PERFORMANCE ANALYSIS OF WORKING VACATION QUEUEING MODELS IN COMMUNICATION SYSTEMS

A Thesis Submitted

for the Award of the Degree of

DOCTOR OF PHILOSOPHY

by

COSMIKA GOSWAMI

(Roll Number: 05612304)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI 781039, INDIA

October 2010

CERTIFICATE

It is certified that the work contained in the thesis entitled **"Performance Analysis of Working Vacation Queueing Models in Communication Systems"** by **Cosmika Goswami** (Roll Number: 05612304), a student in the Department of Mathematics, Indian Institute of Technology Guwahati, for the award of the degree of Doctor of Philosophy, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Guwahati 781039
October 2010

Dr. N. Selvaraju
Thesis Supervisor

Dedicated

to

My Husband

and My Parents

ACKNOWLEDGEMENT

First, I would like to thank my supervisor, Dr. N. Selvaraju, without whose guidance and suggestions this work would not have seen the light of day. I am very grateful to him for giving me many patient hearings and correcting me when I have gone astray. Discipline, scientific method and mathematical rigor are some of the virtues that he has repeatedly imbibed in me.

I am thankful to the members of my Doctoral committee, Dr. Rajen Kumar Sinha, Dr. Natesan Srinivasan and Dr. Sriparna Bandopadhyay for their valuable suggestions given at crucial junctures during the period of the work. I would like to thank Prof. Jyotiprasad Medhi (Gauhati University), Prof. Krishna B. Athreya (Iowa State University, USA), Dr. Aloke Goswami (ISI, Kolkata) and Dr. T. Venkatesh (IIT Guwahati) for their invaluable conceptual guidance.

I would like to thank my many friends, who cannot be named here individually, for helping me throughout. I am thankful to the Indian Institute of Technology Guwahati for being my home for the last five years and providing me with a healthy academic environment for pursuing research. Special thanks to Larry and Sergey for leaving their Ph.Ds at Stanford and setting up Google. I could never hope to imitate them.

I would like to thank my parents, in-laws and dear brother(in-law)s and sisters for having firm belief in me and providing me crucial moral support. I would like to thank my husband for understanding me and for being with me throughout.

Above all, I would like to thank God.

***C**osmika*

ABSTRACT

This thesis studies, both analytically and through numerical experiments, the performance of queueing models with a 'working vacation' policy arising naturally in communication systems, especially in wavelength division multiplexing (WDM) networks. In a queueing system with this vacation policy, the server switches between vacation and non-vacation periods. Unlike in a classical vacation framework, this system serves customers even during vacation periods but at a lower service rate. We study some working vacation queueing models incorporating different features of system characteristics with a view to assisting optimization and guiding the design of new generation systems.

First, we consider a single-server queueing model with working vacations and correlated arrivals as network traffic is seldom similar and this often leads to system congestion and packet loss. The model is studied in discrete-time scale by constructing a quasi-birth-death (QBD) process. Since the matrix-geometric method gives an efficient way to solve homogeneous QBDs, we use this method to analyze and study the performance of the correlated model. The analogous continuous-time model is also outlined. Next, we consider a finite-buffer model with this working vacation policy and correlated arrivals to lay the emphasis on the role of correlation in arrivals on customer loss probabilities. A multi-server model is presented next, where the servers obey asynchronous multiple working vacation policy and the formulated non-homogeneous QBD process is analyzed using the finite truncation method of approximation.

A working vacation model with different priority classes of customers is studied as priority based traffic can enhance network efficiency and ensure quality of service (QoS). Explicit expressions for system performance measures are obtained and also comparisons are made for different classes of customers. Another important model considered is with retrial or repeated attempts of customers. This model, mostly seen in mobile networks, is analyzed to obtain closed-form solutions. Finally, a queue with impatient customers is studied with two types of working vacation policies, multiple and single working vacation policy, and comparisons are made to determine the most effective policy.

Contents

Abbreviations

AMWV	Asynchronous Multiple Working Vacation
ATM	Asynchronous Transfer Mode
DMAP	Discrete-time Markovian Arrival Process
DPH	Discrete-time Phase-type
EAS	Early Arrival System
FCFS	First Come First Served
LAN	Local Area Network
LAS	Late Arrival System
LDQBD	Level-Dependent Quasi-Birth-Death process
LST	Laplace-Steiltjes Transform
MAC	Medium Access Control
MANET	Mobile Ad hoc Network
MAP	Markovian Arrival Process
MMBP	Markov-modulated Bernoulli Process
MMPP	Markov-modulated Poisson Process
MWV	Multiple Working Vacation
OBS	Optical Burst Switching
p.d.f.	Probability density function
p.m.f.	Probability mass function
PGF	Probability Generating Function
PH	Phase-type
QBD	Quasi-Birth-Death process
QoS	Quality of Service
SWV	Single Working Vacation
VI	Vacation Interruption
WDM	Wavelength Division Multiplexing
WV	Working Vacation

List of Figures

List of Tables

Chapter 1

Introduction

Queueing theory and the development of communication systems have had a strong influence on each other. The first queueing theoretic model was developed for the dimensioning of communication systems. Vice versa, queueing theory has developed partly under the stimulus of new problems encountered in communication systems. In this thesis, we analyze some queueing models to study the performance issues that arise in communication networks. This introductory chapter includes a brief description of the basics of queueing modelling and its applications in various fields including communication systems. It also contains a short discussion on certain mathematical concepts such as phase-type distributions, quasi-birth-death processes and queueing modelling and their solution methodologies, with a focus on matrix-geometric method, relevant to this thesis. Motivation for this research work and the literature survey related to the research work are also included in this chapter.

1.1 Queueing theory

Queueing theory embodies the range of models covering all perceivable systems that incorporate queues or waiting lines. Queues occur when the aggregate demand exceeds the available capacity of resources. With the dramatic increase in complexity associated

with the systems of the future, formal performance models are necessary for efficient and reliable design and/or optimization. Queueing theory, with its latest methodologies, continues to be one of the most extensive theories of stochastic models to analyze complex systems and to quantify their performance to high accuracy [67]. This mathematical theory has wide areas of application, ranging from transportation and inventory systems [119], manufacturing and production systems [127], healthcare systems [130], to computer and communication systems [157]. In communication systems, queueing models arise naturally because such systems represent contention for resources and quality of service is a major concern of these systems in terms of layout, capacities and control [46]. Measuring a system performance via simulation is another method of studying a real life scenario.

