QUEUEING MODELS WITH VACATIONS AND WORKING VACATIONS

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

SREENIVASAN C

Department of Mathematics

Cochin University of Science and Technology

Cochin - 682 022

June 2012

TO

MY PARENTS

Certificate

This is to certify that the thesis entitled '**Queueing Models with Vacations and Working Vacations**' submitted to the Cochin University of Science and Technology by Mr. Sreenivasan C for the award of the degree of Doctor of Philosophy under the Faculty of Science is a bona fide record of studies carried out by him 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. A. Krishnamoorthy
(Supervisor)
Emeritus Professor
Department of Mathematics
Cochin University of Science and Technology
Cochin- 682022, Kerala.

Cochin-22
16/06/12

Declaration

I, Sreenivasan C, hereby declare that this thesis entitled '**Queueing Models with Vacations and Working Vacations**' 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.

Sreenivasan C
Research Scholar
Registration No.3365
Department of Mathematics
Cochin University of Science and Technology
Cochin-682022, Kerala.

Cochin-22
16/06/12

Acknowledgement

I would like to express my whole hearted thanks to each and everyone who have helped me to complete my research work, in one way or other. First and foremost I thank my supervisor Prof. A. Krishnamoorthy, for his guidance, kindness, inspiration and motivation. I am also indebted to Prof. Srinivas R Chakravarthy for sharing his knowledge of matrix analytic methods and programming with FORTRAN with me, in addition to his contributions in the fourth chapter of this thesis.

I thankfully acknowledge helps I received from Mr. Varghese Jacob, my fellow research scholar. Thanks are also due to Dr. K.P. Naveena Chandran, and Dr. G.N. Prasanth of my parent Department, for the encouragement given by them. Discussions with Prof. Alexander Dudin, Dr. B. Krishnakumar, Dr. P.K. Pramod, Dr. T.G. Deepak and Dr. Viswanath C Narayanan were fruitful. I thank Dr. B. Lakshmy, Dr. R.S. Chakravarti, Dr. A. Vijayakumar, Dr. M.N.N. Namboodiri and Dr. P.G. Romeo for the support given by them. The office staff and the library staff of the Dept. of Mathematics, Cochin University of Science and Technology, have been very cooperative. The help and support given by the Principals, who have been in the office of Govt. College, Chittur are gratefully acknowledged. The encouragement given by Mr. Chidambaran, Mrs. Mary Shalet, Dr. Reji, Dr. Ali Akbar, Ms. Haseena, Ms. V.P. Lakshmy and Dr. Shine Lal of Department of Mathematics, Govt. College, Chittur was tremendous. Financial support was given by the UGC under the faculty improvement program.

My fellow research scholars have been very sincere to me. They include Mr. Manikandan, Mrs. Deepthi, Mr. Kiran Kumar, Mr. Pravas, Mrs. Anu Varghese, Mr. Jayaprasad, Mr. Tijo James, Dr. Ajayakumar, Dr. Seema Varghese, Mrs. Chithra, Mr. Tonny, Mr. Sajeev S Nair, Mr. Sathyan, Mr. Gopakumar, Mr. Jaison Jacob, Dr. Lalitha, Ms. Anusha, Mrs. Vinitha, Mrs. Jaya, Mr. Santhosh Pandey, Mr. Didimos, Mrs. Raji George, Mrs Pamy Sebastian, Mr. Gireesan, Mr. Manjunath, Mr. Vijayagovind, Mr. Tibin Thomas, Ms. Seethu Varghese, Ms. Binitha Benny, Ms. Dhanya Shajin, Mrs. Reshma and Mr. Vivek.

Throughout my career my parents stood with me in all my ups and downs. Without them I would not have achieved anything. During my days in Cochin University of Science and Technology my wife and kid sorely missed my care and attention. They sacrificed quite a lot for me and took extreme care not to distract me from my studies.

Sreenivasan C

QUEUEING MODELS WITH VACATIONS AND WORKING VACATIONS

CONTENTS

NOTATIONS, SYMBOLS AND ABBREVIATIONS

$\boldsymbol{e}$	-	Column vector consisting of 1's appropriate dimension
$\boldsymbol{e}_r$	-	Column vector of dimension r consisting of 1's
$\boldsymbol{e}_r(j)$	-	Column vector of dimension r with 1 in the j^{th} position and zero elsewhere
I	-	Identity matrix of appropriate dimension
I_r	-	Identity matrix of dimension r
$\otimes$	-	Kronecker product
$\oplus$	-	Kronecker sum
LST	-	Laplace-Stieltjes Transform
$CTMC$	-	Continuous time Markov Chain
MAP	-	Markovian arrival process
$MMAP$	-	Marked Markovian arrival process
QBD	-	Quasi-Birth-and-Death
$LIQBD$	-	Level Independent QBD
PH	-	Phase Type

1. INTRODUCTION

1.1 Queueing theory and Matrix analytic methods

We encounter queues in almost all walks of our life. Some times the queues that we are in are visible while at other times they are not. For instance, when we make a request for some service to a telephone call centre we are not aware of the queue which we may be in. Apparently no one really wants to be in a queue especially when it is too long. However, given the fact that one has to spend enormous amount of time in queues, it is of great significance to analyze these congestion situations using appropriate queueing models.

Until early 1970's queueing theorists all over the world relied heavily on complex analytic tools to tackle problems in queueing theory. Due to this research publications in this area became exceedingly long and had very little impact on those who apply queueing models in engineering and technology. This motivated M.F. Neuts to develop phase type distributions (abbreviated as PH distributions) [46] and matrix analytic methods. Later Neuts developed versatile Markovian point process ($VMPP$) [48] which is now known as batch Markovian arrival process ($BMAP$). These developments triggered a revolution in the field of queueing theory as algorithmic probability emerged to be a very effective tool in solving queueing theoretic problems.

1.2 Phase Type Distributions

Here, we confine our discussion to continuous time phase type distributions. Consider a finite state Markov chain with m transient states and one absorbing state. The infinitesimal generator Q of this Markov chain be partitioned as

$$Q = \begin{pmatrix} T & \boldsymbol{T^0} \\ O & 0 \end{pmatrix},$$

where T is a matrix of order m and $\boldsymbol{T^0}$ is a column vector such that $T\boldsymbol{e} + \boldsymbol{T^0} = \boldsymbol{0}$, $\boldsymbol{e}$ 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 $F(x) = 1 - \boldsymbol{\alpha}\ \exp(\mathrm{Tx})\boldsymbol{e}$, 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 $f(x) = \boldsymbol{\alpha}\ exp(Tx)\,\boldsymbol{T^0}$, for $x \geq 0$. The Laplace Stieltjes transform $\tilde{f}(s)$ of $F(x)$ is given by $\tilde{f}(s) = \alpha_{m+1} + \boldsymbol{\alpha}(sI - T)^{-1}\boldsymbol{T^0}$, for $Re\ s \geq 0$. Hence the k^{th} non central moments of $F(x)$ is given by the formula ${\mu_k}' = (-1)^k\ k!(\boldsymbol{\alpha}T^{-k}\boldsymbol{e})$ for $k \geq 1$. The class of PH distributions include the distributions such as exponential, hyperexponential, Erlang and generalized Erlang 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 [39], [9], [50] and [10].

1.3 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

$$Q^* = \begin{pmatrix} D_0 & D_1 & 0 & 0 & \dots \\ 0 & D_0 & D_1 & 0 & \dots \\ 0 & 0 & D_0 & D_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

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 $(D_0 + D_1)\boldsymbol{e}_m = 0$, $\boldsymbol{e}_m$ 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 $\lambda = \boldsymbol{\delta} D_1 \boldsymbol{e}_m$ 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 nonrenewal process. However, by appropriately choosing the parameters of the MAP the underlying arrival process can be made as a renewal process. It can easily be verified that a renewal process with interarrival times phase type distributed with representation ($\boldsymbol{\alpha}$,T) and the exit rates vector $\boldsymbol{T}^0 = -T\boldsymbol{e}$ can be obtained as a special case of the MAP. To see this it is enough to replace D_0 and D_1 respectively by T and $\boldsymbol{T}^0\boldsymbol{\alpha}$ in the above discussion of MAP. 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 [43], [51], [52] and for a review and recent work on MAP we refer the reader to [10]. Chakravarthy [11] and Krishnamoorthy et al. [37] provide an account of more recent works in this area.

1.4 Quasi-Birth-and-Death Process

A level independent quasi-birth-and-death (QBD) 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

$$\tilde{Q} = \begin{pmatrix} B_0 & A_0 & & & \\ B_1 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Note that 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 less than 1. Although level dependent quasi-birth-and-death process is also there, it is not used in this thesis.

1.5 *Logarithmic Reduction Algorithm for computation of R*

Step 0: $H \leftarrow (-A_1)^{-1}A_0$, $L \leftarrow (-A_1)^{-1}A_2$, $G = L$, and $T = H$.

Step 1:

$$U = HL + LH$$

$$M = H^2$$

$$H \leftarrow (I - U)^{-1}M$$

$$M \leftarrow L^2$$

$$L \leftarrow (I - U)^{-1}M$$

$$G \leftarrow G + TL$$

$$T \leftarrow TH$$

Continue Step 1 until $||\boldsymbol{e} - G\boldsymbol{e}||_\infty < \epsilon$.

Step 2: $R = -A_0(A_1 + A_0G)^{-1}$

1.6 Kronecker Product and Kronecker Sum

Let A be a matrix of order $m \times n$ and B one of order $p \times q$, 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 more details on Kronecker products and sums, we refer the reader to [24] and [44].

1.7 Queues with Vacations and Working Vacations

In the modern world there is tough competition between service providers. So in order to survive service systems have to be managed efficiently and economically. Demand for service often fluctuates. There may be periods of low customer inflow. During such periods, it may not be economical from the system point of view to retain idle servers in the system. At the same time no system can afford to lose its customers and goodwill. So there is a need to strike a balance between the two extreme situations. It is from this stand point, we study queueing models with vacations and working vacations.

Queues with vacations have been extensively studied by several authors. Doshi [19] provides an exhaustive survey of such work through 1985. Since then the vacation models have been studied in different contexts. Among these include stochastic decomposition of queue length and that of stationary waiting time and we refer the reader to the recent book by Tian and Zhang [64] for details. Recently vacation models have gained significance in telecommunication networks. However, compared to continuous time models discrete time models are more appropriate for modelling computer and telecommunication systems. Servi and Finn [55] introduced a working vacation model with the idea of offering services but at a lower rate whenever the server is on vacation. Their model was generalized to the case of $M/G/1$ in ([32], [68]), and to $GI/M/1$ model in [8]. A survey of working vacation models with emphasis on the use of matrix analytic methods is given in Tian and Li [65]. Working vacation models have a number of applications in practice. Two such examples are given in [65].

Recently, Li and Tian [42] studied an $M/M/1$ queue with working vacations in which vacationing server offers services at a lower rate for the first customer arriving during a vacation. Upon completion of the service at a lower rate the server will (a) continue the current vacation (if not already completed) or take another vacation (if the working vacation expired) if there are no customers waiting; or (b) resume at a normal rate (irrespective of whether the vacation expired or not) if there are customers waiting. Resuming services at a normal rate while the vacation is still in progress corresponds to the vacation being interrupted.

M/M/1 retrial queue with working vacations has been discussed by Van Do [18]. In classical retrial queueing systems server idle time is very high. In the modern scenario it is not desirable from the service system's point of view to have a long idle time. To this end Artalejo et al. [4] introduced a notion called orbital search where the server looks out for potential customers from the orbit immediately after a service completion with probability p $(0 \leq p \leq 1)$. Dudin et al. [20] and Krishnamoorthy et al. [34] also consider orbital search with different arrival streams and different service time distributions. Chakravarthy et al. [12] consider the multi server case.

But even with the search option, system may not be able to utilize the entire server idle time. It is from this stand point one explores the possibility of retrial queueing systems with vacations and working vacations. During vacations the idle server may attend some less urgent secondary task. We may also consider the notion of working vacation depending upon the nature of the secondary job attended. In the latter case the server returns to attend the primary job as and when a customer arrives in the system.

1.8 Summary of the thesis

The thesis entitled "Queueing Models with Vacations and Working Vacations" consists of seven chapters including the introductory chapter. In chapters 2 to 7 we analyze different queueing models highlighting the role played by vacations and working vacations. The duration of vacation is exponentially distributed in all these models and multiple vacation policy is followed.

In chapter 2 we discuss an $M/M/2$ queueing system with heterogeneous servers, one of which is always available while the other goes on vacation in the absence of customers waiting for service. Using matrix geometric methods the system is analyzed in the steady-state. Busy period structure is analyzed and the mean waiting time is computed. Conditional stochastic decomposition of queue length is derived. An illustrative example is provided to study the effect of the input parameters on the system performance measures.

Chapter 3 considers a similar setup as chapter 2. However, in this model the vacationing server returns to serve at a lower rate when an arrival finds the other server busy. The model is analyzed in essentially the same way as in chapter 2 and a numerical example is provided to bring out the qualitative nature of the model.

In reality the assumptions like Poisson arrivals and the exponential service times are very restrictive though they make the system analytically more tractable. The traffic in modern communication network is highly irregular. Of late to model systems with repeated calls and bursty arrivals MAP (Markovian arrival process) is used. The MAP is a tractable class of point process which is in general nonrenewal. In spite of its versatility it is highly tractable as well. Phase type distributions are ideally suited for applying matrix analytic methods. In all the remaining chapters we assume the arrival process to be MAP and service process to be phase type.

In chapter 4 we consider a $MAP/PH/1$ queue with working vacations. At a departure epoch, the server finding the system empty, takes a vacation. A customer arriving during a vacation will be served but at a lower rate. Vacation mode service is also phase type distributed. The server continues to serve at this rate until either the vacation clock expires or the queue length hits the threshold value N, $1 \leq N < \infty$. When either of these two occurs the server instantaneously switches over to the normal rate and continues to serve at this rate until the system becomes empty. Conditional mean waiting time of a customer who arrives when the service is in a) vacation mode b)normal mode and then the unconditional mean waiting time of a customer is computed. The slow service mode is analyzed in detail. The mean duration of uninterrupted vacation and the mean number of times the server goes to vacation during the slow service are computed. Numerical illustration has been provided to get an insight into the model.

Chapter 5 discusses a $MAP/PH/1$ retrial queueing system with working vacations. If an arrival finds the server busy he joins a group of retrial customers called orbit. We consider the case of constant retrial rate which has applications in the local area networks and communication protocols. The server takes a vacation if there are no customers in the orbit at a departure epoch. The service offered to a customer who arrives during vacation is slower than the regular service. A number of performance measures are listed with their formulae and illustrative numerical examples have been provided.

In chapter 6 the setup of the model is similar to that of chapter 5. The significant difference in this model is that there is a finite buffer for arrivals. If a departure leaves the buffer empty the server goes on a working vacation.

Each customer in the orbit makes retrial for a place in the server or buffer and the retrial rate is independent of the number of customers in the orbit. The system characteristics are studied with the help of numerical illustrations.

Chapter 7 considers an $MMAP(2)/PH/1$ queueing model with a finite retrial group. High priority customers enjoy infinite waiting space. In the absence of high priority customers the server leaves the service area to proceed on a vacation. During vacation if a customer arrives the server returns to serve. Service offered during vacation has the same distribution as the regular one. If a low priority arrival encounters a busy server he tries to find a place (if any) in the retrial group. If there is no vacancy in the orbit the customer leaves the system forever. Once a low priority customer is taken for service he is not dislodged before service completion.

2. AN $M/M/2$ QUEUEING SYSTEM WITH HETEROGENEOUS SERVERS INCLUDING ONE VACATIONING SERVER

It has been observed by Neuts and Takahashi [49] that queueing systems with more than two heterogeneous servers are analytically intractable. So in order to get some explicit results one has to restrict the domain to systems with two heterogeneous servers. In this chapter we study an M/M/2 queueing system with heterogeneous servers, with one server taking multiple vacations. The other server remains in the system even when the system is empty. In this aspect our model differs from that of Krishna Kumar and Pavai Madheswari [33]. They consider a system of two heterogeneous servers, where both servers go on vacation in the absence of customers waiting for service. Towards the end of this chapter a numerical example is provided to illustrate how the system characteristics behave as the input parameters change.

2.1 Mathematical Model

We consider an M/M/2 queueing model with heterogeneous servers, called server 1 and server 2. Server 1 is always available whereas server 2 goes on

[0] To appear in Calcutta Statistical Association Bulletin

vacation whenever there is no customer waiting for service. Let the service rates of servers 1 and 2 be μ_1 and μ_2, respectively, where $\mu_1 \neq \mu_2$. Customers arrive to the system according to a Poisson process of parameter λ. The duration of vacation is exponentially distributed with parameter η. At the end of a vacation, if there is no customer waiting for service the server goes on another vacation. Otherwise it resumes service. For clarity we assume that if an arriving customer finds a free server he enters service immediately. Else he joins the queue.

2.1.1 The QBD process

The model discussed above can be studied as a level independent quasi-birth-and-death ($LIQBD$) process. First, we set up the necessary notations.
At time t, let $N(t)$ be the number of customers in the system and

$$J(t) = \begin{cases} 0, \textit{if the server 2 is on vacation} , \\ 1, \textit{if it is busy,} \end{cases}$$

Let $X(t) = (N(t), J(t))$; then $(X(t) : t \geq 0)$ is a continuous time Markov Chain (CTMC) with state space

$$\Omega = \{(0,0)\} \bigcup \bigcup_{i=1}^{\infty} l(i)$$

where

$$l(i) = \{(i, j) : i \geq 1, j = 0 \text{ or } 1\}.$$

The infinitesimal generator matrix Q of this Markov chain is given by

$$Q = \begin{bmatrix} B_{00} & B_{01} & & & \\ B_{10} & B_{11} & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

where the block matrices appearing in Q are as follows.

$$B_{00} = -\lambda,\ B_{01} = \begin{bmatrix} \lambda & 0 \end{bmatrix},$$

$$B_{10} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\ B_{11} = \begin{bmatrix} -\lambda - \mu_1 & 0 \\ 0 & -\lambda - \mu_2 \end{bmatrix},\ A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -\lambda - \mu_1 - \eta & \eta \\ 0 & -\lambda - \mu_1 - \mu_2 \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_1 + \mu_2 \end{bmatrix}.$$

2.2 *Steady-state analysis*

In this section we discuss the steady-state analysis of the model under study.

2.2.1 Stability Condition

Theorem 2.2.1. *The queueing system described above is stable if and only if $\rho < 1$ where $\rho = \lambda/(\mu_1 + \mu_2)$.*

Proof. To establish the stability condition we use Pakes' lemma (see [58]). Let N_i be the number of customers in the system immediately after the departure of the i^{th} customer. Then $\{N_i : i \in \mathbb{N}\}$ satisfies the equation

$$N_i = \begin{cases} N_{i-1} - 1 + V_i & if \; N_{i-1} \geq 1 \\ V_i & if \; N_{i-1} = 0 \end{cases}$$

where V_i is the number of arrivals during the service of i^{th} customer. Clearly $\{N_i : i \in \mathbb{N}\}$ is an irreducible aperiodic Markov chain. Pakes' lemma asserts that an aperiodic irreducible Markov chain is ergodic, if there exists an $\epsilon > 0$ such that the mean drift

$$\phi_j = E[(N_{i+1} - N_i)/N_i = j]$$

is finite for all $j \in \mathbb{N}$ and $\phi_j \leq -\epsilon$ for all $j \in \mathbb{N}$ except perhaps for a finite number. In the present model, value of the mean drift is

$$\phi_j = \begin{cases} -1 + \rho & if \; j \geq 1 \\ \rho & if \; j = 0 \end{cases}$$

Thus if $\rho < 1$ the Markov chain $\{N_i : i \in \mathbb{N}\}$ is ergodic and hence the condition is sufficient.

To prove the necessity of the condition assume that $\rho \geq 1$. We use theorem 1 in Sennot et al. [56], which states that $\{N_i : i \in \mathbb{N}\}$ is nonergodic if it satisfies Kaplan's condition; $\phi_j < \infty$, for $j \geq 0$ and there exists a j_0 such that $\phi_j \geq 0$, for $j \geq j_0$. When $\rho \geq 1$ Kaplan's condition is readily satisfied. Hence the Markov chain is not ergodic. □

2.2.2 Steady-state probability vector

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (x_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$, be the steady-state probability vector of Q. Note that x_0 is a scalar and $\boldsymbol{x}_i = (x_{i0}, x_{i1})$, for $i \geq 1$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = \mathbf{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ is a column vector of 1's with appropriate dimension. Apparently when the stability condition is satisfied the sub vectors of $\boldsymbol{x}$, corresponding to the different levels are given by the equation $\boldsymbol{x}_j = x_1 R^{j-1}$, $j \geq 2$, where R is the minimal nonnegative solution of the matrix quadratic equation (see [50])

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{2.1}$$

Knowing the matrix R, x_0 and $\boldsymbol{x}_1$ are obtained by solving the equations

$$x_0 B_{00} + \boldsymbol{x}_1 B_{10} = 0 \tag{2.2}$$

and

$$x_0 B_{01} + \boldsymbol{x}_1 (B_{11} + RA_2) = \mathbf{0} \tag{2.3}$$

subject to the normalizing condition

$$x_0 + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{e} = 1. \tag{2.4}$$

Theorem 2.2.2. *The matrix R of equation (2.1) is given by* $R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$, *where* $R_{11} = \frac{\lambda+\mu_1+\eta-\sqrt{(\lambda+\mu_1+\eta)^2-4\lambda\mu_1}}{2\mu_1}$, $R_{22} = \rho$ *and* $R_{12} = \rho - \mu_1 R_{11}/(\mu_1 + \mu_2)$.

Proof. Since A_0, A_1 and A_2 are upper triangular, R is essentially an upper triangular matrix. The value of R_{11} follows from the assertion that R is the minimal non negative solution of (2.1). The rest of the proof is an easy consequence of the condition $RA_2\boldsymbol{e} = A_0\boldsymbol{e}$. □

Remark: Though R has a nice structure which enables us to make use of the properties like $R^k = \begin{bmatrix} R_{11}^k & R_{12}\sum_{j=0}^{k-1} R_{11}^j R_{22}^{k-j-1} \\ 0 & R_{22}^k \end{bmatrix}$, for $k \geq 1$, due to the form of the expression for R_{11} it may not be easy to carry out the computations required in the forthcoming discussions. Hence we explore the possibility of algorithmic computation of R. The computation of R matrix can be carried out using logarithmic reduction algorithm.

2.2.3 Busy period analysis

For the system under study, busy period is the interval between arrival of a customer to the empty system and the first epoch thereafter when the system becomes empty again. Thus it is precisely the first passage time from the state (1,0) to the state (0,0). For the vacation model, busy cycle for the system is the time interval between two successive departures, which leave the system empty. Thus the busy cycle is the first return time to state $(0, 0)$ with at least one visit to any other state. Before analyzing the busy period structure, we need to introduce the notion of fundamental period. For the QBD process under consideration, it is the first passage time from level i, where $i \geq 2$, to the level $i - 1$. The cases $i = 1$ and $i = 0$ corresponding to the boundary states need to be discussed separately. It should be noted that due to the structure of the QBD process the distribution of the first passage time is invariant in i away from the boudary states.

Let $G_{jj'}(k,x)$ denote the conditional probability that a QBD process, starting in the state (i,j) at time $t=0$ reaches the level $(i-1)$ for the first time no later than time x, after exactly k transitions to the left, and does so by entering the state $(i-1,j')$. For convenience, we introduce the joint transform

$$\tilde{G}_{jj'}(z,s) = \sum_{k=1}^{\infty} z^k \int_0^{\infty} e^{-sx} \mathrm{d}G_{jj'}(k,x) \qquad ; \qquad |z| \leq 1,\ Re(s) \geq 0$$

and the matrix

$$\tilde{G}(z,s) = (\tilde{G}_{jj'}(z,s)).$$

The matrix $\tilde{G}(z,s)$ is the unique solution to the equation (see [50])

$$\tilde{G}(z,s) = z(sI - A_1)^{-1}A_2 + (sI - A_1)^{-1}A_0\tilde{G}^2(z,s). \tag{2.5}$$

The matrix $G = \tilde{G}(1,0)$ takes care of the first passage times except for the boundary states. If we know the R matrix then G matrix can be computed using the result (see [39])

$$G = -(A_1 + RA_2)^{-1}A_2.$$

Otherwise we may use logarithmic reduction method to compute G. For the boundary level states 1 and 0 let $G^{(1,0)}_{jj'}(k,x)$ and $G^{(0,0)}_{jj'}(k,x)$ be the conditional probability discussed above for the first passage time from level 1 to level 0 and the first return time to the level 0 respectively. Then as in (2.5) we get

$$\tilde{G}^{(1,0)}(z,s) = z(sI - B_{11})^{-1}B_{10} + (sI - B_{11})^{-1}A_0\tilde{G}(z,s)\tilde{G}^{(1,0)}(z,s) \quad (2.6)$$

and

$$\tilde{G}^{(0,0)}(z,s) = [\lambda/(s+\lambda), \quad 0]\tilde{G}^{(1,0)}(z,s). \quad (2.7)$$

Note that $\tilde{G}^{(1,0)}(z,s)$ is a 2×1 matrix. Thus the Laplace Stieltjes transform (LST) of the busy period is the first element of $\tilde{G}^{(1,0)}(1,s)$. For convenience use the notations

$$G_{10} = \tilde{G}^{(1,0)}(1,0) \text{ and } G_{00} = \tilde{G}^{(0,0)}(1,0).$$

Due to the positive recurrence of the QBD process, matrices G, G_{10} and G_{00} are all stochastic. If we let

$$C_0 = (-A_1)^{-1}A_2 \text{ and } C_2 = (-A_1)^{-1}A_0,$$

then G is the minimal nonnegative solution (see [50]) to the matrix equation

$$G = C_0 + C_2G^2.$$

From equations (2.6) and (2.7) we get

$$G_{10} = -(B_{11} + A_0G)^{-1}B_{10} \quad (2.8)$$

and

$$G_{00} = [1, \quad 0]G_{10}. \tag{2.9}$$

Equation (2.5) is equivalent to

$$zA_2 - (sI - A_1)\tilde{G}(z,s) + A_0\tilde{G}^2(z,s) = 0. \tag{2.10}$$

Let

$$M = -\left.\frac{\partial \tilde{G}(z,s)}{\partial s}\right|_{z=1,s=0}$$

and

$$\tilde{M} = \left.\frac{\partial \tilde{G}(z,s)}{\partial z}\right|_{z=1,s=0}.$$

Differentiation of (2.10) with respect to s and z followed by setting $z = 1$ and $s = 0$ leads to (see [50])

$$M = -A_1^{-1}G + C_2(GM + MG)$$

and

$$\tilde{M} = C_0 + C_2(G\tilde{M} + \tilde{M}G).$$

With $\mathbf{0}$ as starting value for M and $\tilde{M}$, successive substitutions in the above equations yield the values of M and $\tilde{M}$. Applying an exactly similar reasoning to (2.6) and (2.7), we get

$$M_{10} = -(B_{11} + A_0G)^{-1}(I + A_0M)G_{10} \tag{2.11}$$

and

$$M_{00} = [1/\lambda, \quad 0]G_{10} + [1, \quad 0]M_{10}, \tag{2.12}$$

where

$$M_{10} = -\left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial s}\right|_{z=1,s=0}$$

and

$$M_{00} = -\left.\frac{\partial \tilde{G}^{(0,0)}(z,s)}{\partial s}\right|_{z=1,s=0}.$$

Note that M_{10} is a 2×1 matrix and M_{00} is a scalar. The first element of the vector M_{10} and M_{00} are mean lengths of a busy period and a busy cycle respectively. The second element of M_{10} gives the first passage time from the state (1,1) to the state (0,0). With the notation

$$\tilde{M}_{10} = \left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial z}\right|_{z=1,s=0}.$$

It follows from equation (2.6) that

$$\tilde{M}_{10} = -(B_{11} + A_0 G)^{-1}(B_{10} + A_0 M G_{10}). \tag{2.13}$$

The first component of the vector $\tilde{M}_{10}$ is the mean number of service completions in a busy period.

