^2 \times K b^2},$$

$$H_1 = I_{2b^2} \otimes (\lambda_1 P_1 - (\xi + \theta_1) I_{K+1} - L_2) + I_2 \otimes \Delta^* \otimes I_{K+1},$$

$$H_2 = I_2 \otimes I_{b^2} \otimes (\lambda_1 P_1 - \xi I_{K+1} - L_2),$$

$$V_1 = \begin{bmatrix} H_3 & \theta_1 I_{2b^2 K} \otimes \hat{I}_{M-1} \\ \xi \ddot{I} \otimes I_{b^2 K} \otimes \dot{I}_{M-1} & H_4 \end{bmatrix},$$

$$V_2 = \theta_1 \begin{bmatrix} I_{2b^2K} \otimes \hat{\mathbf{e}}_{M-1} \\ O \end{bmatrix}_{2K(2M-1)ab^2 \times 2Kab^2},$$

$$V_3 = I_{4b^2} \otimes (I_K \otimes T - L_1 \otimes I_a) + \begin{bmatrix} \mathbf{e}_2' \otimes \theta_1 I_{2Kab^2} \\ O \end{bmatrix} + \begin{bmatrix} \Delta^* \otimes I_{Ka} & O \\ O & O \end{bmatrix},$$

$$V_4 = \xi I_{2b^2} \otimes \ddot{I} \otimes I_{Ka},$$

$$V_5 = I_{2b^2} \otimes (\lambda_1 P_1 - \xi I_K - L_1 \otimes I_a),$$

$$E_3 = \begin{bmatrix} 0 \\ I_K \otimes \hat{I}_M \end{bmatrix}_{K(M+1)+1 \times KM},$$

$$E_4 = \begin{bmatrix} 0 \\ I_K \otimes \hat{\mathbf{e}}_M \end{bmatrix}_{K_1 \times K},$$

$$H_3 = I_{2b^2} \otimes (\lambda_1 P_1 - (\xi + \theta_1) I_{KM} - L_1 \otimes I_M) + \Delta * \otimes I_{KM},$$

$$H_4 = I_{2b^2} \otimes (\lambda_1 P_1 - \xi I_{K(M-1)} - L_1 \otimes I_{M-1}).$$

4.6 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study. We first establish the stability condition of the queueing system.

4.6.1 Stability condition

Let $\boldsymbol{\pi}$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{4.10}$$

The vector $\boldsymbol{\pi}$ cannot be obtained explicitly in terms of the parameters of the model.

Define the traffic intensity ρ_2 as

$$\rho_2 = \frac{\boldsymbol{\pi} A_0 \mathbf{e}}{\boldsymbol{\pi} A_2 \mathbf{e}}. \tag{4.11}$$

Note that the stability condition in equation (4.10) is equivalent to $\rho_2 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 4.7

4.6.2 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3, \tag{4.12}$$

where R is the minimal non-negative solution of the matrix equation

$$R^2A_2 + RA_1 + A_0 = 0. \tag{4.13}$$

Knowing the matrix R, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained by solving the equations

$$\begin{aligned} \boldsymbol{x}_0B_0 + \boldsymbol{x}_1B_2 &= 0 \\ \boldsymbol{x}_0B_1 + \boldsymbol{x}_1B_3 + \boldsymbol{x}_2B_5 &= 0 \\ \boldsymbol{x}_1B_4 + \boldsymbol{x}_2(A_1 + RA_2) &= 0 \end{aligned} \tag{4.14}$$

subject to the normalizing condition

$$\boldsymbol{x}_0\mathbf{e} + \boldsymbol{x}_1\mathbf{e} + \boldsymbol{x}_2(I - R)^{-1}\mathbf{e} = 1. \tag{4.15}$$

4.6.3 Expected number of interruptions to a customer at the main server

To compute the expected number of interruptions faced by a customer during his service, we consider the Markov process $Z(t) = \{(N_1(t), \hat{S}(t), J_3(t)) : t \geq 0\}$, where $N_1(t)$ is the number of interruptions already befell to the main server, $\hat{S}(t) = S(t) - \{1, 3, 4\}$ and all other variables are as defined earlier. Z has the state space $\{(i, j, t_1) : 0 \leq i \leq M, 1 \leq t_1 \leq a\} \cup \{\Delta\}$, where Δ is the absorbing state which denotes the customer leaves the system after service completion. Thus the infinitesimal generator $\tilde{V}$ of the process $Z(t)$ takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$$

where

$$V = \begin{bmatrix} G_1 & G_0 & & & & & \\ & G_1 & G_0 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & & & & G_1 & \tilde{G}_0 & \\ & & & & & \tilde{G}_1 & \hat{G}_0 \\ & & & & & & \hat{G}_1 \end{bmatrix}_{M+1 \times M+1} \text{ and } V_0 = \begin{bmatrix} G_2 \\ \vdots \\ \vdots \\ G_2 \\ \tilde{G}_2 \\ \hat{G}_2 \end{bmatrix}.$$

Here

$G_0 = \xi \begin{bmatrix} \underline{\mathbf{0}} & 0 \\ I_2 & \mathbf{0} \end{bmatrix} \otimes I_a$, $\tilde{G}_0 = \xi \begin{bmatrix} \underline{\mathbf{0}} \\ I_2 \end{bmatrix} \otimes I_a$, $\hat{G}_0 = \xi \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes I_a$;

$$G_1 = diag(T, O)_{3a \times 3a} + \begin{bmatrix} -\theta_2 & \theta_2 & \\ & -(\theta_1 + \xi) & \theta_1 \\ & & -\xi \end{bmatrix} \otimes I_a,$$

$\tilde{G}_1 = diag(T, O)_{2a \times 2a} + \begin{bmatrix} -\theta_2 & \theta_2 \\ & -\xi \end{bmatrix} \otimes I_a$, $\hat{G}_1 = T$;

$G_2 = \mathbf{e}_3(1) \otimes T^0$, $\tilde{G}_2 = \mathbf{e}_2(1) \otimes T^0$, $\hat{G}_2 = T^0$.

If z_j is the probability that there are exactly j interruptions during a single service, then

$z_j = \boldsymbol{\eta}(-G_1^{-1}G_0)^j(-G_1^{-1}G_2)\mathbf{e}$, $j = 0, 1, ..., M-2$,

$z_{M-1} = \boldsymbol{\eta}(-G_1^{-1}G_0)^{M-2}(-G_1^{-1}\tilde{G}_0)(-\tilde{G}_1^{-1}\tilde{G}_2)\mathbf{e}$,

$z_M = \boldsymbol{\eta}(-G_1^{-1}G_0)^{M-2}(-\tilde{G}_1^{-1}\tilde{G}_0)(-\tilde{G}_1^{-1}\hat{G}_0)(-\hat{G}_1^{-1}\hat{G}_2)\mathbf{e}$,
where $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \mathbf{0})$.
Expected number of interruptions during a single service is given by

$$E(NI) = \sum_{j=0}^{M} j z_j.$$

4.6.4 Performance measures

Note that $\boldsymbol{x}_0 = (\boldsymbol{x}_{00}, \boldsymbol{x}_{0k}), 1 \leq k \leq K+1$, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i21}, \boldsymbol{x}_{i22}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$, for $i \geq 2$. Here x_{00} is a scalar, $\boldsymbol{x}_{0k}$, $\boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}$, $\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i21}, \boldsymbol{x}_{i22}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$, for $i \geq 2$ are vectors of dimensions $(M+1)a$, $2K_1b$, $2(K+1)b$, $2KMab$, $2Kab$, K_1b^2, $4(K+1)b^2$, $2KMab^2$, $2K(M-1)ab^2$, $4Kab^2$ and $2Kab^2$ respectively.

(1) Expected number of customers in the system

$$ES = \sum_{j=1}^{K} j\boldsymbol{x}_{00j} + \sum_{m=0}^{1}\sum_{t_2=1}^{b}\sum_{j=0}^{K}(1+j)\boldsymbol{x}_{1mt_1j} + \sum_{m=2}^{3}\sum_{t_2=1}^{b}\sum_{j=1}^{K}(1+j)\boldsymbol{x}_{1mt_1j}$$

$$+\sum_{i=2}^{\infty}[\sum_{m=0}^{2}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\sum_{j=0}^{K}(i+j)\boldsymbol{x}_{imt_1t_2j}\mathbf{e}+\sum_{m=3}^{4}\sum_{t_2=1}^{b}\sum_{t_2=1}^{b}\sum_{j=1}^{K}(i+j)\boldsymbol{x}_{imt_1t_2j}\mathbf{e}].$$

(2) Expected number of type 1 customers in the system

$$ES_1 = \sum_{j=1}^{K} j\boldsymbol{x}_{0j}\mathbf{e} + \sum_{m=0}^{3}\sum_{t_1=1}^{b}\sum_{j=1}^{K} j\boldsymbol{x}_{1mt_1j}\mathbf{e}$$

$$+\sum_{i=2}^{\infty}\sum_{m=0}^{4}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\sum_{j=1}^{K} j\boldsymbol{x}_{imt_1t_2j}\mathbf{e}$$

(3) Expected number of type 2 customers in the system

$$ES_2 = \sum_{i=1}^{\infty} i\boldsymbol{x}_i\mathbf{e}.$$

(4) Expected number of type 2 customers in the queue

$$EQ_2 = \sum_{i=3}^{\infty}(i-2)\boldsymbol{x}_i\mathbf{e}.$$

(5) Probability that an arriving Type 1 customer is lost due lack of room in buffer

$$Pro(L) = \boldsymbol{x}_{0K}\mathbf{e} + \sum_{m=0}^{3}\sum_{t_1=1}^{b}\boldsymbol{x}_{1mt_1K}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{m=0}^{4}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\boldsymbol{x}_{imt_1t_2K}\mathbf{e}.$$

(6) Probability that the system is idle

$$Pro(SI) = \boldsymbol{x}_{00}.$$

(7) Probability that the main server is idle

$$Pro(MI) = \boldsymbol{x}_{00} + \sum_{t_1=1}^{b}\boldsymbol{x}_{10t_10}\mathbf{e} + \sum_{i=2}^{\infty}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\boldsymbol{x}_{i0t_1t_20}\mathbf{e}.$$