The ultimate objective of the analysis of queueing systems is to understand the behavior of their underlying processes and to make intelligent decisions in their management. In a queueing model, the servers associated with queues correspond to the resources; and the customers that enter queues correspond to the units or jobs that constitute the workload of the physical system. The analysis of a queueing system involves a stochastic description of the system and measuring of system performances like the distribution of number of customers in the system, the distribution of waiting time in the queue and distribution of busy period of the system. The analysis is set about a steady-state solution or a transient (time-dependent) solution, to study a long term behavior of a queue or to study a time-dependent one respectively [128]. For steady-state systems, Little [112] gave a famous proof on the relation between mean number of customers and the expected waiting times in the queue which is called *'Little's Law'*. Whitt [162] and Newell [123] studied the queue length behavior of a system where the input process is the superposition of multiple independent renewal processes. These studies lead to the property called 'PASTA' (Poisson Arrival See Time Averages).

To provide an adequate description of a queueing system, queueing models are represented with their basic characteristics: arrival/input pattern of customers, service pattern of servers, number of servers, system capacity and queue discipline. Input pattern refers to the manner of arrivals which can be in groups or individually. The input pattern may

be stationary or non-stationary depending on whether it changes with time or not. The uncertainties involved in the service mechanism are the number of servers, the number of customers being served at any time, and the duration and mode of service. Networks of queues are systems which contain more than one server arranged in several patterns like in series or in parallel. System capacity refers to the number of customers that can wait at a time in a queueing system. If the waiting room is large, one can assume that, for all practical purposes, it is infinite. If there is a limitation of space, the system is said to be finite where an arrival is forced to abandon the system when the space is filled to capacity. Queue discipline refers to the rule followed by the server in accepting customers for service. In this context, rules such as 'first come first served' (FCFS), 'last come first served' (LCFS) and 'random selection for service' (RSS) are among the important ones. In many situations, customers in some classes have priority in service over others. Also, there are other factors of customer behavior, such as balking, reneging and jockeying, that require consideration as well. The characteristics of a queueing model are represented by the notation introduced by Kendall in the form A/B/c/K/E, where A stands for distribution of interarrival times (or arrival process), B for service time distribution, c for number of servers, K for capacity of the servers and E for queue discipline. Symbols used to denote some of the common formulations are M (Exponential distribution), E_k (Erlang-k distribution), HE_k (Hyperexponential-k distribution), D (Deterministic or constant), Geo (Geometric distribution), G or GI (Arbitrary or general distribution), PH (Phase-type distribution), MAP (Markovian arrival process). Also, depending on various arrival patterns, types of customers, service policies and queue behaviors, the models are categorized into different classes. Important classes include polling models, vacation queues, queues with impatient customers, priority queues, queueing networks and retrial queues.

Queueing theory was pioneered by A.K. Erlang with his fundamental work on congestion in telephone traffic, where he analytically formulated several practical problems arising in telephony. Analysis on the application of this theory of telephony was expanded by Fry [64] and Molina [118]. Gradually, several researchers became interested in these problems and developed general models which could be used in more complex situations.

Some of the authors with important contributions are Crommelin, Feller, Jensen, Khintchine, Kolmogorov, Palm and Pollaczek. Many researchers developed various types of models of queues including Markovian and non-Markovian arrivals, bulk arrivals, bulk services, multiple servers, finite and infinite capacities and also various queueing disciplines. Some of the important references on queueing theory are Cohen [43], Dshalalow [54], Gross et al. [67], Kijima [92], Kleinrock [93, 94], Medhi [117], Syski [142] and Saaty [137]. The performance measures of queues having PH-distributions with their basic properties are extensively given by Neuts [120, 121] and Latouche and Ramaswami [102].

One of the realistic factors that affect the system performance is the vacation of a server. In a vacation queueing system, the server becomes unavailable to the customers for a certain period of time. In this period, the server may do some supplementary works, may be under repair or may simply take a break. A fundamental property of vacation models is their stochastic decomposition property. In a model with this property, the stationary queue length or the stationary waiting time can be decomposed into sum of two independent variables one of which corresponds to a system without vacation and the other is the additional queue length or delay due to vacation. Many vacation policies have been presented and analyzed in literature. The monograph by Tian and Zhang [153] gives a comprehensive account of models of M/G/1 and GI/M/l type, with vacation policies like exhaustive, non-exhaustive, multiple adaptive vacations, N-policy and T-policy. Work by Doshi [52] and Takagi [143] also gives detail of queues with vacations. Recently, Fiems et al. [60] considered M/G/1 queueing systems with combined disruptive and non-disruptive renewal-type server interruptions. Along with renewal arrival models, vacation queues with non-renewal arrivals have also been studied. Lucantoni et al. [114] analyzed a single server vacation queue with Markovian arrival process (MAP). Using the matrix-geometric method, Kasahara et al. [85] studied the MAP/G/1 queue under N-policy with and without vacations. Lee et al. [103] studied the model under multiple and single vacations with N-policy.

In discrete-time systems, the measurement of time is discrete in nature and changes of system states can occur only at discrete epochs of time. Takagi [144] gave a detailed

analysis of a Geo/G/1 queue with various vacation policies. Later, Zhang and Tian [172] added the multiple adaptive vacations to the model. A discrete-time GI/Geo/1 queue with server vacation is presented by Tian and Zhang [150]. Fiems et al. [61] investigated a discrete-time single-server queue subjected to server interruptions. Server interruptions are modeled as an on/off process with geometrically distributed on-periods and generally distributed off-periods. Alfa [4] gave a survey on PH/PH/1 queues in discrete-time. Frigui and Alfa [63] considered a discrete-time cyclic polling MAP arrival system in which each queue was visited according to exhaustive time-limited service discipline. Gupta et al. [69] treated a finite-buffer discrete-time MAP/G/1 queue with exhaustive single and multiple vacations; and used embedded Markov chain method to obtain various performance measures. Using matrix-geometric method, Alfa [2, 5] analyzed a discrete-time MAP/PH/1 queue with exhaustive and non-exhaustive service where the vacations follow PH-distributions. Discrete NT-policy single-server queue with Markovian arrival process and PH-service is also analyzed by Alfa and Frigui [6]. Various vacation models have also been studied with $BMAP$ arrival processes.