2.2.4 Stationary waiting time in the queue

Let $W(t)$ be the distribution function of the waiting time in the queue of an arriving (tagged) customer. Note that if there is no customer in the system, the arrival receives service immediately. This happens with probability x_0.

Also when the only customer in the system is receiving service from server 2, the tagged customer receives service from the server 1 without any delay. This event occurs with probability x_{11}. Thus with probability $1 - x_0 - x_{11}$ the customer has to wait before getting the service. The waiting time may be viewed as the time until absorption in a Markov chain with state space

$$\Omega_1 = \{\mathbf{*}\} \bigcup \{\mathbf{1}, \mathbf{2}, \mathbf{3}, ...\}$$

Here $\mathbf{*}$ is the absorbing state which corresponds to taking the tagged customer into service and is obtained by lumping together the states $(0, 0)$ and $(1, 1)$. Further $\mathbf{1} = \{(1, 0)\}$ and $\mathbf{i} = \{(i, j), i \geq 2, j = 0 \text{ or } 1\}$. The states other than the absorbing state correspond to the number of customers present in the system as the tagged customer arrives. Once the tagged customer joins the queue, the subsequent arrivals will not affect his waiting time in the queue. Hence the parameter λ does not show up in the generator matrix $\tilde{Q}$ of this Markov process, given by

$$\tilde{Q} = \begin{array}{c} \\ * \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} * & 1 & 2 & 3 & \ldots \end{array} \\ \begin{pmatrix} & & & & \\ C_{10} & C_{11} & & & \\ C_{20} & C_{21} & B_1 & & \\ & & B_2 & B_1 & \\ & & & \ddots & \ddots \end{pmatrix} \end{array}, \text{ where}$$

$C_{10} = \mu_1$, $C_{11} = -\mu_1$, and $C_{20} = \begin{bmatrix} 0 \\ \mu_1 + \mu_2 \end{bmatrix}$,

$C_{21} = \begin{bmatrix} \mu_1 \\ 0 \end{bmatrix}$, $B_1 = \begin{bmatrix} -\mu_1 - \eta & \eta \\ 0 & -\mu_1 - \mu_2 \end{bmatrix}$, and $B_2 = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_1 + \mu_2 \end{bmatrix}$.

Define

$$\boldsymbol{y}(t) = (y_*(t), y_1(t), \boldsymbol{y}_2(t), \boldsymbol{y}_3(t), \ldots),$$

where

$$\boldsymbol{y}_i(t) = (y_{i0}(t), y_{i1}(t)), \text{ for } i \geq 2.$$

The components of $\boldsymbol{y}_i(t)$ are the probabilities that at time t the $CTMC$ with generator $\tilde{Q}$, is in the respective states of level i. Note that $y_1(t)$ and $y_*(t)$ respectively, determine the probability that the process is in state (1,0) and absorbing state at time t. By the PASTA property we may write

$$\boldsymbol{y}(0) = (x_0 + x_{11}, x_{10}, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots).$$

Clearly

$$W(t) = y_*(t), \text{ for } t \geq 0. \tag{2.14}$$

The Markov process for finding the waiting time distribution has the initial probability vector $\boldsymbol{y}(0)$. Then the matrix differential equation $\boldsymbol{y}'(t) = \mathbf{y}(t)\tilde{Q}$ for $t \geq 0$ reduces to

$$y_*'(t) = y_1(t)C_{10} + \boldsymbol{y}_2(t)C_{20},$$

$$y_1'(t) = y_1(t)C_{11} + \boldsymbol{y}_2(t)C_{21}$$

and for $i \geq 2$,

$$\boldsymbol{y}_i'(t) = \boldsymbol{y}_i(t)B_1 + \boldsymbol{y}_{i+1}(t)B_2.$$

For $i \geq 2$ and $j = 0$ or 1, the LST of first passage time to a state $(2, j)$ in the level 2 is the $(j+1)^{th}$ element of the vector $\psi(s)$ (see [50]) given by

$$\psi(s) = \sum_{i=2}^{\infty} \boldsymbol{y}_i(0)[(sI - B_1)^{-1}B_2]^{i-2}. \tag{2.15}$$

Now starting from the state $(i, j), i = 1, 2$ the LST of the time until absorption, $\phi_j(i, s)$, is the $(j+1)^{th}$ component of the column vector $\phi(i, s)$. From $\tilde{Q}$, we get

$$\phi(1, s) = (sI - C_{11})^{-1}C_{10} \tag{2.16}$$

and

$$\phi(2, s) = (sI - B_1)^{-1}C_{21}\phi(1, s) + (sI - B_1)^{-1}C_{20}. \tag{2.17}$$

Therefore, the LST of the waiting time distribution is given by

$$\tilde{W}(s) = \psi(s)\phi(2, s) + y_1(0)\phi(1, s) \tag{2.18}$$

The mean waiting time can be obtained from $\tilde{W}(s)$ as

$$E(W) = -\tilde{W}'(0) = \frac{x_{10}}{\mu_1} - \psi'(0)\boldsymbol{e} - \psi(0)\phi'(2, 0). \tag{2.19}$$

The only term in the expression for $E(W)$ given by equation (2.19), which needs serious computation is the second one. For this we make use of the ideas in [47], [49] and [33]. It can be verified that

$$\psi'(0) = -\sum_{i=1}^{\infty} \boldsymbol{y}_{2+i}(0) \sum_{j=0}^{i-1} U^j(-B_1)^{-1}U^{i-j} \tag{2.20}$$

where $U = (-B_1)^{-1}B_2$ is a stochastic matrix. Hence $U^{i-j}\boldsymbol{e} = \boldsymbol{e}$. Thus we get

$$-\psi'(0)\boldsymbol{e} = \sum_{i=1}^{\infty} \boldsymbol{y}_{2+i}(0) \sum_{j=0}^{i-1} U^j(-B_1)^{-1}\boldsymbol{e}. \tag{2.21}$$

Now consider the matrix $U_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ which has the property that

$$UU_2 = U_2U = U_2.$$

Then we get

$$\sum_{j=0}^{i-1} U^j(I - U + U_2) = I - U^i + iU_2 \text{ for } i \geq 1.$$

By the classical theorem on finite Markov chains the matrix $(I - U + U_2)$ is nonsingular (see [31]). In view of the last equation, equation (2.21) becomes

$$-\psi'(0)\boldsymbol{e} = [\sum_{i=1}^{\infty} \boldsymbol{y}_{2+i}(0)(I - U^i + iU_2)](I - U + U_2)^{-1}(-B_1)^{-1}\boldsymbol{e}. \tag{2.22}$$

With this simplification for $-\psi'(0)\boldsymbol{e}$, we get

$$-\psi'(0)\boldsymbol{e} = [\boldsymbol{x}_2R(I-R)^{-1} - \psi(0) + I + \boldsymbol{x}_2R(I-R)^{-2}U_2](I-U+U_2)^{-1}(-B_1)^{-1}\boldsymbol{e}. \tag{2.23}$$

The fact that

$$\psi(0)\boldsymbol{e} = 1 - x_0 - x_{10} - x_{11}$$

enables us to compute the value of $\psi(0)$ to any desired degree of accuracy. This completes computation of $E(W)$.

2.2.5 Conditional stochastic decomposition of queue length

In this section we provide a stochastic decomposition of queue length in the stationary regime, subject to the condition that both servers are busy. Note that from equations (2.2) and (2.3) we get

$$x_{10} = \frac{\lambda x_0}{(\lambda + \mu_1 - \mu_1 R_{11})} \tag{2.24}$$

and

$$x_{11} = \frac{(\lambda - \mu_1 R_{11})\lambda x_0}{(\lambda + \mu_1 - \mu_1 R_{11})\mu_2} \tag{2.25}$$

The last two equations, along with the equation (2.4) determine x_0, x_{10} and x_{11}. Let Q_v be the queue length of the vacation model under study, subject to the condition that both servers are busy. Then we have

Theorem 2.2.3. *If $\rho < 1$, then $Q_v = Q_0 + Q_d$, where Q_0 and Q_d are two independent random variables. Q_0 is the queue length of the M/M/2 queueing model with heterogeneous servers without vacation and Q_d can be interpreted as the additional queue length due to vacation, subject to the condition that both servers are busy.*

Proof. Let P_b denote the Probability that both servers are are busy. Then

$$P_b = \sum_{n=2}^{\infty} x_{n1} = \sum_{n=2}^{\infty} (x_{10} R_{12} \sum_{j=0}^{n-2} R_{11}^j \rho^{n-j-2} + x_{11} \rho^{n-1})$$

.

$$= x_{10} R_{12} \sum_{k=0}^{\infty} R_{11}^k \sum_{k=0}^{\infty} \rho^k + x_{11} \rho \sum_{k=0}^{\infty} \rho^k \quad ; \; k = n-2$$

$$= (1-\rho)^{-1} (x_{10} R_{12} (1 - R_{11})^{-1} + x_{11} \rho),$$

so that

$$\frac{1}{P_b} = (1-\rho)(x_{10} R_{12} (1 - R_{11})^{-1} + x_{11} \rho)^{-1} = (1-\rho)\delta,$$

where

$$\delta = (x_{10} R_{12} (1 - R_{11})^{-1} + x_{11} \rho)^{-1}.$$

$Q_v(z)$, the generating function of the queue length subject to the condition that both servers are busy, is given by

$$Q_v(z) = \frac{1}{P_b} \sum_{n=2}^{\infty} x_{n1} z^{n-2} = \frac{1}{P_b} \sum_{n=2}^{\infty} (x_{10} R_{12} \sum_{n=2}^{\infty} R_{11}^j \rho^{n-j-2} + x_{11} \rho^{n-1}) z^{n-2}.$$

By following a computational procedure similar to that of P_b, we arrive at

$$Q_v(z) = \frac{1-\rho}{1-\rho z} \left\{ \delta \left(\frac{\mathrm{x}_{10} \mathrm{R}_{12}}{1 - R_{11} z} + \mathrm{x}_{11} \rho \right) \right\}$$

$$= Q_0(z) Q_d(z),$$

where

$$Q_0(z) = \frac{1-\rho}{1-\rho z} \tag{2.26}$$

and

$$Q_d(z) = \delta\left(\frac{\mathrm{x}_{10}\mathrm{R}_{12}}{1-R_{11}z} + \mathrm{x}_{11}\rho\right). \tag{2.27}$$

From (2.26) it follows that $Q_0(z)$ is the generating function of an M/M/2 heterogeneous queueing model without vacations, which is precisely the case $\beta = 1$ in [57]. Equation (2.27) suggests that Q_d has a geometric distribution with parameter $1 - R_{11}$. □

Remark: Due to the algorithmic approach used in the derivation of stationary waiting time distribution, a similar decomposition result for the waiting time is far from reality.

2.2.6 Key system performance measures

In this section we list a number of key system performance measures along with their formulae in addition to the busy period structure and the mean waiting time discussed above.

1. Probability that the system is empty: $P_{EMP} = x_0$.

2. Probability that the server 1 is idle: $P_{IDL} = x_0 + x_{11}$.

3. Probability that the server 2 is on vacation:
   $P_{VAC} = x_0 + \sum_{i=1}^{\infty} x_{i0} = x_0 + \frac{x_{10}}{(1-R_{11})}$.

4. Mean number of customers in the system: $\mu_{NS} = \boldsymbol{x}_1(I-R)^{-2}\boldsymbol{e}$.

5. Mean number of customers in the system when server 2 is on vacation:

   $\mu_{NSV} = \frac{x_{10}}{(1-R_{11})^2}$.

2.3 Numerical Results

ILLUSTRATIVE EXAMPLE 2.1: We analyze the effect of the parameters λ and η on the key performance measures. To this end we use the following abbreviations in addition to the notations used in section 2.2.6.

μ_{WTQ}: Mean waiting time in the queue.

μ_{LBP}: Mean length of a busy period.

μ_{LBC}: Mean length of a busy cycle.

μ_{NSBP}: Mean number of service completions in a busy period.

Table 2.1

case A : $\mu_1 = 10$, $\mu_2 = 5$, and $\eta = 1$.

case B : $\mu_1 = 5$, $\mu_2 = 10$, and $\eta = 1$.

λ	A/B	P_{EMP}	P_{VAC}	P_{IDL}	μ_{NS}	μ_{NSV}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
	A	0.799	0.989	0.806	0.248	0.232	0.027	0.126	0.626	1.252
2	B	0.637	0.976	0.648	0.536	0.432	0.096	0.285	0.785	1.569
	A	0.601	0.950	0.625	0.629	0.535	0.064	0.166	0.416	1.665
4	B	0.370	0.900	0.399	1.455	1.186	0.194	0.426	0.676	2.704
	A	0.418	0.873	0.463	1.224	0.911	0.107	0.232	0.399	2.394
6	B	0.206	0.777	0.246	2.879	1.971	0.268	0.644	0.811	4.863
	A	0.264	0.750	0.325	2.170	1.315	0.156	0.349	0.474	3.795
8	B	0.114	0.623	0.155	4.802	2.541	0.326	0.968	1.093	8.746
	A	0.148	0.579	0.211	3.720	1.595	0.222	0.577	0.677	6.765
10	B	0.062	0.452	0.095	7.294	2.630	0.393	1.509	1.608	16.087
	A	0.069	0.366	0.117	6.634	1.485	0.355	1.119	1.202	14.429
12	B	0.030	0.274	0.052	11.028	2.098	0.529	2.707	2.791	33.489
	A	0.018	0.127	0.037	17.717	0.707	1.016	3.815	3.886	54.408
14	B	0.008	0.092	0.016	22.818	0.878	1.193	8.564	8.636	120.901

Table 2.2

Case A : $\lambda = 12$, $\mu_1 = 10$ and $\mu_2 = 5$.

Case B : $\lambda = 12$, $\mu_1 = 5$ and $\mu_2 = 10$.

η	A/B	P_{EMP}	P_{VAC}	P_{IDL}	μ_{NS}	μ_{NSV}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
0.1	A	0.021	0.538	0.031	25.540	12.904	1.380	3.823	3.906	46.871
	B	0.004	0.297	0.007	73.887	20.965	1.674	20.232	20.315	243.782
0.2	A	0.034	0.498	0.051	15.232	6.704	0.787	2.373	2.457	29.481
	B	0.008	0.293	0.013	38.908	10.470	1.052	10.506	10.590	127.075
0.3	A	0.043	0.468	0.066	11.716	4.583	0.597	1.872	1.955	23.463
	B	0.011	0.290	0.019	27.261	6.974	0.842	7.262	7.346	88.148
0.4	A	0.049	0.446	0.077	9.928	3.501	0.506	1.613	1.696	20.537
	B	0.015	0.288	0.025	21.445	5.228	0.735	5.639	5.722	68.667
0.5	A	0.054	0.427	0.087	8.842	2.841	0.453	1.454	1.537	18.444
	B	0.018	0.285	0.030	17.961	4.182	0.669	4.664	4.747	56.967
0.6	A	0.058	0.411	0.094	8.112	2.395	0.419	1.345	1.428	17.139
	B	0.020	0.283	0.035	15.643	3.486	0.624	4.013	4.096	49.157
0.7	A	0.062	0.398	0.101	7.586	2.073	0.396	1.266	1.349	16.189
	B	0.023	0.280	0.039	13.991	2.989	0.591	3.548	3.631	43.571
0.8	A	0.065	0.386	0.107	7.191	1.830	0.379	1.205	1.289	15.464
	B	0.025	0.278	0.044	12.754	2.617	0.566	3.198	3.281	39.376

- Referring to table 2.1, an increase in λ naturally leads to a decrease in P_{EMP}, P_{VAC} and P_{IDL}. As λ increases traffic intensity ρ increases. Consequently μ_{NS}, μ_{WTQ}, μ_{LBP} and μ_{NSBP} also increase with λ. But due to the decrease in P_{VAC}, μ_{LBC} initially shows a downward trend and reaches a minimum. However, as the increase in μ_{LBP} becomes more dominant, the value of μ_{LBC} starts to increase. Both λ and P_{VAC} affect μ_{NSV}. As λ Increases the number of customers accumulated in the system rises. But the increase in λ lowers P_{VAC}, which in turn lowers μ_{NSV}. So the dominant of these two decides the direction of the change of μ_{NSV}. This is the reason for the pattern of behavior of μ_{NSV}. It is worth comparing the results of the tables corresponding to the sets A and B of the input parameters. Even though the net service rate $\mu_1 + \mu_2 = 15$ in both cases the effect of the vacation parameter

η becomes more predominant when $\mu_1 < \mu_2$. Due to this the measures P_{EMP}, P_{VAC} and P_{IDL} take smaller values and the measures μ_{NS}, μ_{WTQ}, μ_{LBP}, μ_{LBC} and μ_{NSBP} take larger values in case B, compared to their values in case A.

- Next let us analyze the results of table 2.2. When η is small, the mean duration of vacation $1/\eta$ is large. Hence it is natural to expect P_{EMP} and P_{IDL} to be small and P_{VAC} to be large. The effect of vacation parameter yields large values for μ_{NS}, μ_{NSV}, μ_{WTQ}, μ_{LBP}, μ_{LBC} and μ_{NSBP}. But as η increases the mean duration of vacation decreases. Consequently P_{EMP} and P_{IDL} increase and P_{VAC} decreases. μ_{NS}, μ_{NSV}, μ_{WTQ}, μ_{LBP}, μ_{LBC} and μ_{NSBP} decrease as η increases. The argument given in the previous paragraph holds here also for the difference in magnitude of the measures for the cases A and B.

3. AN $M/M/2$ QUEUEING SYSTEM WITH HETEROGENEOUS SERVERS INCLUDING ONE WITH WORKING VACATION

In this chapter we modify the model discussed in chapter 2 by replacing pure vacation of server 2 by a working vacation. This will ensure a better utilization of the servers by the system, there by reducing the waiting time of the customers in the system. A comparison of the two models (chapter 2 and chapter 3) is provided towards the end of this chapter.

Model discussed here also have two heterogeneous servers but the vacationing server returns to serve at a lower rate when an arrival finds the other server busy. To be precise, we consider an M/M/2 queueing model with heterogeneous servers, server 1 and server 2. Server 1 is always available whereas server 2 goes on vacation whenever there is no customer waiting for service. Let the service rates of servers 1 and 2 be μ_1 and μ_2 respectively, where $\mu_1 \neq \mu_2$. Customers arrive to the system according to a Poisson process of parameter λ. The duration of vacation is exponentially distributed with parameter η. At the end of a vacation, service commences if there is a customer waiting for service. Otherwise the server goes on another vacation. During vacation if an arrival finds server 1 busy, server 2 returns to serve the

[0] To appear in International Journal of Stochastic Analysis

customer but at a lower rate. To be precise, the server 2 serves this customer at the rate $\theta\mu_2$, $0 < \theta < 1$. As this vacation gets over, server 2 instantaneously switches over to the normal service rate μ_2 if there is at least one customer waiting for service. Upon completion of a service at lower rate, the server will (a) continue the current vacation if it is not finished and no customer is waiting for service; (b) continue the slow service if the vacation has not expired and if there is at least one customer waiting for service. For clarity we assume that if an arriving customer finds a free server he enters service immediately. Else he joins the queue.

3.1 The QBD process

The model discussed above can be studied as a level independent quasi-birth-and-death ($LIQBD$) process. First, we set up the necessary notations.

At time t, let $N(t)$ be the number of customers in the system and

$$J(t) = \begin{cases} 0, \textit{if the server 2 is on vacation }, \\ 1, \textit{if the server 2 is working in vacation mode}, \\ 2, \textit{if the server 2 is working in normal mode}, \end{cases}$$

Let $X(t) = (N(t), J(t))$. Then $(X(t) : t \geq 0)$ is a continuous time Markov Chain (CTMC) with states space

$$\Omega = \{(0,0), (1,0), (1,1), (1,2)\} \bigcup \bigcup_{i=2}^{\infty} l(i)$$

where

$$l(i) = \{(i, j) : i \geq 2, j = 1 \text{ or } 2\}.$$

The infinitesimal generator matrix Q of this Markov chain is given by

$$Q = \begin{bmatrix} B_{00} & B_{01} & & & & \\ B_{10} & B_{11} & B_{12} & & & \\ & B_{21} & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix},$$

where the block matrices appearing in Q are as follows.

$$B_{00} = -\lambda,\ B_{01} = \begin{bmatrix} \lambda & 0 & 0 \end{bmatrix},$$

$$B_{10} = \begin{bmatrix} \mu_1 \\ \theta\mu_2 \\ \mu_2 \end{bmatrix},\ B_{11} = \begin{bmatrix} -\lambda-\mu_1 & 0 & 0 \\ 0 & -\lambda-\theta\mu_2-\eta & \eta \\ 0 & 0 & -\lambda-\mu_2 \end{bmatrix},$$

$$B_{12} = \begin{bmatrix} \lambda & 0 \\ \lambda & 0 \\ 0 & \lambda \end{bmatrix},\ B_{21} = \begin{bmatrix} \theta\mu_2 & \mu_1 & 0 \\ \mu_2 & 0 & \mu_1 \end{bmatrix},\ A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -\lambda-\mu_1-\theta\mu_2-\eta & \eta \\ 0 & -\lambda-\mu_1-\mu_2 \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} \mu_1+\theta\mu_2 & 0 \\ 0 & \mu_1+\mu_2 \end{bmatrix}.$$

3.2 Steady-state analysis

In this section we discuss the steady-state analysis of the model under study.

3.2.1 Stability Condition

Theorem 3.2.1. *The queueing system described above is stable if and only if $\rho < 1$ where $\rho = \lambda/(\mu_1 + \mu_2)$.*

Proof. To establish the stability condition we use Pakes' lemma (see [58]). Let N_i be the number of customers in the system immediately after the departure of the i^{th} customer. Then $\{N_i : i \in \mathbb{N}\}$ satisfies the equation

$$N_i = \begin{cases} N_{i-1} - 1 + V_i & if\ N_{i-1} \geq 1 \\ V_i & if\ N_{i-1} = 0 \end{cases}$$

where V_i is the number of arrivals during the service of i^{th} customer. Clearly $\{N_i : i \in \mathbb{N}\}$ is an irreducible aperiodic Markov chain. Pakes' lemma asserts that an aperiodic irreducible Markov chain is ergodic, if there exists an $\epsilon > 0$ such that the mean drift

$$\phi_j = E[(N_{i+1} - N_i)/N_i = j]$$

is finite for all $j \in \mathbb{N}$ and $\phi_j \leq -\epsilon$ for all $j \in \mathbb{N}$ except perhaps for a finite number. In the present model, value of the mean drift is

$$\phi_j = \begin{cases} -1 + \rho & if\ j \geq 1 \\ \rho & if\ j = 0 \end{cases}$$

Thus if $\rho < 1$ the Markov chain $\{N_i : i \in \mathbb{N}\}$ is ergodic and hence the condition is sufficient.

To prove the necessity of the condition assume that $\rho \geq 1$. We use theorem 1 in Sennot et al. [56], which states that $\{N_i : i \in \mathbb{N}\}$ is nonergodic if it satisfies Kaplan's condition, $\phi_j < \infty$, for $j \geq 0$ and there is a j_0 such that $\phi_j \geq 0$, for $j \geq j_0$. When $\rho \geq 1$ Kaplan's condition is readily satisfied. Hence the Markov chain is not ergodic. □

3.2.2 Steady-state probability vector

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (x_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$, be the steady-state probability vector of Q. Note that x_0 is a scalar, $\boldsymbol{x}_1 = (x_{10}, x_{11}, x_{12})$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ for $i \geq 2$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = \mathbf{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$. Apparently when the stability condition is satisfied the sub vectors of $\boldsymbol{x}$, corresponding to the different levels are given by the equation $\boldsymbol{x}_j = \boldsymbol{x}_2 R^{j-2}, j \geq 3$, where R is the minimal non negative solution of the matrix quadratic equation (see [50])

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{3.1}$$

Knowing the matrix R, x_0 , $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are obtained by solving the equations

$$x_0 B_{00} + \boldsymbol{x}_1 B_{10} = 0, \tag{3.2}$$

$$x_0 B_{01} + \boldsymbol{x}_1 B_{11} + \boldsymbol{x}_2 B_{21} = 0 \tag{3.3}$$

and

$$\boldsymbol{x}_1 B_{12} + \boldsymbol{x}_2(A_1 + RA_2) = 0 \tag{3.4}$$

subject to the normalizing condition

$$x_0 + \boldsymbol{x}_1 \boldsymbol{e} + \boldsymbol{x}_2 (I - R)^{-1} \boldsymbol{e} = 1. \tag{3.5}$$

Theorem 3.2.2. *The matrix R of eqution (3.1) is given by*

$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$, *where* $R_{11} = \frac{(\lambda+\mu_1+\theta\mu_2+\eta-\sqrt{(\lambda+\mu_1+\theta\mu_2+\eta)^2-4\lambda(\mu_1+\theta\mu_2)})}{2(\mu_1+\theta\mu_2)}$,

$R_{12} = \rho - \frac{(\mu_1+\theta\mu_2)R_{11}}{(\mu_1+\mu_2)}$ *and* $R_{22} = \rho$.

Proof. Since A_0, A_1 and A_2 are upper triangular, R is essentially an upper triangular matrix. The value of R_{11} follows from the assertion that R is the minimal non negative solution of (3.1). The rest of the proof is an easy consequence of the condition $RA_2\boldsymbol{e} = A_0\boldsymbol{e}$. □

Though the matrix R has a nice structure it may not be easy to carry out the computations required in the forthcoming discussions. Hence we explore the possibility of algorithmic computation of R. The computation of R matrix can be carried out using a number of well known methods such as logarithmic reduction algorithm.

3.2.3 Busy period analysis

For the system under study, busy period is the interval between arrival of a customer to the empty system and the first epoch thereafter when the system becomes empty again. Thus it is precisely the first passage time from the state $(1,0)$ to the state $(0,0)$. For the working vacation model, busy cycle for the system is the time interval between two successive departures, which leave the system empty. Thus the busy cycle is the first return time to state $(0,0)$ with at least one visit to any other state. Before analyzing the busy period structure, we need to introduce the notion of fundamental period. For the QBD process under consideration, it is the first passage time from level i, where $i \geq 3$, to the level $i-1$. The cases $i=2$, $i=1$ and $i=0$ corresponding to the boundary states need to be discussed separately. It should be noted that due to the structure of the QBD process the distribution of the first passage time is invariant in i away from the boundary states.