(8) Probability that both the regular servers are idle

$$Pro(RI) = \boldsymbol{x}_0\mathbf{e}.$$

(9) Probability that all the servers are busy

$$Pro(AB) = \sum_{i=2}^{\infty}\sum_{m=0}^{4}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\sum_{j=1}^{K} \boldsymbol{x}_{imt_1t_2j}\mathbf{e}.$$

(10) Probability that the service completion takes place at the main server without any interruption

$$Pro(WI) = \sum_{i=1}^{\infty}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\sum_{j=1}^{K}\sum_{t=1}^{a} \boldsymbol{x}_{i0t_1t_2j0t}T_t^0.$$

(11) Probability that the service completion takes place at the main server with at least one interruption

$$Pro(WAI) = \sum_{i=1}^{\infty}\sum_{t_1=1}^{b}\sum_{t_2=1}^{b}\sum_{j=1}^{K}\sum_{l=1}^{M}\sum_{t=1}^{a} \boldsymbol{x}_{i0t_1t_2jlt}T_t^0.$$

4.6.5 An optimization problem

In this section we propose an optimization problem and discuss it through an illustrative example. To construct an objective function we assume that the service produces revenue to the system whereas idle servers and waiting spaces involve expenditure to the system. Thus we produce per unit time revenue and cost as follows:

1. r_1 be the revenue per customer leaving the system after service completion

2. c be the holding cost of type 1 customers in the system

3. r_2 be revenue loss due to buffer is full

4. r_3 be revenue due to consultation obtained by interrupting a customer

The problem of interest is to find an optimum value of the number of servers to be employed so that the expected total profit ETP is maximum. The objective function is as follows:

$$ETP = r_1 ESR - cES_1 - r_2 Pro(L) + r_3 \times E(NI). \tag{4.16}$$

Here $ESR = \boldsymbol{\pi} A_2 \boldsymbol{e}$.

4.7 Numerical results

In this section, we present some numerical examples that describe the performance characteristics under study.

Example 4.7.1. The purpose of this example to study the effect of M and K on the expected total profit ETP. Fix
$T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$, $U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$, $\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$,

$\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$, $\theta_1 = 3$, $\theta_2 = 2$, $\xi = 2$, $\lambda_1 = 10$, $\lambda_2 = 4$.

The above data is so chosen that the stability condition $\rho_2 < 1$ is satisfied.
Fix $r_1 = 100$, $c = 50$, $r_2 = 30$ and $r_3 = 20$.

Table 4.4: Effect of M on various performance measures

$K = 3$

M	3	4	5	6	7
ESR	4.2706	4.0691	3.9482	3.8716	3.8214
ES_1	1.6728	1.7143	1.7577	1.7849	1.8036
$Prob(L)$	0.5291	0.5451	0.5564	0.5645	0.5705
$E(NI)$	0.2902	0.3932	0.4832	0.5588	0.6206
ETP	333.35	312.71	299.91	292.17	287.26

Table 4.5: Effect of K on various performance measures

$M = 4$

K	4	5	6	7	8
ESR	4.0691	4.0241	3.9252	3.8539	3.7963
ES_1	2.3071	2.9200	3.5362	4.1460	4.7602
$Prob(L)$	0.5396	0.5366	0.5347	0.5334	0.5322
$E(NI)$	0.3932	0.3932	0.3932	0.3932	0.3932
ETP	283.23	248.18	207.53	169.95	133.52

Table 4.4 shows that as number of interruptions possible to the main server during the service of a customer M increases, naturally, the effective service rate ESR will decrease. This results in a faster accumulation of customers at the main server until the buffer is full and so ES_1 increases. Thus the probability for loss of customers at the main server due to buffer is full also increases. Since M increases, expected number of interruptions

$E(NI)$ also increases as is to be expected. Thus the ETP decreases with increase in M.

Table 4.5 shows that $E(NI)$ remains a constant as the buffer size K increases. This is so because $E(NI)$ does not depend on K. As K increases ES_1, the expected number of type 1 customers increases. Thus those customers get more space at the main server. So $Pro(L)$ decreases and ESR decreases. Here also ETP decreases with increase in K.

Example 4.7.2. In this section we analyse the effect of the parameters λ_1 and λ_2 on the performance measures. Let T, U, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and ξ are as in example **4.7.1.** Choose $\theta_1 = 1$, $\theta_2 = 1$, $M = 2$ and $K = 3$.
The stability condition $\rho_2 < 1$ is satisfied for the above matrices, vectors and values.

Table 4.6: Effect of λ_1 on various performance measures

$\lambda_2 = 2$

λ_1	1	1.5	2	2.5	3	3.5
ρ_2	0.4627	0.5406	0.6184	0.6962	0.7740	0.8518
ES_2	0.7094	0.7100	0.7105	0.7109	0.7113	0.7116
EQ_2	0.1268	0.1272	0.1275	0.1278	0.1280	0.1282
$Pro(L)$	0.0074	0.0208	0.0395	0.0626	0.0895	0.1197
$Pro(SI)$	0.4522	0.3994	0.3484	0.3000	0.2551	0.2144
$Pro(MI)$	0.6050	0.5465	0.4873	0.4290	0.3728	0.3201
$Pro(RI)$	0.5570	0.5569	0.5568	0.5567	0.5567	0.5566

Table 4.6 shows that as λ_1 increases the buffer will be filled in an increased rate, so $Prob(L)$ increases, as is to be expected. Since the system

Table 4.7: Effect of λ_2 on various performance measures

$$\lambda_1 = 2$$

λ_2	1	1.5	2	2.5	3	3.5	4
ρ_2	0.4638	0.5411	0.6184	0.6957	0.7730	0.8503	0.9276
ES_2	0.2984	0.4820	0.7105	1.0087	1.4125	1.972	2.7438
EQ_2	0.0137	0.0497	0.1275	0.2725	0.5213	0.9251	1.5447
$Pro(L)$	0.0169	0.0271	0.0395	0.0542	0.0708	0.089	0.1077
$Pro(SI)$	0.4876	0.4136	0.3484	0.2909	0.2399	0.1947	0.1546
$Pro(MI)$	0.6004	0.5408	0.4873	0.4390	0.395	0.3546	0.3175
$Pro(RI)$	0.7538	0.6501	0.5568	0.4723	0.3953	0.3249	0.2605

is fed with more type 1 customers, the main server has to serve more customers and so the main server's idle time $Pro(MI)$ reduces. Then there will be a slight delay for the regular server to get consultations. Thus the expected number of type 2 customers ES_2 increases slightly even if ES_2 does not depend on λ_1 directly. So $Pro(RI)$ has a slight decrease. As a whole system's idle time reduces.

From table 4.7 we can see that as λ_2 increases more type 2 customers accumulate in the system and in the queue. Therefore ES_2 and EQ_2 increase. Thus main server has to spend more time in consultation to the regular servers. By this time type 1 customers accumulate at main server in a faster rate and this increases $Pro(L)$, the probability for loss of type 1 customers due to buffer is full. Any how the busy time at all the servers increases and thus the idle times $Pro(SI)$, $Pro(MI)$ and $Pro(RI)$ decrease.

Chapter 5

A multi-server queue with consultations

5.1 Introduction

We analysed queueing models with two and three servers in the previous chapters. The interruptions to the main server and consultations to the regular servers are controlled by certain parameters such as upper bounds on number of interruptions, number of consultations and total duration of service interruption of a customer at the main server. Service times at all the servers follow phase type distributions. In this chapter, we analyse a multi-server queueing model with $c+1$ servers, namely a main server and c identical regular servers. There is a common queue of customers. Service time at the main server follows phase type distribution and that at the regular servers follow i.i.d. exponential distribution. There are no upper bounds on the number of interruptions to the main server and of consultations to the regular servers. We derive an explicit expression for the system stability. An expression for expected number of interruptions to a customer at the main server is derived. We discuss an optimization problem to determine the number of regular servers to be employed to maximize the expected total profit ETP. Some important performance measures are studied numerically.

5.1.1 Model description

Customers arrive according to a Poisson process with rate λ. An arriving customer enters into service immediately if at least one of the servers is free. Whenever the main server is free, the arriving customer will be served by the main server. The service time of the customers at the main server has phase type distribution with representation $(\boldsymbol{\alpha}, T)$. The service times of the c regular servers are i.i.d. exponential random variables with parameter μ. If there are i regular servers busy $(1 \leq i \leq c)$, then the rate of requirement of consultation is $i\theta$. The request for consultation will be attended immediately. Then the main server (and hence the customer's service at the main server) is said to be interrupted. However, if the main server is busy offering consultation to a regular server, any other regular server requiring consultation will be queued up. Thus, the regular servers are offered consultation on a FIFO basis by the main server. Note that at any given time there can be a maximum of c servers requiring consulting work. The interrupted customer's service will be resumed by the main server only after all consultations in the queue are completed. We assume that the duration of consultation is exponentially distributed

with parameter ξ. Here the service of customers at the regular servers whose servers need consultation during their services is not considered to be interrupted since such consultations are considered to be a part of their services.
The main server in the system can be in any one of the following states:

(1) Main server together with none, one or more regular servers are busy serving customers

(2) Main server is idle and none, one or more regular servers are busy

(3) Main server is giving consultation only

(4) Main server is giving consultation with one interrupted customer

Notations: We use the following notations in this chapter.