Vacation models with finite-buffer space represent many realistic systems. The analysis of the loss probability is the main concern in finite-buffer models. Takagi [144] gave details on finite-buffer M/G/1 type vacation models. Karaesman and Gupta [84] examined the GI/M/1/K queue with exponentially distributed multiple vacations. Also, Chaudhry et al. [38] obtained explicitly the system measures for M/G/1/K and GI/M/1/K models. Ke and Wang [90] studied a single removable server in a G/M/1 queueing system with finite capacity operating under N-policy. Choi et al. [41] considered an MMPP/G/1/K queue where arrival rates depend on system queue length. Blondia [29] presented a MAP/G/1/K queue with multiple vacations for exhaustive and limited service discipline. A more general study of MAP/G/1/K queue with single and multiple vacations along with setup and close-down time can be found in Nui and Takahashi [125]. Further, Niu et al. [124] have extended the analysis for a BMAP/G/1/K queue. The analysis of MAP/G/1/K queue with limited service discipline is carried out by Gupta et al. [68] and the same queue is studied with E-limited with limit variation (ELV) service by Banik et al. [25]. Also,

Gupta and Sikdar [70] presented queue length distribution of the model with PH-service times.

Multi-server queueing systems with vacations have been investigated only in a limited number of studies. Levy and Yechiali [104] first discussed a M/M/c queue with exponentially distributed vacations. Igaki [75] provided a detailed analysis on a M/M/2 queue in which one and only one server can take vacations when the system becomes empty. Igaki developed the stationary distributions for the queue length and gave a proof of the conditional stochastic decomposition property in such a system. Tian and Li [146], Tian et al. [147], Tian and Zhang [149] studied a variety of vacation models with multiple servers. They established the conditional stochastic decomposition properties on the steady-state queue length and the waiting time when all servers are busy and obtained the stationary distributions for queue length and waiting times. Ausina et al. [21] considered the problem of designing a GI/M/c queueing system with the objective to choose the optimal number of servers to minimize an expected cost function. Tian and Zhang [151] solved a GI/M/c type queueing system with PH-vacations in which all servers take vacations together when the system becomes empty. Recently, Tian and Zhang [152] considered a two-threshold vacation policy in the context of a multi-server queueing model M/M/c. A multi-server queueing system with identical unreliable servers with PH-distributed service times is considered by Yang and Alfa [168]. Chakravarthy [36] studied a MAP/M/c queueing system, in which a group of servers take a simultaneous PH-vacation.

In a priority queue, different classes of customers are present and each class of customers is served according to its priority over the other classes. Both the service discipline and the service time distribution may vary with the customer type. In a priority queue, the priority discipline followed may be either nonpreemptive or preemptive in nature. In nonpreemptive discipline, the customer in service is allowed to complete normally even if a higher priority customer arrives in the queue while its service is going on. In the preemptive case, the service to the ongoing customer will be preempted by the arrival of a higher priority customer. If the priority discipline is preemptive resume, then the service to the interrupted customer resumes from the point at which the service was interrupted.

For preemptive non-resume case, service already provided to the interrupted customer is forgotten and its service is started from the beginning. For exponentially distributed service times, which satisfy memoryless property, the preemptive resume and preemptive non-resume cases have same results. Priority queues have important uses in the modelling and analysis of computer systems and communication networks. Jaiswal [78] and Takagi [143] have given details of M/G/1 type priority queues with and without different vacation policies. Other references in this area are Kleinrock [93, 94], Medhi [117]. Houdt and Blondia [73] analyzed priority customers in a tree-like process. Recently, Jin and Min [79] have presented a comprehensive performance model that can obtain the closed-form expressions for queue length distribution and loss probability of priority queuing systems subject to heterogeneous traffic. Katayama [86] studied M/G/1 priority queue with semi-exhaustive service and with multiple/single server vacations. Applying the delay cycle analysis they derive the LST of waiting time distributions for each class. Katayama and Kobayashi [87] studied nonpreemptive services controlled by an exponential timer and multiple vacations. An arrival time approach of finding the queue length distributions for M/G/1 and MAP/G/1 priority queues with generalized vacations is given by Chae et al. [35]. Recently, Lui and Wu [115] presents MAP/G/1 queue with possible preemptive resume service discipline and multiple vacations wherein arrivals of negative customers follows the Markovian arrival process.

Retrial queueing systems have been introduced to study the situation of repeated calls. Such systems are generally characterized by the feature that the arriving customer, who finds all the servers and waiting positions (if any) occupied, joins a trial queue called orbit and retries for service after a period of time. Retrial queues have been widely used to model many problems in telephone switching systems, communication networks and computer systems. The basics of retrial queues can be found in [17], [59] and [67]. Artalejo gave a detailed survey on retrial queues in [12, 14]. Artalejo and Falin [16] have taken the M/G/1 single-server retrial queue with non-exhaustive vacations and derived the stochastic decomposition property and mean number of customers in orbit. Later, Artalejo [11] analyzed an M/G/1 retrial queue with exhaustive server vacations. Kumar

and Arivudainambi [99] dealt with an M/G/1 retrial queue where the server operates according to a Bernoulli vacation policy and generally distributed retrial times. Li and Yang [105] have also given an M/G/1 retrial queue with Bernoulli vacations and have obtained system performances. Atenika et al. [19, 20] studied the model with impatient customers and batch arrivals respectively. Falin [58] has taken an M/M/1 retrial queue, where he assumed that if the server fails during service of a customer, the customer leaves the server, joins a retrial group and in random intervals repeats attempts to get service. Kumar and Arumuganathan [100] studied the steady-state behavior of an M/G/1 retrial queue with non-persistent customers and two phases of heterogeneous service and different vacation policies. Recently, Ke and Chang [89] presented a batch arrival retrial queue with general retrial times, where the server is subject to starting failures under Bernoulli vacation schedule.