Let $G_{jj'}(k,x)$ denote the conditional probability that a QBD process, starting in the state (i,j) at time $t=0$ reaches the level $i-1$ for the first time no later than time x, after exactly k transitions to the left, and does so by entering the state $(i-1,j')$. For convenience we introduce the joint transform

$$\tilde{G}_{jj'}(z,s) = \sum_{k=1}^{\infty} z^k \int_0^{\infty} e^{-sx} \mathrm{d}G_{jj'}(k,x) \qquad ; \quad |z| \leq 1, Re(s) \geq 0$$

and the matrix

$$\tilde{G}(z,s) = (\tilde{G}_{jj'}(z,s)).$$

The matrix $\tilde{G}(z,s)$ is the unique solution to the equation (see [50])

$$\tilde{G}(z,s) = z(sI - A_1)^{-1}A_2 + (sI - A_1)^{-1}A_0\tilde{G}^2(z,s). \tag{3.6}$$

The matrix $G = \tilde{G}(1,0)$ takes care of the first passage times, except for the boundary states. If we know the R matrix then G matrix can be computed using the result (see [39])

$$G = -(A_1 + RA_2)^{-1}A_2.$$

Otherwise we may use logarithmic reduction method to comput G. For the boundary level states 2, 1 and 0 let $G_{jj'}^{(2,1)}(k,x)$, $G_{jj'}^{(1,0)}(k,x)$ and $G_{jj'}^{(0,0)}(k,x)$ be the conditional probability discussed above for the first passage times from level 2 to level 1, level 1 to level 0 and the first return time to the level 0 respectively. Then as in (3.6) we get

$$\tilde{G}^{(2,1)}(z,s) = z(sI - A_1)^{-1}B_{21} + (sI - A_1)^{-1}A_0\tilde{G}(z,s)\tilde{G}^{(2,1)}(z,s), \tag{3.7}$$

$$\tilde{G}^{(1,0)}(z,s) = z(sI - B_{11})^{-1}B_{10} + (sI - B_{11})^{-1}B_{12}\tilde{G}^{(2,1)}(z,s)\tilde{G}^{(1,0)}(z,s) \tag{3.8}$$

and

$$\tilde{G}^{(0,0)}(z,s) = [\lambda/(s+\lambda), \quad 0, \quad 0]\tilde{G}^{(1,0)}(z,s). \tag{3.9}$$

Note that $\tilde{G}^{(1,0)}(z,s)$ is a 3×1 matrix. Thus the Laplace Stieltjes transform (LST) of the busy period is the first element of $\tilde{G}^{(1,0)}(1,s)$. For convenience, we use the notations

$$G_{21} = \tilde{G}^{(2,1)}(1,0), G_{10} = \tilde{G}^{(1,0)}(1,0) \text{ and } G_{00} = \tilde{G}^{(0,0)}(1,0).$$

Due to the positive recurrence of the QBD process, matrices G, G_{21}, G_{10} and G_{00} are all stochastic. If we let

$$C_0 = (-A_1)^{-1}A_2 \text{ and } C_2 = (-A_1)^{-1}A_0,$$

then G is the minimal non negative solution (see [50]) to the matrix equation

$$G = C_0 + C_2G^2.$$

From equations (3.7), (3.8) and (3.9), we get

$$G_{21} = -(A_1 + A_0G)^{-1}B_{21}, \tag{3.10}$$

$$G_{10} = -(B_{11} + B_{12}G_{21})^{-1}B_{10} \tag{3.11}$$

and

$$G_{00} = [1, \quad 0, \quad 0]G_{10} \tag{3.12}$$

respectively. Equation(3.6) is equivalent to

$$zA_2 - (sI - A_1)\tilde{G}(z,s) + A_0\tilde{G}^2(z,s) = 0. \tag{3.13}$$

Let

$$M = -\left.\frac{\partial \tilde{G}(z,s)}{\partial s}\right|_{z=1,s=0}$$

and

$$\tilde{M} = \left.\frac{\partial \tilde{G}(z,s)}{\partial z}\right|_{z=1,s=0}$$

Differentiation of (3.13) with respect to s and z followed by setting $z = 1$ and $s = 0$ leads to (see [50])

$$M = -A_1^{-1}G + C_2(GM + MG)$$

and

$$\tilde{M} = C_0 + C_2(G\tilde{M} + \tilde{M}G)$$

With **0** as starting value for M and $\tilde{M}$,successive substitutions in the above equations yields the values of M and $\tilde{M}$. Applying an exactly similar reasoning to (3.7), (3.8) and (3.9), we get

$$M_{21} = -(A_1 + A_0G)^{-1}(I + A_0M)G_{21}, \tag{3.14}$$

$$M_{10} = -(B_{11} + B_{12}G_{21})^{-1}(I + B_{12}M_{21})G_{10} \tag{3.15}$$

and

$$M_{00} = [1/\lambda, \quad 0, \quad 0]G_{10} + [1, \quad 0]M_{10} \tag{3.16}$$

where

$$M_{21} = -\left.\frac{\partial \tilde{G}^{(2,1)}(z,s)}{\partial s}\right|_{z=1,s=0},$$

$$M_{10} = -\left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial s}\right|_{z=1,s=0}$$

and

$$M_{00} = -\left.\frac{\partial \tilde{G}^{(0,0)}(z,s)}{\partial s}\right|_{z=1,s=0}.$$

Note that M_{10} is a 3×1 matrix and M_{00} is a scalar. The first element of the matrix M_{10} and M_{00} are mean lengths of a busy period and a busy cycle respectively. The second and third elements of the matrix M_{10} are the first passage times to the state (0,0) from (1,1) and (1,2) respectively. With the notations

$$\tilde{M}_{21} = \left.\frac{\partial \tilde{G}^{(2,1)}(z,s)}{\partial z}\right|_{z=1,s=0}$$

and

$$\tilde{M}_{10} = \left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial z}\right|_{z=1,s=0}$$

It follows from equations (3.7) and (3.8) that

$$\tilde{M}_{21} = -(A_1 + A_0 G)^{-1}(B_{21} + A_0 M G_{21}) \tag{3.17}$$

and

$$\tilde{M}_{10} = -(B_{11} + B_{12}G_{21})^{-1}(B_{10} + B_{12}M_{21}G_{10}). \tag{3.18}$$

The first component of the vector $\tilde{M}_{10}$ is the mean number of service completions in a busy period.

3.2.4 Stationary waiting time in the queue

Let $W(t)$ be the distribution function of the waiting time in the queue of an arriving (tagged) customer. Note that if there is no customer in the system, the arrival receives service immediately. If either of the two servers is not busy then also there would be no delay in getting service. Thus the probability that the customer gets service without waiting is $x_0 + x_{10} + x_{11} + x_{12}$. Hence, with probability $1 - x_0 - x_{10} - x_{11} - x_{12}$, the customer has to wait before getting service. The waiting time may be viewed as the time until absorption in a Markov chain with state space

$$\Omega_1 = \{*\} \bigcup \{\mathbf{2}, \mathbf{3}, ...\}$$

Here * is the absorbing state, which corresponds to taking the tagged customer into service and is obtained by lumping together the level states $\mathbf{0} = \{(0,0)\}$ and $\mathbf{1} = \{(1,0),(1,1),(1,2)\}$. For $i \geq 2$, the level $\mathbf{i}$ is given by $\mathbf{i} = \{(i,j), j = 1 \text{ or } 2\}$. The states other than the absorbing state correspond to the number of customers present in the system as the tagged customer arrives. Once the tagged customer joins the queue, the subsequent arrivals will not affect his waiting time in the queue. Hence the parameter λ does not show up in the generator matrix $\tilde{Q}$ of this Markov process, given by

$$\tilde{Q} = \begin{array}{c} \\ * \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} * & 2 & 3 & \cdots \end{array} \\ \left(\begin{array}{cccc} & & & \\ A_2\boldsymbol{e} & D & & \\ & A_2 & D & \\ & & \ddots & \ddots \end{array} \right) \end{array}, \text{ where } D = \begin{bmatrix} -\mu_1 - \theta\mu_2 - \eta & \eta \\ 0 & -\mu_1 - \mu_2 \end{bmatrix}.$$

Define

$$\boldsymbol{y}(t) = (y_*(t), \boldsymbol{y}_2(t), \boldsymbol{y}_3(t), \ldots),$$

where

$$\boldsymbol{y}_i(t) = (y_{i1}(t), y_{i2}(t)), \text{ for } i \geq 2.$$

The components of $\boldsymbol{y}_i(t)$ are the probabilities that at time t, the CTMC with generator $\tilde{Q}$ is in the respective states of level i. Note that $y_*(t)$ is the probability that the process is in the absorbing state at time t. By the PASTA property we may write

$$\boldsymbol{y}(0) = (x_0 + x_{11} + x_{10} + x_{12}, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots).$$

Clearly

$$W(t) = y_*(t), \text{for } t \geq 0. \tag{3.19}$$

The LST of $y_*(t)$ is given by (see [50])

$$\tilde{W}(s) = \sum_{i=2}^{\infty} \boldsymbol{y}_i(0)[(sI - D)^{-1}A_2]^{i-2}(sI - D)^{-1}A_2\boldsymbol{e}. \tag{3.20}$$

The mean waiting time can be obtained from *W(s)* as

$$E(W) = -\tilde{W}'(0) = \sum_{i=1}^{\infty} \boldsymbol{x}_{2+i} \sum_{j=0}^{i-1} U^j(-D)^{-1}U^{i-j}U\boldsymbol{e} + \sum_{i=0}^{\infty} \boldsymbol{x}_{2+i}U^i(-D)^{-2}A_2\boldsymbol{e}. \tag{3.21}$$

where $U = (-D)^{-1}A_2$ is a stochastic matrix. Hence the expression for $E(W)$ given by (3.21) can be simplified as

$$E(W) = -\tilde{W}'(0) = \sum_{i=1}^{\infty} \boldsymbol{x}_{2+i} \sum_{j=0}^{i-1} U^j(-D)^{-1}\boldsymbol{e} + \sum_{i=0}^{\infty} \boldsymbol{x}_{2+i}U^i(-D)^{-1}\boldsymbol{e} \tag{3.22}$$

Let

$$H = \sum_{i=0}^{\infty} \boldsymbol{x}_{2+i}U^i$$

Since U is stochastic, we get

$$H\boldsymbol{e} = \boldsymbol{x}_2(I - R)^{-1}\boldsymbol{e} = 1 - x_0 - x_{10} - x_{11} - x_{12}.$$

This result can be used to find an approximate value of H and hence that of the second term in the expression for $E(W)$, given by equation (3.22) to any desired degree of accuracy. Thus only the first term in equation (3.22) demands serious computation. For this we make use of the ideas in [47], [49] and [33]. Now consider the matrix

$$U_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

which has the property that

$$UU_2 = U_2U = U_2.$$

Then we get

$$\sum_{j=0}^{i-1} U^j(I - U + U_2) = I - U^i + iU_2, \text{ for } i \geq 1.$$

By the classical theorem on finite Markov chains, the matrix $(I - U + U_2)$ is nonsingular (see [31]). In view of the last equation, the first term in equation (3.22) becomes $[\sum_{i=1}^{\infty} \boldsymbol{x}_{2+i}(I - U^i + iU_2)](I - U + U_2)^{-1}(-D)^{-1}\boldsymbol{e}$.

With this simplification, we get

$$E(W) = [\boldsymbol{x}_2(R(I-R)^{-1} + I + R(I-R)^{-2}U_2) - H](I-U+U_2)^{-1}(-D)^{-1}\boldsymbol{e}+$$

$$H(-D)^{-1}\boldsymbol{e}$$

3.2.5 Conditional stochastic decomposition of queue length

In this section we provide a stochastic decomposition of queue length in the stationary regime, subject to the condition that both servers are busy. Note that from equations (3.2)-(3.5) we get x_0, x_{10}, x_{11}, x_{12}, x_{21} and x_{22}. Let Q_v be the queue length of the vacation model under study, subject to the condition that both servers are busy. Then we have

Theorem 3.2.3. *If $\rho < 1$, then $Q_v = Q_0 + Q_d$, where Q_0 and Q_d are two independent random variables. Q_0 is the queue length of the M/M/2 queueing model with heterogeneous servers without vacation and Q_d can be interpreted as the additional queue length due to vacation and consequent slow service, subject to the condition that both servers are busy.*

Proof. Let P_b denote the Probability that both servers are are busy. Then

$$P_b = \sum_{n=2}^{\infty} x_{n2} = \sum_{n=2}^{\infty} x_{22}\rho^{n-2} + \sum_{n=3}^{\infty} x_{21}R_{12}\sum_{j=0}^{n-3} R_{11}^{j}\rho^{n-j-3}$$

.

$$= x_{22}\rho\sum_{k=0}^{\infty}\rho^k + x_{21}R_{12}\sum_{k=0}^{\infty}R_{11}^{k}\sum_{k=0}^{\infty}\rho^k \quad ; \; k = n-3$$

$$= (1-\rho)^{-1}(x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1})$$

so that

$$\frac{1}{P_b} = (1-\rho)(x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1})^{-1} = (1-\rho)\delta,$$

where

$$\delta = (x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1})^{-1}.$$

$Q_v(z)$, the generating function of the queue length subject to the condition that both servers are busy, is given by

$$Q_v(z) = \frac{1}{P_b}\sum_{n=2}^{\infty} x_{n1}z^{n-2} = \frac{1}{P_b}\sum_{n=2}^{\infty} x_{22}\rho^{n-2}z^{n-2} + \frac{1}{P_b}\sum_{n=3}^{\infty}(x_{21}R_{12}\sum_{j=0}^{n-3} R_{11}^{j}\rho^{n-j-3})z^{n-3}.$$

By following a computational procedure similar to that of P_b, we arrive at

$$Q_v(z) = \frac{1-\rho}{1-\rho z}\left\{\delta\left(\mathrm{x}_{22}\rho z + \frac{\mathrm{x}_{21}\mathrm{R}_{12}}{1-R_{11}z}\right)\right\}$$

$$= Q_0(z)Q_d(z)$$

where

$$Q_0(z) = \frac{1-\rho}{1-\rho z} \tag{3.23}$$

and

$$Q_d(z) = \delta\left(\mathrm{x}_{22}\rho z + \frac{\mathrm{x}_{21}\mathrm{R}_{12}}{1-R_{11}z}\right). \tag{3.24}$$

From (3.23) it follows that $Q_0(z)$ is the generating function of an M/M/2 heterogeneous queuing model without vacations, which is precisely the case $\beta = 1$ in [57]. Relation (3.24) suggests that Q_d has a geometric distribution with parameter $1 - R_{11}$. □

Remark: Due to the algorithmic approach used in the derivation of stationary waiting time distribution, a similar decomposition result for the waiting time is far from reality.

3.2.6 Key system performance measures

In this section we list a number of key system performance measures along with their formulae in addition to the busy period structure and the mean waiting time discussed above.

1. Probability that the system is empty: $P_{EMP} = x_0$.
2. Probability that the server 1 is idle: $P_{IDL} = x_0 + x_{11} + x_{12}$.
3. Probability that the server 2 is on vacation: $P_{VAC} = x_0 + x_{10}$.
4. Probability that the server 2 is working in vacation mode:

   $P_{SLOW} = \sum_{j=1}^{\infty} x_{j1} = x_{11} + \frac{x_{21}}{(1-R_{11})}$.
5. Probability that the server 2 is working in normal mode:

   $P_{NORM} = 1 - x_0 - P_{SLOW}$
6. Mean number of customers in the system:

   $\mu_{NS} = \sum_{j=1}^{\infty} j\boldsymbol{x}_j\boldsymbol{e} = x_{10} + x_{11} + x_{12} + \boldsymbol{x}_2(I-R)^{-2}R^{-1}\boldsymbol{e} - \boldsymbol{x}_2R^{-1}\boldsymbol{e}$

3.3 Numerical Results

ILLUSTRATIVE EXAMPLE 3.1: We analyze the effect of the parameters λ, η and θ on the key performance measures in tables 3.1, 3.2 and 3.3 respectively. To this end we use the following abbreviations.

μ_{WTQ} : Mean waiting time in the queue.

μ_{LBP} : Mean length of a busy period.

μ_{LBC} : Mean length of a busy cycle.

μ_{NSBP} : Mean number of service completions in a busy period.

Table 3.1

case A : $\mu_1 = 10$, $\mu_2 = 5$, $\eta = 1$ and $\theta = 0.6$.

case B : $\mu_1 = 5$, $\mu_2 = 10$, $\eta = 1$ and $\theta = 0.6$.

λ	A/B	P_{VAC}	P_{IDL}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
	A	0.913	0.829	0.065	0.022	1.676	0.002	0.080	0.496	1.167
2	B	0.916	0.707	0.070	0.014	2.176	0.005	0.125	0.482	1.17
	A	0.745	0.687	0.168	0.087	1.934	0.009	0.067	0.246	1.128
4	B	0.755	0.520	0.188	0.057	2.301	0.016	0.097	0.236	1.156
	A	0.568	0.550	0.246	0.186	2.189	0.017	0.062	0.166	1.121
6	B	0.574	0.374	0.296	0.130	2.60	0.031	0.088	0.164	1.203
	A	0.405	0.414	0.285	0.311	2.547	0.027	0.062	0.132	1.172
8	B	0.399	0.251	0.363	0.238	3.178	0.047	0.094	0.143	1.362
	A	0.262	0.282	0.278	0.460	3.193	0.036	0.072	0.122	1.340
10	B	0.245	0.150	0.365	0.390	4.282	0.062	0.120	0.153	1.755
	A	0.141	0.158	0.218	0.641	4.702	0.043	0.106	0.144	1.849
12	B	0.123	0.074	0.282	0.595	6.626	0.069	0.191	0.216	2.818
	A	0.042	0.049	0.092	0.866	11.885	0.047	0.294	0.324	4.659
14	B	0.034	0.020	0.113	0.853	16.043	0.064	0.566	0.585	8.418

Table 3.2

Case A : $\lambda = 12$, $\mu_1 = 10$, $\mu_2 = 5$ and $\theta = 0.6$.

Case B : $\lambda = 12$, $\mu_1 = 5$, $\mu_2 = 10$ and $\theta = 0.6$.

η	A/B	P_{VAC}	P_{IDL}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
	A	0.092	0.112	0.333	0.575	5.575	0.059	0.137	0.174	2.20
0.1	B	0.044	0.028	0.451	0.506	14.857	0.209	0.536	0.56	6.952
	A	0.104	0.125	0.306	0.589	5.314	0.054	0127	0.165	2.086
0.2	B	0.064	0.04	0.408	0.528	10.582	0.137	0.365	0.39	4.907
	A	0.113	0.133	0.287	0.600	5.169	0.051	0.121	0.159	2.019
0.3	B	0.078	0.048	0.379	0.543	9.044	0.111	0.301	0.326	4.137
	A	0.119	0.140	0.272	0.608	5.066	0.049	0.117	0.155	1.974
0.4	B	0.088	0.054	0.357	0.555	8.232	0.097	0.266	0.290	3.714
	A	0.125	0.144	0.26	0.615	4.983	0.047	0.114	0.152	1.941
0.5	B	0.097	0.059	0.340	0.564	7.724	0.088	0.243	0.268	3.442
	A	0.129	0.148	0.250	0.622	4.914	0.046	0.112	0.150	1.915
0.6	B	0.103	0.063	0.325	0.572	7.374	0.082	0.227	0.252	3.248

Table 3.3

Case A : $\lambda = 12$, $\mu_1 = 10$, $\mu_2 = 5$ and $\eta = 1$.

Case B : $\lambda = 12$, $\mu_1 = 5$, $\mu_2 = 10$ and $\eta = 1$.

θ	A/B	P_{VAC}	P_{IDL}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
0.1	A	0.073	0.112	0.167	0.761	4.402	0.044	0.158	0.195	2.439
	B	0.042	0.034	0.188	0.770	8.45	0.105	0.507	0.531	6.601
0.2	A	0.085	0.121	0.178	0.737	4.504	0.044	0.144	0.182	2.286
	B	0.053	0.039	0.207	0.740	8.126	0.098	0.420	0.444	5.556
0.3	A	0.098	0.13	0.190	0.713	4.581	0.044	0.133	0.171	2.152
	B	0.066	0.046	0.226	0.708	7.762	0.091	0.345	0.370	4.664
0.4	A	0.112	0.139	0.200	0.689	4.638	0.044	0.123	0.161	2.036
	B	0.082	0.054	0.246	0.673	7.379	0.084	0.283	0.308	3.918
0.5	A	0.126	0.149	0.209	0.664	4.678	0.044	0.114	0.152	1.936
	B	0.101	0.063	0.265	0.635	6.995	0.076	0.232	0.257	3.307
0.6	A	0.141	0.158	0.218	0.641	4.702	0.043	0.106	0.144	1.849
	B	0.123	0.074	0.282	0.595	6.626	0.069	0.191	0.216	2.818

- Let us first examine table 1. Since μ_1 and μ_2 are fixed, the traffic intensity ρ increases with λ. Due to this P_{NORM}, μ_{NS} and μ_{WTQ} increase and P_{VAC} and P_{IDL} decrease as λ increases. Note that the busy period starts with the Markov chain in the state $(1, 0)$; i.e. with server 2 on vacation. Hence initially P_{SLOW} increases with λ. For this reason μ_{LBP}, μ_{LBC} and μ_{NSBP} show an early downward trend. But as λ further increases P_{SLOW} declines as expected due to the high traffic intensity. Hence μ_{LBP} and μ_{NSBP} reverse the direction of change. Due to the effect of P_{VAC} and P_{IDL} this reversal occurs only at a later stage for μ_{LBC}. It is worth comparing the values of the measures in cases A and B. Even though the net service rate $\mu_1 + \mu_2 = 15$ in both cases, the effect of the vacation parameter η becomes more predominant when $\mu_1 < \mu_2$. Due to this the measures P_{VAC} and P_{IDL} take smaller values and the measures μ_{NS}, μ_{WTQ}, μ_{LBP} and μ_{NSBP} take larger values in case B, compared to their values in case A.

- Next let us analyze the results shown in table 2. As η increases, the mean duration of vacation decreases. This reduces the probability P_{SLOW} of service in vacation mode. Thus chance of an early expiry of vacation always results in an increase in P_{NORM} and P_{VAC}. Note that $P_{VAC} + P_{SLOW}$ decreases as η increases and $P_{VAC} + P_{SLOW} < P_{NORM}$ for any value of η in the given range. So P_{IDL} increases with η. Thus the proportion of time in which both servers work at the normal rate increases as η increases. Hence the measures μ_{NS}, μ_{WTQ}, μ_{LBP}, μ_{LBC} and μ_{NSBP} decrease as η increases. The argument given in the previous paragraph holds here also for the difference in magnitude of the measures in cases A and B.

- Finally we consider table 3 to study the effect of the parameter θ. As θ increases, the service rate $\theta\mu_2$ of the second server in vacation mode of service, increases. As a result server 2 clears out customers at an increased rate in slow service mode. This produces an increase in P_{VAC}, P_{SLOW} and P_{IDL} and a decrease in P_{NORM} as expected. Consequently μ_{LBP}, and μ_{LBC} and μ_{NSBP} decrease as θ increases. The huge difference in the value of net service rate $\mu_1 + \theta\mu_2$ between cases A and B in vacation mode of service, is the reason for the pattern of behavior of μ_{NS} in these two cases. Increase in θ does not affect μ_{WTQ} significantly in case A, but it affects the measure in case B. This is because the effect of θ becomes significant only when μ_2 is large compared to μ_1.

3.4 Comparison of models discussed in chapters 2 and 3

In chapter 2 we discussed an $M/M/2$ queueing system with heterogeneous servers, where one server takes multiple vacations in the absence of customers waiting for service. This server would be available in the system only if there is a customer waiting for service on expiry of a vacation. But in the model discussed in chapter 3 the vacation of the server is interrupted the moment an arrival finds the other server busy. Thus under the working vacation policy, the vacationing server is made available in the system as and when there is a demand for service. As a result the waiting time of a customer is very less in the model discussed in chapter 3 compared to that in chapter 2. The numerical illustrations provided in these two chapters justify our arguments. Further it distributes the customers more evenly among the two servers and hence manages the system more efficiently.

4. $MAP/PH/1$ QUEUE WITH WORKING VACATIONS, VACATION INTERRUPTIONS AND N POLICY

In chapters 2 and 3, we considered the case of a two server system where the second server goes on a vacation, whenever no customer is found waiting at the end of a service. This server followed a simple vacation policy in the model discussed in chapter 2 and a working vacation policy in the model of chapter 3. These two queueing models were with Poisson arrivals and exponential service times. In reality these assumptions are very restrictive though they make the system analytically more tractable. The traffic in modern communication network is highly irregular. Of late to model systems with repeated calls and bursty arrivals MAP (Markovian arrival process) is used. The MAP is a tractable class of point process which is in general nonrenewal. However by choosing the parameters of the MAP appropriately the underlying arrival process can be made a renewal process. The MAP can represent a variety of processes which includes, as special cases, the Poisson process, the phase-type renewal processes, the Markov modulated Poisson process and superpositions of these.

Here we consider a single server queueing model in which customers arrive according to a Markovian arrival process with representation (D_0, D_1) of

[0] To appear in Applied Mathematical Modelling

order m. The service times are assumed to be of phase type with representation $(\boldsymbol{\alpha}, T)$ of order n. At a service completion epoch the server, finding the system empty, takes a vacation. The duration of vacation is assumed to be exponentially distributed with parameter η. A customer arriving during a vacation will be served at a lower rate. To be precise, the service time during vacation follows phase type distribution with representation $(\boldsymbol{\alpha}, \theta T)$, $0 < \theta < 1$. Thus $\mu = [\boldsymbol{\alpha}(-T)^{-1}\boldsymbol{e}]^{-1}$ is the normal service rate and $\theta\mu$ is the rate of the vacation mode of service. The server continues to serve at this rate until either the vacation clock expires or the queue length hits the threshold value N, $1 \leq N < \infty$. When either of these two occurs the server instantaneously switches over to the normal rate and continues to serve at this rate until the system becomes empty.

Let $Q^* = D_0 + D_1$ be the generator matrix of the arrival process and $\boldsymbol{\pi}$ be the stationary probability vector of the Markov process with generator Q^*. That is, $\boldsymbol{\pi}$ is the unique (positive) probability vector satisfying

$$\boldsymbol{\pi} Q^* = \mathbf{0}, \quad \boldsymbol{\pi}\boldsymbol{e} = 1. \tag{4.1}$$

The constant $\lambda = \boldsymbol{\pi} D_1 \boldsymbol{e}$, referred to as the ***fundamental rate***, gives the expected number of arrivals per unit of time in the stationary version of the MAP. Often, in model comparisons, it is convenient to select the time scale of the MAP so that 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.

4.1 The QBD process

The model described in Section 1 can be studied as a quasi-birth-and-death (QBD) process. First, we set up necessary notations.

Define $N(t)$ to be the number of customers in the system at time t,

$$S_1(t) = \begin{cases} 0, \textit{if the service is in vacation mode}, \\ 1, \textit{if the service is normal}, \end{cases}$$

$S_2(t)$ is the phase of the service process when the server is busy and $M(t)$ to be the phase of the arrival process at time t. It is easy to verify that $\{(N(t), S_1(t), S_2(t), M(t)) : t \geq 0\}$ is a level independent quasi-birth-and-death process ($LIQBD$) with state space

$$\Omega = \bigcup_{i=0}^{\infty} l(i)$$

where

$$l(0) = \{(0,1), (0,2), \ldots (0,m)\}$$

and for $i \geq 1$,

$$l(i) = \{(i, j_1, j_2, k) : j_1 = 0 \text{ or } 1; 1 \leq j_2 \leq n; 1 \leq k \leq m\}.$$

Note that when $N(t) = 0$, server will be on vacation and so $S_1(t)$ and $S_2(t)$ do not play any role and will not be tracked. The only other component in the state vector would be $M(t)$.

The generator, Q, of the QBD process under consideration is of the form

$$Q = \begin{pmatrix} D_0 & C_0 & & & & & & & \\ C_2 & B_1 & I \otimes D_1 & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & B_2 & B_1 & I \otimes D_1 & & & & \\ & & & B_2 & B_1 & \boldsymbol{e} \otimes I \otimes D_1 & & & \\ & & & & \boldsymbol{e}_2'(2) \otimes \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I & A_1 & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the (block) matrices appearing in Q are as follows.

$$C_0 = [\boldsymbol{\alpha} \otimes D_1 \ O], \quad C_2 = \begin{bmatrix} \theta \boldsymbol{T}^0 \otimes I \\ \boldsymbol{T}^0 \otimes I \end{bmatrix},$$

$$B_1 = \begin{bmatrix} \theta T \oplus D_0 - \eta I & \eta I \\ O & T \oplus D_0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} \theta \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I & O \\ O & \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I \end{bmatrix}$$

$$A_0 = I \otimes D_1, \quad A_1 = T \oplus D_0, \quad A_2 = \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I.$$

The boundary blocks B_1 and B_2 are of order $2mn \times 2mn$, C_0 and C_2 are of orders $m \times 2mn$ and $2mn \times m$ respectively. A_0, A_1 and A_2 are square matrices of order mn.