- $\mathbf{f}_i = \mathbf{e}_i(1)$
- $\tilde{I}_i = \begin{bmatrix} I_i & \mathbf{0} \end{bmatrix}_{i \times (i+1)}$
- $\omega = \theta + \mu$

Consider the queueing model $X = \{X(t), t \geq 0\}$,
where $X(t) = \{N(t), S(t), J(t), K(t)\}$, where

- $N(t)-$ the number of customers in the system
- $J(t)-$ number of regular servers in the queue for consultation

- $K(t)-$phase of service of the customer at the main server (the service of that customer may be under interruption)

Here $S(t)$ denotes the status of the servers at time t such that

$$S(t) = \begin{cases} \tilde{0}, & \text{if the regular server(s) are busy and main server is idle} \\ 0, & \text{the main along with (or without) regular server is busy} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation} \\ & \text{with one interrupted customer at the main server} \end{cases}$$

$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as

$$\begin{aligned} \psi(1) &= \{(1,0,t_1) : 1 \leq t_1 \leq a\} \cup \{(1,\tilde{0}) \cup (1,1)\}, \\ \psi(i) &= \{(i,0,t_1) : 1 \leq t_1 \leq a\} \cup \{(i,\tilde{0})\} \cup \{(i,1,j) : 1 \leq j \leq i\} \\ & \quad \cup \{(i,2,j,t_1) : 1 \leq j \leq i-1, 1 \leq t_1 \leq a\}, \text{ for } 2 \leq i \leq c, \\ \psi(i) &= \{(i,0,t_1) : 1 \leq t_1 \leq a\} \cup \{(i,1,j) : 1 \leq j \leq c\} \\ & \quad \cup \{(i,2,j,t_1) : 1 \leq j \leq c, 1 \leq t_1 \leq a\}, \text{ for } i \geq c+1. \end{aligned}$$

The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} -\lambda & C_0 & & & & & & & & \\ D_1 & B_1 & C_1 & & & & & & & \\ & D_2 & B_2 & C_2 & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & & D_c & B_c & C_c & & & \\ & & & & & D_{c+1} & A_1 & A_0 & & \\ & & & & & & A_2 & A_1 & A_0 & \\ & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & \end{bmatrix} \tag{5.1}$$

where $C_0 = \lambda \begin{bmatrix} \boldsymbol{\alpha} & \mathbf{0} \end{bmatrix}_{1\times a+2}$;

$$C_1 = \lambda \begin{bmatrix} I_a & O & O \\ \boldsymbol{\alpha} & 0 & O \\ \mathbf{0} & 0 & \tilde{I}_1 \end{bmatrix}_{(a+2)\times(2a+3)} ; \text{ where}$$

$$C_i = \lambda \begin{bmatrix} C_{i1} & \mathbf{0} & C_{i2} \end{bmatrix}_{(1+i+ia)\times(i+2+(i+1)a)}, \text{ for } 2 \le i \le c-1, \text{ with}$$

$$C_{i1} = \begin{bmatrix} I_a \\ \boldsymbol{\alpha} \\ O \end{bmatrix}_{(1+i+ia)\times a} \text{ and } C_{i2} = \begin{bmatrix} O \\ diag(\tilde{I}_i, \tilde{I}_{i-1} \otimes I_a) \end{bmatrix}_{(1+i+ia)\times(i+1+ia)};$$

$$C_c = \lambda \begin{bmatrix} I_a \\ diag(\boldsymbol{\alpha}, I_c, \tilde{I}_{c-1} \otimes I_a) \end{bmatrix}_{(1+c+ca)\times r};$$

$$A_0 = \lambda I;$$

$$B_1 = \begin{bmatrix} T & & \\ & -\omega & \theta \\ & \xi & -\xi \end{bmatrix}_{a+2\times a+2} - \lambda I;$$

$$B_i = \begin{bmatrix} T-(i-1)\omega I_a & & & (i-1)\theta\mathbf{f}_{i-1}^{'}\otimes I_a \\ & -i\omega & i\theta\mathbf{f}_i^{'} & \\ & \xi\mathbf{f}_i & F_i & \\ \xi\mathbf{f}_{i-1}\otimes I_a & & & F_{i-1}\otimes I_a \end{bmatrix} - \lambda I,$$

for $2 \leq i \leq c$;

$$A_1 = \begin{bmatrix} T-c\omega I_a & O & c\theta\mathbf{f}_c^{'}\otimes I_a \\ \xi\mathbf{f}_c\otimes\boldsymbol{\alpha} & F_c & \\ \xi\mathbf{f}_c\otimes I_a & & F_c\otimes I_a \end{bmatrix} - \lambda I, \text{ where}$$

$$F_i = \begin{bmatrix} -(i-1)\omega & (i-1)\theta & & & & & \\ \xi & -(i-2)\omega & (i-2)\theta & & & & \\ & \xi & -(i-3)\omega & (i-3)\theta & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & \xi & -\omega & \theta \\ & & & & & \xi & 0 \end{bmatrix}_{i\times i} - \xi I_i;$$

$$D_1 = \begin{bmatrix} T^0 \\ \mu \\ \underline{\mathbf{0}} \end{bmatrix}_{(a+2)\times 1}; \; D_2 = \begin{bmatrix} \mu I_a & T^0 & \\ & 2\mu & \\ & & E_1 \\ O & O & O \end{bmatrix}_{(3+2a)\times(a+2)};$$

$$D_i = \begin{bmatrix} (i-1)\mu I_a & T^0 & & \\ & i\mu & & \\ & & E_{i-1} & \\ & & & E_{i-2}\otimes I_a \end{bmatrix}_{(i+1+ia)\times(i+(i-1)a)},$$

for $3 \leq i \leq c$;

$$D_{c+1} = \begin{bmatrix} c\mu I_a & T^0 & & \\ & & A_{21} & \\ & & & E_{c-1} \otimes I_a \end{bmatrix}_{(a+c+ca)\times(c+1+ca)}, \text{ where}$$

$$E_i = \begin{bmatrix} diag(i\mu, (i-1)\mu ..., \mu) \\ \mathbf{0} \end{bmatrix}_{(i+1)\times i}, \text{ for } 2 \leq i \leq c;$$

$$A_2 = \begin{bmatrix} T^0 \otimes \boldsymbol{\alpha} + c\mu I_a & & \\ & A_{21} & \\ & & A_{21} \otimes I_a \end{bmatrix}, \text{ where}$$

$A_{21} = diag((c-1)\mu, (c-2)\mu, ..., \mu, 0)$.

Here A_0, A_1 and A_2 are matrices of order $c + (c+1)a$.

5.2 Steady state analysis

In this section we discuss the steady-state analysis of the model under study. We first establish the stability condition of the queueing system.

5.2.1 Stability condition

Let the steady-state probability vector of the generator $A = A_0 + A_1 + A_2$ be denoted by $\boldsymbol{\pi}$. That is,

$$\boldsymbol{\pi} A = 0 \tag{5.2}$$

$$\boldsymbol{\pi}\mathbf{e} = 1 \tag{5.3}$$

The following theorem gives the stability of the queueing system under study.

Theorem 5.3.1 : The Markov Chain X is stable if and only if

$$\lambda < \frac{1}{\zeta}[\mu_1 + \mu \sum_{i=1}^{c} \frac{c!}{(i-1)!}(\frac{\theta}{\xi})^{c-i}] \tag{5.4}$$

where μ_1 and μ are the service rates of the main and regular servers respectively and

$$\zeta = \sum_{i=0}^{c} \frac{c!}{(c-i)!}(\frac{\theta}{\xi})^{i}. \tag{5.5}$$

Proof. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}. \tag{5.6}$$

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, where
$\boldsymbol{\pi}_1 = (\pi_{11}, ..., \pi_{1c})$ and $\boldsymbol{\pi}_2 = (\boldsymbol{\pi}_{21}, ..., \boldsymbol{\pi}_{2c})$.
Using the structure of A and equation (5.2), it is easy to verify that

$$\boldsymbol{\pi}_1 = \underline{\mathbf{0}} \ ;$$

$$\xi \boldsymbol{\pi}_{2i} = (c+1-i)\theta \boldsymbol{\pi}_{2i-1}; \text{for } 1 \leq i \leq c. \tag{5.7}$$

Using equation (5.7) and the normalizing condition (5.3), it follows that

$$\zeta \boldsymbol{\pi}_0 = 1 \tag{5.8}$$

where ζ is given in (5.5). Also we have

$$\boldsymbol{\pi} A_2 \mathbf{e} = \boldsymbol{\pi}_0[\mu_1 + \mu \sum_{i=1}^{c} \frac{c!}{(i-1)!} (\frac{\theta}{\xi})^{c-i}]$$

Then the stability condition (5.6) implies the result (5.4).

5.2.2 Expected number of interruptions to a customer at the main server

Since we are not imposing any upper bounds to the number of interruptions to a customer at the main server, we intend to find the expected number of interruptions before the service completion of a customer at the main server. For this, we consider the Markov process

$$Y(t) = \{(N_1(t), \hat{S}(t), J(t), K(t)) : t \geq 0\},$$

where $N_1(t)$ is the number of interruptions already befell to a customer at the main server.
$\hat{S}(t) = S(t) - \{\tilde{0}, 1\}$ and all other variables are as defined earlier.
The state space is

$$\{(i, 0, t_1) \cup (i, 2, j, t_1) : 1 \leq j \leq c, 1 \leq t_1 \leq a\} \cup \{\Delta\}.$$

The absorbing state Δ denote the customer at the main server leaves the system after service. Thus the infinitesimal generator $\tilde{V}$ of the process

$Y(t)$ takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}.$$

Here

$$V = \begin{bmatrix} G_1 & G_0 & & & \\ & G_1 & G_0 & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & \ddots \end{bmatrix}, V_0 = \begin{bmatrix} G_2 \\ G_2 \\ \vdots \\ \vdots \end{bmatrix}$$

where $G_2 = \mathbf{f}_{c+1} \otimes T^0$, $G_1 = diag(T, O)_{(c+1)a \times (c+1)a} + \tilde{G}_1 \otimes I_a$ and
$G_0 = \xi \begin{bmatrix} \underline{\mathbf{0}} & 0 \\ I_c & \mathbf{0} \end{bmatrix} \otimes I_a$, with

$$\tilde{G}_1 = \begin{bmatrix} -c\theta & c\theta & & & & \\ & -(c-1)\theta & (c-1)\theta & & & \\ & & \ddots & \ddots & & \\ & & & & -\theta & \theta \\ & & & & & 0 \end{bmatrix}_{c+1 \times c+1} - \xi diag(0, I_c).$$

If z_j is the probability that there are exactly j interruptions during the service of a customer at the main server, then

$$z_j = \boldsymbol{\eta}(-G_1^{-1}G_0)^j(-G_1^{-1}G_2), j = 0, 1, ...$$

where $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \mathbf{0})$.