Customer impatience phenomenon is commonly observed in queueing systems, where customers leave a service system, before receiving service, due to high waiting time or due to uncertainty of receiving service. Impatient customers can be of three types. The first is balking, the reluctance of a customer to join a queue upon arrival; the second reneging, the reluctance to remain in line after joining and waiting; and the third jockeying between lines when each of a number of parallel lines has its own queue [67]. Customer reneging represents loss in revenues and customer goodwill to the service provider. In order to reveal impacts of impatient customers on performance measures, research efforts on queues with impatient customers (or simply queues with impatience) have been conducted over a long period. The traditional approach to dealing with reneging is to vary the service capacity according to the amount of work or the number of customers in waiting, which may not always be physically feasible or cost effective. The problem of queues with impatient customers was first analyzed by Palm [126]. A bibliography can be found in Gross et al. [67]. Barrer [27] analyzed the M/M/1 + D system, where D stands for the deterministic distribution of impatient times. Baccelli and Hebuterne [24] analyzed the waiting time distribution in M/M/c queue with general impatience bound on queueing times, by constructing a simple Markov process and also gave the waiting time distribution in the

M/G/1 queue. Kok and Tijms [95] and later, Xiong et al. [164] studied the M/G/1+D queue. Finch [62] analyzed GI/M/1+D queue with constant limitation on queueing times and found the distribution of waiting time in the queue. Doshi and Jagerman [53] studied the M/G/1 queue with class dependent reneging. Daley [47] analyzed the GI/G/1+G queue by setting up an integral equation for the waiting time distribution and focused on M/G/1+D and M/G/1+M queues. The GI/G/1+G queue was also studied by Baccelli et al. [23], Stanford [140, 141] and Jouini et al. [80]. In this work, a stability condition was established for the general case while for the M/GI/1+GI queue the virtual waiting time was studied. Gans el al. [65] and Zohar [175] gave a tutorial and research prospects on impatient customers in call centers. Recently, Perel and Yechiali [129] considered a two-phase service impatient model where the customers become impatient if the server is in slow service phase. There are situations where customer's impatience is due to the absence of the server, more precisely, due to the server being on vacation and is independent of the customers in system. Altman and Yechiali [7, 8] studied the impatient customers in classical vacation model and system with additional task respectively. Recently, Economou and Kapodistria [56] considered a unreliable queue where the customers leave the system at system failure times. Multi-server queues with impatience, however, have attracted much attention in queueing literature possibly because of explosive demands to efficiently design and manage call or contact centers. Barrer [27] analyzed M/M/c queues with customer impatience of constant duration. Jurkevic [82, 83] treated a M/M/c queue with impatience having exponentially distributed time; and with generally distributed impatience time respectively. Baccelli and Hebuterne [23] studied the waiting time distribution in M/M/c queue with general impatience bound on queueing times by constructing a simple Markov process and also gave the waiting time distribution in the M/G/1 queue with general impatience on queueing times. Yechiali [169] considered a M/M/c system which as a whole suffers occasionally a disastrous breakdown, upon which all present customers (waiting and served) are cleared from the system and lost.

1.2 Queues in communication systems

Queueing models arise in communication systems because they represent contention for resources. Performance modelling of communication systems has been carried out for many years with a view to assisting optimization and guiding the design of new generation systems. In a communication system, the messages or packets are transmitted through links from a source node to a destination node. In a queueing context, the messages are referred to as customers, channels as servers, message transmission times through a channel as service times and the number of links (from the source node to the destination node) as the number of servers. The emphasis on analysis of queueing models with application in communication systems is laid by many researchers including Boxma and Syski [31], Daigle [46], Gebali [66], Koole [96], Trivedi [157] among others.

Performance analysis of a communication network deals with the evaluation of the level of efficiency the network achieves, and the level of (dis)satisfaction of its users. A key element is the characterization of the impact of 'user parameters' on the performance offered by the network. Performance analysis is a probabilistic discipline, as the main underlying assumption of user behavior is that it is inherently random. Therefore, such analysis is described by a queueing model which defines the probabilistic properties of the traffic at the network. The performance measures in communication networks include the tuning range, processing requirements, propagation delay with respect to the packet transmission time, waiting time before packet transmission and channel allocation [48, 116]. Use of queueing theory in performance analysis of ATM networks [98], performance analysis of telephone systems [174], queueing analysis of IEEE 802.11 MAC based wireless networks [34, 155] and many others can be found in literature.

For high-speed networks, data traffic is seldom uniform and is characterized by periods of burstiness. Traffic bursts tax the network resources and lead to network congestion and data loss. Burstiness in sources like, voice, coded video and bulk data transfers; and correlation between interarrival times of packets or cells, are very important factors in the analysis of system performance. A Bernoulli/Poisson process or an independent renewal

arrival process is not an appropriate assumption for an arrival process of network traffic as it cannot capture the correlation of packets in a network, However, processes like Poisson process can capture burstiness in network traffic. Thus a generalized model is needed to study the performance of networks where arrivals are correlated. A Markov-modulated process or more generally a Markov arrival process (MAP) is widely used for non-renewal type arrival processes [45]. In a Markov-modulated traffic model, states are introduced to the model where the source changes its characteristic, depending on the current state. The state of the source could represent its data rate, its packet length etc. When the Markov process represents data rate, the source can be in any of several active states and generates traffic with a rate that is determined by the state.

Packet losses are common in packet networking. They are caused by the limited buffering space in network devices. Packet loss probability depends on the type of network since this determines the number of paths that can be simultaneously established to reduce the amount of buffer occupancy. Wavelengths or channels are used to transfer packets within the network. The wavelength independence assumption, in which wavelength usage on each link is characterized by a fixed probability, independent of other wavelengths and links, makes it possible to study the blocking performance of networks quantitatively [132, 145]. The packet loss process is an important task as it enables better network design in terms of buffer sizing and management, congestion control mechanisms, protocols etc. and can seriously influence the performance of the network.

Multiple services are required for high efficiency of bandwidth-intensive applications. Different services may require different channel capacities and capacity of a channel depends upon the number of resources allocated to it. A wavelength division multiplexing (WDM) network divides the available fiber bandwidth into WDM channels [71, 158]. This division of bandwidth or channel allocation is based on the capacities required for various services. For a high performance system, WDM channel allocation should lead to optimized resource utilization in a given network, which is physically feasible and cost-effective.

In a computer network, jobs can be divided into different classes. This enables quality of service (QoS) support. For instance, there may be a natural distinction between data, voice and video packets; and different classes may require different services. In such cases, it is common to implement service discipline at the networks that treat the jobs according to their 'priority'. Priority based channel assignment ensures the transmission of a high priority packet prior to a low priority one. The highest priority queue does not need a buffer to store incoming data when preemptive static priority is employed. If a nonpreemptive scheme is employed, the highest priority queue will require a buffer to store incoming data until it is sent. Lack of priorities in the current channel assignment techniques can severely limit the viability of networks [55, 116].