4.1.1 The steady-state probability vector

Defining $A = A_0 + A_1 + A_2$ and $\boldsymbol{\delta}$ to be the steady-state probability vector of the irreducible matrix A, it is easy to verify that the vector $\boldsymbol{\delta}$ satisfying

$$\boldsymbol{\delta} A = \boldsymbol{0}, \quad \boldsymbol{\delta e} = 1,$$

is given by

$$\boldsymbol{\delta} = (\mu\boldsymbol{\alpha}(-T)^{-1} \otimes \boldsymbol{\pi}), \tag{4.2}$$

where $\boldsymbol{\pi}$ as given in (4.1).

The condition $\boldsymbol{\delta} A_0 \boldsymbol{e} < \boldsymbol{\delta} A_2 \boldsymbol{e}$, required for the stability of the queueing model under study (see [50]) reduces to $\lambda < \mu$.

Let $\boldsymbol{x}$ be the steady-state probability vector of Q. Partitioning this vector as

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \ldots, \boldsymbol{x}_N, \boldsymbol{x}_{N+1}, \ldots),$$

where $\boldsymbol{x}_0$ is of dimension m; $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N$ are of dimension $2mn$; and $\boldsymbol{x}_{N+1}$, $\boldsymbol{x}_{N+2}, \ldots$ are of dimension mn. Under the condition that $\lambda < \mu$, the steady-state probability vector $\boldsymbol{x}$ is obtained as follows.

$$\boldsymbol{x}_{N+i} = \boldsymbol{x}_{N+1} R^{i-1}, i \geq 1, \tag{4.3}$$

where the matrix R is the minimal nonnegative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{4.4}$$

and the vectors $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{N+1}$ are obtained by solving

$$\boldsymbol{x}_0 D_0 + \boldsymbol{x}_1 C_2 = \boldsymbol{0},$$

$$\boldsymbol{x}_0 C_0 + \boldsymbol{x}_1 B_1 + \boldsymbol{x}_2 B_2 = \boldsymbol{0},$$

$$\boldsymbol{x}_{i-1}(I \otimes D_1) + \boldsymbol{x}_i B_1 + \boldsymbol{x}_{i+1} B_2 = \boldsymbol{0}, \; 2 \leq i \leq N-1,$$

$$\boldsymbol{x}_{N-1}(I \otimes D_1) + \boldsymbol{x}_N B_1 + \boldsymbol{x}_{N+1}(\boldsymbol{e}_2'(2) \otimes \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I) = \boldsymbol{0},$$

$$\boldsymbol{x}_N(\boldsymbol{e} \otimes I \otimes D_1) + \boldsymbol{x}_{N+1}(A_1 + RA_2) = \boldsymbol{0},$$

subject to the normalizing condition

$$\sum_{i=0}^{N} \boldsymbol{x}_i \boldsymbol{e} + \boldsymbol{x}_{N+1}(I - R)^{-1} \boldsymbol{e} = 1.$$

The computation of the vectors $\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{N+1}$ can be carried out by exploiting the special structure of the coefficient matrices and the details are omitted. For use in the sequel, we partition $\boldsymbol{x}_i = (\boldsymbol{u}_i, \boldsymbol{v}_i), 1 \leq i \leq N$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are of dimension mn.

4.1.2 The stationary waiting time distribution in the Queue

The stationary waiting time distribution in the queue of a customer is derived here. We obtain this by conditioning on the fact that at an arrival epoch the server is serving in normal mode or in vacation mode. First note that an arriving customer will enter into service immediately (at a lower service rate) when the server is on vacation. Otherwise, the customer has to wait before getting into service (either at a lower rate or normal rate).

4.1.3 Conditional waiting time in the queue (Normal mode)

Here we condition that an arriving customer finds the server busy serving in normal mode. First note that in this case, the waiting time is always positive. We now define $\boldsymbol{z}_{i,j}$ to be the steady-state probability that an arrival finds the server busy in normal mode with the current service in phase j, and the number of customers in the system including the current arrival to be i, for $1 \leq j \leq n, i \geq 2$. Let $\boldsymbol{z}_i = (\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,n})$ and $\boldsymbol{z} = (0, \boldsymbol{z}_2, \boldsymbol{z}_3, \ldots)$. Then it is easy to verify that

$$\boldsymbol{z}_i = \begin{cases} \boldsymbol{v}_{i-1}(I \otimes \frac{D_1}{\lambda}\boldsymbol{e}), & 2 \leq i \leq N, \\ (\boldsymbol{u}_N + \boldsymbol{v}_N)(I \otimes \frac{D_1}{\lambda}\boldsymbol{e}), & i = N+1, \\ \boldsymbol{x}_{i-1}(I \otimes \frac{D_1}{\lambda}\boldsymbol{e}), & i \geq N+2. \end{cases}$$

The waiting time may be viewed as the time until absorption in a Markov chain with a highly sparse structure. The state space (that includes the arriving customer in its count) of this Markov chain is given by $\Omega_1 = \{*\} \cup \{(i,j) : i \geq 2, 1 \leq j \leq n\}$. The state $*$ corresponds to the absorbing state indicating the completion of waiting for the service. It is easy to verify that the generator, $\tilde{Q}_1$, of this Markov process is of the form

$$\tilde{Q}_1 = \begin{pmatrix} 0 & O & & & \\ \boldsymbol{T}^0 & T & & & \\ & \boldsymbol{T}^0\boldsymbol{\alpha} & T & & \\ & & \boldsymbol{T}^0\boldsymbol{\alpha} & T & \\ & & & \ddots & \ddots \end{pmatrix}$$

Define $W(t), t > 0$ to be the probability that an arriving customer will enter into service no later than time t conditioned on the fact that the service is in normal mode. Let $\tilde{W}_{normal}(s)$ denote the Laplace-Stieltjes transform of the conditional stationary waiting time in the queue of an arriving customer during the normal service mode. Using the structure of $\tilde{Q}_1$ it can readily be verified that the following result holds good.

Theorem 4.1.1. *The LST of the conditional waiting time distribution of an arriving customer, finding the server busy in normal mode, is given by*

$$\tilde{W}_{normal}(s) = c\sum_{i=2}^{\infty} \boldsymbol{z}_i(sI - T)^{-1}\boldsymbol{T}^0[\boldsymbol{\alpha}(sI - T)^{-1}\boldsymbol{T}^0]^{i-2}, \quad Re(s) \geq 0, \quad (4.5)$$

where the normalizing constant c is given by

$$c = \left[\sum_{i=2}^{\infty} \boldsymbol{z}_i \boldsymbol{e}\right]^{-1}. \quad (4.6)$$

Note: The conditional mean waiting time, μ'_{normal} in the queue of an arrival finding the server to be busy in normal mode soon after the arrival is calculated as

$$\mu'_{normal} = -\tilde{W}'_{normal}(0) = c\sum_{i=2}^{\infty} \boldsymbol{z}_i(-T)^{-1}\boldsymbol{e} + \frac{c}{\mu}\sum_{i=2}^{\infty}(i-2)\boldsymbol{z}_i\boldsymbol{e}.$$

Substituting for $\boldsymbol{z}_i$ in the last equation, we get μ'_{normal} in the simplified form as

$$\mu'_{normal} = \tfrac{c}{\lambda}[\textstyle\sum_{i=1}^{N} v_i + \boldsymbol{u}_N + \boldsymbol{x}_{N+1}(I-R)^{-1}][(-T)^{-1}\boldsymbol{e} \otimes D_1\boldsymbol{e}]$$

$$+\tfrac{c}{\lambda\mu}[\textstyle\sum_{i=1}^{\infty} \boldsymbol{v}_i + (N-1)\boldsymbol{u}_N + N\boldsymbol{x}_{N+1}(I-R)^{-1} + \boldsymbol{x}_{N+1}R(I-R)^{-2}][\boldsymbol{e} \otimes D_1\boldsymbol{e}].$$

4.1.4 Conditional waiting time in the queue (vacation mode)

The conditional stationary waiting time in the queue of an arriving customer given that the server is busy in vacation mode at that instant is derived here. Let $w_{i,j_2,k}; 1 \leq i \leq N; 1 \leq j_2 \leq n; 1 \leq k \leq m$ denote the steady-state probability that a customer immediately after arrival, finds the server busy in vacation mode with the service in phase j_2 and the number of customers in the system (including the current arrival) to be i and the arrival process is in phase k. Let $\boldsymbol{w}_i = (w_{i,1,1}, \cdots, w_{i,n,m})$. It is easy to verify that

$$\boldsymbol{w}_i = \begin{cases} \boldsymbol{x}_0(\boldsymbol{\alpha} \otimes \frac{D_1}{\lambda}), \ i = 1, \\ \\ \boldsymbol{u}_{i-1}(I \otimes \frac{D_1}{\lambda}), \ 2 \leq i \leq N. \end{cases}$$

Observe that the conditional waiting time in the queue of an arriving customer, finding the server busy in vacation mode, depends on the future arrivals due to threshold N placed on the system for bringing the service rate to normal. Also note that with probability $d\boldsymbol{w}_1\boldsymbol{e}$ (the normalizing constant d is given below) an arriving customer will enter into service immediately with service in vacation mode. Thus, for the case of positive waiting time in the queue for an arriving customer, we need to keep track of the phase of the arrival process until the service rate comes to normal mode either due to meeting the threshold N or due to the vacation getting completed. Towards this end, we define the following set of states.

Let $(i, j, j_2, k) : 1 \leq i \leq N-1; 1 \leq j \leq i; 1 \leq j_2 \leq n; 1 \leq k \leq m$, denote the state that corresponds to the server being in vacation mode with i customers in the queue; the arriving customer's position in the queue is j; the current service is in phase j_2 and the arrival process is in phase k. Define $(i^*, j_2) : 1 \leq i^* \leq N-1; 1 \leq j_2 \leq n$, to be the state that corresponds to the server serving in normal mode with the position of the tagged customer in the queue being i^* and the current service in phase j_2.

Let $\boldsymbol{i} = \{(i, j, j_2, k) : 1 \leq j \leq i; 1 \leq j_2 \leq n; 1 \leq k \leq m\}, 1 \leq i \leq N-1$, and $\boldsymbol{i}^* = \{(i^*, j_2), 1 \leq j_2 \leq n\}, 1 \leq i^* \leq N-1$.

Before we formally state the result we need the following notations.

- $\tilde{I}_r$ is a matrix of dimension $r \times r+1$ of the form

$$\tilde{I}_r = \begin{pmatrix} I_r & O \end{pmatrix}, \ 1 \leq r \leq N-2.$$

- $\hat{I}_r$ is a matrix of dimension $r \times N-1$ of the form

$$\hat{I}_r = \begin{pmatrix} I_r & O \end{pmatrix}, \ 1 \leq r \leq N-1.$$

- $\bar{I}_r$ is a matrix of dimension $r \times r-1$ of the form

$$\bar{I}_r = \begin{pmatrix} O \\ I_{r-1} \end{pmatrix}, \; 2 \leq r \leq N-1.$$

- I_r is the identity matrix of dimension r

- d is the normalizing constant given by $d = \left[\sum_{i=1}^{N} \boldsymbol{w}_i \boldsymbol{e}\right]^{-1}$.

Let

$$L_{1,1} = \begin{pmatrix} T & & & & \\ \boldsymbol{T}^0\boldsymbol{\alpha} & T & & & \\ & \boldsymbol{T}^0\boldsymbol{\alpha} & T & & \\ & \ddots & \ddots & & \\ & & & \boldsymbol{T}^0\boldsymbol{\alpha} & T \end{pmatrix}, \; L_{2,1} = \begin{pmatrix} \eta\hat{I}_1 \otimes I \otimes \boldsymbol{e} \\ \eta\hat{I}_2 \otimes I \otimes \boldsymbol{e} \\ \vdots \\ \eta\hat{I}_{N-2} \otimes I \otimes \boldsymbol{e} \\ I_{N-1} \otimes (\eta I \otimes \boldsymbol{e} + I \otimes D_1\boldsymbol{e}) \end{pmatrix},$$

$$L_{2,2} = \begin{pmatrix} \tilde{B}_1 & \tilde{I}_1 \otimes I \otimes D_1 & & & \\ F_2 & I_2 \otimes \tilde{B}_1 & \tilde{I}_2 \otimes I \otimes D_1 & & \\ & F_3 & I_3 \otimes \tilde{B}_1 & \tilde{I}_3 \otimes I \otimes D_1 & \\ & & \ddots & \ddots & \\ & & & F_{N-1} & I_{N-1} \otimes \tilde{B}_1 \end{pmatrix},$$

and

$$\tilde{B}_1 = (\theta T \oplus D_0) - \eta I; F_K = \theta \bar{I}_K \otimes \boldsymbol{T}^0\boldsymbol{\alpha} \otimes I, \; 2 \leq K \leq N-1. \qquad (4.7)$$

Under this setup, it can readily be verified that

Theorem 4.1.2. *The conditional waiting time distribution in the queue of a customer, finding the server in vacation mode on arrival, is of phase type with representation* $(\boldsymbol{\gamma}, L)$ *of order* $[(N-1)n + \frac{1}{2}N(N-1)mn]$ *where*

$$\boldsymbol{\gamma} = d(\boldsymbol{0}, \boldsymbol{w}_2, \boldsymbol{e}'_2(2) \otimes \boldsymbol{w}_3, \boldsymbol{e}'_3(3) \otimes \boldsymbol{w}_4, \cdots, \boldsymbol{e}'_{N-1}(N-1) \otimes \boldsymbol{w}_N),$$

and

$$L = \begin{pmatrix} L_{1,1} & 0 \\ L_{2,1} & L_{2,2} \end{pmatrix}.$$

Note: The conditional mean waiting time, $\mu'_{vacation}$, in the queue of an arrival finding the server busy in vacation mode on arrival is calculated as $\mu'_{vacation} = \boldsymbol{\gamma}(-L)^{-1}\boldsymbol{e}$. The computation of this mean is achieved by exploiting the special structure of $\boldsymbol{\gamma}$ and L. We will briefly present the steps involved in this.

Define

$$\boldsymbol{\gamma}(-L)^{-1} = (\boldsymbol{a}, \boldsymbol{b}),$$

and partition the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as

$$\boldsymbol{a} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{N-1}),$$

$$\boldsymbol{b} = (\boldsymbol{b}_{1,1}, \boldsymbol{b}_{2,1}, \boldsymbol{b}_{2,2}, \cdots, \boldsymbol{b}_{N-1,1}, \cdots, \boldsymbol{b}_{N-1,N-1}),$$

where $\boldsymbol{a}_i, 1 \le i \le N-1$, is of dimension n and $\boldsymbol{b}_{i,j}, 1 \le j \le i, 1 \le i \le N-1$, is of dimension of mn. The vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_{i,j}$ are ideally suited for solving using any of the well-known methods such as (block) Gauss-Seidel. The necessary equations are as follows.

$$\boldsymbol{a}_1 = \boldsymbol{a}_2\,\boldsymbol{T}^0\boldsymbol{\alpha}(-T)^{-1} + \eta\sum_{r=1}^{N-1}\boldsymbol{b}_{r,1}(-T^{-1}\otimes\boldsymbol{e}) + \boldsymbol{b}_{N-1,1}(-T^{-1}\otimes D_1\boldsymbol{e}),$$

$$\boldsymbol{a}_i = \boldsymbol{a}_{i+1}\,\boldsymbol{T}^0\boldsymbol{\alpha}(-T)^{-1} + \eta\sum_{r=i}^{N-1}\boldsymbol{b}_{r,i}(-T^{-1}\otimes\boldsymbol{e}) + \boldsymbol{b}_{N-1,i}(-T^{-1}\otimes D_1\boldsymbol{e}),$$
$$2\le i\le N-2,$$

$$\boldsymbol{a}_{N-1} = \eta\boldsymbol{b}_{N-1,N-1}(-T^{-1}\otimes\boldsymbol{e}) + \boldsymbol{b}_{N-1,N-1}(-T^{-1}\otimes D_1\boldsymbol{e}),$$

$$\boldsymbol{b}_{1,1} = [\boldsymbol{w}_2 + \theta\boldsymbol{b}_{2,2}(\boldsymbol{T}^0\boldsymbol{\alpha}\otimes I)](-\tilde{B}_1)^{-1},$$

$$\boldsymbol{b}_{i,1} = [\boldsymbol{b}_{i-1,1}(I\otimes D_1) + \theta\boldsymbol{b}_{i+1,2}(\boldsymbol{T}^0\boldsymbol{\alpha}\otimes I)](-\tilde{B}_1)^{-1},\ 2\le i\le N-2,$$

$$\boldsymbol{b}_{i,j} = [\boldsymbol{b}_{i-1,j}(I\otimes D_1) + \theta\boldsymbol{b}_{i+1,j+1}(\boldsymbol{T}^0\boldsymbol{\alpha}\otimes I)](-\tilde{B}_1)^{-1},$$
$$2\le j\le i-1;\ 2< i\le N-2,$$

$$\boldsymbol{b}_{i,i} = [\boldsymbol{w}_{i+1} + \theta\boldsymbol{b}_{i+1,i+1}(\boldsymbol{T}^0\boldsymbol{\alpha}\otimes I)](-\tilde{B}_1)^{-1},\ 2\le i\le N-2,$$

$$\boldsymbol{b}_{N-1,j} = \boldsymbol{b}_{N-2,j}(I\otimes D_1)(-\tilde{B}_1)^{-1},\ 1\le j\le N-2,$$

$$\boldsymbol{b}_{N-1,N-1} = \boldsymbol{w}_N(-\tilde{B}_1)^{-1},$$

subject to the condition

$$\boldsymbol{a}_1\,\boldsymbol{T}^0 + \theta\sum_{i=1}^{N-1}\boldsymbol{b}_{i,1}(\boldsymbol{T}^0\otimes\boldsymbol{e}) = 1 - d\boldsymbol{w}_1\boldsymbol{e}.$$

Once $\boldsymbol{a}_i, 1 \leq i \leq N-1$, and $\boldsymbol{b}_{i,j}, 1 \leq j \leq i; 1 \leq i \leq N-1$, are extracted from the above equations, the mean $\mu'_{vacation}$ is given by

$$\mu'_{vacation} = \sum_{i=1}^{N-1} \left[\boldsymbol{a}_i \boldsymbol{e} + \sum_{j=1}^{i} \boldsymbol{b}_{i,j} \boldsymbol{e} \right].$$

The stationary waiting time in the queue

From the knowledge of conditional stationary waiting time in the queue, one can get the (unconditional) stationary waiting time in the queue; the details are omitted.

Note: The (unconditional) mean, μ'_{WTQ}, waiting time of a customer in the queue is obtained as

$$\mu'_{WTQ} = \tfrac{1}{\lambda} \left[\textstyle\sum_{i=1}^{N} \boldsymbol{v}_i + \boldsymbol{u}_N + x_{N+1}(I-R)^{-1} \right] [(-T)^{-1} \boldsymbol{e} \otimes D_1 \boldsymbol{e}]$$

$$+ \tfrac{1}{\lambda\mu} \left[\textstyle\sum_{i=1}^{\infty} \boldsymbol{v}_i + (N-1)\boldsymbol{u}_N + N x_{N+1}(I-R)^{-1} + x_{N+1} R (I-R)^{-2} \right] [\boldsymbol{e} \otimes D_1 \boldsymbol{e}]$$

$$+ \tfrac{1}{d} \textstyle\sum_{i=1}^{N-1} \left[\boldsymbol{a}_i \boldsymbol{e} + \sum_{j=1}^{i} \boldsymbol{b}_{i,j} \boldsymbol{e} \right].$$

4.2 Analysis of slow service mode

In this section we will discuss the duration of the server spending in slow service mode as well as the number of visits to level 0 before hitting normal service mode.

4.2.1 Distribution of a slow service mode

The duration, T_{slow}, in slow service mode is defined as the time the server starts in slow service mode (through initiating a working vacation) until either the server takes another vacation or the server gets back to normal mode through the working vacation expiring or the working vacation is interrupted as the queue length hits the threshold value N. In this section we will show that the random variable T_{slow} can be studied as the time until absorption in a finite state continuous time Markov chain with two absorbing states. We first define

$$\boldsymbol{\gamma}_M = c_1(\boldsymbol{\alpha} \otimes \boldsymbol{x}_0 D_1, \boldsymbol{0}),$$

$$M = \begin{pmatrix} \tilde{B}_1 & I \otimes D_1 & & & \\ \theta(\boldsymbol{T}^0\boldsymbol{\alpha} \otimes I) & \tilde{B}_1 & I \otimes D_1 & & \\ & \theta(\boldsymbol{T}^0\boldsymbol{\alpha} \otimes I) & \tilde{B}_1 & I \otimes D_1 & \\ & \ddots & \ddots & & \\ & & & \theta(\boldsymbol{T}^0\boldsymbol{\alpha} \otimes I) & \tilde{B}_1 \end{pmatrix},$$

$$\boldsymbol{M}_1^0 = \begin{pmatrix} \theta(\boldsymbol{T}^0 \otimes \boldsymbol{e}) \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{M}_2^0 = \begin{pmatrix} \eta\boldsymbol{e} \\ \eta\boldsymbol{e} \\ \vdots \\ \eta\boldsymbol{e} \\ \eta\boldsymbol{e} + (\boldsymbol{e} \otimes D_1\boldsymbol{e}) \end{pmatrix},$$

where $c_1 = [\boldsymbol{x}_0 D_1 \boldsymbol{e}]^{-1}$ is the normalizing constant and $\tilde{B}_1$ is as given in (4.7). The matrix M is of dimension Nmn. First note that the probability, p_{slow}, that the server will serve only in slow mode before taking another vacation is given by $p_{slow} = \boldsymbol{\gamma}_M(-M)^{-1}\boldsymbol{M}_1^0$. We now have the following result.

Theorem 4.2.1. *The (conditional) probability density function of T_{slow}, conditioned on the fact that the slow service mode ends through the server taking another vacation, is given by*

$$f_{T_{slow}}(y) = \frac{1}{p_{slow}}\boldsymbol{\gamma}_M e^{My}\boldsymbol{M}_1^0, \ y \geq 0. \tag{4.8}$$

Given that the slow service mode ends through the server taking another vacation the (conditional) mean time spent in slow mode can be calculated as

$$\mu'_{SM} = \frac{1}{p_{slow}}\boldsymbol{\gamma}_M(-M)^{-2}\boldsymbol{M}_1^0. \tag{4.9}$$

Note: 1. The special structure of $\boldsymbol{\gamma}_M$, M, and $\boldsymbol{M}_1^0$ is to be exploited when computing this mean. The details are similar to the computation of $\mu'_{vacation}$ and hence omitted.

2. By a similar argument we can get the (conditional) probability density function of T_{slow} and the mean, conditioned on the fact that the server ends the slow service mode by entering into the normal rate. The details are omitted.

4.2.2 Distribution of the number of visits to level 0 before hitting normal service mode

We consider the queueing system at an arrival epoch that finds the server in vacation. At this instant the service will start in slow mode. The quantity that is of interest here is the probability mass function $\{p_k, k \geq 0\}$, of the number of visits to level **0** before hitting normal service mode. This mass function and its associated measures such as mean and standard deviation, play an important role in the qualitative study of the model under consideration. Using the set up in 4.2.1 it can easily be verified that

$$p_k = \boldsymbol{\gamma}_M(-M)^{-1}B^k\boldsymbol{M}_2^0,\ k \geq 0, \tag{4.10}$$

where

$$B = \theta\left[(\boldsymbol{e}_N(1)\boldsymbol{e}'_N(1) \otimes \boldsymbol{T}^0\boldsymbol{\alpha} \otimes (-D_0)^{-1}D_1)\right](-M)^{-1}. \tag{4.11}$$

Note: It is easy to see that the mean number of visits to level **0** before hitting level $N+1$, μ_{NVZ}, is obtained as

$$\mu_{NVZ} = \boldsymbol{\gamma}_M(-M)^{-1}B(I-B)^{-2}\boldsymbol{M}_2^0. \tag{4.12}$$

The computation of μ_{NVZ} can be carried out by exploiting the special structure of $\boldsymbol{\gamma}_M, M$, and B. Below, we will outline the main steps. Towards this end, we first define

$$\boldsymbol{\gamma}_M(-M)^{-1} = (\boldsymbol{d}_1, \cdots, \boldsymbol{d}_N), \tag{4.13}$$

where the vectors $\boldsymbol{d}_i, 1 \leq i \leq N$, are of dimension nm, and their computation is very similar to the one discussed in finding $\mu'_{vacation}$. From (4.11) it is clear that B is of the form

$$B = \begin{pmatrix} B_1 & B_2 & & B_N \\ 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \end{pmatrix},$$

where the matrices $B_i, 1 \leq i \leq N$, of order nm are obtained by solving the following equations that are ideally suited for any of the well-known methods such as (block) Gauss-Seidel.

$$B_1 = \theta[B_2(\boldsymbol{T}^0\boldsymbol{\alpha} \otimes I) + (\boldsymbol{T}^0\boldsymbol{\alpha} \otimes (-D_0)^{-1}D_1)](-\tilde{B}_1)^{-1},$$

$$B_i = [B_{i-1}(I \otimes D_1) + \theta B_{i+1}(\boldsymbol{T}^0\boldsymbol{\alpha} \otimes I)](-\tilde{B}_1)^{-1}, \ 2 \leq i \leq N-1,$$

$$B_N = B_{N-1}(I \otimes D_1)(-\tilde{B}_1)^{-1},$$

subject to the condition

$$\theta B_1(\boldsymbol{T}^0 \otimes \boldsymbol{e}) + B_N(\boldsymbol{e} \otimes D_1\boldsymbol{e}) + \eta \sum_{i=1}^{N} B_i\boldsymbol{e} = \theta(\boldsymbol{T}^0 \otimes \boldsymbol{e}),$$

and $\tilde{B}_1$ is as given in (4.7). Using the facts that

$$p_{slow} = \theta \boldsymbol{d}_1(\boldsymbol{T}^0 \otimes \boldsymbol{e}) \ \text{ and } \ \mu_{NVZ} = \boldsymbol{\gamma}_M(-M)^{-1}(I-B)^{-2}\boldsymbol{M}_2^0 - 1,$$

and the special form of B, it can easily be verified that

$$\mu_{NVZ} = \theta \boldsymbol{d}_1 (I - B_1)^{-1} (\boldsymbol{T^0} \otimes \boldsymbol{e}).$$

4.2.3 *The uninterrupted duration of a vacation*

The duration of the time the server is in uninterrupted vacation(s) is the interval between the epoch at which the server goes on vacation and the next arrival epoch. It is easy to verify that this duration is of phase type with representation $(\boldsymbol{\xi}, D_0)$ of dimension m, where $\boldsymbol{\xi} = c_2(\theta \boldsymbol{u}_1 + \boldsymbol{v}_1)(\boldsymbol{T^0} \otimes I)$ and c_2 is the normalizing constant given by $c_2 = [(\theta \boldsymbol{u}_1 + \boldsymbol{v}_1)(\boldsymbol{T^0} \otimes \boldsymbol{e})]^{-1}$. The mean, μ_{UIV}, is calculated as $\mu_{UIV} = \boldsymbol{\xi}(-D_0)^{-1}\boldsymbol{e}$.

4.2.4 *Key system 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. The measures are listed below along with their formulae for computation.

1. Probability that the server is on vacation: $P_{VAC} = \boldsymbol{x}_0 \boldsymbol{e}$.

2. Probability that the server is serving at a lower rate: $P_{LR} = \sum_{i=1}^{N} \boldsymbol{u}_i \boldsymbol{e}$.

3. Probability that the server is serving at a normal rate rate:
   $P_{NR} = \sum_{i=1}^{N} \boldsymbol{v}_i \boldsymbol{e} + \boldsymbol{x}_{N+1}(I - R)^{-1}\boldsymbol{e}$.