Expected number of interruptions during the service of a customer at the main server is given by,

$$E(NI) = \sum_{j=0}^{\infty} j z_j.$$

5.2.3 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{1\tilde{0}}, \boldsymbol{x}_{11})$, and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i\tilde{0}}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, for $2 \leq i \leq c$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, for $i \geq c+1$.
Here $\boldsymbol{x}_{10}$ is an $a-$dimensional vector whereas $\boldsymbol{x}_{1\tilde{0}}$ and $\boldsymbol{x}_{11}$ are scalars.
$\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i\tilde{0}}$, $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$, for $2 \leq i \leq c$ are vectors of dimensions $a, 1, i, (i-1)a$, for $2 \leq i \leq c$, and $\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}$, for $i \geq c+1$ are vectors of dimensions a, c, ca, respectively.
The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_{c+1} R^{j-(c+1)}, j \geq c+1, \tag{5.9}$$

where R is the minimal non-negative solution of the matrix equation $R^2 A_2 + R A_1 + A_0 = 0$. Knowing the matrix R, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$, ... $\boldsymbol{x}_{c+1}$ are obtained by solving the equations

$$-\lambda \boldsymbol{x}_0 + \boldsymbol{x}_1 D_1 = 0$$

$$\boldsymbol{x}_{i-1} C_{i-1} + \boldsymbol{x}_i B_i + \boldsymbol{x}_{i+1} D_{i+1} = 0, \text{ for i=1,...,c-1} \tag{5.10}$$

$$\boldsymbol{x}_c C_c + \boldsymbol{x}_{c+1}(A_1 + R A_2) = 0$$

subject to the normalizing condition

$$\boldsymbol{x}_0 + \boldsymbol{x}_1 \mathbf{e} + ... + \boldsymbol{x}_c \mathbf{e} + \boldsymbol{x}_{c+1}(I - R)^{-1} \mathbf{e} = 1. \tag{5.11}$$

5.2.4 Performance measures

Now we compute some performance measures.

(1) Expected number of customers in the system

$$ES = \sum_{i=1}^{\infty} i \ \boldsymbol{x}_i \mathbf{e}.$$

(2) Expected number of customers in the queue

$$EQ = \sum_{i=c+1}^{\infty} (i-c) \ \boldsymbol{x}_{i1}\mathbf{e} + \sum_{i=c+2}^{\infty} (i-c-1)(\ \boldsymbol{x}_{i0}\mathbf{e} + x_{i2}\mathbf{e}).$$

(3) Expected number of idle regular servers

$$E(IS) = \sum_{i=1}^{c-1}(c-i) \ \boldsymbol{x}_{i\tilde{0}}\mathbf{e} + \sum_{i=1}^{c}(c+1-i)\boldsymbol{x}_{i0}\mathbf{e}.$$

(4) Effective rate of interruption

$$EI = \sum_{i=2}^{c}(i\theta \ \boldsymbol{x}_{i0}+\sum_{j=1}^{i-1} (i-j)\theta \ \boldsymbol{x}_{i2j})\mathbf{e}+ \sum_{i=c+1}^{\infty} (c\theta \ \boldsymbol{x}_{i0}+\sum_{j=1}^{c-1}(c-j)\theta \ \boldsymbol{x}_{i2j})\mathbf{e}.$$

(5) Effective rate of consultation

$$ECo = EI + \theta \ \boldsymbol{x}_{1\tilde{0}} + \sum_{i=2}^{c}\sum_{j=1}^{i-1} (i-j)\theta \ \boldsymbol{x}_{i1j} + \sum_{i=c+1}^{\infty}\sum_{j=1}^{c-1}(c-j)\theta \ \boldsymbol{x}_{i1j}.$$

(6) Fraction of time the main server is idle

$$F_{mi} = \boldsymbol{x}_0 + \sum_{i=1}^{c} \boldsymbol{x}_{i\tilde{0}}\mathbf{e}.$$

(7) Fraction of time all the servers are busy serving customers

$$F_{ab} = \sum_{i=c+1}^{\infty} \boldsymbol{x}_{i0}\mathbf{e}.$$

5.2.5 An optimization problem

In this section we propose an optimization problem and discuss it through an illustrative example. To construct an objective function, we assume that the service completion produces a revenue to the system whereas idle regular servers and waiting spaces involve expenditure to the system. Thus we produce per unit time revenue and cost as follows:

1. r be the revenue per customer leaving the system after service completion
2. c_1 be the holding cost monetary of customers in the system
3. c_2 holding cost of idle regular servers

The problem of interest is to find an optimum value of the number of regular servers to be employed so that the expected total profit ETP is maximum. The objective function is given below:

$$ETP = r \times ESR - c_1 \times ES - c_2 \times E(IS) \tag{5.12}$$

where $ESR = \boldsymbol{\pi} A_2 \mathbf{e}$.

5.3 Numerical examples

Now we present numerical results for implementing the qualitative nature of the model under study. The purpose of this example is to see the impact of parameter c. Here we consider that the service rate of the main server and that of the regular servers are equal. That is, we choose T and $\boldsymbol{\alpha}$ so that $[\boldsymbol{\alpha}(-T)^{-1}\mathbf{e}]^{-1} = \mu$.

Let $\lambda = 5$, $\theta = 9$, $\mu = 2$, $\xi = 3$, $T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$, $\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$.

The above data of matrices, vectors and values satisfy the stability condition (5.4). Fix $r = 25$, $c_1 = 15$ and $c_2 = 100$.

Table 5.1: Effect of number of c on cost function

c	3	4	5	6	7	8
ESR	9.8130	15.2409	20.3885	25.1093	29.5694	33.8892
ES	11.7206	3.9603	2.6137	2.1678	1.9604	1.8370
$E(IS)$	0.6429	2.1892	3.6003	4.8710	6.0460	7.1640
ETP	5.226	102.698	**110.477**	108.116	105.229	103.275

From the table 5.1 we can see that as c increases the effective service rate ESR increases. Since the services are done in a faster rate, accumulation of customers becomes less. Thus ES decreases. If the number of regular servers increases, for a fixed λ, number of idle regular servers $E(IS)$ also increases as is to be expected.

Thus ETP has an optimum value **110.477** when the number of regular servers $c = 5$.

5.3.1 More numerical examples

In this section, we present some numerical examples that describe the performance characteristics of the queueing model under study.
Let $c = 3$ and λ, θ, μ, ξ, T and $\boldsymbol{\alpha}$ are as given in the above example.

Table 5.2: Effect of θ on various performance measures

θ	3	4	5	6	7
ES	0.8597	1.0946	1.4578	2.0588	3.1573
EQ	0.0531	0.1264	0.2809	0.6087	1.3436
F_{mi}	0.6444	0.5966	0.5401	0.4727	0.3915
F_{ab}	0.0081	0.0106	0.0137	0.0177	0.0227

Table 5.3: Effect of λ on various performance measures

$\theta = 3$

λ	3	3.5	4	4.5	5
ES	1.9332	2.8234	4.1764	6.4072	10.266
EQ	0.4030	0.8879	1.8212	3.6097	6.9135
F_{mi}	0.4658	0.3805	0.2985	0.2197	0.1440
F_{ab}	0.0386	0.0667	0.1045	0.1516	0.2038

From the table 5.2, we see that as θ increases rate of consultation (and hence interruption) increases. So the customers have to spend longer time to get their services completed. Thus accumulation of customers in system

Table 5.4: Effect of μ on various performance measures

$\theta = 3$

μ	2	2.5	3	3.5	4
ES	4.1764	2.7497	2.0696	1.6839	1.4400
EQ	1.8212	0.8359	0.4485	0.2660	0.1693
F_{mi}	0.2985	0.3683	0.4130	0.4434	0.4651
F_{ab}	0.1045	0.0625	0.0400	0.0271	0.0192

Table 5.5: Effect of ξ on various performance measures

$\theta = 6, \mu = 2.5$

ξ	3	3.5	4	4.5	5
ES	11.7064	11.0572	6.1958	4.2548	3.3207
EQ	8.6051	7.6826	3.4795	1.8598	1.1410
F_{mi}	0.0283	0.1303	0.2095	0.2662	0.3087
F_{ab}	0.0556	**0.1018**	0.0909	0.0813	0.0743

and queue happens in a faster rate, which results in increased number of ES and EQ. So the fraction of time all servers are busy serving customers F_{ab} increases. Naturally, F_{mi} will decrease.

Table 5.3 shows that as λ increases, the server is fed with customers more frequently and so ES and EQ increase. Busy time of each server increases, therefore F_{ab} increases and thus idle time of main server F_{mi} decreases.

From table 5.4, we see that as μ increases, regular servers serve customers in a faster rate. Thus there is a slow accumulation of customers in system and in queue which results in decrease in ES and in EQ. As the customers get served in a faster rate at the regular servers, less number of customers approach the main server. So the main server gets more idle time, ie, F_{mi} increases. So as a whole, the fraction of time all servers are busy F_{ab} decreases.

We see from table 5.5 that an increase in ξ results in a faster rate of consultation completion. So the servers get more time to serve customers and so the accumulation of customers in the system and in the queue decrease. Thus ES and EQ decrease. As larger number of customers are served by the regular servers, main server gets more idle time. So F_{mi} increases. Now we consider F_{ab}. We can see that F_{ab} increases until ξ=3.5. If again ξ increases, since the value of λ is fixed, the servers need less amount of time for service completion and therefore the fraction of time all servers are busy serving customers F_{ab} will decrease.

Chapter 6

A two-server queue with mutual consultations

In the previous chapters, we discussed queueing models in which one of the servers (namely, main server) offers consultation to the fellow servers (regular servers). These are seen to occur in banks (with the manager in addition to providing service to customers, helping other bank staff in their work also), hospitals (where the chief physician treats patients and clarifies the doubts of the fellow doctors), super markets, etc. If there are more than one server, the servers can consult among themselves whenever necessary. They can clarify their doubts with other's help. This type of queueing systems are common in banks, hospitals, railway ticket counters, super markets and petrol pumps. Certainly, these consultations will improve the quality of the service. For example, if the doctors in a hospital discuss (consult) with each other, the patients are sure to get a better diagnosis of their problems and hence a better treatment. Thus consultation provided by a main server can be extended to mutual consultations among servers, in pairs or even in larger numbers. However in this chapter we restrict the mutual consultation of servers in pairs only.