Mobile ad hoc network (MANET) is a self-configuring network of mobile devices in which mobile subscribers are connected to a base station by wireless links where the retrial is a very common issue. When a call request comes to a base station, it assigns the mobile subscriber a link to the destination. Due to traffic congestion, if no links are available, the call either retries till a link is allocated successfully or it balks the system. Also in an optical access network, when a traffic request arrives, the network operator executes the routing and wavelength assignment (RWA) procedure which is responsible to find a working path from the source node to the destination node and assign an available wavelength to this connection as the working wavelength to carry data along the connection [134]. In case the path is not found or there is no available wavelength, the issued request is blocked [160, 71].

Optical Burst Switching (OBS) is a technology for reducing the gap between transmissions and switching speeds. In OBS, incoming traffic from clients at the edge of the network is aggregated and further transmitted through WDM links [131, 171]. The operation of an OBS controller can be seen as a queue with reneging or impatience. When a path is not assigned to a request, the burst control packet is accepted by the queue and is kept waiting for a path. If its delay budget is lower than the effective processing delay, it becomes impatient and leaves the system unserved. To have more efficient use of the network, the loss of packet burst has to be reduced and enhance its performance [30, 133].

1.3 WDM networks and working vacation queues

Wavelength division multiplexing (WDM) technique has emerged as a promising solution to meet the rapidly growing demands on bandwidth in present communication systems. WDM is a method of transmitting packets from different sources, over the same fiber optic link, to the destination. WDM divides the available fiber bandwidth into WDM channels, called wavelengths. When a packet arrives at a source node, a specific wavelength is assigned to it before being transmitted to the destination node. As the system load increases, packets are queued at the source nodes and to control the congestion, reconfiguration of wavelength is done. The reconfiguration problem consists of determining which source node should be connected to the destination node using which wavelength. During reconfiguration, a token attends a channel and assigns wavelengths with a faster rate than the usual one. After assigning wavelengths to all the packets in queue, the token moves to another channel to reconfigure. In a queueing context, the wavelengths are the servers and the packets are the customers who join a queue if the server is busy. When a particular server possesses a token, customers are served at a faster rate till the queue becomes empty, whereas the other queues are serving at their normal rates. After the token leaves the queue, the service rate is reduced to normal for subsequent arrivals. The token moves from one queue to another in a cyclic manner. From the point of view of a queue, the period of time during which the queue holds the token can be termed as a normal busy period as, its server serves at a faster rate during this period. As soon as the queue becomes empty, the token moves out and the server starts serving at a slower rate until the token comes back again to that particular queue. This period will be called a working vacation (WV) period as, during this period the server serves at a slow rate. Thus, a reconfigurable WDM system can be modelled as a queue with WVs. This vacation cannot be put in a classical vacation framework because here, unless the system is empty, the service do not cease completely. Servi and Finn [139] were the first to model such a WDM network into a working vacation queueing model.

Servi and Finn [139] considered multi-queue generalization of a cyclic service queueing model arise in WDM network. They have obtained the transform formulae for the distribution of number of customers in the system and sojourn time in steady-state. As an example of the WDM optical access network they applied those results to performance analysis of a gateway router in fiber communication network and showed that the use of WV model for WDM access networks can improve performance over the alternative of having no reconfiguration or reconfiguring all wavelengths. They have modelled the WDM by assuming that, multiple wavelengths who serve as a router, can be accurately approximated by a single-server or, alternately, by assuming that the multiple wavelengths are engineered to indeed perform as if they are a single-server which is called continuous bandwidth mode [138].

The difference between a classical vacation model and a WV model can be outlined as follows. In a classical vacation queueing model, a server completely stops serving the customers during its vacation period and made all the customers wait for service till the vacation period ends. But in a WV model, the server is always available to the customers and rather than cease the service completely during its vacation period, the server serves at a lower rate. Also, during a vacation, customers in the former can only depart the system unserved however, departing customers in the later may complete its service before leaving.

Apart from WDM systems, application of WV models are also seen in other communication networks. Some examples of practical systems which can be viewed as a WV model are provided:

1. Mobile ad hoc network (MANET) is a self-configuring network of mobile devices in which mobile subscribers (MS) are connected to a base station (BS) by wireless links. Wireless hosts are usually powered by batteries which provide a limited amount of energy. To reduce the energy consumption, power control schemes are used which suitably vary the transmit power and save energy. When a call request comes to a BS, it assigns the MS a link to the destination. If no links are available, the call retries till a link is allocated successfully or balks the system. During transmission,

the source node sends out a RTS ('Request to Send') packet. The receiver node replies with a packet called CTS ('Cleared to Send') packet. After the transmitter node receives the CTS packet, it transmits the data packets. The source node may transmit DATA using a lower power level. The destination node transmits an ACK (acknowledgement) using the minimum required power to reach the source node. In Power Control MAC (PCM) protocol [81] different power levels are used for DATA transmission. After the RTS-CTS handshake using higher power say p_{max}, suppose the source and destination nodes decide to use power level p_1 for DATA and ACK. But to avoid a potential collision with ACK, the source node transmits DATA at a higher power level p_{max}, periodically, for just enough time, so that nodes in the carrier sensing zone can sense it. If the DATA is allowed to transmit only with p_{max}, more power will be consumed and also it cannot be transferred with p_1 which causes DATA fading. Thus during the sequence of an RTS-CTS-DATA-ACK transmission, the PCM protocol uses a power level changes between p_1 and p_{max} with a maximum power savings without causing throughput degradation. We can model this network as a queue with WVs where the RTS-CTS-DATA-ACK transmission with power p_1 is a WV period, transmission with power p_{max} is non-vacation period. The repeated call attempts are retrials.

2. Edge devices, which are primarily comprised of desktop computers, are the largest Internet-related energy consumers. A desktop computer or a PC is connected to a first-level LAN (Local Area Network) switch through an Ethernet link. In computer networking, Ethernet or IEEE 802.11h, is a digital data transmission technology for LANs. The power consumption of such links are higher for high transmission rates. Switching to lower link rates during low utilization periods result in reducing the energy consumption. Medium Access Control (MAC) layer is proposed and analyzed for dynamically changing the link rate in the Network Interface Card (NIC), adapting to network utilization. MAC handshake protocol [9] helps to decrease the average power consumption without causing any user-perceivable delays. The protocol uses a dual threshold policy for dictating when to change link rates. Whenever

the buffer occupancy drops below the low threshold or rises above the high threshold, the handshake mechanism is activated. The link rate is reduced to a low rate when buffer occupancy is below the low threshold value and the rate switched to a higher rate as the buffer congestion reaches the high threshold. This system can also be modeled as a WV queue where the WV (non-vacation) duration is the one when the link rate is low (high).