4. Mean number of customers in the system:
   $\mu_{NS} = \sum_{i=1}^{N} i(\boldsymbol{u}_i + \boldsymbol{v}_i)\boldsymbol{e} + N\boldsymbol{x}_{N+1}(I - R)^{-1}\boldsymbol{e} + \boldsymbol{x}_{N+1}(I - R)^{-2}\boldsymbol{e}$.

4.3 Numerical Results

For the arrival process we consider the following five sets of matrices for D_0 and D_1.

1. Erlang (ERA)

$$D_0 = \begin{pmatrix} -5 & 5 & & & \\ & -5 & 5 & & \\ & & -5 & 5 & \\ & & & -5 & 5 \\ & & & & -5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 5 & & & & \end{pmatrix}$$

2. Exponential (EXA)

$D_0 = (-1), D_1 = (1)$

3. Hyperexponential (HEA)

$$D_0 = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 9 & 1 \\ 0.9 & 0.1 \end{pmatrix}$$

4. MAP with negative correlation (MNA)

$$D_0 = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.02 & 0 & 1.98 \\ 445.995 & 0 & 4.505 \end{pmatrix}$$

5. MAP with positive correlation (MPA)

$$D_0 = \begin{pmatrix} -2 & -2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.98 & 0 & 0.02 \\ 4.505 & 0 & 445.995 \end{pmatrix}$$

All these five MAP processes are normalized so as to have an arrival rate of 1. However, these are qualitatively different in that they have different variance and correlation structure. 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.

For the service time distribution we consider the following two phase type distributions.

1. Erlang (ERS)
   $$\boldsymbol{\alpha} = (1, 0) \; T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}$$
2. Hyperexponential (HES)
   $$\boldsymbol{\alpha} = (0.9, 0.1) \; T = \begin{pmatrix} -1.90 & 0 \\ 0 & -0.19 \end{pmatrix}$$

The above two distributions will be normalized to have a specific mean in our illustrative example. Note that these are qualitatively different in that they have different variances. The ratio of the standard deviation of HES to that of ERS is 3.1745.

ILLUSTRATIVE EXAMPLE 4.1: The purpose of this example is to see how various system performance measures behave under different scenarios. We fix $\lambda = 1, \mu = 1.1$, and $\theta = 0.6$. First we look at the effect of varying

N and η on the performance measures: (conditional) mean duration of service in slow mode which ends in the server taking another vacation and the mean number of visits to level zero before hitting the normal service mode. In the following we summarize the observations based on the graphs of these performance measures.

- Consider figures 4.1 and 4.2. An increase in η leads to a decrease in the mean duration of vacation. Hence a switching from the lower service rate to the normal one occurs more frequently. Once the service rate is brought back to normal, the server clears out the customers at a faster rate. So the measure P_{VAC} appears to increase as η increases. This is true for all values of N and for all combinations of arrival and service processes under study. As N increases the duration of vacation mode of service gets extended, as is expected. Due to the slow service rate the customers get accumulated faster. So P_{VAC} decreases until the service rate gets to normal. Also note that the probability, P_{LR}, that the server is serving at a low rate increases as N is increased (for fixed η) for all combinations of arrival and service distributions. This in turn will cause the probability, P_{NR}, of the server serving under normal mode to decrease as N increases. As expected, the measure P_{NR} appears to increase with increasing η. When comparing the mean duration of service in slow mode, we notice(for fixed N and η) that HES yield a lower value as opposed to ERS. This is the case for all five arrival processes considered.

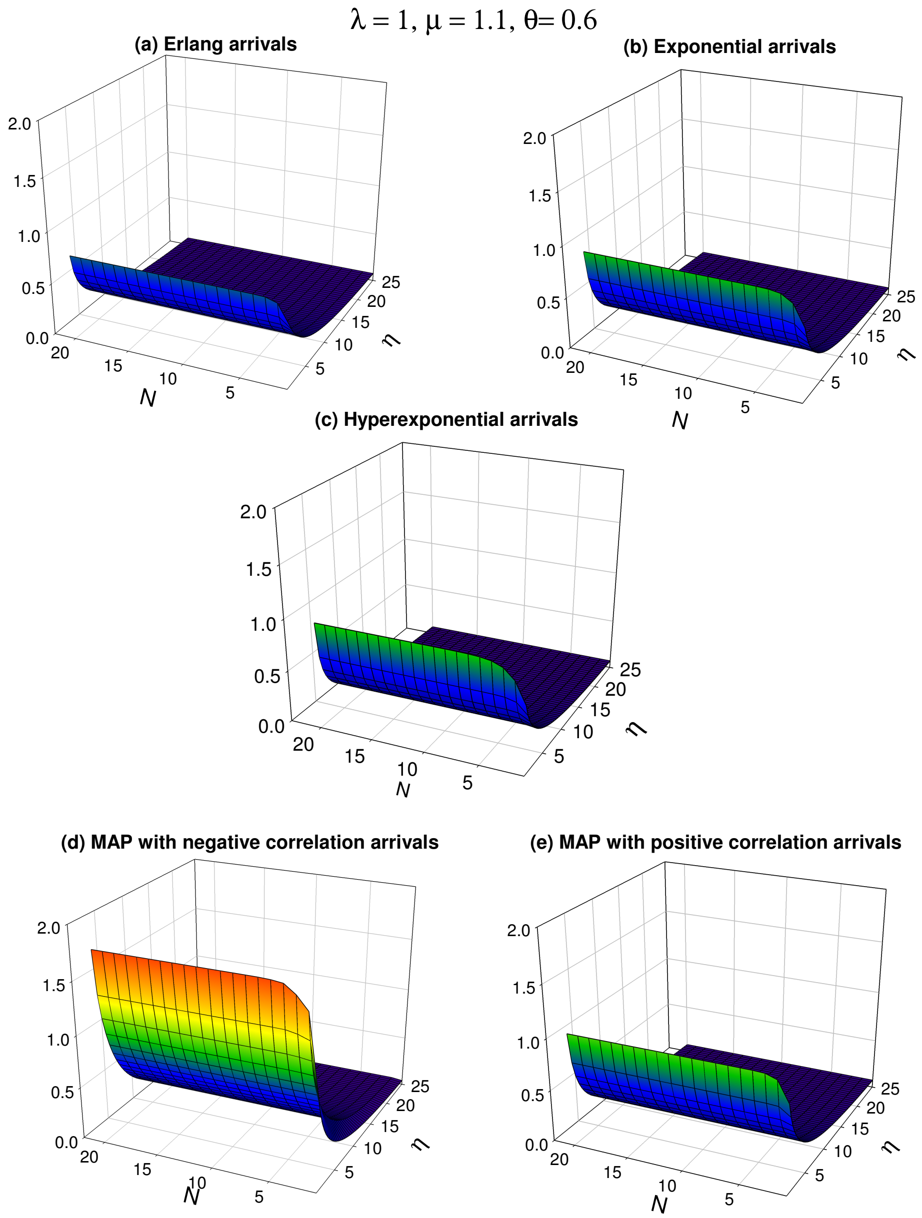

Fig. 4.1: Mean duration in slow mode - Erlang services

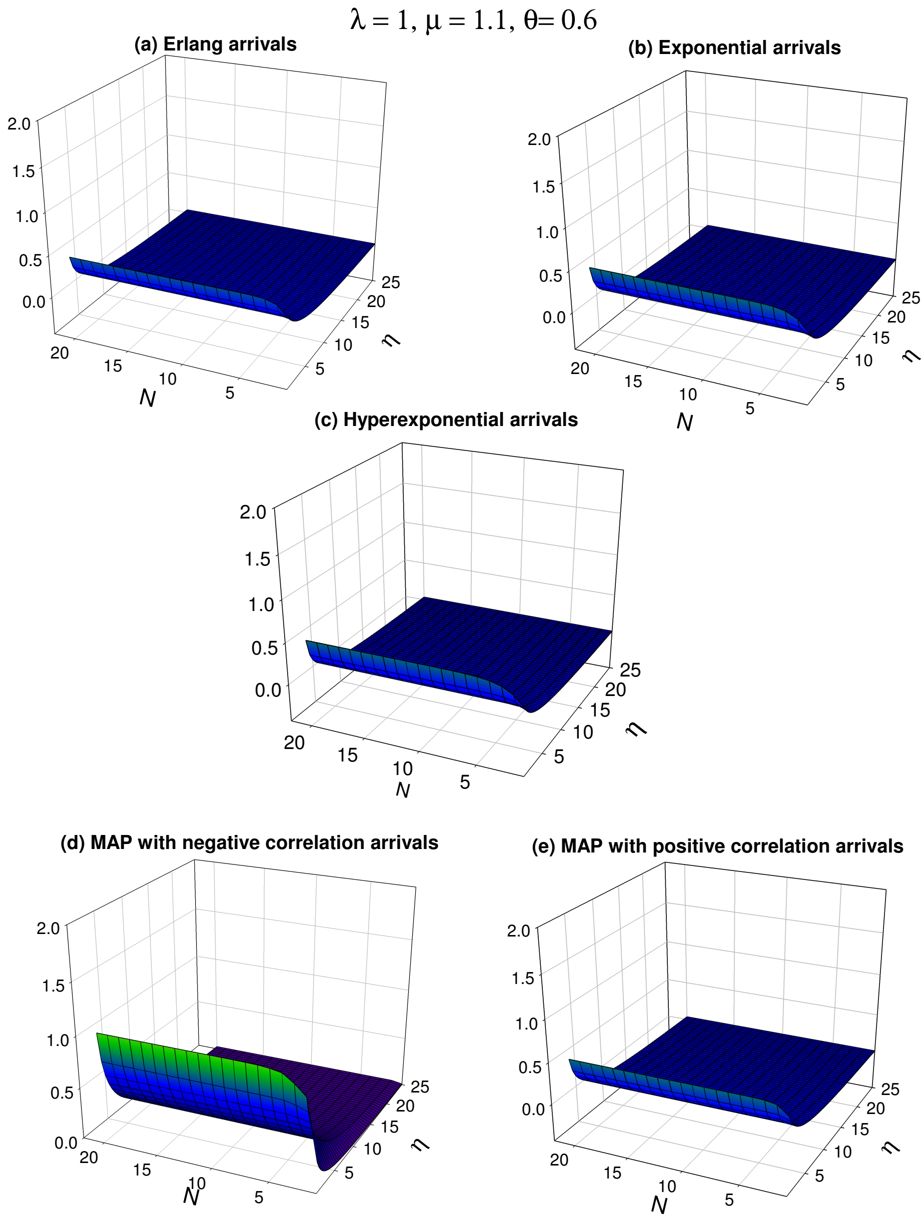

Fig. 4.2: Mean duration in slow mode - hyperexponential services

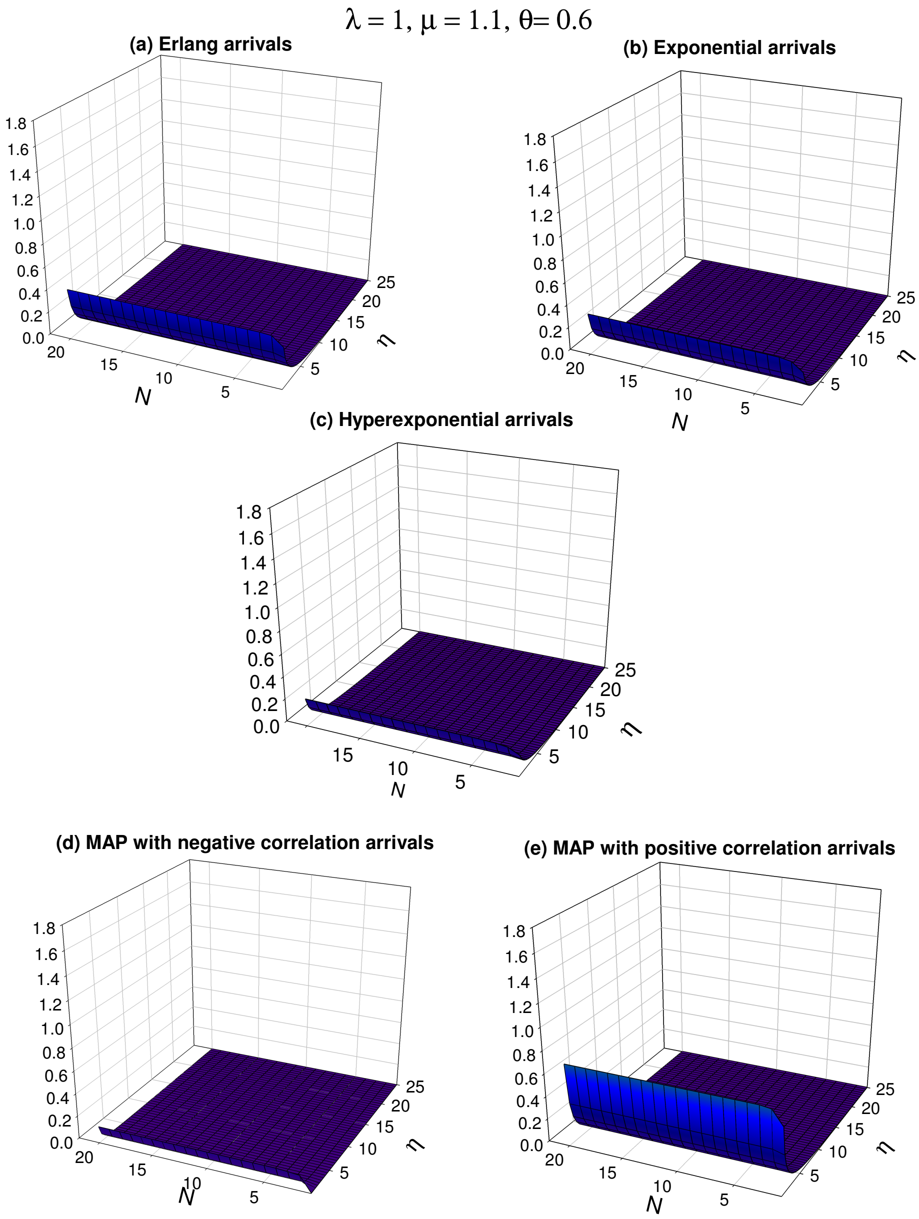

Fig. 4.3: Mean number of visits to level zero - Erlang services

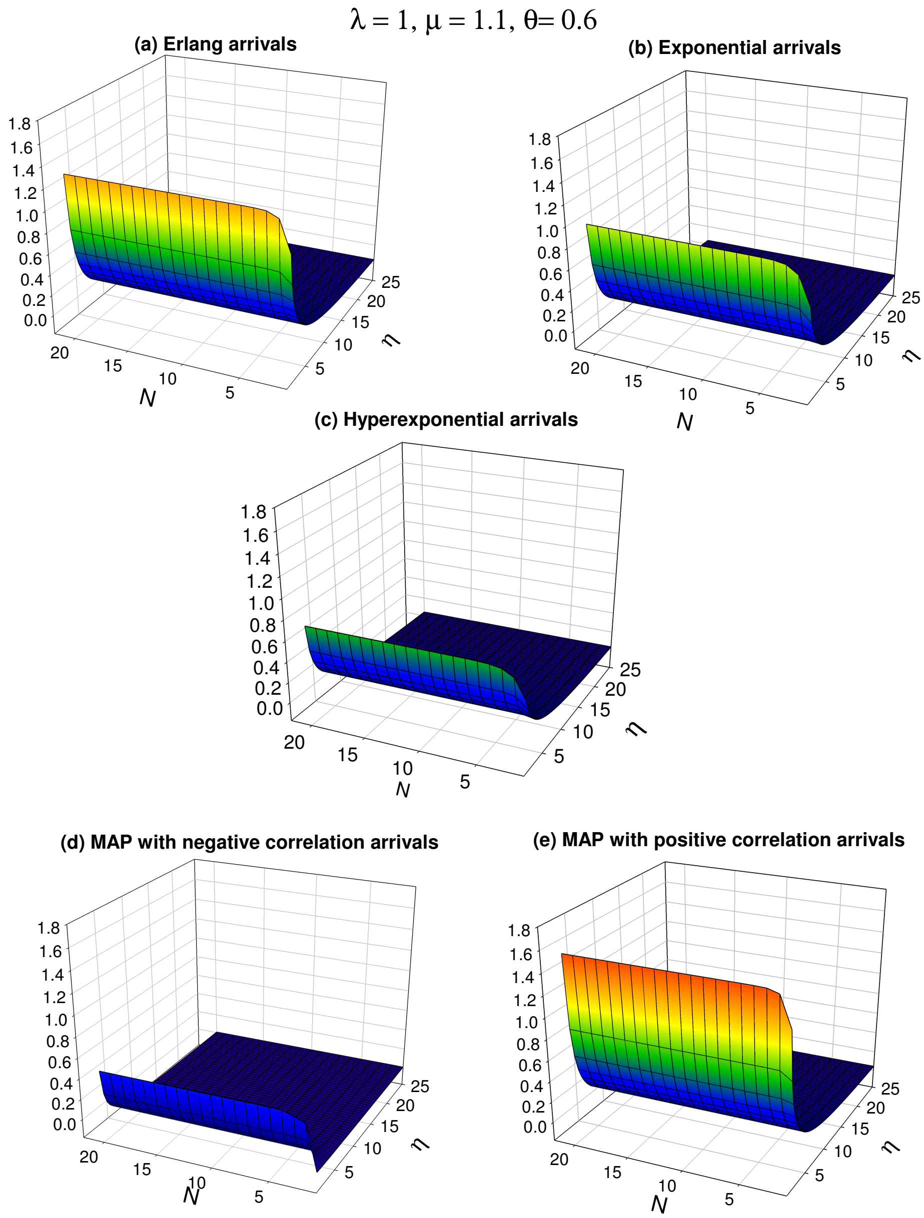

Fig. 4.4: Mean number of visits to level zero - hyperexponential services

- Referring to Figures 4.3 and 4.4, we note that as η increases, the measure μ_{NVZ} appears to decrease in all cases, as expected, for any fixed N. Among renewal arrivals, those with larger variation yields a smaller value for this measure. That is, HEA has a smaller value compared to EXA and EXA has a smaller value compared to ERA. Among correlated arrivals, MPA has a higher value than MNA. It is worth pointing out that both MNA and MPA processes have the same mean and variance, but MPA has a positive correlation while MNA has a negative correlation. This indicates the significant role played by correlation. As N increases, this measure appears to increase monotonically to a limiting value (which depends on η as well as on the arrival and service time distributions). It should be noted that the rate of approach is higher for larger values of η. That is, the impact of N on this measure decreases as η increases. We notice that this measure appears to have a larger value when services are changed from Erlang to hyperexponential. When comparing this measure for various distributions (for fixed N and η), we notice that HES yield a higher value as opposed to ERS. This is the case for all five arrival processes considered.

Now we look at the unconditional mean waiting time, μ'_{WTQ}, in the queue of a customer. The values of this measure as functions of N and η under different scenarios are displayed in Table 4.1. Some key observations are as follows.

- As is to be expected, the mean is a non-increasing function of η (for fixed N) and is a non-decreasing function of N (for fixed η). This is the case for all combinations of arrival and service processes. However, the rate of change is much smaller in the case of MPA as compared to the other arrivals.

- The mean is significantly larger for MPA case indicating the role played by the (positively) correlated arrivals.

- For all arrivals except MPA arrivals, we notice the mean changes significantly as a function of η when N becomes large. This is due to the fact that for large N the mean waiting time can only be reduced through an increase in η (which will decrease the duration of the slow service period).

The unconditional mean waiting time in the queue (μ'_{WTQ})

Table 4.1

		Erlang services					Hyperexponential services				
N	η	ERA	EXA	HEA	MNA	MPA	ERA	EXA	HEA	MNA	MPA
	0.1	3.21	6.97	25.21	7.07	497.45	23.82	27.64	46.32	27.63	518.10
	0.2	3.20	6.96	25.21	7.07	497.44	23.81	27.63	46.31	27.63	518.08
1	0.3	3.19	6.96	25.21	7.07	497.44	23.80	27.62	46.31	27.63	518.07
	0.4	3.18	6.95	25.20	7.07	497.43	23.79	27.61	46.30	27.63	518.06
	0.5	3.18	6.95	25.20	7.07	497.43	23.78	27.61	46.30	27.63	518.05
	0.1	3.50	7.19	25.39	7.42	497.57	24.13	27.90	46.51	28.05	518.27
	0.2	3.46	7.16	25.37	7.40	497.55	24.08	27.86	46.48	28.02	518.22
2	0.3	3.42	7.14	25.36	7.38	497.52	24.03	27.82	46.46	27.99	518.18
	0.4	3.39	7.12	25.34	7.36	497.50	23.99	27.79	46.44	27.96	518.15
	0.5	3.36	7.10	25.33	7.35	497.49	23.96	27.77	46.43	27.94	518.12
	0.1	3.83	7.45	25.59	7.57	497.68	24.43	28.17	46.72	28.24	518.40
	0.2	3.73	7.38	25.55	7.52	497.61	24.31	28.08	46.66	28.17	518.31
3	0.3	3.64	7.32	25.51	7.48	497.57	24.22	28.00	46.62	28.11	518.24
	0.4	3.56	7.27	25.48	7.44	497.54	24.14	27.94	46.58	28.06	518.19
	0.5	3.50	7.23	25.45	7.41	497.52	24.08	27.90	46.55	28.02	518.15
	0.1	4.16	7.73	25.81	7.90	497.74	24.70	28.43	46.93	28.57	518.50
	0.2	3.96	7.59	25.72	7.79	497.65	24.50	28.27	46.83	28.42	518.36
4	0.3	3.80	7.49	25.65	7.70	497.59	24.36	28.15	46.76	28.31	518.27
	0.4	3.68	7.40	25.60	7.63	497.55	24.24	28.06	46.70	28.23	518.21
	0.5	3.58	7.34	25.55	7.57	497.53	24.15	27.99	46.65	28.16	518.16
	0.1	4.46	8.00	26.03	8.14	497.79	24.94	28.68	47.15	28.79	518.58
	0.2	4.14	7.78	25.89	7.96	497.67	24.66	28.44	46.99	28.57	518.40
5	0.3	3.92	7.62	25.78	7.82	497.60	24.46	28.27	46.88	28.42	518.29
	0.4	3.75	7.50	25.70	7.71	497.56	24.31	28.15	46.79	28.30	518.22
	0.5	3.63	7.41	25.63	7.63	497.53	24.19	28.05	46.72	28.21	518.17
	0.1	5.54	9.10	26.99	9.29	497.84	25.82	29.61	48.06	29.76	518.77
	0.2	4.61	8.36	26.48	8.59	497.69	25.08	28.94	47.55	29.11	518.46
10	0.3	4.12	7.95	26.17	8.19	497.62	24.67	28.56	47.24	28.73	518.32
	0.4	3.85	7.70	25.96	7.95	497.57	24.42	28.32	47.04	28.50	518.23
	0.5	3.68	7.54	25.81	7.79	497.54	24.25	28.16	46.90	28.34	518.18

5. $MAP/PH/1$ RETRIAL QUEUE WITH CONSTANT RETRIAL RATE AND WORKING VACATIONS

In this chapter we study a $MAP/PH/1$ retrial queueing model in which the server is subject to taking vacations and serving at a lower rate during those times. The service returns to normal rate whenever the vacation gets completed. If an arriving customer finds the server busy it joins a pool of unsatisfied customers called orbit. Inter retrial times are exponentially distributed with intensity independent of the number of customers in the orbit.

5.1 A brief review of research on retrial queues

In the retrial queuing system customers arriving to a busy service system, join a group of blocked customers called orbit. From the orbit each unit tries to access a free server, after a random amount of time. Such situations occur in communication and computer networks. For a nearly exhaustive account of developments in this area up to 2000, we refer to Yang and Templeton [69], Falin [21], Artalejo ([2], [3]) and Falin and Templeton [22]. For recent developments in this area we refer the reader to Artalejo and Gomez-Corral [6].

In classical retrial queueing systems server idle time is very high. This is because every service is preceded and followed by an idle period in the absence of a buffer for the customers to wait. In the modern scenario, it is not desirable from the service system's point of view, to have a long idle time. To this end Artalejo et al. [4] introduced a concept called orbital search, where the server looks out for potential customers from the orbit immediately after every service completion with a positive probability. Dudin et al. [20] and Krishnamoorthy et al. [34] also consider orbital search with different arrival streams and different service time distributions. Chakravarthy et al. [12] consider orbital search in the multi server case. But even with the search option, system may not be able to utilize the entire server idle time. It is from this stand point, one explores the possibility of retrial queueing systems with vacations and working vacations. During vacations the idle server may attend some less urgent secondary task. We may also consider the notion of working vacation depending upon the nature of the secondary job attended. In the latter case the server returns to attend the primary job as and when a customer arrives to the system.

However, to the best of our knowledge, there has been no attempt so far to analyze a $MAP/PH/1$ retrial queuing model with working vacations. Further in most of the works on retrial queues the retrial rate depends on the number of customers in the orbit. However, recent applications to communication protocols and local area networks show that there are queueing situations in which the retrial rate is independent of the number of customers in the orbit. Hence, in this model we prefer the constant retrial policy.

5.2 Mathematical Model

We consider a single server retrial queueing system in which customers arrive according to a Markovian arrival process (MAP) with parameter matrices D_0 and D_1 of dimension m. An arriving primary customer who finds the server free, immediately occupies the server and obtains service. On the other hand if the arriving unit finds the server busy, it joins an orbit of infinite size. From the orbit the unit makes retrial at the rate β, which is independent of the number of customers in the orbit. The service times follow phase type distribution with representation $(\boldsymbol{\alpha}, T)$ of order n. The server takes vacation when the customer being served depart from the system and no customer is left in the orbit. Duration of vacation is exponentially distributed with parameter η. During a vacation if a customer arrives, the server returns to attend that customer. However, the customers are served during vacation only at a lower rate compared to the regular service. Precisely the vacation mode service times are also phase type distributed with representation $(\boldsymbol{\alpha}, \theta T)$, with $0 < \theta < 1$. Even when the vacation is interrupted by a customer arrival and consequent service commencement, vacation clock continues to tick so that on completion of this service if the vacation clock has not expired, the server continues to be on vacation irrespective of whether there are customers in the orbit. At the end of each vacation, the server takes another vacation if the orbit is empty and remains idle otherwise.

5.2.1 The QBD process

The model discussed in Section 5.2 can be studied as a level independent QBD process. First, we set up necessary notations. Let μ denote the regular service rate; then it is easy to verify that $\mu = [\boldsymbol{\alpha}(-T)^{-1}\boldsymbol{e}]^{-1}$. Let $\theta, 0 < \theta < 1$, denote the factor by which the normal service rate will be reduced, when the server is serving in the vacation mode. That is, when the server is serving in the vacation mode, the rate of service is given by $\theta\mu$.