In this chapter, we analyse a two-server queueing model in which the servers provide mutual consultations. They provide consultation with a preemptive priority over the customers being served. Thus the customers at the servers undergo interruptions during their services. The arrivals to the system follows Poisson process and requirement of consultations by the servers follow mutually independent Poisson processes. Duration of consultations follow mutually independent exponential distributions. The service times of customers at these servers are assumed to follow mutually independent phase type distributions. An explicit expression for the system stability is obtained. Towards the end of the chapter we consider two particular cases of this model. A comparison of the respective performance measures of the three models is also presented.

6.1 Model description

Here we consider a service system with two servers to which customers arrive according to a Poisson process with rate λ. Let these servers be denoted by $\Re_1$ and $\Re_2$. An arriving customer enters into service immediately if at least one server is free whereas that customer waits in a queue, otherwise. The service times of the customers at these servers follow phase type distributions with representations ($\boldsymbol{\alpha}$,T) and ($\boldsymbol{\beta}$,U) respectively. Write $T^0 = -T\mathbf{e}$ and $U^0 = -U\mathbf{e}$ where $\mathbf{e}$ is a column vector of 1' s of appropriate order. The servers offer consultation to the fellow server whenever it is required. Requirement of consultations for $\Re_i$ follow mutually independent Poisson processes with rates θ_i, $(i = 1, 2)$. The request for consultation of one server is attended immediately by the other server even when a customer is being served at the latter one and that customer has to wait until the consultation is completed. At this stage the service of the customer at the second server is said to be interrupted. The duration of consultation for the i^{th} server by the other is exponentially distributed with parameter ξ_i, $i = 1, 2$.

Notations :- We use the following notations in this model.

- $\Theta = diag(\theta_1 I_a, \theta_2 I_b)$, $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}$
- $\Phi_1 = -diag(\xi_1 I_a, \xi_2 I_b)$, $\Phi_2 = \begin{bmatrix} \xi_1 I_a \otimes \boldsymbol{\beta} \\ \xi_2 \boldsymbol{\alpha} \otimes I_b \end{bmatrix}$
- $\nabla = diag(\xi_1, \xi_2)$

Consider the queueing model

$$X = \{X(t), t \geq 0\},$$

where

$$X(t) = \{N(t), S(t), D(t), J_1(t), J_2(t)\}.$$

Here $N(t)$ is the number of customers in the system and $J_i(t)$ is the phase of the server $\Re_i$, $i = 1, 2$ and

$$S(t) = \begin{cases} 0, & \text{if either one or both servers are busy} \\ 1, & \text{if a server is giving consultation to the other} \\ & \text{without any interrupted customer at the former} \\ 2, & \text{if a server is giving consultation to the other} \\ & \text{with an interrupted customer at the former} \end{cases}$$

If $N(t) = 1$ and $S(t) = 0$, then $D(t) = \{1, 2\}$ according to $\Re_1$ or $\Re_2$ is busy.
If $N(t) = 1$ and $S(t) = 1$ or $N(t) \geq 2$ and $S(t) = \{1, 2\}$, then $D(t) = \{1, 2\}$ according to $\Re_1$ or $\Re_2$ is getting consultation.
$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).$$

The terms $\psi(i)$'s are defined as
$\psi(1) = \{(1, 0, j, l_j)\} \cup \{(1, 1, j, l_j)\}$ and
$\psi(i) = \{(i, 0, l_1, l_2)\} \cup \{(i, 1, j, l_j)\} \cup \{(i, 2, j, l_1, l_2)\}$, for $i \geq 2$,
with $j = 1, 2$; $1 \leq l_1 \leq a$ and $1 \leq l_2 \leq b$.
The infinitesimal generator Q is given by

$$Q = \begin{bmatrix} -\lambda & B_1 & & & & \\ B_2 & B_3 & B_4 & & & \\ & B_5 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix} \tag{6.1}$$

Here the matrices B_1, B_2, B_4 and B_5 are of orders $1 \times 2(a+b)$, $2(a+b) \times 1$, $2(a+b) \times (3ab + a + b)$ and $(3ab + a + b) \times 2(a + b)$ respectively. B_3 is a square matrix of order $2(a+b)$ and A_0, A_1 and A_2 are square matrices of order $3ab+a+b$. These matrices are described as follows:

$$B_1 = \frac{\lambda}{2}\begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} & \underline{\mathbf{0}} \end{bmatrix},\ B_2 = \begin{bmatrix} T^0 \\ U^0 \\ \mathbf{0} \end{bmatrix},$$

$$B_3 = \begin{bmatrix} diag(T,U) & O \\ O & O \end{bmatrix} + \begin{bmatrix} -\Theta & \Theta \\ -\Phi_1 & \Phi_1 \end{bmatrix} - \lambda I,$$

$$B_4 = \lambda\begin{bmatrix} B_{41} & O \end{bmatrix},\ \text{where } B_{41} = \lambda\begin{bmatrix} I_a \otimes \boldsymbol{\beta} & & \\ \boldsymbol{\alpha} \otimes I_b & & \\ & I_a & \\ & & I_b \end{bmatrix}_{2(a+b)\times(ab+a+b)},$$

$$B_5 = \begin{bmatrix} B_{51} & O \\ O & O \end{bmatrix},\ \text{where } B_{51} = \begin{bmatrix} I_a \otimes U^0 & T^0 \otimes I_b \end{bmatrix}_{ab\times(a+b)},$$

$$A_0 = \lambda I,\ A_1 = \begin{bmatrix} T \oplus U - (\theta_1+\theta_2)I & & \boldsymbol{\theta} \otimes I_{ab} \\ \Phi_2 & \Phi_1 & \\ \nabla \otimes \mathbf{e}_2 \otimes I_{ab} & & \nabla \otimes I_{ab} \end{bmatrix} - \lambda I,$$

$$A_2 = \begin{bmatrix} T^0 \otimes \boldsymbol{\alpha} \oplus U^0 \otimes \boldsymbol{\beta} & O \\ O & O \end{bmatrix}.$$

6.2 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study. We first establish the stability condition of the queueing system.

6.2.1 Stability condition

Let the steady-state probability vector of the generator $A = A_0 + A_1 + A_2$ be denoted by $\boldsymbol{\pi}$. That is, $\boldsymbol{\pi} A = 0; \boldsymbol{\pi}\mathbf{e} = 1$.
The following theorem gives the stability of the queueing system under study.
Theorem 6.2.1 : The Markov Chain X is stable if and only if

$$\lambda < \frac{1}{\zeta}(\mu_1 + \mu_2) \tag{6.2}$$

where $\zeta = 1 + \frac{\theta_1}{\xi_1} + \frac{\theta_2}{\xi_2}$; μ_1 and μ_2 are the respective service rates at the servers $\Re_1$ and $\Re_2$.

Proof. The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if $\boldsymbol{\pi} A_0 \mathbf{e} < \boldsymbol{\pi} A_2 \mathbf{e}$. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_{11}, \boldsymbol{\pi}_{12}, \boldsymbol{\pi}_{21}, \boldsymbol{\pi}_{22})$.

The matrix A is given by

$$A = \begin{bmatrix} B - (\theta_1 + \theta_2) I_{ab} & & & \theta_1 I_{ab} & \theta_2 I_{ab} \\ \xi_1 I_a \otimes \boldsymbol{\beta} & -\xi_1 I_a & & & \\ \xi_2 \boldsymbol{\alpha} \otimes I_b & & -\xi_2 I_b & & \\ \xi_1 I_{ab} & & & -\xi_1 I_{ab} & \\ \xi_2 I_{ab} & & & & -\xi_2 I_{ab} \end{bmatrix} \tag{6.3}$$

where $B = T \oplus U + T^0 \otimes \boldsymbol{\alpha} \oplus U^0 \otimes \boldsymbol{\beta}$.

It is easy to verify that

$$\boldsymbol{\pi}_{11} = \boldsymbol{\pi}_{12} = 0;$$

$$\xi_1 \boldsymbol{\pi}_{21} = \theta_1 \boldsymbol{\pi}_0; \xi_2 \boldsymbol{\pi}_{22} = \theta_2 \boldsymbol{\pi}_0. \tag{6.4}$$

Using equation (6.4) and the normalizing condition, it follows that

$$(1 + \frac{\theta_1}{\xi_1} + \frac{\theta_2}{\xi_2})\boldsymbol{\pi}_0 \mathbf{e} = 1. \tag{6.5}$$

Then stability condition (5.11) implies that

$$\lambda < (1 + \frac{\theta_1}{\xi_1} + \frac{\theta_2}{\xi_2})^{-1}[\frac{1}{\boldsymbol{\alpha}(-T)^{-1}\mathbf{e}} + \frac{1}{\boldsymbol{\beta}(-U)^{-1}\mathbf{e}}]. \tag{6.6}$$

Putting $\zeta = 1 + \frac{\theta_1}{\xi_1} + \frac{\theta_2}{\xi_2}$, $\mu_1 = (\boldsymbol{\alpha}(-T)^{-1}\mathbf{e})^{-1}$ and $\mu_2 = (\boldsymbol{\beta}(-U)^{-1}\mathbf{e})^{-1}$, we get the required result.