In literature, the study of WV queueing systems has been carried out by many researchers after Servi and Finn. Liu et al. [113] analyzed the M/M/1/WV model using QBD process and matrix-geometric method to obtain explicit expressions of the performance measures and their stochastic decompositions. Tian et al. [154] have presented the M/M/1 queue with single WV. Xu and Tian [167] and Xiu et al. [165] analyzed this model with single WV and setup times. Wang et al. [161] observed the M/M/1 machine repair problem with WV in which the server works with different repair rates and used Newton's method to compute steady-state probabilities and several system performance measures. Wu and Takagi [163] extended Servi and Finn's work to M/G/1/WV model with generally distributed service times and vacation duration times, taking the Laplace-Stieltjes Transforms for the distribution of vacation length to be a rational function. Jain and Agrawal [76] dealt with state dependent M/E_k/1 queueing system with server breakdown. Baba [22] considered the GI/M/1/WV system with general independent arrival process where the distribution of the vacation duration times and service times are exponential. He formulated the queueing system as a embedded Markov chain and solved it using matrix-geometric method. Recently, Jain et al. [77] investigated a single-server WV queueing model with multiple types of server breakdowns where each type requires a finite random number stages of repair before service is restored. Wang [158] studied the M/M/1 queue with WV and non-zero switching times and studied a cyclic service system in WDM-based access networks with reconfigurable delay.

The finite-buffer model GI/M/1/K/WV was presented by Banik et al. [26] with multiple WV policy as an extension of Baba's infinite buffer model. Chen et al. [39, 40] proposed N-policy WV and cyclic polling system for WDM taking the service times as

exponential and PH respectively. Lin and Ke [111] considered a multi-server M/M/c queue and a cost model is derived to determine the optimal values of the number of servers and the WV rate simultaneously, in order to minimize the total expected cost per unit time. Li et al. [109] analyzed a bulk input $M^{[X]}/M/1$ queue with single WV. Dutta and Chaubey [55] presented priority assignment to incoming call connection requests for all optical WDM communication, using queuing theory concept. Do [51] studied a retrial M/M/1/WV queue which was motivated by the performance analysis of a Media Access Control (MAC) function in wireless systems and derived the close-form solutions. Connecting vacation interruption(VI)s and the WV policy, Li and Tian [107] was first to introduce VIs in the basic M/M/1/WV model and derived the stochastic decomposition structure. They further extended the work to GI/M/1/WV and VI queue with multiple exponential vacations [108] and Zhao et al.[173] considered the model with setup period.

In discrete-time case, Tian [148] analyzed the Geom/Geom/1/WV with geometrically distributed vacation duration as an extension of Servi's work but using quasi birth-death process and matrix-geometric solution method. Subsequently, Li et al. [106] presented the GI/Geo/1 queue with multiple WV, with the same geometrically distributed vacation duration. Considering the two variations of the arrival and departure schemes, early arrival system (EAS) and late arrival system (LAS) with delayed access (LAS-DA) and immediate access (LAS-IA), they obtained the uniform results on the stationary distributions and stochastic decomposition properties under both the schemes. Yi et al. [170] considered a Geo/G/1 queue with disasters that remove all workloads from the system upon their occurrence and presented the steady-state queue length distribution. Xu et al. [166] studied a bulk input $Geom^{[X]}/Geom/1$ queue with single WV and found the bi-parameter addition theorems of the conditional negative Binomial distribution. Thus the WV queues are analyzed in continuous-time and also in discrete-time considering different aspects of the system under study.

Motivation

The WV queues have been studied considering various features of system characteristics. Queues with WVs are seen to arise in several communication systems including the WDM networks. Communication systems carry various types of data packets or voice packets which have special properties and accordingly they need special attentions. For example, packets are bursty and correlation between the packets is high and has to be maintained; some packets need priority over the others and have to be kept separate to allocate different sets of wavelengths on priority basis; some packets make repeated attempts for services, mostly in mobile networks, and have to be assigned a buffer space for their temporary storage, until they receive the requested service; also, packets may abandon the system before receiving service because of impatience and have to be refrained from leaving the system, to maintain the system efficiency. Such features need earnest attention as they are very common in communication networks. Communication systems along with these specific features of packets can be modelled as queues with WVs. In literature, such queues have not been the focus of study. In this thesis, we consider some WV queueing models which incorporate these characteristics of packets. We model these modified WV queues as Markov processes and analyze the performance of those queueing systems.

In the following sections, some mathematical concepts relevant to this thesis are discussed.

1.4 Phase-type distribution

An elementary and analytically tractable model is M/M/1, where the arrival process is a Poisson process and the distribution of service times is exponential. But the growing importance of qualitative modelling demands generalized arrival process or generalized service process for more accuracy. Elementary model closely resemble the behavior of numerous physical phenomena in which arrivals involve independent and similar users. These Markovian models are appropriate in certain applications but may be inadequate

in instances where the demands for general arrival or service process arises. Analytic approaches to such models under general distributional assumptions become complicated or essentially intractable. To study such complicated models, a phase-type (PH) distribution is introduced, which is capable to maintain the mathematical tractability of the models, which reflects their essential qualitative features and provides useful information on their physical behaviors. PH-distributions provide a simple framework to demonstrate how one can extend simple results on exponential distributions to more complex models, without losing computational tractability. They are based on the method of phases. The primary theoretical utility of PH-distribution is that, any distribution on the non-negative real numbers can be approximated by a PH-distribution and the resulting queueing models, such as M/PH/1, PH/PH/1, can be analyzed by means of Markovian environments. In practice, such an approximation is strictly limited to study those inferences which are warranted by their precise mathematical statements but not recommended to study a feature that is highly sensitive [135, 136].

To define a PH-distribution, let us consider a queueing model with finite capacity. If upon arrival of a customer the system is filled, the new arrival will not be allowed to join the system and the customer is said to be lost. Let us define the state when the new arrival has to leave the system as an absorbing state. Now, from an system administrator's point of view, the loss of an event (customer) is regarded as a bad event and shows a poor performance of the system. To keep track of the loss of a customer, we need to know about the distribution of the time up and till the first loss, *i.e.,* when the system enters the absorbing state. This problem can be addressed by the concept of PH-distribution [32].