Defining $N(t)$ to be the number of customers in the orbit at time t,

$$S_1(t) = \begin{cases} 0, & \textit{if the server is not working,} \\ j, & \textit{if the server is busy in phase } j,\ 1 \leq j \leq n, \end{cases}$$

$$S_2(t) = \begin{cases} 0, & \textit{if the server is on (working) vacation,} \\ 1, & \textit{otherwise,} \end{cases}$$

and $M(t)$, the phase of the arrival process at time t. Note that the case $S_1(t) = 0$, $S_2(t) = 0$ corresponds to server on vacation and the case $S_1(t) = 0$, $S_2(t) = 1$ indicates that the server is idle. It is easy to verify that $\{(N(t), S_1(t), S_2(t), M(t)) : t \geq 0\}$ is a level independent QBD process with state space

$$\Omega = \bigcup_{i=0}^{\infty} l(i)$$

where

$$l(i) = \{(i, j_1, j_2, k) : i \geq 0; 0 \leq j_1 \leq n; j_2 = 0 \text{ or } 1; 1 \leq k \leq m\}.$$

The generator matrix Q of the QBD process under consideration is of the form

$$Q = \begin{pmatrix} B_1 & B_0 & & \\ B_2 & A_1 & A_0 & \\ & A_2 & A_1 & A_0 \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the (block) matrices appearing in Q are as follows:

$$B_0 = \begin{bmatrix} O & O & O & O \\ O & O & I \otimes D_1 & O \\ O & O & O & I \otimes D_1 \end{bmatrix}, \quad B_1 = \begin{bmatrix} D_0 & \boldsymbol{\alpha} \otimes D_1 & O \\ \theta \boldsymbol{T}^0 \otimes I & \theta T \oplus D_0 - \eta I & \eta I \\ \boldsymbol{T}^0 \otimes I & O & T \oplus D_0 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} O & \beta(\boldsymbol{\alpha} \otimes I) & O \\ O & O & \beta(\boldsymbol{\alpha} \otimes I) \\ O & O & O \\ O & O & O \end{bmatrix}, \quad A_0 = \begin{bmatrix} O & O & O & O \\ O & O & O & O \\ O & O & I \otimes D_1 & O \\ O & O & O & I \otimes D_1 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} D_0 - \eta I - \beta I & \eta I & \boldsymbol{\alpha} \otimes D_1 & O \\ O & D_0 - \beta I & O & \boldsymbol{\alpha} \otimes D_1 \\ \theta \boldsymbol{T}^0 \otimes I & O & \theta T \oplus D_0 - \eta I & \eta I \\ O & \boldsymbol{T}^0 \otimes I & O & T \oplus D_0 \end{bmatrix} \text{ and}$$

$$A_2 = \begin{bmatrix} O & O & \beta(\boldsymbol{\alpha} \otimes I) & O \\ O & O & O & \beta(\boldsymbol{\alpha} \otimes I) \\ O & O & O & O \\ O & O & O & O \end{bmatrix}.$$

5.3 Steady-state analysis

In this section we analyze the the model under the condition that the system is stable.

5.3.1 Stability condition

Define $A = A_0 + A_1 + A_2$. Let $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \boldsymbol{\pi_3}, \boldsymbol{\pi_4})$ be the steady-state probability vector of A, where $\boldsymbol{\pi_1}$, $\boldsymbol{\pi_2}$ are of dimension m and $\boldsymbol{\pi_3}$, $\boldsymbol{\pi_4}$ are of dimension mn. For the stability of the queueing model we must have $\boldsymbol{\pi} A_0 \boldsymbol{e} < \boldsymbol{\pi} A_2 \boldsymbol{e}$, (see [50]) which simplifies to $(\boldsymbol{\pi_3} + \boldsymbol{\pi_4})(\boldsymbol{e}_n \otimes \mathrm{D}_1 \boldsymbol{e}_m) < \beta(\boldsymbol{\pi_1} + \boldsymbol{\pi_2})\boldsymbol{e}_m$. The last inequality suggests that for stability of the queueing system discussed here, it is required that the rate of inflow in to the orbit is less than the effective retrial rate.

5.3.2 Steady-state probability vector

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$, be the steady-state probability vector of Q. Note that $\boldsymbol{x}_0$ is of dimension $m + 2mn$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ are of dimension $2m + 2mn$. $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = \mathbf{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$. Apparently when the stability condition is satisfied the sub vectors of $\boldsymbol{x}$ except $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, corresponding to the different level sates are given by the equation

$\boldsymbol{x}_j = \boldsymbol{x}_1 R^{j-1},\ j \geq 2$, where R is the minimal non negative solution of the matrix quadratic equation (see [50])

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{5.1}$$

The sub vectors $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are obtained by solving the equations

$$\boldsymbol{x}_0 B_0 + \boldsymbol{x}_1 B_1 = \mathbf{0} \tag{5.2}$$

$$\boldsymbol{x}_0 B_0 + \boldsymbol{x}_1 (A_1 + R A_2) = \mathbf{0} \tag{5.3}$$

subject to the normalizing condition

$$\boldsymbol{x}_0 \boldsymbol{e}_{(m+2mn)} + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{e}_{2(m+n)} = 1. \tag{5.4}$$

The computation of R matrix can be carried out using a number of well known methods such as logarithmic reduction algorithm.

5.3.3 Key system 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. The measures are listed below along with their formulae for computation.

1. Probability that the orbit is empty: $P_{EMPTY} = \boldsymbol{x}_0 \boldsymbol{e}$.

2. Probability that the server is on vacation: $P_{VACN} = \sum_{i=0}^{\infty} \sum_{k=1}^{m} \boldsymbol{x}_{i00k}$.

3. Probability that the server is idle: $P_{IDLE} = \sum_{i=1}^{\infty} \sum_{k=1}^{m} \boldsymbol{x}_{i01k}$.

4. Probability that the server is busy in vacation mode:

   $P_{BVM} = \sum_{i=0}^{\infty} \sum_{j_1=1}^{n} \sum_{k=1}^{m} \boldsymbol{x}_{ij_1 0k}$.

5. Probability that the server completes a service in vacation mode : $P_{SCSLO} = P$(service time in slow mode $<$ an exponentially distributed random variable with parameter $\eta) = \boldsymbol{\alpha}(\eta I - \theta T)^{-1}\theta \boldsymbol{T}^0$

6. Probability that the server is busy in normal mode:

   $P_{BNM} = \sum_{i=0}^{\infty} \sum_{j_1=1}^{n} \sum_{k=1}^{m} \boldsymbol{x}_{ij_1 1k}$.

7. Probability that the server is busy: $P_B = P_{BVM} + P_{BVM}$.

8. Mean number of customers in the orbit:

   $\mu_{OBT} = \sum_{i=1}^{\infty} i\boldsymbol{x}_i \boldsymbol{e} = \boldsymbol{x}_1(I - R)^{-2}\boldsymbol{e}$

9. Mean number of customers in the system: $\mu_{NS} = \mu_{OBT} + P_B$

10. Probability of a successful retrial :

    $P_{SRT} = \beta/(\beta + \lambda) \sum_{i=1}^{\infty} \sum_{j_2=0}^{1} \sum_{k=1}^{m} \boldsymbol{x}_{i0j_2k}$.

11. Mean number of successful retrials :

    $\mu_{SRT} = \beta/(\beta + \lambda) \sum_{i=1}^{\infty} \sum_{j_2=0}^{1} \sum_{k=1}^{m} i\boldsymbol{x}_{i0j_2k}$.

5.4 Numerical Results

In order to bring out the qualitative nature of the model under study, we present a few representative examples in this section. For the arrival process we consider the following five sets of matrices for D_0 and D_1.

1. Erlang (ERA)

$$D_0 = \begin{pmatrix} -5 & 5 & & & \\ & -5 & 5 & & \\ & & -5 & 5 & \\ & & & -5 & 5 \\ & & & & -5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 5 & & & & \end{pmatrix}$$

2. Exponential (EXA)

$D_0 = (-1), D_1 = (1)$

3. Hyperexponential (HEA)

$$D_0 = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 9 & 1 \\ 0.9 & 0.1 \end{pmatrix}$$

4. MAP with negative correlation (MNA)

$$D_0 = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.02 & 0 & 1.98 \\ 445.995 & 0 & 4.505 \end{pmatrix}$$

5. MAP with positive correlation (MPA)

$$D_0 = \begin{pmatrix} -2 & -2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.98 & 0 & 0.02 \\ 4.505 & 0 & 445.995 \end{pmatrix}$$

All these five MAP processes are normalized so as to have an arrival rate of 1. However, these are qualitatively different in that they have different variance and correlation structure. The first three arrival processes, namely ERA, EXA, and HEA correspond to renewal processes and so the correlation is 0. The arrival process labeled 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.

For the service time distribution we consider the following three phase type distributions.

1. Erlang (ERS)

   $\boldsymbol{\alpha} = (1, 0)\ T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}$

2. Exponential (EXS)

   $\boldsymbol{\alpha} = 1.0, T = -1.0$

3. Hyperexponential (HES)

   $\boldsymbol{\alpha} = (0.9, 0.1)\ T = \begin{pmatrix} -1.90 & 0 \\ 0 & -0.19 \end{pmatrix}$

The above three distributions will be normalized to have a specific mean in our illustrative examples. Note that these are qualitatively different in that they have different variances. The ratio of the standard deviations of these two service distributions with respect to ERS are, respectively, 1, 1.4142, and 3.1745.

ILLUSTRATIVE EXAMPLE 5.1: We analyze the effect of the parameter β on the measure mean number of customers, μ_{NS}, in the system for different arrival and service processes. Table 5.1 analyzes the effect of β with Erlang service, table 5.2 explains the effect of β with exponential service and table 5.3 examines the effect of β with hyperexponential service process. We fix $\lambda = 1$, $\mu = 1.4$, $\eta = 0.5$, $\theta = 0.6$ and get the following results.

Table 5.1: With Erlang Service Process

β	ERA	EXA	HEA	MNA	MPA
3	5.7545	17.2162	84.0281	144.2722	2503.3365
4	3.5757	7.6028	26.9322	10.2311	702.2934
5	2.9482	5.6673	18.2386	6.684	421.82
10	2.2154	3.7114	10.0957	4.0226	211.4321
20	1.9829	3.1431	7.819	3.3778	160.8913
30	1.9178	2.9868	7.1963	3.2093	147.7319
40	1.8872	2.9135	6.905	3.1317	141.6795
50	1.8695	2.871	6.7361	3.087	138.201

Table 5.2: With Exponential Service Process

β	ERA	EXA	HEA	MNA	MPA
3	8.4593	20.7893	88.4049	172.1747	2525.252
4	4.9365	9.0907	28.5759	12.1387	704.7186
5	3.9594	6.7375	19.4243	7.8697	422.713
10	2.8415	4.3627	10.857	4.6653	212.0658
20	2.4924	3.6744	8.4683	3.8884	161.4055
30	2.3951	3.4852	7.8163	3.6852	148.215
40	2.3495	3.3960	7.5115	3.5917	142.1482
50	2.323	3.3452	7.3349	3.5379	138.3973

Table 5.3: With Hyperexponential Service Process

β	ERA	EXA	HEA	MNA	MPA
3	30.5737	49.6303	123.8119	430.8089	2607.088
4	16.1295	21.0991	41.387	27.7478	721.168
5	12.3164	15.3777	28.5051	17.5934	432.4837
10	8.0737	9.6319	16.4436	9.2434	217.3843
20	6.778	7.9792	13.1039	7.4015	165.7004
30	6.4193	7.5269	12.1966	6.9201	152.2438
40	6.2513	7.3155	11.7733	6.6984	146.0543
50	6.1539	7.193	11.5282	6.5708	142.497

- For fixed values of other parameters, as β increases μ_{NS} decreases as expected. This is because as β increases P_B increases. Thus the server is fed with customers more frequently and hence more and more customers leave the system after completing the service. But the above tables suggest that the magnitude of μ_{NS} not only depends on β but also the characteristics of the inter arrival and service time distributions. For a given value of β and for a given service process, among the renewal arrivals those with larger variance yield larger values for μ_{NS}. That is HEA has highest value for μ_{NS}, EXA has the next highest value and ERA has the smallest value . Among the correlated arrivals MPA has larger value for this measure compared to MNA. Note that both MNA and MPA have the same mean and variance but MPA has a positive correlation and MNA has a negative correlation. This explains the effect of correlation. Again for a fixed value of β and for a given arrival process, μ_{NS} increases as the variance of the service time distribution increases. It is least for ERS and greatest for HES. However, as β increases beyond a limit P_{EMPTY} and P_{VACN} approach their maximum values. As a result P_{SRT} becomes negligible. Hence no significant change is observed in the value of μ_{NS} in any case.

ILLUSTRATIVE EXAMPLE 5.2: We examine the effect of the parameter β on probability of successful retrials (P_{SRT}) for different arrival and service processes. Again we fix $\lambda = 1$, $\mu = 1.4$, $\eta = 0.5$, $\theta = 0.6$ and get the following graphs;

- Examine figures 5.1, 5.2, 5.3. As β increases P_B increases and hence P_{SRT} decreases. From the figures it is clear that P_{SRT} increases with variance of the inter arrival time distributions. Note that the graphs of MNA and MPA almost coincide for all service time distributions discussed here. This establishes the fact that P_{SRT} does not depend on the correlation of the inter arrival time distributions. MNA and MPA have the greatest variance and they have the greatest value for P_{SRT}.

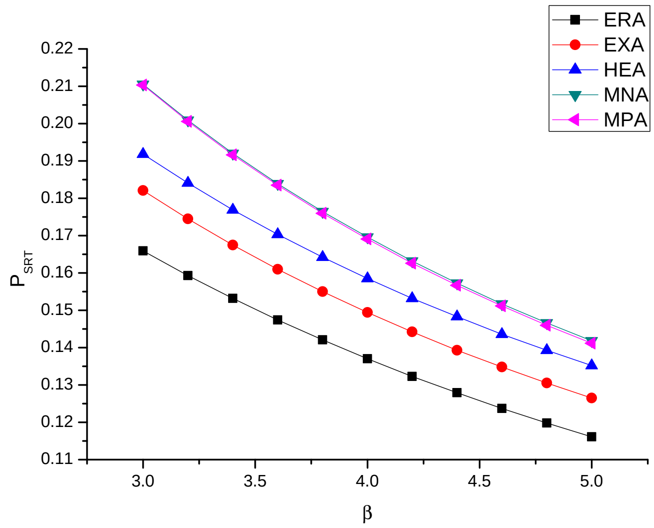

Fig. 5.1: Probability of successful retrials - Erlang services

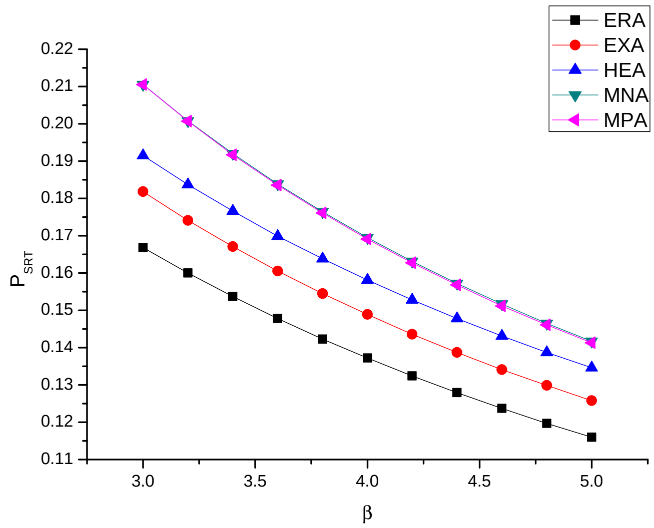

Fig. 5.2: Probability of successful retrials - Exponential services

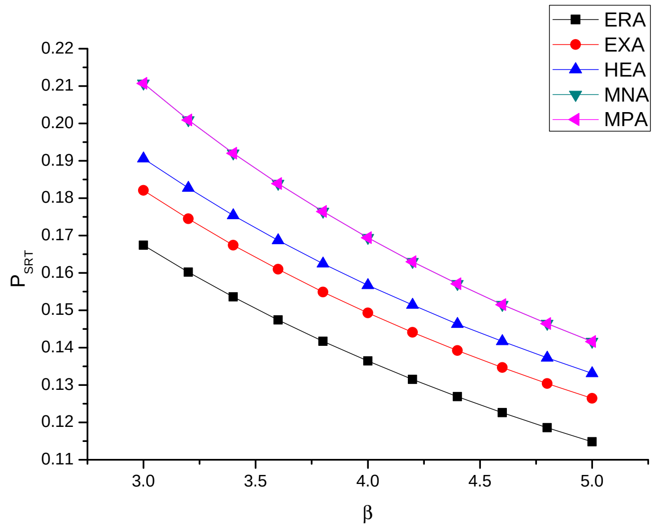

Fig. 5.3: Probability of successful retrials - Hyperexponential services

ILLUSTRATIVE EXAMPLE 5.3: In this example we study the effect of the parameter η on the measure probability of a service completion in slow mode (P_{SCSLO}). Fix $\lambda = 1$, $\mu = 1.4$, $\beta = 3$ and $\theta = 0.6$.

- From the expression for P_{SCSLO}, it is clear that this measure is independent of the inter arrival time distributions and that it decreases as η increases. So we compare the values for P_{SCSLO} for the three service time distributions. Figure 5.4 suggests that P_{SCSLO} increases with the variance of the service time distributions.

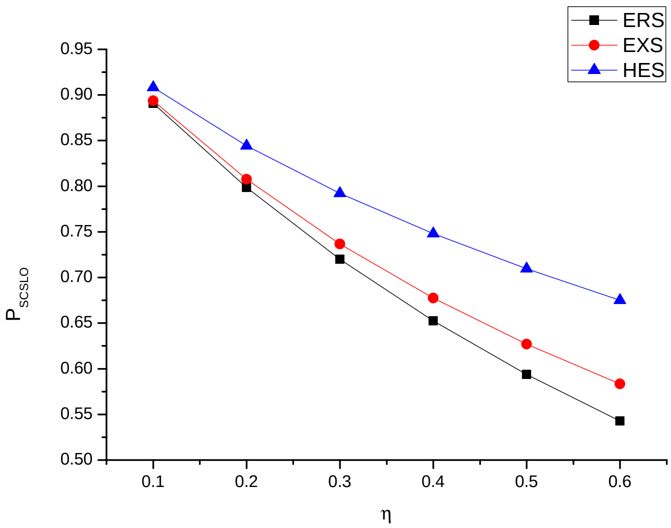

Fig. 5.4: Probability of a service completion in slow mode

6. $MAP/PH/1$ RETRIAL QUEUE WITH CONSTANT RETRIAL RATE, WORKING VACATIONS AND A FINITE BUFFER FOR ARRIVALS

In the last chapter we analyzed a $MAP/PH/1$ retrial queueing model with constant retrial rate and working vacations to the server. In such a model, whenever the server is busy retrial does not make any difference in the state of the system and any primary arrival will be redirected to the orbit. Hence, the objective of minimising the server idle time cannot be achieved beyond a certain extent. Keeping this in mind, we make some changes in the setup of the model discussed in chapter 5. Here we introduce a finite buffer for the customers (primary and orbital), which improves the chance of a customer getting service with reduced waiting time. This also enhances the server utilization to the extent that server in this retrial model has no idle time at all. In practice, we can see many situations which can be modelled like this. Detailed description of the present model is as given below.

Here we consider a single server retrial queueing system in which customers arrive according to a Markovian arrival process (MAP) with parameter matrices D_0 and D_1 of dimension m. An arriving primary customer who finds the server free, immediately occupies the server and starts getting service. On the other hand if the arriving unit finds the server busy it joins a finite buffer of capacity L. If an arrival finds the buffer also full, it moves to an orbit of infinite size. From the orbit the unit makes retrial at the rate β, which is independent of the number of customers in the orbit, for a place in the server or buffer. The service times follow phase type distribution with representation $(\boldsymbol{\alpha}, T)$ of order n. The server takes vacation when the customer being served depart from the system and no customers are left in the buffer. Duration of vacation is exponentially distributed with parameter η. During a vacation if a customer (primary or orbital) arrives, the server returns from vacation. However customers are served during vacation only at a lower rate compared to the regular service. Precisely the vacation mode service times are also phase type distributed with representation $(\boldsymbol{\alpha}, \theta T)$, with $0 < \theta < 1$. Even when the vacation is interrupted by a customer, vacation clock continues to tick so that on completion of this service if the vacation clock has not expired, the server continues on vacation in the absence of a customer in the buffer. At the end of each vacation the server takes another vacation if the buffer is empty.

6.1 The QBD process

The model discussed above can be studied as a QBD process. First, we set up necessary notations. Let μ denote the regular service rate. Then it is easy to verify that $\mu = [\boldsymbol{\alpha}(-T)^{-1}\boldsymbol{e}]^{-1}$. Let $\theta, 0 < \theta < 1$, denote the factor by which the normal service rate is reduced when the server is serving in vacation mode. That is, when the server is serving in vacation mode, the rate of service is given by $\theta\mu$.

At time t, let

$$N_1(t) = \textit{The number of customers in the orbit,}$$

$$N_2(t) = \textit{The number of customers in the buffer,}$$

$$S_1(t) = \begin{cases} 0, & \textit{if the server is not working,} \\ j, & \textit{if the server is busy in phase } j,\ 1 \leq j \leq n, \end{cases}$$

If $S_1(t) \neq 0$, then

$$S_2(t) = \begin{cases} 0, & \textit{if the service is in vacation mode,} \\ 1, & \textit{if the service is in normal mode,} \end{cases}$$

and $M(t)$ to be the phase of the arrival process at time t. It is easy to verify that $\{(N_1(t), N_2(t), S_1(t), S_2(t), M(t)) : t \geq 0\}$ is a level independent QBD process with state space

$$\Omega = \bigcup_{i_1=0}^{\infty} l(i_1)$$

where

$$l(i_1) = \{(i_1, i_2, j_1, j_2, k) : i_1 \geq 0; 0 \leq i_2 \leq L; 0 \leq j_1 \leq n; j_2 = 0 \text{ or } 1; 1 \leq k \leq m\}.$$

Note that when $S_1(t) = 0$, $S_2(t)$ does not play any role and will not be tracked. In this case we need to track only the component $M(t)$.

The generator Q of the QBD process under consideration is of the form

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the (block) matrices appearing in Q are as follows.

$$B_0 = \begin{bmatrix} D_0 & \boldsymbol{\alpha} \otimes D_1 & O & O & O \\ \theta \boldsymbol{T^0} \otimes I & \theta T \oplus D_0 - \eta I & \eta I & C_1 & O \\ \boldsymbol{T^0} \otimes I & O & T \oplus D_0 & O & C_1 \\ O & \boldsymbol{e}_L \otimes \theta \boldsymbol{T^0} \boldsymbol{\alpha} \otimes I & O & C_2 & O \\ O & O & \boldsymbol{e}_L \otimes \boldsymbol{T^0} \boldsymbol{\alpha} \otimes I & O & C_3 \end{bmatrix} \text{ with}$$

$C_1 = \begin{bmatrix} I \otimes D_1 & O \end{bmatrix}$; C_2 has the block matrix $\theta T \oplus D_0$ along the diagonal, $I \otimes D_1$ along the superdiagonal and O matrices elsewhere; and the matrix C_3 has the block matrix $T \oplus D_0$ along the diagonal, $I \otimes D_1$ along the superdiagonal and O matrices elsewhere.

$$A_0 = \begin{bmatrix} O & O & O & O & O \\ O & O & O & O & O \\ O & O & O & O & O \\ O & O & O & \boldsymbol{e}_L(L)\boldsymbol{e}'_L(L) \otimes I \otimes D_1 & O \\ O & O & O & O & \boldsymbol{e}_L(L)\boldsymbol{e}'_L(L) \otimes I \otimes D_1 \end{bmatrix};$$

$$A_1 = \begin{bmatrix} D_0 - \beta I & \boldsymbol{\alpha} \otimes D_1 & O & O & O \\ \theta \boldsymbol{T}^0 \otimes I & F_1 & \eta I & F_4 & O \\ \boldsymbol{T}^0 \otimes I & O & F_2 & O & F_4 \\ O & F_3 & O & E_1 & O \\ O & O & F_5 & O & E_2 \end{bmatrix} ; \text{ where}$$

$F_1 = \theta T \oplus D_0 - \eta I - \beta I$, $F_2 = T \oplus D_0 - \beta I$, $F_3 = \boldsymbol{e}_L \otimes \theta \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I$,

$F_4 = \boldsymbol{e}'_L(1) \otimes I \otimes D_1$, $F_5 = \boldsymbol{e}_L \otimes \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I$.

The matrix E_1 has the block $\theta T \oplus D_0 - \beta I$ along the diagonal, $I \otimes D_1$ along the superdiagonal and O matrices elsewhere. E_2 has $T \oplus D_0 - \beta I$ along the diagonal, $I \otimes D_1$ along the superdiagonal and O blocks elsewhere.

$$A_2 = \begin{bmatrix} O & \beta(\boldsymbol{\alpha} \otimes I) & O & O & O \\ O & O & O & H_1 & O \\ O & O & O & O & H_1 \\ O & O & O & H_2 & O \\ O & O & O & O & H_2 \end{bmatrix} \text{ with}$$

$H_1 = \begin{bmatrix} \beta I & O \end{bmatrix}$; H_2 has the block matrix βI along the superdiagonal and O blocks elsewhere.

6.2 Steady-state analysis

In this section we will discuss the steady-state analysis of the model under study.

6.2.1 Stability condition

Define $A = A_0 + A_1 + A_2$. Let $\boldsymbol{\pi} = (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \boldsymbol{\pi_3}, \boldsymbol{\pi_4})$ be the steady-state probability vector of A, where $\boldsymbol{\pi_0}$ is of dimension m, $\boldsymbol{\pi_1}$, $\boldsymbol{\pi_2}$ are of dimension mn and $\boldsymbol{\pi_3}$ and $\boldsymbol{\pi_4}$ are of dimension Lmn. Also let π_{ij} denote the components of the vector $\boldsymbol{\pi_i}$, $0 \leq i \leq 4$. Note that $\boldsymbol{\pi}$ is the unique vector satisfying the condition $\boldsymbol{\pi} A = \mathbf{0}$ and $\boldsymbol{\pi e} = 1$. For stability of the queueing model we must have $\boldsymbol{\pi} A_0 \boldsymbol{e} < \boldsymbol{\pi} A_2 \boldsymbol{e}$, (see [50]) which simplifies to $(\boldsymbol{\pi_3} + \boldsymbol{\pi_4})\boldsymbol{e}_L(L) \otimes (\boldsymbol{e}_n \otimes \mathrm{D}_1 \boldsymbol{e}_m) < \beta(\boldsymbol{\pi_0} \boldsymbol{e}_m + (\boldsymbol{\pi_1} + \boldsymbol{\pi_2})\boldsymbol{e}_{mn} + \sum_{j=1}^{(L-1)mn}(\pi_{3j} + \pi_{4j}))$. The last inequality suggests that for stability of the queueing system discussed here it is required that the rate of inflow in to the orbit is less than the effective retrial rate.

6.2.2 Steady-state probability vector

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$, be the steady-state probability vector of Q. Note that $\boldsymbol{x}_j$ is of dimension $m + 2mn + 2Lmn$ for $j \geq 0$. The vector $\boldsymbol{x}$ satisfies the condition $\boldsymbol{x}Q = \mathbf{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$. When the stability condition is satisfied the sub vectors of $\boldsymbol{x}$, corresponding to the different level sates are given by the equation $\boldsymbol{x}_j = \boldsymbol{x}_0 R^j, j \geq 1$,where R is the minimal non negative solution of the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{6.1}$$

The sub vector $\boldsymbol{x}_0$ is obtained by solving the equations

$$\boldsymbol{x}_0 (B_1 + R A_2) = \mathbf{0} \tag{6.2}$$

subject to the normalizing condition

$$\boldsymbol{x}_0(I-R)^{-1}\boldsymbol{e}=1. \tag{6.3}$$

The computation of R matrix can be carried out using methods such as logarithmic reduction algorithm.

6.2.3 Key system 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. The measures are listed below along with their formulae for computation.

1. Probability that the orbit is empty:$P_{OTY}=\boldsymbol{x}_0\boldsymbol{e}$.

2. Probability that the buffer is empty:$P_{BUFTY}=\sum_{i_1=0}^{\infty}\boldsymbol{x}_{i_10}\boldsymbol{e}_{m+2mn}$.

3. The probability that the server is on vacation:
   $P_{VACN}=\sum_{i_1=0}^{\infty}\sum_{k=1}^{m}\boldsymbol{x}_{i_100.k}$.