6.2.2 Steady state probability vector

Let $\boldsymbol{x}$, partitioned as, $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1 = (\boldsymbol{x}_{10}, \boldsymbol{x}_{111}, \boldsymbol{x}_{112})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i11}, \boldsymbol{x}_{i12}, \boldsymbol{x}_{i21}, \boldsymbol{x}_{i22})$, for $i \geq 2$. Here $\boldsymbol{x}_{10}$, $\boldsymbol{x}_{111}$, $\boldsymbol{x}_{112}$, $\boldsymbol{x}_{i0}$, $\boldsymbol{x}_{i11}$, $\boldsymbol{x}_{i12}$, $\boldsymbol{x}_{i21}$, $\boldsymbol{x}_{i22}$ are vectors of dimensions $a + b, a, b$, ab, a, b, ab and ab respectively.
The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = 0$ and $\boldsymbol{x}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is

satisfied, the sub-vectors of $\boldsymbol{x}$ are given by the equation

$$\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3, \tag{6.7}$$

where R is the minimal non-negative solution of the matrix equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{6.8}$$

Knowing the matrix R, the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained by solving the equations

$$\begin{aligned} -\lambda \boldsymbol{x}_0 + \boldsymbol{x}_1 B_2 &= 0 \\ \boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_3 + \boldsymbol{x}_2 B_5 &= 0 \\ \boldsymbol{x}_1 B_4 + \boldsymbol{x}_2 (A_1 + R A_2) &= 0 \end{aligned} \tag{6.9}$$

subject to the normalizing condition

$$\boldsymbol{x}_0 + \boldsymbol{x}_1 \mathbf{e} + \boldsymbol{x}_2 (I - R)^{-1} \mathbf{e} = 1. \tag{6.10}$$

6.2.3 Performance characteristics

Now we compute some performance measures.

(1) Expected number of customers in the system

$$ES = \sum_{i=1}^{\infty} i \boldsymbol{x}_i \mathbf{e}. \tag{6.11}$$

(2) Expected number of customers in the queue

$$EQ = \sum_{i=2}^{\infty}(i-1)\boldsymbol{x}_{i1}\mathbf{e} + \sum_{i=3}^{\infty}(i-2)(\boldsymbol{x}_{i0}\mathbf{e} + \boldsymbol{x}_{i2}\mathbf{e}). \tag{6.12}$$

(3) Fraction of time both servers are busy serving customers

$$\Gamma = \sum_{i=2}^{\infty} \boldsymbol{x}_{i0}\mathbf{e}. \tag{6.13}$$

(4) Fraction of time $\Re_j$ $(j = 1, 2)$ is getting consultation

$$C_j = \boldsymbol{x}_{11j} + \sum_{i=2}^{\infty}(\boldsymbol{x}_{i1j}\mathbf{e} + \boldsymbol{x}_{i2j}\mathbf{e}) \tag{6.14}$$

(5) Fraction of time $\Re_j$ $(j = 1, 2)$ is under interruption

$$K_j = \sum_{i=2}^{\infty}(\boldsymbol{x}_{i2}\mathbf{e} - \boldsymbol{x}_{i2j}\mathbf{e}) \tag{6.15}$$

(6) Effective rate of interruption to $\Re_j$ $(j = 1, 2)$

$$\nu_j = \Gamma(\boldsymbol{\theta} \otimes \mathbf{e}_2 - \theta_j) \tag{6.16}$$

(7) Effective rate of consultation for $\Re_j$ $(j = 1, 2)$

$$\sigma_j = \theta_j(\Gamma + \boldsymbol{x}_{10j}\mathbf{e}) \tag{6.17}$$

6.3 Numerical examples

In this section, we present some numerical examples that describe the performance characteristics of the queueing model under study.
We choose $T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$, $U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}$,

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \xi_1 = 2, \xi_2 = 3.$$

We choose the above matrices, vectors and values so that the stability condition given in equation (6.2) is satisfied.

Table 6.1: Effect of λ on various performance measures

$\theta_1 = 1, \theta_2 = 2$

λ	1	2	3	4	5
ES	0.3666	1.0281	2.3935	6.1480	30.1399
EQ	0.0586	0.3590	1.3280	4.6701	28.3250
Γ	0.0224	0.0828	0.1732	0.2878	0.4041
C_1	0.0426	0.0861	0.1301	0.1743	0.2097
C_2	0.0602	0.1189	0.1770	0.2347	0.2805
K_1	0.0149	0.0552	0.1155	0.1919	0.2689
K_2	0.0112	0.0414	0.0866	0.1439	0.2015
ν_1	0.0448	0.1655	0.3464	0.5756	0.8082
ν_2	0.0224	0.0828	0.1732	0.2878	0.4041
σ_1	0.0851	0.1722	0.2601	0.3485	0.4206
σ_2	0.1805	0.3566	0.5309	0.7041	0.8430

From table 6.1 we see that as λ increases the system is fed with more customers and then ES and EQ increase. This results in an increase in busy time of both servers, ie, Γ increases. Since ν_j and σ_j directly depends

Table 6.2: Effect of θ_1 on various performance measures

$\lambda = 3, \theta_2 = 2$

θ_1	1	1.5	2	2.5	3
ES	2.3935	3.2311	4.4343	6.3138	9.6633
EQ	1.3280	2.0423	3.1159	4.8593	8.0656
Γ	0.1732	0.1867	0.2002	0.2137	0.2272
C_1	0.1301	0.1945	0.2590	0.3239	0.3894
C_2	0.1770	0.1775	0.1778	0.1777	0.1773
K_1	0.1155	0.1245	0.1335	0.1425	0.1515
K_2	0.0866	0.1401	0.2002	0.2671	0.3408
ν_1	0.3464	0.3735	0.4004	0.4274	0.4545
ν_2	0.1732	0.2801	0.4004	0.5343	0.6817
σ_1	0.2601	0.3891	0.5181	0.6478	0.7788
σ_2	0.5309	0.5326	0.5333	0.5331	0.5320

Table 6.3: Effect of θ_2 on various performance measures

$\lambda = 3, \theta_1 = 1$

θ_2	3	3.5	4	4.5	5
ES	3.4594	4.2239	5.2493	6.6985	8.9023
EQ	2.2171	2.8913	3.8251	5.1808	7.2894
Γ	0.1913	0.2003	0.2092	0.2182	0.2272
C_1	0.1312	0.1315	0.1318	0.1320	0.1320
C_2	0.2630	0.3059	0.3488	0.3918	0.4352
K_1	0.1913	0.2336	0.2790	0.3273	0.3787
K_2	0.0956	0.1001	0.1046	0.1091	0.1136
ν_1	0.5738	0.7009	0.8369	0.9819	1.1361
ν_2	0.1913	0.2003	0.2092	0.2182	0.2272
σ_1	0.2623	0.2631	0.2636	0.2640	0.2641
σ_2	0.7890	0.9176	1.0463	1.1755	1.3055

on Γ increase in Γ results in an increase in ν_j and σ_j, for $j = 1, 2$. Therefore the duration of time the servers getting consultations C_j and hence the servers under interruption K_j increase.

In table 6.2, we can see that as θ_1 increases, σ_1 and hence ν_2 increase. Thus C_1 and K_2 increase. As a result of these changes customers stay in the system and in the queue for longer time and thus ES and EQ increase rapidly. So the servers have to serve the customers for a longer time. Thus there is an increase in Γ also. Since the values C_2, K_1, ν_1 and σ_2 do not depend on θ_1 directly, there are only slow increase in these values as θ_1 increases.

Table 6.3 shows that corresponding to an increase in θ_2 there are increases in the performance measures σ_2 and ν_1 and hence an increase in C_2 and K_1. This results in a faster accumulation of customers in the system and in the queue and so ES and EQ increase as θ_2 increases. Thus the fraction of time all the servers are busy serving customers Γ increases. There are slow increase in the other measures C_1, K_2, ν_2 and σ_1 because these values do not depend on θ_2 directly.

6.4 Particular cases

In this section we take some particular values for θ_1 and θ_2. The present queueing model consisting of two servers with mutual consultations reduces to two distinct two server queueing models. We will get a two server queueing model with consultation by main server if either θ_1 or θ_2 equal to zero (but not both), whereas the problem reduces to the case of an $M/(PH, PH)/2$ queue if both θ_1 and θ_2 are allowed to zero. A brief discussion of the two cases are given below.

6.4.1 case 1: $\theta_1 = 0,\ \theta_2 = \theta$

Let us assume that $\theta_1 = 0$ (obviously, ξ_1 does not exist), $\theta_2 = \theta$ and $\xi_2 = \xi$, then we get a two server queueing model with $\Re_1$ as the main server which provides consultation to the regular server $\Re_2$. Note that this model can also be deduced from model 1 of chapter 2 by omitting the concepts of the upper bounds on the number of interruptions and consultations, super clock and threshold clock and by considering an exponentially distributed consultation duration instead of a phase type distributed consultation.

Here also the infinitesimal generator Q takes the same form given in (6.1) where the sub-matrices are different. It can be seen that the matrices B_1, B_2, B_4 and B_5 are of orders $1 \times a+2b$, $a+2b \times 1$, $a+2b \times 2ab+b$ and $2ab+b \times a+2b$ respectively. B_3 is a square matrix of order $a+2b$ and A_0, A_1 and A_2 are square matrices of order $2ab+b$.

These matrices are described as follows:

$$B_1 = \lambda \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix},\ B_2 = \begin{bmatrix} T^0 \\ U^0 \\ \mathbf{0} \end{bmatrix},\ B_3 = \begin{bmatrix} T & O & O \\ O & U-\theta I_b & \theta I_b \\ O & \xi I_b & -\xi I_b \end{bmatrix} - \lambda I,$$

$$B_4 = \lambda \begin{bmatrix} I_a \otimes \boldsymbol{\beta} & \\ \boldsymbol{\alpha} \otimes I_b & \\ & I_b \end{bmatrix}, \; B_5 = \begin{bmatrix} I_a \otimes U^0 & T^0 \otimes I_b & O \\ O & O & O \end{bmatrix}, \; A_0 = \lambda I,$$

$$A_1 = \begin{bmatrix} T \oplus U - \theta I & & \theta \otimes I_{ab} \\ \xi \boldsymbol{\alpha} \otimes I_b & -\xi I_b & \\ \xi I_{ab} & & -\xi I_{ab} \end{bmatrix} - \lambda I, \; A_2 = \begin{bmatrix} T^0 \otimes \boldsymbol{\alpha} \oplus U^0 \otimes \boldsymbol{\beta} & O \\ O & O \end{bmatrix}.$$

The stability condition is obtained from theorem (6.2.1) by choosing $\theta_1 = 0$, $\theta_2 = \theta$ and $\xi_2 = \xi$. Thus the condition for stability of the present system is as given below:

Theorem 6.4.1 : The queueing system is stable if and only if

$$\lambda < \frac{1}{\tilde{\zeta}}(\mu_1 + \mu_2) \tag{6.18}$$

where $\tilde{\zeta} = 1 + \frac{\theta}{\xi}$, μ_1 and μ_2 are the respective service rates at the main and regular servers.