Let $\chi = \{X_t, t \geq 0\}$ be a time-homogeneous continuous-time Markov chain with a finite state-space $\{1, 2, ..., m+1\}$ and with infinitesimal generator matrix $Q = \begin{pmatrix} T & T^0 \\ \mathbf{0} & 0 \end{pmatrix}$, where T is a square matrix of dimension m, T^0 is a column vector and $\mathbf{0}$ is the zero row vector of appropriate dimension. The first m states $\{1, 2, ..., m\}$ are transient, while the state $m+1$ is absorbing. The matrix T is nonsingular (Lemma 1.4.1 below) and describes the

transitions until absorption, where $(T)_{ii} < 0,\ 1 \leq i \leq m$ and $(T)_{ij} \geq 0,\ 1 \leq i \neq j \leq m$. The matrix T^0 is a non-negative, m-dimensional column vector, grouping the absorption rates from any state to the absorbing one, and satisfies $T^0 + T\mathbf{1} = 0$, where $\mathbf{1}$ is the column vector of appropriate dimension with all the entries equal to one. Let the initial distribution of χ be the row vector $\bar{\alpha} = (\alpha, \alpha_{m+1})$, with α being a row vector of dimension m and $\alpha_{m+1} = 1 - \alpha\mathbf{1}$.

Lemma 1.4.1. *The states $\{1, 2, \ldots, m\}$ of a the Markov chain are transient if and only if its transition rate matrix T is nonsingular, [120].*

Now, let us give the definition of a PH-distribution.

Definition 1.4.1. *Let $Z = \inf\{t \geq 0, X_t = m + 1\}$ be the random variable denoting the time until absorption in state $m + 1$ of the Markov chain χ. Then Z is said to follow a PH-distribution with parameters (α, T), denoted by PH(α, T), and with distribution function*

$$F(t) = P(Z \leq t) = 1 - \alpha e^{Tt}\mathbf{1}, \quad t \geq 0. \tag{1.1}$$

The density function of Z is given by

$$f(t) = \alpha e^{Tt} T^0, \quad t \geq 0, \tag{1.2}$$

where the matrix exponential is defined for a square matrix T by

$$e^{Tt} = \sum_{n=0}^{\infty} \frac{t^n}{n!} T^n. \tag{1.3}$$

The dimension m is called the **order** of the distribution PH(α, T) and the states $\{1, 2, ..., m\}$ are called **phases**. The vector T^0 is said to be the **exit vector**. The distribution $F(t)$ has a jump of height α_{m+1} at $t = 0$. We can assume, without loss of generality that $\alpha_{m+1} = 0$.

The infinitesimal generator $Q^* = T + T^0\alpha$, describes a new Markov chain with state-space $\{1, 2, \ldots, m\}$ in which the absorption state $m + 1$ is an instantaneous return state.

That is, upon absorption in the Markov chain χ, the chain is instantaneously restarted by selecting a new initial state (independent of the past), using the same probability vector α. This resetting is repeated indefinitely. The epochs of successive visits to the 'instantaneous' state $m+1$ form a renewal process with interrenewal time distribution $F(t)$. This means, the interrenewal times of the new Markov chain with generator Q^* follows PH-distribution.

Markov chains are classified depending on different properties of its states.

Definition 1.4.2. *A Markov chain in which every state leads back to itself and also to every other state is called an* ***irreducible*** *Markov chain; otherwise, the Markov chain is said to be* ***reducible****.*

Definition 1.4.3. *Let p_{ii}^n be the probability that a Markov chain starting from state i, i belongs to its state-space, will be again in state i after n additional transitions. The state i of the Markov chain has a period d_i if $d_i = g.c.d\ \{n, p_{ii}^n > 0\}$, that is, d_i is the greatest common divisor of such n's.*

In an irreducible Markov chain all states has a common period d and if $d > 1$ the chain is said to be **periodic** whereas if $d = 1$, the chain is called **aperiodic**.

Definition 1.4.4. *Let f_{ii} be the probability that a Markov chain starting at state i will ever return to i. Then the state i is called a recurrent state if $f_{ii} = 1$ and transient if $f_{ii} < 1$. A Markov chain is called a* ***recurrent*** *chain, if all of its states are recurrent and a* ***transient*** *chain, if all of its states are transient.*

An irreducible Markov chain is necessarily either a transient chain or a recurrent chain. A state of a Markov chain is an absorbing state if the probability of moving out of that state is zero.

Definition 1.4.5. *A non-absorbing recurrent state is said to* ***positive recurrent*** *or* ***null recurrent*** *according as the mean return time to that state is finite or infinite. A Markov chain is a positive recurrent chain if all its states are positive recurrent and a null recurrent chain if all its states are null recurrent.*

The corresponding transition matrices of these Markov chains are named accordingly. We have a result that an irreducible Markov chain in finite state-space is always positive recurrent. Another result says that, an irreducible and positive recurrent Markov chain has a unique stationary distribution [72].

For a PH-distribution, PH(α, T), the generator Q^* is assumed to be irreducible. The Markov chain with generator Q^* is positive recurrent, as its state-space is finite, and so its stationary distribution exists. The stationary probability vector ξ of Q^* is obtained by solving the equations $\xi Q^* = \mathbf{0}$, $\xi \mathbf{1} = 1$. The rate of arrival or rate of successive visits to the instantaneous state is given by $\lambda = \xi T^0$, which is always positive. Following are some important properties of a PH-distribution.

Lemma 1.4.2. *The Laplace-Stieltjes transform of $PH(\alpha, T)$ with distribution function $F(t)$ is given by*

$$\Phi(s) = \int_0^\infty e^{st} dF(t) = \alpha(sI - T)^{-1} T^0, \quad \text{for } Re(s) \geq 0. \tag{1.4}$$

Lemma 1.4.3. *If Z follows $PH(\alpha, T)$, the mean and the raw moments of Z are respectively, given by*

$$E(Z) = -\alpha T^{-1} \boldsymbol{1} \tag{1.5}$$

$$\text{and } E(Z^n) = (-1)^n n! \alpha T^{-n} \boldsymbol{1}, \quad \text{for } n = 2, 3, \ldots. \tag{1.6}$$

Another expression for mean, in terms of stationary vector ξ, is given by

$$E(Z) = \left(\xi T^0\right)^{-1}. \tag{1.7}$$

Lemma 1.4.4. *For a $PH(\alpha, T)$ distribution, the remaining time Z until absorption, given that the current phase is 'i' can be expressed by*

$$P(Z \leq t | X_0 = i) = 1 - e_i e^{Tt} \boldsymbol{1}, \tag{1.8}$$

*which has the same form as $PH(e_i, T)$ distribution, with e_i denoting the i^{th} canonical row base vector. This is called the **conditional memoryless property** of PH-distribution.*