4. The probability that the server is busy in vacation mode:
   $P_{BVM}=\sum_{i_1=0}^{\infty}\sum_{i_2=0}^{L}\sum_{j_1=1}^{n}\sum_{k=1}^{m}\boldsymbol{x}_{i_1i_2j_10k}$.

5. Probability that the server completes a service in vacation mode : $P_{SCSLO}=P$(service time in slow mode $<$ an exponentially distributed random variable with parameter $\eta)=\boldsymbol{\alpha}(\eta I-\theta T)^{-1}\theta\boldsymbol{T}^0$

6. The probability that the server is busy in normal mode:
   $P_{BNM}=\sum_{i_1=0}^{\infty}\sum_{i_2=0}^{L}\sum_{j_1=1}^{n}\sum_{k=1}^{m}\boldsymbol{x}_{i_1i_2j_11k}$.

7. The mean number of customers in the orbit:

   $\mu_{MNOBT} = \sum_{i_1=1}^{\infty} i_1 \boldsymbol{x}_{i_1} \boldsymbol{e} = \boldsymbol{x}_0 R(I-R)^{-2}\boldsymbol{e}$

8. The mean number of customers in the buffer:

   $\mu_{BUF} = \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{L} i_2 \boldsymbol{x}_{i_1 i_2} \boldsymbol{e}_{2mn}$

9. Probability of a successful retrial:

   $P_{SRT} = \beta/(\beta+\lambda) \sum_{i_1=1}^{\infty} \sum_{k=1}^{m} (\sum_{i_2=1}^{L-1} \sum_{j_1=1}^{n} \sum_{j_2=0}^{1} \boldsymbol{x}_{i_1 i_2 j_1 j_2 k} + \boldsymbol{x}_{i_1 00.k})$.

10. Mean number of successful retrials:

    $\mu_{SRT} = \beta/(\beta+\lambda) \sum_{i_1=1}^{\infty} i_1 \sum_{k=1}^{m} (\sum_{i_2=1}^{L-1} \sum_{j_1=1}^{n} \sum_{j_2=0}^{1} \boldsymbol{x}_{i_1 i_2 j_1 j_2 k} + \boldsymbol{x}_{i_1 00.k})$.

6.3 Numerical Results

In order to bring out the qualitative nature of the model under study, we present a few representative examples in this section. For the arrival process we consider the following five sets of matrices for D_0 and D_1.

1. Erlang (ERA)

$$D_0 = \begin{pmatrix} -5 & 5 & & & \\ & -5 & 5 & & \\ & & -5 & 5 & \\ & & & -5 & 5 \\ & & & & -5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 5 & & & & \end{pmatrix}$$

2. Exponential (EXA)

   $D_0 = (-1), D_1 = (1)$

3. Hyperexponential (HEA)

$$D_0 = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 9 & 1 \\ 0.9 & 0.1 \end{pmatrix}$$

4. MAP with negative correlation (MNA)

$$D_0 = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.02 & 0 & 1.98 \\ 445.995 & 0 & 4.505 \end{pmatrix}$$

5. MAP with positive correlation (MPA)

$$D_0 = \begin{pmatrix} -2 & -2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \qquad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.98 & 0 & 0.02 \\ 4.505 & 0 & 445.995 \end{pmatrix}$$

These five MAP processes are qualitatively different in that they have different variance and correlation structure. 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 $ERLA$ are, respectively, 1, 2.2361, 5.0194, 3.1518, and 3.1518.

For the service time distribution we consider the following three phase type distributions.

1. Erlang (ERS)

   $$\boldsymbol{\alpha} = (1, 0) \; T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}$$

2. Exponential (EXS)

   $\boldsymbol{\alpha} = 1.0, T = -1.0$

These two phase type distributions have a service rate of 1. Note that these are qualitatively different in that they have different variances. The ratio of the standard deviation of EXS to that of ERS is 1.4142.

ILLUSTRATIVE EXAMPLE 6.1: We analyze the effect of change in the buffer size on the measure 'probability of successful retrials P_{SRT}', for different arrival and service processes. Figure 6.1 analyzes the effect of the buffer size with Erlang service and figure 6.2 explains its effect with exponential service. We fix $\lambda = 0.9$, $\mu = 1$, $\eta = 0.5$, $\theta = 0.6$ and $\beta = 1$.

- As the buffer size increases more primary arrivals occupy the buffer. This reduces the flow of customers to the orbit and the chance of successful retrial. From the figures it is clear that P_{SRT} increases with variance of the inter arrival time distributions. Note that both MNA and MPA have the same variance but this measure is higher for MPA compared to MNA. Observe that MPA has a positive correlation and MNA has a negative correlation. This shows the effect of correlation on this measure.

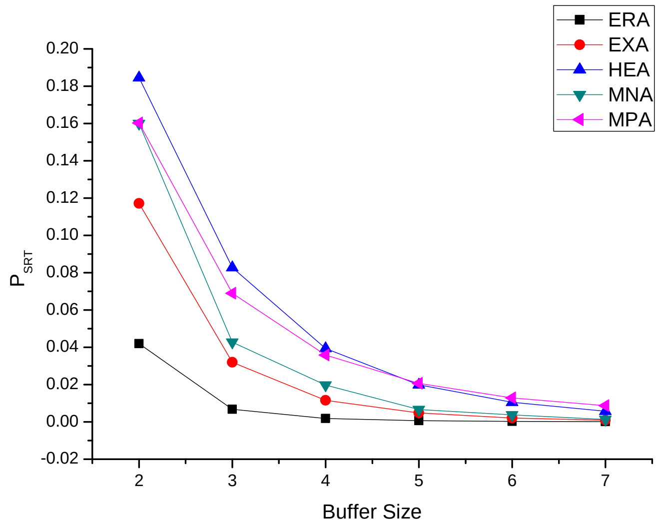

Fig. 6.1: Probability of successful retrials - Erlang services

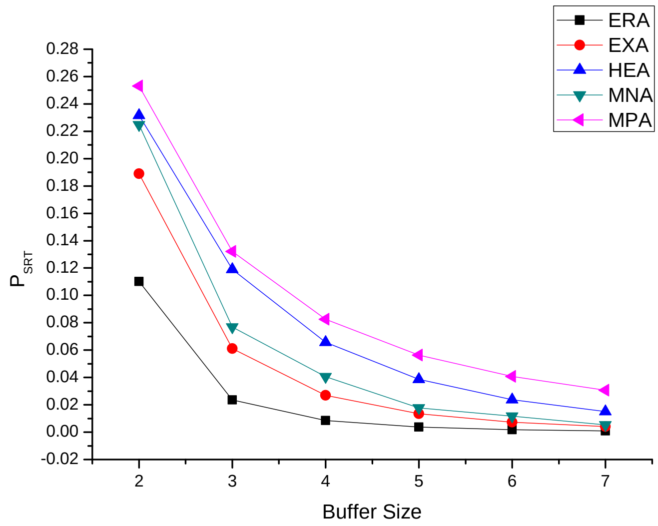

Fig. 6.2: Probability of successful retrials - Exponential services

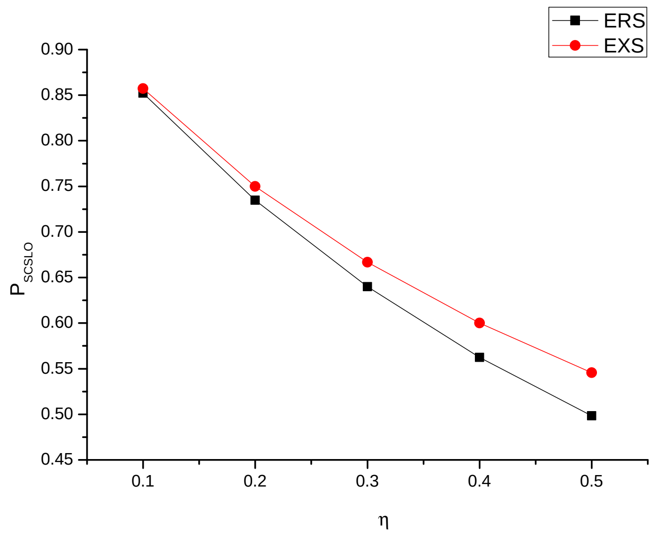

Fig. 6.3: Probability of a service completion in slow mode

ILLUSTRATIVE EXAMPLE 6.2: In this example we study the effect of the parameter η on the measure probability of a service completion in slow mode (P_{SCSLO}). Fix $\lambda = 0.9$, $\mu = 1$, $\beta = 1$ and $\theta = 0.6$ and $L = 3$.

- From the expression for P_{SCSLO}, it is clear that this measure is independent of the inter arrival time distributions and that it decreases as η increases. So we compare the values for P_{SCSLO} for the two service time distributions. From figure 6.3, it is clear that P_{SCSLO} increases with the variance of the service time distributions.

7. $MMAP(2)/PH/1$ RETRIAL QUEUE WITH A FINITE RETRIAL GROUP AND WORKING VACATIONS

In the last two chapters we considered retrial queueing models with working vacations. In these models, only one type of arrivals figured. however, congestions in modern communication and other service systems are very complex and have to be modelled taking all possible aspects into consideration to manage the systems efficiently and economically. Very often, the system will have to deal with different types of arrivals. Some of them require immediate attention (priority) while others could wait till all the more urgent calls are attended. In this chapter we study an $MMAP(2)/PH/1$ retrial queueing model in which the server takes working vacations. There are two types of arrivals, type 1 and type 2. While type 1 customers enjoy infinite waiting space, type 2 have to move to find a place (if any) in an orbit of size L when the server is busy. But once taken for service, a type 2 customer completes his service and leaves the system. In the absence of type 1 customers, sever goes on vacation but returns, when a customer arrives for service.

7.1 Model description

We consider a single server retrial queueing system in which customers arrive according to a marked Markovian arrival process ($MMAP$) with parameter matrices D_0, D_1 and D_2. The matrix D_0 governs transitions without an arrival. D_1 and D_2 respectively contain transition rates with an arrival of class 1 (high priority) and that of class 2 (low priority). $D = D_0 + D_1 + D_2$ is the infinitesimal generator matrix of the arrival process. The matrices D_k, (k=0, 1, 2) are square matrices of order m. Let $\boldsymbol{\delta}(1)$ denote the stationary probability vector of D. The stationary arrival rate of class k (k=1, 2), is given by $\lambda_k = \boldsymbol{\delta}(1)D_k\boldsymbol{e}$. Upon arrival if a customer finds the server free, immediately he occupies the server and obtains service. Type 1 customers have a waiting space of infinite capacity. If a type 2 customer encounters a busy server he proceeds to a group of retrial customers, called orbit. This orbit has only a finite capacity L. When the orbit is full a type 2 arrival proceeding to the orbit is forced to leave the system forever. Inter retrial times are exponentially distributed with parameter γ. The service time for both category of customers follows phase type distribution with representation $(\boldsymbol{\alpha}, T)$ of order n. The server goes on vacation when no priority customer is waiting for service at a departure epoch. Duration of vacation is exponentially distributed with parameter η. During vacation if a customer (primary or orbital) arrives, he interrupts the vacation of the server. The vacation mode service has the same distribution as the regular service time. Even when the vacation is interrupted by a customer vacation clock continues to tick so that on completion of this service if the vacation clock has not expired, the server continues on vacation if there are no priority customers in the system. While in service we do not distinguish the type of customer. From the above description it is clear that, a type 2 customer gets a chance for being served only during vacations.

7.1.1 The QBD process

The model discussed in Section 2 can be studied as a level independent QBD process. It is easy to verify that the service rate μ is given by $\mu = [\boldsymbol{\alpha}(-T)^{-1}\boldsymbol{e}]^{-1}$ and the invariant probability vector of the finite Markov process with generator $T+T^0\boldsymbol{\alpha}$ by $\boldsymbol{\delta}(2) = \mu\boldsymbol{\alpha}(-T)^{-1}$. First, we set up necessary notations.

At time t, let

$N(t)$= *The number of priority customers in the queue and with the server*

$$I(t) = \begin{cases} 0, & \textit{if the server is on vacation }, \\ 1, & \textit{if service is provided during vacation}, \\ 2, & \textit{if the service is regular}, \end{cases}$$

$M(t)$= *The number of customers in the orbit,*

$S(t)$= *Phase of the service process,*

and

$A(t)$= *Phase of the arrival process*

It is easy to verify that $\{(N(t), I(t), M(t), S(t), A(t)) : t \geq 0\}$ is a quasi-birth-and-death process (QBD) with state space

$$\Omega = \bigcup_{i=0}^{\infty} l(i_1)$$

where

$$l(i_1) = \{(i_1, i_2, j_1, j_2, k) : i_1 \geq 0; i_2 = 0, 1 \text{ or } 2; 0 \leq j_1 \leq L; 1 \leq j_2 \leq n; 1 \leq k \leq m\}.$$

The generator matrix Q of the QBD process under consideration is of the form

$$Q = \begin{pmatrix} B_1 & B_0 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the (block) matrices appearing in Q are as follows.

The boundary block B_0 is of order $(L+1)mn \times 2(L+1)mn$ given by

$$B_0 = \begin{bmatrix} B_{01} & & & & & \\ \gamma I & B_{01} & & & & \\ & 2\gamma I & B_{01} & & & \\ & & \ddots & \ddots & & \\ & & & L\gamma I & B_{01} & \end{bmatrix},$$

where $B_{01} = \boldsymbol{\alpha} \otimes (D_1 + D_2)$;

B_1 is a square matrix of order $(L+1)mn$ and is given by

$B_1 = diag(D_0, D_0 - \gamma I, D_0 - 2\gamma I, \ldots, D_0 - L\gamma I)$;

$B_2 = \boldsymbol{e}_2 \otimes I_{L+1} \otimes \boldsymbol{T}^0 \otimes I_m, \quad A_0 = I_{2(L+1)} \otimes I_n \otimes D_1$

and $\quad A_2 = I_{2(L+1)} \otimes \boldsymbol{T}^0\boldsymbol{\alpha} \otimes I_m$;

$$A_1 = \begin{bmatrix} A_{10} & A_{11} & & & & \eta I & & & & \\ & A_{10} & A_{11} & & & & \eta I & & & \\ & & \ddots & \ddots & & & & \ddots & & \\ & & & A_{10} & A_{11} & & & & & \\ & & & & A_{111} & & & & & \eta I \\ & & & & & A_{12} & A_{11} & & & \\ & & & & & & A_{12} & A_{11} & & \\ & & & & & & & \ddots & \ddots & \\ & & & & & & & & A_{12} & A_{11} \\ & & & & & & & & & A_{121} \end{bmatrix},$$

where,

$A_{10} = T \oplus D_0 - \eta I, \quad A_{11} = I \otimes D_2\ , \quad A_{111} = A_{10} + A_{11}, \quad A_{12} = T \oplus D_0$

and $A_{121} = A_{12} + A_{11}$.

7.2 Steady-state analysis

In this section we will discuss the steady-state analysis of the model under study.

7.2.1 Stability condition

Define $A = A_0 + A_1 + A_2$. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, ..., \boldsymbol{\pi}_{2(L+1)})$ be the steady-state probability vector of A, where each $\boldsymbol{\pi}_j$ is of dimension mn. $\boldsymbol{\pi}$ is the unique probability vector which satisfies the conditions $\boldsymbol{\pi} A = \boldsymbol{0}$ and $\boldsymbol{\pi e} = 1$. These equations are equivalent to

$$\boldsymbol{\pi}_1 F = 0 \tag{7.1}$$

$$\boldsymbol{\pi}_i K + \boldsymbol{\pi}_{i+1} F = 0 \qquad 1 \le i \le L-1 \tag{7.2}$$

$$\boldsymbol{\pi}_L K + \boldsymbol{\pi}_{L+1}(F + K) = 0 \tag{7.3}$$

$$\boldsymbol{\pi}_1 K + \boldsymbol{\pi}_{L+2}(F + E) = 0 \tag{7.4}$$

$$\boldsymbol{\pi}_j E + \boldsymbol{\pi}_{j+L} K + \boldsymbol{\pi}_{j+L+1}(F + E) = 0 \qquad 2 \le j \le L \tag{7.5}$$

$$\boldsymbol{\pi}_{L+1} E + \boldsymbol{\pi}_{2L+1} K + \boldsymbol{\pi}_{2(L+1)}(F + E + K) = 0 \tag{7.6}$$

$$\sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i \boldsymbol{e}_{mn} = 1 \tag{7.7}$$

where $F = T \oplus D_0 - \eta I + I \otimes D_1 + \boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I, \quad E = \eta I$ and $\quad K = I \otimes D_2$

Adding equations from (7.1) to (7.6) we get,

$$\sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i (F + E + K) = 0 \tag{7.8}$$

That is

$$\sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i [(T + \boldsymbol{T}^0 \boldsymbol{\alpha}) \oplus (D_0 + D_1 + D_2)] = 0 \tag{7.9}$$

Obviously the steady-state probability vector of $(T + \boldsymbol{T}^0 \boldsymbol{\alpha}) \oplus (D_0 + D_1 + D_2)$

is $\boldsymbol{\delta}(2) \otimes \boldsymbol{\delta}(1)$. In view of equations (7.7) and (7.9) and by the uniqueness of the steady-state probability vector we get,

$$\sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i = \boldsymbol{\delta}(2) \otimes \boldsymbol{\delta}(1) \tag{7.10}$$

$\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots\ldots, \boldsymbol{\pi}_{2(L+1)}$ can be obtained recursively from equations (7.1) to (7.7). Note that

$$\boldsymbol{\pi} A_0 \boldsymbol{e} = \sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i (I \otimes D_1) \boldsymbol{e}_{mn} = (\boldsymbol{\delta}(2) \otimes \boldsymbol{\delta}(1))(\boldsymbol{e}_n \otimes D_1 \boldsymbol{e}_m) = \boldsymbol{\delta}(1) D_1 \boldsymbol{e}_m = \lambda_1$$

$$\boldsymbol{\pi} A_2 \boldsymbol{e} = \sum_{i=1}^{2(L+1)} \boldsymbol{\pi}_i (\boldsymbol{T}^0 \boldsymbol{\alpha} \otimes I) \boldsymbol{e}_{mn} = (\boldsymbol{\delta}(2) \otimes \boldsymbol{\delta}(1))(\boldsymbol{T}^0 \otimes \boldsymbol{e}_m) = \boldsymbol{\delta}(2) \boldsymbol{T}^0 = \mu$$

Thus for the stability of the queueing model it is necessary and sufficient that $\lambda_1 < \mu$.

7.2.2 Steady-state probability vector

Let $\boldsymbol{x}$, partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots)$, be the steady-state probability vector of Q. Note that $\boldsymbol{x}_0$ is of dimension $(L+1)m$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ are of dimension $2(L+1)mn$. $\boldsymbol{x}$ satisfies the conditions $\boldsymbol{x}Q = \boldsymbol{0}$ and $\boldsymbol{x}\boldsymbol{e} = 1$. Apparently when the stability condition is satisfied the sub vectors of $\boldsymbol{x}$, except $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, corresponding to the different level states are given by the equation $\boldsymbol{x}_j = \boldsymbol{x}_1 R^{j-1},\ j \geq 2$, where R is the minimal non negative solution of the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{7.11}$$

The sub vectors $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are obtained by solving the equations

$$\boldsymbol{x}_0 B_1 + \boldsymbol{x}_1 B_2 = \mathbf{0} \tag{7.12}$$

$$\boldsymbol{x}_0 B_0 + \boldsymbol{x}_1 (A_1 + RA_2) = \mathbf{0} \tag{7.13}$$

subject to the normalizing condition

$$\boldsymbol{x}_0 \boldsymbol{e}_{(L+1)m} + \boldsymbol{x}_1 (I - R)^{-1} \boldsymbol{e}_{2(L+1)mn} = 1. \tag{7.14}$$

The computation of R matrix can be carried out using a number of well known methods such as logarithmic reduction algorithm.

7.2.3 Stationary waiting time of a priority customer in the queue

First note that an arriving type 1 customer enters service immediately with probability $z_0 = \boldsymbol{x}_0 \boldsymbol{e}$. Thus with probability $1 - z_0$ he has to wait before getting into service. Let $\boldsymbol{z}_{i,j}$ denote the steady-state probability that an arrival will find the server busy with the service in phase j and the number of customers in the system including the current arrival is i, for $1 \le j \le n; i \ge 2$. Define $\boldsymbol{z}_i = (\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,n})$ and $\boldsymbol{z} = (z_0, \boldsymbol{z}_2, \boldsymbol{z}_3 \ldots)$. It is easy to verify that

$$\boldsymbol{z}_i = \boldsymbol{x}_{i-1} (\boldsymbol{e}_{2(L+1)} \otimes I \otimes \frac{D_1}{\lambda_1} \boldsymbol{e}_m), \quad i \ge 2. \tag{7.15}$$

The waiting time may be viewed as the time until absorption in a Markov chain with a highly sparse structure. The state space (that includes the arriving customer in its count) of this Markov chain is given by $\Omega_1 = \{*\} \cup$

$\{(i, j) : i \geq 2; 1 \leq j \leq n\}$. The state $*$ corresponds to the server being on vacation when the customer arrives. Note that once a customer joins the queue, the subsequent arrivals do not contribute to his waiting time. Hence we do not consider the arrival process while computing the waiting time. The generator matrix of this Markov chain is

$$\tilde{Q} = \begin{pmatrix} 0 & O & & & & \\ \boldsymbol{T}^0 & T & & & & \\ & \boldsymbol{T}^0\boldsymbol{\alpha} & T & & & \\ & & \boldsymbol{T}^0\boldsymbol{\alpha} & T & & \\ & & & \ddots & \ddots & \end{pmatrix}$$

Define $W(t)$ for $t > 0$ as the probability that an arriving customer enters into service no later than time t. Let $\tilde{W}(s)$ denote the Laplace Stieltjes transform of the stationary waiting time in the queue of an arriving customer. Using the structure of $\tilde{Q}$ it can readily be verified that

$$\tilde{W}(s) = c \sum_{i=2}^{\infty} \boldsymbol{z}_i (sI - T)^{-1} \boldsymbol{T}^0 [\boldsymbol{\alpha}(sI - T)^{-1} \boldsymbol{T}^0]^{i-2}, \quad Re(s) \geq 0,$$

where the normalizing constant c is given by

$$c = \left[\sum_{i=2}^{\infty} \boldsymbol{z}_i \boldsymbol{e} \right]^{-1}. \tag{7.16}$$

The mean waiting time, μ'_W in the queue of an arrival, finding the server busy, is calculated as

$$\mu_W' = -\tilde{W}'(0) = c\sum_{i=2}^{\infty} \boldsymbol{z}_i(-T)^{-1}\boldsymbol{e} + \frac{c}{\mu}\sum_{i=2}^{\infty}(i-2)\boldsymbol{z}_i\boldsymbol{e}. \tag{7.17}$$

The equations (7.15) and (7.17) lead us to

$$\mu_W' = \frac{c}{\lambda}[\boldsymbol{x}_1(I-R)^{-1}(\boldsymbol{e}_{2(L+1)}\otimes(-T)^{-1}\boldsymbol{e}_n\otimes D_1\boldsymbol{e}_m) + \frac{1}{\mu}\boldsymbol{x}_1R(I-R)^{-2}(\boldsymbol{e}\otimes D_1\boldsymbol{e}_m)], \tag{7.18}$$

where $\boldsymbol{e}$ in equation 7.18 is of dimension $2(L+1)n$.

7.2.4 The uninterrupted duration of a vacation

The duration of the time the server is in uninterrupted vacation(s) is the interval between the epoch at which the server goes on vacation and the next arrival epoch. Clearly this duration is of phase type with representation $(\boldsymbol{\beta}, B_1)$ of dimension $(L+1)m$, where $\boldsymbol{\beta} = d\boldsymbol{x}_1B_2$ and the normalizing constant d is given by $d = [\boldsymbol{x}_1B_2\boldsymbol{e}]^{-1}$. Hence the mean duration of uninterrupted vacation, $\mu_{UIV} = \boldsymbol{\beta}(-B_1)^{-1}\boldsymbol{e}$. But $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots.., \boldsymbol{\beta}_L)$, each $\boldsymbol{\beta}_i$ being of dimension m. Exploiting the structre of the matrix B_1 we can simplify the expression for μ_{UIV} as $\mu_{UIV} = \sum_{i=0}^{L}\boldsymbol{\beta}_i[-(D_0 - i\gamma)]^{-1}\boldsymbol{e}_m$.

7.2.5 Busy period analysis

In this section we analyze the structure of a busy period of the model discussed in section 7.1. A busy period is the interval between the arrival of a customer to the empty system and the first epoch thereafter the system becomes empty again. Thus it is the first passage time from level 1 to level

0. A busy cycle is defined as the first return time to level 0 with at least one visit to a state in any other level. Before analyzing busy period we need to introduce the notion of fundamental period. For the QBD process under consideration it is the first passage time from level i to level $i-1$, $i \geq 2$. The cases $i = 1, 0$ corresponding to the boundary states need to be discussed separately. Note that for each level i, $i \geq 1$, there corresponds $2(L+1)mn$ states. Thus by the state (i, j) of level i we mean the j^{th} state of level i when the states are arranged in the lexicographic order. Let $G_{jj'}(h, x)$ denote the conditional probability that starting in the state (i, j) at time $t = 0$, the QBD process visits the level $i-1$, for the first time no later than time x, after exactly h transitions to the left and does so by entering the state $(i-1, j')$.

For convenience we introduce the joint transform

$$\tilde{G}_{jj'}(z, s) = \sum_{h=1}^{\infty} z^h \int_0^{\infty} e^{-sx} \mathrm{d}G_{jj'}(h, x) \qquad ; \quad |z| \leq 1, Re(s) \geq 0$$

and the matrix

$$\tilde{G}(z, s) = (\tilde{G}_{jj'}(z, s)).$$

The matrix $\tilde{G}(z, s)$ satisfies the equation (see [50])

$$\tilde{G}(z, s) = z(sI - A_1)^{-1} A_2 + (sI - A_1)^{-1} A_0 \tilde{G}^2(z, s). \tag{7.19}$$

The matrix $G = (G_{jj'}) = \tilde{G}(1, 0)$ takes care of the first passage times, except for the boundary states. If we know the R matrix then G matrix can be computed using the result (see [39])

$$G = -(A_1 + RA_2)^{-1}A_2.$$

Otherwise we may use logarithmic reduction method to compute G. For the boundary level states 1 and 0 let $G_{jj'}^{(1,0)}(h,x)$ and $G_{jj'}^{(0,0)}(h,x)$ be the conditional probability discussed above for the first passage times from level 1 to level 0 and the first return time to the level 0 respectively. For the boundary levels 1 and 0 we get

$$\tilde{G}^{(1,0)}(z,s) = z(sI - A_1)^{-1}B_2 + (sI - A_1)^{-1}A_0\tilde{G}(z,s)\tilde{G}^{(1,0)}(z,s) \qquad (7.20)$$

and

$$\tilde{G}^{(0,0)}(z,s) = (sI - B_1)^{-1}B_0\tilde{G}^{(1,0)}(z,s) \qquad (7.21)$$

Since the first passage time from level i to level $i-1$ is independent of i, we may conveniently use the following notations.