Let $\frac{1}{\tilde{\zeta}}(\mu_1 + \mu_2)$ be denoted by δ_2.

Note : This result can also be obtained directly by applying the stability condition on the generator matrix $A = A_0 + A_1 + A_2$.

6.4.2 Performance measures

The steady state probability vector $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, ...)$ for this model can be obtained by a procedure similar to that described in section 6.2.2.

Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1$ is a vector of dimension $a + 2b$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$, for $i \geq 2$, where $\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ are vectors of dimensions ab, b and ab, respectively.

After computing the steady state probability vector, the performance measures such as expected number of customers in the system, ES_2, expected number of customers in the queue, EQ_2 and fraction of time both servers are busy serving customers, Γ_2 can be calculated using the equations (6.11), (6.12), (6.13). C_1, K_2, ν_2 and σ_1 has no relevance in the present aspect. We get the following performance measures also.

(1) Fraction of time the regular server is getting consultation

$$F_{rc} = \sum_{i=1}^{\infty} \boldsymbol{x}_{i1}\mathbf{e}. \tag{6.19}$$

(2) Fraction of time the main server is under interruption

$$F_{min} = \sum_{i=2}^{\infty} \boldsymbol{x}_{i2}\mathbf{e}. \tag{6.20}$$

(3) Effective rate of interruption to the main server

$$EI = \theta\Gamma_2. \tag{6.21}$$

(4) Effective rate of consultation by the regular server

$$ECo = EI + \theta \sum_{i=a+1}^{a+b} \boldsymbol{x}_{1i}. \tag{6.22}$$

6.4.3 case 2: $\theta_1 = \theta_2 = 0$

If both θ_1 and θ_2 assume the value zero, then the queueing model reduces to an $M/(PH, PH)/2$ queue.

The infinitesimal generator Q takes the same form given in (6.1) where the sub-matrices are as given below:

$$B_1 = \lambda \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix}_{1\times a+b}, \; B_2 = \begin{bmatrix} T^0 \\ U^0 \end{bmatrix}_{a+b\times 1}, \; B_3 = diag(T, U) - \lambda I,$$
$$B_4 = \lambda \begin{bmatrix} I_a \otimes \boldsymbol{\beta} \\ \boldsymbol{\alpha} \otimes I_b \end{bmatrix}_{a+b\times ab}, \; B_5 = \begin{bmatrix} I_a \otimes U^0 & T^0 \otimes I_b \end{bmatrix}_{ab\otimes a+b},$$
$$A_0 = \lambda I, \; A_1 = T \oplus U - \lambda I, \; A_2 = T^0 \otimes \boldsymbol{\alpha} \oplus U^0 \otimes \boldsymbol{\beta}.$$

B_3 is a square matrix of order $a + b$ and A_0, A_1 and A_2 are square matrices of order ab.
The stability condition is obtained from theorem (6.2.1) by putting $\theta_1 = \theta_2 = 0$. Thus the stability of the present system is as given below:

Theorem 6.4.2 : The queueing system is stable if and only if

$$\lambda < \mu_1 + \mu_2 \tag{6.23}$$

where μ_1 and μ_2 are the service rates of the servers.
Let $\mu_1 + \mu_2$ be denoted by δ_3.
Note: This result is same as the stability condition for the $M/(PH, PH)/2$ queue.

6.4.4 Performance measures

The steady state probability vector $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, ...)$ for this model can be obtained by a procedure similar to that described in section 6.2.2. Note that $\boldsymbol{x}_0$ is a scalar, $\boldsymbol{x}_1$ and $\boldsymbol{x}_i$ are vectors of dimensions $a+b$ and ab, respectively.

After computing the steady state probability vector, the performance measures such as expected number of customers in the system, ES_3 and fraction of time both servers are busy serving customers, Γ_3 can be calculated using the equations (6.11) and (6.13), respectively.

Expected number of customers in the queue is given by

$$EQ_3 = \sum_{i=3}^{\infty}(i-2)\boldsymbol{x}_{i0}\mathbf{e}.$$

6.4.5 Numerical example

In this section we present a comparison of the performance measures in the two cases with the model having mutual consultations. Here we consider the model with mutual consultation as model 1 and the models in cases 1 and 2 as model 2 and model 3, respectively. The measures ES, EQ and Γ in model 1 are denoted by ES_1, EQ_1 and Γ_1, respectively for the purpose of comparison.
We choose T, U, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and ξ_1 as in section 6.3.
For tables 6.4 and 6.5 we choose $\theta_1 = 1$, $\theta_2 = \theta = 2$, $\xi_2 = \xi = 3$.

For table 6.6 we choose $\lambda = 3$, $\theta_1 = 1$, $\xi_2 = \xi = 3$.

Table 6.4: Effect of λ on various performance measures

$$\delta_1 = 5.2715, \delta_2 = 6.853, \delta_3 = 11.4217$$

λ	1	2	3	4	5
ES_1	0.3666	1.0281	2.3935	6.1480	30.1399
ES_2	0.2075	0.5393	1.1041	2.1568	4.4829
ES_3	0.1704	0.3523	0.5540	0.7877	1.0725
EQ_1	0.0586	0.3590	1.3280	4.6701	28.3250
EQ_2	0.0110	0.0975	0.3811	1.1260	3.1253
EQ_3	0.0013	0.0108	0.0382	0.0966	0.2055
Γ_1	0.0224	0.0828	0.1732	0.2878	0.4041
Γ_2	0.0269	0.1078	0.2351	0.3990	0.5909
Γ_3	0.0136	0.0509	0.1073	0.1793	0.2640

From table 6.4, we see that as λ increases all the performance measures in all the three models increase as is to be expected. There is no consultation (and interruption) at all in model 3, consultation by the main server only in model 2 and consultation (and interruption) by both servers in model 1. So the expected number of customers in the system follows the inequality $ES_3 < ES_2 < ES_1$. A similar inequality for the expected number of customers in the queue $EQ_3 < EQ_2 < EQ_1$. Since ES_3 is the least among all the values of ES, the Γ_3 is the least among the fraction of time both servers are busy. Since there are consultations and interruptions by both servers in model 1, the servers get less time to serve customers. So Γ_1 is less than Γ_2 because there is only one consultation in model 2. Thus $\Gamma_3 < \Gamma_1 < \Gamma_2$.

Table 6.5: Effect of λ on various performance measures

$$\delta_1 = 5.2715, \delta_2 = 6.853, \delta_3 = 11.4217$$

λ	1	2	3	4	5
C_2	0.0602	0.1189	0.1770	0.2347	0.2805
F_{rc}	0.0128	0.0379	0.0597	0.0692	0.0619
K_1	0.0149	0.0552	0.1155	0.1919	0.2689
F_{min}	0.0100	0.0399	0.0881	0.1520	0.2290
ν_1	0.0448	0.1655	0.3464	0.5756	0.8082
EI	0.0537	0.2156	0.4702	0.7980	1.1817
σ_2	0.1805	0.3566	0.5309	0.7041	0.8430
ECo	0.0757	0.2754	0.5598	0.8982	1.2691

Table 6.5 shows that $F_{rc} < C_2$, $F_{min} < K_1$ and $ECo < \sigma_2$ as is to be expected, since there are two consultations in model 1 and only one consultation in model 2. Since ν_1 and EI directly depends on Γ_1 and Γ_2 respectively, $\nu_1 < EI$ because $\Gamma_1 < \Gamma_2$.

From table 6.6 it is seen that as θ increases, all the performance measures in model 1 and model 2 increase as is to be expected. Also $\delta_1 < \delta_2$, $ES_2 < ES_1$, $EQ_2 < EQ_1$, $\Gamma_1 < \Gamma_2$, $F_{rc} < C_2$ and $F_{min} < K_1$. These are all expected. Since ν_1 and σ_2 depends directly on Γ_1, EI and ECo depends directly on Γ_2 and since $\Gamma_1 < \Gamma_2$, we get $\nu_1 < EI$ and $\sigma_2 < ECo$.

Table 6.6: Effect of $\theta_2 = \theta$ on various performance measures

$\lambda = 3$

θ	3	3.5	4	4.5	5
δ_1	4.5687	4.2831	4.0312	3.8072	3.6068
δ_2	5.7108	5.2715	4.8950	4.5687	4.2831
ES_1	3.4594	4.2239	5.2493	6.6985	8.9023
ES_2	1.5627	1.8729	2.2626	2.7634	3.4245
EQ_1	2.2171	2.8913	3.8251	5.1808	7.2894
EQ_2	0.7135	0.9537	1.2683	1.6883	2.2623
Γ_1	0.1913	0.2003	0.2092	0.2182	0.2272
Γ_2	0.3119	0.3543	0.3997	0.4484	0.5007
C_2	0.2630	0.3059	0.3488	0.3918	0.4352
F_{rc}	0.0583	0.0568	0.0546	0.0518	0.0484
K_1	0.1913	0.2336	0.2790	0.3273	0.3787
F_{min}	0.1462	0.1795	0.2159	0.2557	0.2992
ν_1	0.5738	0.7009	0.8369	0.9819	1.1361
EI	0.9358	1.2401	1.5987	2.0176	2.5034
σ_2	0.7890	0.9176	1.0463	1.1755	1.3055
ECo	1.0524	1.3656	1.7298	2.1509	2.6353

Concluding remarks and suggestions for further study

In this work we studied multi-server queueing models with consultations. Consultation is an important aspect which enhance the reliability of the services provided by the trainees by accepting timely advices and clarifications from the experienced servers. In chapters 2 and 3, we discussed two-server queueing models with consultations by the main server to the regular server. In chapter 4, three-server queueing models were considered. A multi-server queueing system was analysed in chapter 5. A different aspect of consultation was discussed in chapter 6, consultation between a pair of servers.