Let m_{1j} be the mean first passage time from the level i to level $i-1$, given that the process is in the state (i,j) at time $t=0$. Also let $\tilde{\boldsymbol{m}}_{\mathbf{1}}$ be the column vector with entries m_{1j}. Let m_{2j} be the mean number of customers served during the first passage time from level i to level $i-1$, given that the first passage time started in the state (i,j) and $\tilde{\boldsymbol{m}}_{\mathbf{2}}$ be the column vector with elements m_{2j}. Then

$$\tilde{\boldsymbol{m}}_{\mathbf{1}} = -\left.\frac{\partial \tilde{G}(z,s)}{\partial s}\right|_{z=1,s=0} \boldsymbol{e} = -(A_1 + A_0(I+G))^{-1}\boldsymbol{e}$$

and

$$\tilde{\boldsymbol{m}}_{\boldsymbol{2}} = \left.\frac{\partial \tilde{G}(z,s)}{\partial z}\right|_{z=1,s=0} \boldsymbol{e} = -(A_1 + A_0(I+G))^{-1} A_2 \boldsymbol{e}$$

Similar to $\tilde{\boldsymbol{m}}_{\boldsymbol{1}}$ and $\tilde{\boldsymbol{m}}_{\boldsymbol{2}}$ we define ${\tilde{\boldsymbol{m}}_{\boldsymbol{1}}}^{(1,0)}$ and ${\tilde{\boldsymbol{m}}_{\boldsymbol{2}}}^{(1,0)}$ respectively, to be the vectors giving the mean first passage times from level 1 to level 0 and mean number of service completions during this first passage times. Vectors ${\tilde{\boldsymbol{m}}_{\boldsymbol{1}}}^{(0,0)}$ and ${\tilde{\boldsymbol{m}}_{\boldsymbol{2}}}^{(0,0)}$ respectively give the first return times to level 0 and the mean umber of service completions during these return times. Stochastic nature of the matrices G, $\tilde{G}^{(1,0)}(1,0)$ and $\tilde{G}^{(0,0)}(1,0)$ enables us to compute

$${\tilde{\boldsymbol{m}}_{\boldsymbol{1}}}^{(1,0)} = -\left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial s}\right|_{z=1,s=0} \boldsymbol{e} = -(A_1 + A_0 G)^{-1}(A_0 \tilde{\boldsymbol{m}}_{\boldsymbol{1}} + \boldsymbol{e})$$

$${\tilde{\boldsymbol{m}}_{\boldsymbol{1}}}^{(0,0)} = -\left.\frac{\partial \tilde{G}^{(0,0)}(z,s)}{\partial s}\right|_{z=1,s=0} \boldsymbol{e} = -B_1^{-1}(B_0 {\tilde{\boldsymbol{m}}_{\boldsymbol{1}}}^{(1,0)} + \boldsymbol{e})$$

$${\tilde{\boldsymbol{m}}_{\boldsymbol{2}}}^{(1,0)} = \left.\frac{\partial \tilde{G}^{(1,0)}(z,s)}{\partial z}\right|_{z=1,s=0} \boldsymbol{e} = -(A_1 + A_0 G)^{-1}(A_0 \tilde{\boldsymbol{m}}_{\boldsymbol{2}} + B_2 \boldsymbol{e})$$

$${\tilde{\boldsymbol{m}}_{\boldsymbol{2}}}^{(0,0)} = \left.\frac{\partial \tilde{G}^{(0,0)}(z,s)}{\partial z}\right|_{z=1,s=0} \boldsymbol{e} = -B_1^{-1} B_0 {\tilde{\boldsymbol{m}}_{\boldsymbol{2}}}^{(1,0)}$$

7.2.6 Key system 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. The measures are listed below along with their formulas for computation.

1. Probability that the server is on vacation: $P_{VACN} = \boldsymbol{x}_0 \boldsymbol{e}$.

2. Probability that the server is working in vacation :
$P_{BV} = \sum_{i_1=1}^{\infty} \sum_{j_1=0}^{L} \sum_{j_2=1}^{n} \sum_{k=1}^{m} \boldsymbol{x}_{i_1 1 j_1 j_2 k}.$

3. Probability that the server is busy: $P_B = \sum_{i_1=1}^{\infty} \boldsymbol{x}_{i_1} \boldsymbol{e}_{2(L+1)mn}.$

4. Mean number of type 1 customers in the system:
$\mu_{NS} = \sum_{i_1=1}^{\infty} i_1 \boldsymbol{x}_{i_1} \boldsymbol{e}_{2(L+1)mn} = \boldsymbol{x}_1 (I-R)^{-2} \boldsymbol{e}.$

5. Mean number of customers in the orbit:
$\mu_{OBT} = \sum_{j_1=1}^{L} j_1 \boldsymbol{x}_{00j_1.} \boldsymbol{e}_m + \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{2} \sum_{j_1=1}^{L} j_1 \boldsymbol{x}_{i_1 i_2 j_1 j_2 k} \boldsymbol{e}_{2mn}$

6. Probability that the orbit is full :$P_F = P_{FV} + P_{FB}$, where
$P_{FV} = \sum_{k=1}^{m} \boldsymbol{x}_{00L.k}$ and $P_{FB} = \sum_{i_1=1}^{\infty} \sum_{i_2=1}^{2} \sum_{j_2=1}^{n} \sum_{k=1}^{m} \boldsymbol{x}_{i_1 i_2 L j_2 k}.$

7. Probability that a type 2 customer is lost :$P_{LOST} = P_{FB} \frac{\lambda_2}{(\lambda_1+\lambda_2)}$

8. Probability of a successful retrial : $PSRT = \sum_{i=1}^{L} \sum_{k=1}^{m} \boldsymbol{x}_{00i.k} \frac{i\gamma}{(\ i\gamma+\lambda_1+\lambda_2)}.$

9. Mean number of successful retrials : $\mu_{SRT} = \sum_{i=1}^{L} \sum_{k=1}^{m} i \boldsymbol{x}_{00i.k} \frac{i\gamma}{(\ i\gamma+\lambda_1+\lambda_2)}.$

7.3 Numerical Results

For the arrival process we consider the following five sets of matrices for D_0, D_1 and D_2.

1. Erlang (ERA)

$$D_0 = \begin{pmatrix} -5 & 5 & & & \\ & -5 & 5 & & \\ & & -5 & 5 & \\ & & & -5 & 5 \\ & & & & -5 \end{pmatrix} \quad D_1 = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 3 & & & & \end{pmatrix} \quad D_2 = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ 2 & & & & \end{pmatrix}$$

2. Exponential (EXA)

$$D_0 = (-1), \quad D_1 = (0.6), \quad D_2 = (0.4)$$

3. Hyperexponential (HEA)

$$D_0 = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix} \quad D_1 = \begin{pmatrix} 5.4 & 0.6 \\ 0.54 & 0.06 \end{pmatrix} \quad D_2 = \begin{pmatrix} 3.6 & 0.4 \\ 0.36 & 0.04 \end{pmatrix}$$

4. MAP with negative correlation (MNA)

$$D_0 = \begin{pmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \quad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.012 & 0 & 1.188 \\ 267.597 & 0 & 2.703 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0.008 & 0 & 0.792 \\ 178.398 & 0 & 1.802 \end{pmatrix}$$

5. MAP with positive correlation (MPA)

$$D_0 = \begin{pmatrix} -2 & -2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{pmatrix} \quad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.188 & 0 & 0.012 \\ 2.703 & 0 & 267.597 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0.792 & 0 & 0.008 \\ 1.802 & 0 & 178.398 \end{pmatrix}$$

These five MAP processes are qualitatively different in that they have different variance and correlation structure. 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.

For the service time distribution we consider the following two phase type distributions.

1. Erlang (ERS)

$$\boldsymbol{\alpha} = (1, 0) \; T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}$$

2. Hyperexponential (HES)

$$\boldsymbol{\alpha} = (0.9, 0.1) \; T = \begin{pmatrix} -1.90 & 0 \\ 0 & -0.19 \end{pmatrix}$$

The above two distributions will be normalized to have a specific mean in our illustrative example. Note that that these are qualitatively different in that they have different variances. The ratio of the standard deviation of HES to that of ERS is 3.1745.

ILLUSTRATIVE EXAMPLE: 7.1 We fix $\lambda_1 = 9$, $\lambda_2 = 6$, $\mu = 10$, $L = 6$, $\gamma = 2$ and let η vary. We analyze how the change in η affects some system performances. First let us examine its effect on the measure P_{LOST}. We make the following observations from the figures 7.1 and 7.2.

- As the value of η increases the mean duration of vacation decreases and hence the server clears out more customers. Hence the measure P_{LOST} tends to decrease as expected. Among the renewal arrival processes the value of P_{LOST} is maximum for HEA and minimum for ERA. Note that HEA has the greatest variance and ERA has the least variance among these arrival processes. Among the correlated arrival processes the value of this measure is greater for MPA compared to that of MNA. This shows the effect of standard deviation among the renewal processes and the effect of correlation among the correlated arrival processes. These arguments applicable to Erlang and hyperexponential services though the measure has a slightly higher values for hyperexponential services.

- Next let us discuss how long vacation remains uninterrupted on the average. We let η vary keeping other parameters fixed, exactly as before. Figures 7.3 and 7.4 suggests that larger the value of η smaller the probability of a vacation being interrupted. Hence the value of the measure μ_{UIV} increases as η increases. This is the case with both Erlang and hyperexponential services. The role played by the standard deviation and correlation of the arrival processes is obvious here as well.

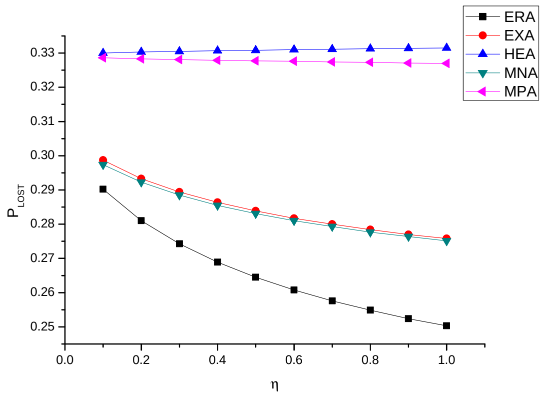

Fig. 7.1: P_{LOST} when the service is Erlang

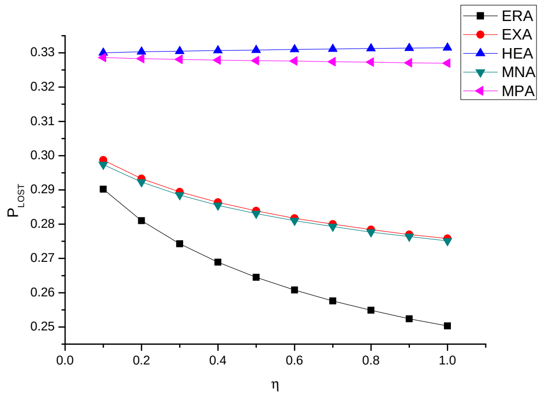

Fig. 7.2: P_{LOST} when the service is hyperexponential

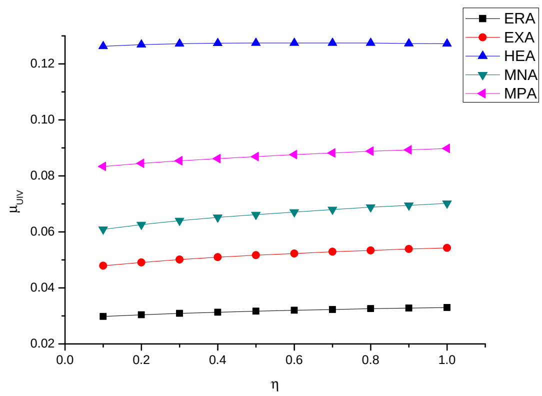

Fig. 7.3: μ_{UIV} when the service is Erlang

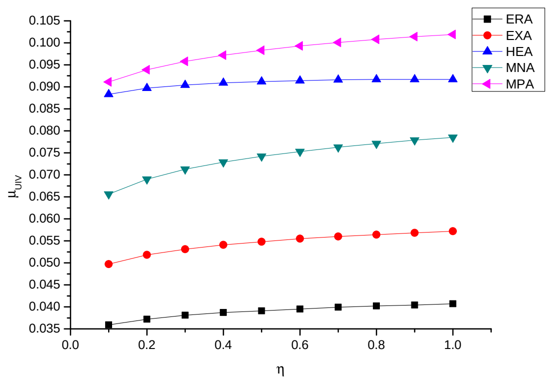

Fig. 7.4: μ_{UIV} when the service is hyperexponential

- We now examine how the waiting time is affected by η, the values of other parameters being fixed as we did earlier (see tables 7.1 and 7.2). As η increases the probability of a vacation being interrupted by an arrival (primary or orbital) decreases. This results in a decrease in mean waiting time in the queue though by a small quantity. It can be seen from the tables that among renewal arrivals the mean waiting time increases as we move from Erlang to hyperexponential through exponential arrival process. This shows the effect of standard deviation of the renewal arrivals in the mean waiting time. Though MNA and MPA have the same standard deviation MNA has a negative correlation and MPA has a positive correlation. Mean waiting time has a very high value for MPA compared to that of MNA. This shows the effect of correlation on the mean waiting time. Note that the entries in the second table are higher than the corresponding entries in the first table. This shows the effect of the standard deviation of service processes on the mean waiting time.

Mean waiting time in the queue

Table 7.1: With Erlang Service :

η	ERA	EXA	HEA	MNA	MPA
0.2	0.889	1.2888	1.6018	1.1803	27.7819
0.4	0.8309	1.2275	1.5991	1.1611	27.6995
0.6	0.7936	1.1882	1.5976	1.1436	27.6311
0.8	0.767	1.1605	1.5964	1.1303	27.5717
1	0.7471	1.1399	1.5954	1.1203	27.519

Table 7.2: With hyperexponential Service:

η	ERA	EXA	HEA	MNA	MPA
0.2	3.8917	4.1159	5.2179	4.117	31.5975
0.4	3.8217	4.0495	5.1888	4.0429	31.3926
0.6	3.7763	4.007	5.173	3.9945	31.2629
0.8	3.7427	3.976	5.1633	3.9584	31.1698
1	3.7164	3.952	5.157	3.9298	31.0981

CONCLUSION AND FUTURE WORK

The objective of the study of “Queueing models with vacations and working vacations” was two fold; to minimize the server idle time and improve the efficiency of the service system. Keeping this in mind we considered queueing models in different set up in this thesis.

Chapter 1 introduced the concepts and techniques used in the thesis and also provided a summary of the work done. In chapter 2 we considered an $M/M/2$ queueing model, where one of the two heterogeneous servers takes multiple vacations. We studied the performance of the system with the help of busy period analysis and computation of mean waiting time of a customer in the stationary regime. Conditional stochastic decomposition of queue length was derived. To improve the efficiency of this system we came up with a modified model in chapter 3. In this model the vacationing server attends the customers, during vacation at a slower service rate. Chapter 4 analyzed a working vacation queueing model in a more general set up. The introduction of N policy makes this $MAP/PH/1$ model different from all working vacation models available in the literature. A detailed analysis of performance of the model was provided with the help of computation of measures such as mean waiting time of a customer who gets service in normal mode and vacation mode.

Recognizing the importance of systems with repeated attempts, a retrial queueing system with working vacation was introduced in chapter 5, again with MAP arrivals and PH service. A minor draw back of this model was that the server had to remain idle (when not on vacation) in the system, when there was no demand for service. In chapter 6 we overcame this handicap by introducing a finite buffer for arrivals (primary as well as orbital). This brought down the server idle time to zero. In chapter 7 we considered a more versatile retrial model, with two different types of arrivals. This $MMAP(2)/PH/1$ model offered an infinite buffer for high priority customers and forced a low priority arrival to join a finite retrial group, when met with a busy server. The performance of the model was analyzed computing measures such as mean waiting time of a high priority customer.

It should be remarked that over the years single server working vacation models have been studied extensively. Though we considered a two server working vacation model in chapter 3, only one server takes working vacations in that model. It would indeed be challenging task to analyze the multiserver working vacation models.

BIBLIOGRAPHY

[1] *Alfa, A.S. and Sapna Isotupa, K.P.*: An M/PH/k retrial queue with finite number of sources, Computers and Operations Research, 31, 1455-1464, 2004.

[2] *Artalejo, J.R.*: Accessible bibliography on retrial queues, Mathematical and Computer Modelling, 30, 1-6, 1999.

[3] *Artalejo, J.R.*: A Classified bibliography of research on retrial queues : Progress in 1990-1999, TOP, 7, 187-211, 1999.

[4] *Artalejo, J.R., Joushua, V.C. and Krishnamoorthy, A.*: An M/G/1 retrial queue with orbital search by the server, in: Advances in Stochastic Modelling, eds., J.R. Artalejo and A. Krishnamoorthy, 41-54, Notable Publications, New Jersey, 2002.

[5] *Artalejo, J.R. and Chakravarthy, S.R.*: Algorithmic Analysis of the MAP/PH/1 retrial queue, Top, 14(2), 293-332, 2006.

[6] *Artalejo, J.R. and Gomez-Corral, A.* : Retrial Queueing Systems: A Computational Approach, Springer Verlag, 2008.

[7] *Artalejo, J.R. and Lopez-Herrero, M.J.*: The single server retrial queue with finite population; a BSDE approach, Operational Research: An International Journal, 12, 109-131, 2012.

[8] *Baba, Y.* : Analysis of a GI/M/1 queue with multiple working vacations, Operations Research Letters, 33, 201-209, 2005.

[9] *Breuer, L. and Baum, D.* : An Introduction to Queueing Theory and Matrix-Analytic Methods, Springer, 2005.

[10] *Chakravarthy, S.R.*: The Batch Markovian Arrival Process: A Review and Future Work, in: Advances in Probability Theory and Stochastic Processes, Eds., A. Krishnamoorthy et al., 21-49, 2001, Notable Publications, Inc., New Jersey.

[11] *Chakravarthy, S.R.*: Markovian arrival processes. Wiley Encyclopedia of Operations Research and Management Science, 15 JUN 2010.

[12] *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, 776-798, 2006.

[13] *Chao, X. and Zhao, Y.* : Analysis of Multiserver Queues with Station and Server Vacations, European Journal of Operational Research, 110, 392-406, 1997.

[14] *Cinlar, E.* : Introduction to stochastic processes, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1975.

[15] *Choi, B.D and Chang, Y.* : $MAP_1, MAP_2/M/c$ Retrial Queue with the retrial group of finite finite capacity and geometric loss, Mathematical and Computer Modelling, 30, 99-113, 1999.

[16] *Diamond, J.E. and Alfa, A.S.*: The MAP/PH/1 Retrial Queue, Stochastic Models, 14, 1151-1177, 1998.

[17] *Diamond, J.E. and Alfa, A.S.*: Matrix analytic methods for a multi-server retrial queue with buffer, Top, 7(2), 249-266.

[18] *Do, T.V.*: M/M/1 retrial queue with working vacations, Acta Informatica, 47, 67-75, 2010.

[19] *Doshi, B.T.* : Queueing systems with vacations- a survey, Queueing Systems, 1, 29-66, 1986.

[20] *Dudin, A.N., Krishnamoorthy, A., Joushua, V.C and Tsarenkov, G.V.*: Analysis of the BMAP/G/1 retrial system with search of customers from the orbit, European Journal of Operational Research, 157, 169-179, 2004.

[21] *Falin, G.I.*: A Survey of retrial queues, Queueing Systems, 7, 127-168, 1990.

[22] *Falin, G.I. and Templeton, J.G.C.*: Retrial Queues, Chapman and Hall, London, 1997.

[23] *Gomez-Corral, A., Krishnamoorthy, A. and Viswanath C Narayanan*: Self generation of Priorities on Multiserver queues with Finite Capacity, Stochastic Models, 21, 427-447, 2005.

[24] *Graham, A.*: Kronecker Products and Matrix calculus with Applications, Ellis Horwood Ltd., Chichester, 1981.

[25] *Gross, D. and Harris, C.M.* : Fundamentals of Queueing theory, Third Edition, John Wiley and sons, Inc., 1998.

[26] *He, Q.M. and Alpha, A.S.*: The MMAP[K]/PH[K]/1 queue with a last-come-first-served preemptive service discipline, Queueing Systems, 28, 269-291, 1998.

[27] *Jaiswal, N.K.*: Priority queues, Academic Press, New York, 1968.

[28] *Kao, E.P.C. and Narayanan, K.S.* : Analysis of an M/M/N Queue with Servers' Vacations, European Journal of Operational Research, 54(2), 256-266, 1991.

[29] *Karlin, S. and Taylor, H.M.* : A first course in stochastic processes, Academic press, Inc., 1975.

[30] *Karlin, S. and Taylor, H.M.* : A second course in stochastic processes, Academic press, Inc., 1981.

[31] *Kemeny, J.G. and Snell, J.L.* : Finite Markov Chains, The University Series in Undergraduate Mathematics, D. Van Nostrand, Princeton, NJ, USA, 1960.

[32] *Kim, J., Choi, D. and Chae, K.* : Analysis of queue-length distribution of the M/G/1 queue with working vacations, International Conference on Statistics and Related Fields, Hawaii, 2003.

[33] *Krishna Kumar, B. and Pavai Madheswari, S.* : An M/M/2 Queueing System with Heterogeneous Servers and Multiple Vacations, Mathematical and Computer Modelling, 41, 1415-1429, 2005.

[34] *Krishnamoorthy, A., Deepak, T.G. and Joushua, V.C.*: An M/G/1 Retrial Queue with Non persistent Customers and Orbital Search, Stochastic Analysis and Applications, 23 , 975-997, 2005.

[35] *Krishnamoorthy, A., Babu, S. and Viswanath C Narayanan* : The MAP/(PH/PH)/c queue with self generation of priorities and non-preemptive service, Stochastic Analysis and Applications, 26, 1250-1266, 2008.

[36] *Krishnamoorthy, A., Babu, S. and Viswanath C Narayanan*: The MAP/(PH/PH)/1 queue with self generation of priorities and non-preemptive service, European Journal of Oprerational Research, 195, 174-185, 2009.

[37] *Krishnamoorthy, A., Pramod, P.K. and Chakravarthy, S.R.*: Queues with interruptions - a survey : To appear in Top, 2012.

[38] *Langaris, C.*: Waiting time analysis of a two stage queueing system with priorities, Queueing Systems, 14, 457-473, 1993.

[39] *Latouche, G. and Ramaswami, V.* : Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, 1999.

[40] *Levy, Y. and Yechiali, U.*: Utilization of idle time in an M/G/1 queueing system, Management Science, 22, 202-211, 1975.

[41] *Levy, Y. and Yechiali, U.* : An M/M/S queue with servers' vacations, Information Systems and Operational Research, 14(2), 153-163, 1976.

[42] *Li, J. and Tian, N.* : The $M/M/1$ queue with working vacations and vacation interruptions, Journal of Systems Science and Systems Engineering, 16(1), 121-127, 2007.

[43] *Lucantoni, D.M.*: New results on the single server queue with a batch Markovian arrival process, Stochastic Models, 7, 1-46, 1991.

[44] *Marcus, M. and Minc, H.*: A Survey of Matrix Theory and Matrix Inequalities, Allyn and Bacon, Boston, MA, 1964.

[45] *Miller, R.G.*: Priority queues, Annals of Mathematical Statistics, 31, 86-103, 1960.

[46] *Neuts, M.F.*: Probability distributions of phase type. Liber Amicorum Professor Emeritus H. Florin, University of Louvain, Belgium, 173-206, 1975.

[47] *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.

[48] *Neuts, M.F.* : A versatile markovian point process, Journal of Applied Probability, 16, 764-779, 1979.

[49] *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.

[50] *Neuts, M.F.* : Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, MD, 1981. [1994 version is Dover Edition]

[51] *Neuts, M.F.*: Structured Stochastic Matrices of M/G/1 type and their Applications, Marcel Dekker, NY, 1989.

[52] *Neuts, M.F.*: Models based on the Markovian arrival process, IEICE Transactions on Communications, E75B, 1255-1265, 1992.

[53] *Neuts, M.F.*: Algorithmic Probability: A collection of problems, Chapman and Hall, NY, 1995.

[54] *Ross, S.M.*: Stochastic processes; 2^{nd} edn. John Wiley and sons, Inc., New York, 1996.

[55] *Servi, L. and Finn, S.* : M/M/1 queue with working vacations (M/M/1/WV), Performance Evaluation, 50, 41-52, 2002.

[56] *Sennott, L.I., Humblet, P.A. and Tweedie, R.L.* : Mean drifts and the non-ergodicity of Markov chains, Operations Research, 31, 783-789, 1983.

[57] *Singh, V.P.* : Two-Server Markovian Queues with Balking: Heterogeneous vs. Homogeneous Servers, Operation Research, 18(1), 145-159, 1970.

[58] *Szpankowski, W.* : Stability conditions for multidimensional queueing systems with computer applications, Operations Research, 36(6), 944-957, 1988.

[59] *Takacs, L.*: Priority queues, Operations Research, 12, 63-74, 1960.

[60] *Takagi, H.* : Queueing Analysis Volume 1: Vacations and Priority System, North Holland, Amsterdam, 1991.

[61] *Takine, T.*: The nonpreemptive priority MAP/G/1queue, Operations Research, 47, 966-973, 1993.

[62] *Takine, T. and Hasegawa, T.*: The work load in the MAP/G/1 queue with state dependent services its application to a queue with preemptive resume priority, Stochastic Models, 10, 183-204, 1994.

[63] *Tian, N., Li, Q. and Cao, J.* : Conditional stochastic decompositions in the M/M/C queue with server vacations, Stochastic Models, 15(2), 367-377, 1999.

[64] *Tian, N. and Zhang, Z.G* : Vacation Queueing Models: Theory and applications, Springer Publishers, New York, 2006.

[65] *Tian, N. and Li, J.* : Matrix Analytic Method and Working Vacations Queues-A Survey. International Journal of Information and Management Sciences, 20, 603-633, 2009.

[66] *Tijms, H.C.*: A first course in stochastic models, John Wiley and sons, 2003.

[67] *Vinod, B.* : Exponential Queues with Servers' Vacations, Journal of Operational Research Society, 37(10), 1007-1014, 1986.

[68] *Wu, D. and Takagi, H.* : M/G/1 queue with multiple working vacations, Performance Evaluation, 63, 654-681, 2006.

[69] *Yang, T. and Templeton, J.G.C.*: A Survey on retrial queues, Queueing Systems, 2, 201-233, 1987.

[70] *Zhang, M. and Hou, Z.* : Performance analysis of MAP/G/1 queue with working vacations and vacation interruption, Applied Mathematical Modelling, 35, 1551-1560, 2011.

LIST OF PUBLICATIONS

1. A. Krishnamoorthy and C. Sreenivasan : An $M/M/2$ Queueing system with Heterogeneous Servers including one with Working Vacation ; To appear in International Journal of Stochastic Analysis, Hindawi Publishing Corporation.

2. C. Sreenivasan, Srinivas R Chakravarthy and A. Krishnamoorthy : $MAP/PH/1$ Queue with working vacations, vacation interruptions and N Policy ; To appear in Applied Mathematical Modelling, Elsevier.

3. A. Krishnamoorthy and C. Sreenivasan : An $M/M/2$ Queueing system with Heterogeneous Servers including one Vacationing Server ; To appear in Calcutta Statistical Association Bulletin.

4. A. Krishnamoorthy and C. Sreenivasan : $MAP/PH/1$ Retrial Queue with Constant Retrial Rate and working vacations ; Communicated.

5. A. Krishnamoorthy and C. Sreenivasan : $MAP/PH/1$ Retrial Queue with constant retrial rate, working vacations and a finite buffer for arrivals ; Communicated.

6. A. Krishnamoorthy and C. Sreenivasan : $MMAP(2)/PH/1$ Retrial Queue with a finite retrial group and working vacations ; Communicated.

CURRICULUM VITAE

Name :	Sreenivasan C
Present Address :	Department of Mathematics, Cochin University of Science & Technology, Cochin, Kerala - 682022, India.
Official Address:	Assistant Professor, Department of Mathematics, Govt. College, Chittur, Kerala - 678104, India.
Permanent Address :	Govindam, Velloli Lane (1) Puthur, Palakkad Kerala - 678001, India.
Email :	sreenikoonathara@gmail.com
Qualifications :	B.Sc. (Mathematics), 1992, University of Calicut, Calicut, Kerala, India. M.Sc. (Mathematics), 1994, University of Calicut, Calicut, Kerala, India.
Research Interest :	Queueing Theory.