It would indeed be a challenging task to extend the multi-server queueing models discussed in chapter 5 by introducing an upper bound on the number of interruptions and a super clock. M/G/2 models can be considered with exponentially distributed service time at main server and general service time at regular server.

It will be interesting to deal with an an infinite server queue with consultations by a consultant (not a server). The servers will queue up for consultations. As an extension of the model discussed in chapter 6, we can consider consultation among the servers in a multi-server system.

Bibliography

[1] *Avi-Itzhak, B. and Naor, P.* : Some queueing problems with the service station subject to breakdowns, Operations Research, 11(3), 303-320, 1963.

[2] *Bellman, R.* : Introduction to matrix analysis, McGraw Hill book Co., New York, 1960.

[3] *Bhaskar Sengupta*: A queue with service interruptions in an alternating random environment, Operations Research, 38(2), 308-318, 1990.

[4] *Boxma, O., Mandjes, M., and Kella, O.* : On a queueing model with service interruptions, Probability in the Engineering and Informational Sciences, 22(4), 537-555, 2008.

[5] *Breuer, L. and Baum, D.* : An introduction to queueing theory and matrix-analytic methods, Springer, 2005.

[6] *Chakravarthy, S. R.* : A multi-server queueing model with server consultations, European Journal of Operational Research, 233(3), 625-639, 2014.

[7] *Chakravarthy, S.R.*: The Batch Markovian Arrival Process: A Review and Future Work, Advances in Probability Theory and Stochastic Processes, Eds., A. Krishnamoorthy et al., Notable Publications, New Jersey, 21-49, 2001.

[8] *Chakravarthy, S.R.* : Analysis of a multi-server queue with Markovian arrivals and synchronous phase type vacations, Asia-Pacific Journal of Operational Research, 26(1), 85-113, 2009.

[9] *Chakravarthy, S.R.*: Markovian arrival processes, Wiley Encyclopedia of Operations Research and Management Science, 2010.

[10] *Chakravarthy, S.R. and Alfa, A.S.* : A finite capacity queue with Markovian arrivals and two servers with group services, Journal of Applied Mathematics and Stochastic Analysis, 7(2), 161-178, 1994.

[11] *Chakravarthy, S.R., Krishnamoorthy, A. and Jousua, V.C.*: Analysis of a multi-server retrial queue with search of customers from the orbit, Performance Evaluation, 63(8), 776-798, 2006.

[12] *Chakravarthy, S.R. and Agnihothri, S. R.* : A server backup model with Markovian arrivals and phase type services, European Journal of Operational Research, 184(2), 584-609, 2008.

[13] *Chao, X. and Zhao, Y.* : Analysis of multi-server queues with station and server vacations, European Journal of Operational Research, 110(2), 392-406, 1998.

[14] *Chung K. L. and AitSahlia F.* : Elementary probability theory with Stochastic Processes and an introduction to Mathematical finance, Springer, New York, 2003.

[15] *Cinlar, E.* : Introduction to stochastic processes, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1975.

[16] *Erlang, A. K.* : The theory of probabilities and telephone conversations, Nyt Tidsskrift Matematik, 20(B), 33-39, 1909.

[17] *Fiems, D., Maertens, T. and Bruneel, H.* : Queueing systems with different types of interruptions, European Journal of Operational Research, 188(3), 838-845, 2008.

[18] *Gaver, D. P.* : A waiting line with interrupted service including priorities, Journal of Royal Statistical Society, B 24(1), 73-90, 1962.

[19] *Gaver, D.P., Jacobs, P.A. and Latouche, G.* : Finite birth and death models in randomly changing environments, Advances in Applied Probability, 16(4), 715-731, 1984.

[20] *Gomez-Corral, A., Krishnamoorthy, A. and Viswanath C Narayanan:* The impact of self generation of priorities on multi-server queues with finite capacity, Stochastic Models, 21, 427-447, 2005.

[21] *Graham, A.*: Kronecker products and matrix calculus with applications, Ellis Horwood Ltd., Chichester, 1981.

[22] *Gross, D. and Harris, C.M.* : Fundamentals of queueing theory, Third Edition, John Wiley and sons, New York, 1998.

[23] *Ibe, O. C. and Trivedi, K. S.*: Two queues with alternating service and server breakdown, Queueing Systems, 7(3), 253-268, 1990.

[24] *Karlin, S. and Taylor, H.M.* : A first course in stochastic processes, Academic press, New York, 1975.

[25] *Karlin, S. and Taylor, H.M.* : A second course in stochastic processes, Academic press, New York, 1981.

[26] *Keilson J.*: Queues subject to service interruptions, The Annals of Mathematical Statistics, 33(4), 1314-1322, 1962.

[27] *Kemeny, J.G. and Snell, J.L.* : Finite markov chains, The University Series in Undergraduate Mathematics, D. Van Nostrand, Princeton, USA, 1960.

[28] *Klimenok, V. and Dudin, A. N.*: A BMAP/PH/N queue with negative customers and partial protection of service, Communications in Statistics- Simulation and Computation, 41(7), 1062-1082, 2012.

[29] *Klimenok, V., Kim, C. S. and Kuznetsov,V.*: A multi-server queue with negative customers and partial protection of service, Proceedings of 13^{th} International Conference on analytical and stochastic Modelling Techniques and applications (ASMTA 06), Bonn, Germany; Eds. K Al-Begain; 143-148, 28-31 May 2006 .

[30] *Krishna Kumar, B. and Pavai Madheswari, S.* : An M/M/2 queueing system with heterogeneous servers and multiple vacations, Mathematical and Computer Modelling, 41(13), 1415-1429, 2005.

[31] *Krishnamoorthy, A., Babu, S. and Viswanath C Narayanan* : The MAP/(PH/PH)/c queue with self generation of priorities and nonpreemptive service, Stochastic Analysis and Applications, 26(6), 1250-1266, 2008.

[32] *Krishnamoorthy, A and Gopakumar,B. and Viswanath,C. N.*: An M/Em/1 queue with protected and unprotected phases from interrup-

tions, 5^{th} International conference on Queueing Theory and Network Applications, Beijing, China, July 24-26, 2010.

[33] *Krishnamoorthy, A., Pramod, P. K. and Chakravarthy, S. R.* : A Note on characterizing service interruptions with phase type distribution, Stochastic Analysis and Applications , 31(4), 671-683, 2013.

[34] *Krishnamoorthy, A., Pramod, P.K. and Deepak, T.G.* : On a queue with interruptions and repeat/resumption of service, Non-linear Analysis, Theory, Methods and Applications, 71, e1673-e1683, Elsevier, 2009.

[35] *Latouche, G. and Ramaswami, V.* : Introduction to matrix analytic methods in stochastic modelling, ASA/SIAM series on Statistics and Applied Probability, 1999.

[36] *Lucantoni, D.M.*: New results on the single server queue with a batch Markovian arrival process, Stochastic Models, 7(1), 1-46, 1991

[37] *Marcus, M. and Minc, H.*: A survey of matrix theory and matrix inequalities, Allyn and Bacon, Boston, 1964.

[38] *Medhi, J.* : Stochastic Processes, New age international, New Delhi, 1984.

[39] *Medhi, J.* : Stochastic models in queueing theory, Academic press, An imprint of Elsevier, USA, 2003.

[40] *Neuts, M.F.*: Probability distributions of phase type, Liber Amicorum Professor Emeritus H. Florin, University of Louvain, Belgium, 173-206, 1975.

[41] *Neuts, M.F.* : A versatile Markovian point process, Journal of Applied Probability, 16(4), 764-779, 1979.

[42] *Neuts, M.F. and Lucantoni, D.M.* : A Markovian queue with N servers subject to breakdowns and repairs, Management Science, 25(9), 849-861, 1979.

[43] *Neuts, M.F. and Takahashi, Y.* : Asymptotic behavior of the stationary distribution in the GI/PH/C queue with heterogeneous servers, Zeitschrift fur Wahrscheinlichkeitstheorie, 57(4), 441-452, 1981.

[44] *Neuts, M.F.* : Matrix-geometric solutions in stochastic models, An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, 1981.

[45] *Neuts, M.F.*: Structured stochastic matrices of M/G/1 type and their applications, Marcel Dekker, New York, 1989.

[46] *Neuts, M.F.*: Models based on the Markovian arrival process, IEICE Transactions on Communications, E75-B (12), 1255-1265, 1992.

[47] *Neuts, M.F.*: Algorithmic probability: A collection of problems, Chapman and Hall, New York, 1995.

[48] *Ross, S.M.*: Stochastic processes, 2nd edn., John Wiley and sons, New York, 1996.

[49] *Szpankowski, W.* : Stability conditions for multidimensional queueing systems with computer applications, Operations Research, 36(6), 944-957, 1988.

[50] *Tian, N., Li, Q. and Gao, J.* : Conditional stochastic decompositions in the M/M/C queue with server vacations, Stochastic Models, 15(2), 367-377, 1999.

[51] *Takine, T. and Sengupta, B.* : A single server queue with service interruptions, Queueing System, 26(3), 285-300, 1997.

[52] *Tijms, H.C.*: A first course in stochastic models, John Wiley and sons, Chichester, 2003.

[53] *White, H. and Christie, L. S.* : Queueing with preemptive priorities or with breakdown, Operations Research, 6(1), 79-95, 1958.

[54] *William J. Stewart.* : Probability, Markov Chains, Queues and Simulation: The Mathematical Basics of Performance Modelling, Princeton University Press, Princeton and Oxford, 2009.

CURRICULUM VITAE

Name :	Resmi. T
Present Address :	Department of Mathematics, Cochin University of Science and Technology, Cochin, Kerala, India – 682 022.
Official Address :	Assistant Professor, Department of Mathematics, K.K.T.M. Govt. College, Pullut Kerala, India
Permanent Address :	Kanat (H) Anandapuram(P.O) Nellayi (Via) Thrissus(Dt) Kerala, India – 680 323.
Email :	resmitkktm@gmail.com
Qualifications :	**B.Sc.** (Mathematics), 1998, Calicut University **M.Sc.** (Mathematics), 2000, Calicut University
Research Interest :	Queueing theory.