Lemma 1.4.5. *Let Z_i follows PH$\big(\alpha^{(i)}, T^{(i)}\big)$ of order m_i, for $i = 1, 2$. Then the closure property gives that $Z = Z_1 + Z_2$ will follow distribution PH(α, T) of order $m = m_1 + m_2$ with representations*

$$\alpha_k = \begin{cases} \alpha_k^{(1)}, & 1 \le k \le m_1; \\ \alpha_{m_1+1}^{(1)} \alpha_{k-m_1}^{(2)}, & m_1 + 1 \le k \le m \end{cases}$$

and $T = \begin{pmatrix} T^{(1)} & T^{0(1)}\alpha^{(2)} \\ \boldsymbol{0} & T^{(2)} \end{pmatrix}$, *where* $T^{0(1)} = -T^{(1)}\boldsymbol{1}$.

Lemma 1.4.6. *Let Z_i follows PH$\big(\alpha^{(i)}, T^{(i)}\big)$ of order m_i for $i = 1, 2$, and $p \in [0, 1]$. Then $Z = pZ_1 + (1-p)Z_2$ will follow distribution PH(α, T) of order $m = m_1 + m_2$ with representations*

$$\alpha = \big(p\alpha_{(1)}, (1-p)\alpha_{(2)}\big) \quad \text{and} \quad T = \begin{pmatrix} T^{(1)} & \boldsymbol{0} \\ \boldsymbol{0} & T^{(2)} \end{pmatrix}.$$

Lemma 1.4.7. *Let Z_i follows PH$\big(\alpha^{(i)}, T^{(i)}\big)$ of order m_i for $i = 1, 2$, and define $Z = min(Z_1, Z_2)$. Then Z will have PH(α, T) of order $m = m_1 m_2$ with representation*

$$\alpha = \alpha^{(1)} \otimes \alpha^{(2)} \quad \text{and} \quad T = T^{(1)} \oplus T^{(2)},$$

in terms of the Kronecker compositions '$\otimes$' and '$\oplus$' (given below, after the special cases).

The proofs of the above Lemmas can be found in Breuer [32]. PH-representation of some special distributions are given as follows:

- **Exponential distribution:** A continuous random variable X follows an exponential distribution with parameter $\lambda > 0$, denoted by Exp(λ), if its probability density function (p.d.f.) is

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \ge 0; \\ 0, & t < 0. \end{cases} \tag{1.9}$$

  Exponential distribution has the 'memoryless property' *i.e.*, if X defines interarrival times, the time we must wait for a new arrival is statistically independent of how long we have already waited for it. PH-representation of Exp(λ) is

$$\alpha = 1, \; T = -\lambda \text{ and } \; T^0 = \lambda. \tag{1.10}$$

- **Erlang-k distribution:** A continuous random variable X is said to follow an Erlang-k distribution with parameters (λ, k), denoted by Erl-$k(\lambda)$, for $\lambda > 0$ and $k \in \{1, 2, \ldots\}$, if its density function is

$$f(t) = \begin{cases} \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda t}, & t \geq 0; \\ 0, & t < 0. \end{cases} \tag{1.11}$$

The exponential distribution is a special case of the Erlang-k distribution with $k = 1$. Many processes in nature can be divided into sequential phases. If the time spent by the process in each of k sequential phases have independent and identical exponential distributions, then the overall time has Erlang-k distribution. The PH-representation of Erl-$k(\lambda)$ is given by

$$\alpha = (1, 0, \ldots, 0),\ T = \begin{pmatrix} -\lambda & \lambda & & \\ & \ddots & \ddots & \\ & & -\lambda & \lambda \\ & & & -\lambda \end{pmatrix}_{k \times k} \text{ and } T^0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \lambda \end{pmatrix}_{k \times 1}. \tag{1.12}$$

- **Hyperexponential-k distribution:** A continuous random variable X follows a k-phase hyperexponential distribution with parameters (α_i, λ_i), denoted by Hyp-$k(\alpha_i, \lambda_i)$, for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} \alpha_i = 1$, if it has p.d.f. as

$$f(t) = \begin{cases} \sum_{i=1}^{k} \alpha_i \lambda_i e^{-\lambda_i t}, & t \geq 0, \alpha_i > 0, \lambda_i > 0; \\ 0, & t < 0. \end{cases} \tag{1.13}$$

If a process consists of alternate phases and it experiences one and only one of many alternate phases, instead of sequential phases, each of which have exponential distributions, then the resulting distribution is hyperexponential. Its PH-representation is given by

$$\alpha = (\alpha_1, \ldots, \alpha_k),\ T = \begin{pmatrix} -\lambda_1 & & \\ & \ddots & \\ & & -\lambda_k \end{pmatrix} \text{ and } T^0 = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix}. \tag{1.14}$$

Kronecker compositions

Here we give some details on Kronecker compositions of matrices, mainly on Kronecker products and Kronecker sums. Kronecker product of two matrices $A = (a_{ij})_{n\times m}$ and $B = (b_{ij})_{r\times s}$ is given by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & \vdots & \vdots \\ a_{n1}B & \cdots & a_{nm}B \end{pmatrix}_{nr\times ms.}$$

If A and B are square matrices, *i.e.*, $n = m$ and $r = s$, then the Kronecker sum is given by $A \oplus B = A \otimes I_{n\times n} + I_{r\times r} \otimes B$, where I_n is the identity matrix of dimension n. For matrices A, B, C and D, some properties of Kronecker products are given below.

1. $A \otimes B \neq B \otimes A$, if $A_{n\times m}$, $B_{p\times q}$;
2. $A \otimes (B \pm C) = A \otimes B \pm A \otimes C$, if $A_{n\times m}$, $B_{p\times q}$ and $C_{p\times q}$;
3. $(A \pm B) \otimes C = A \otimes C \pm B \otimes C$, if $A_{n\times m}$, $B_{n\times m}$, $C_{p\times q}$;
4. $(kA) \otimes B = A \otimes (kB) = k(A \otimes B)$, if k scalar and $A_{n\times m}$, $B_{p\times q}$;
5. $(A \otimes B) \otimes C = A \otimes (B \otimes C) = A \otimes B \otimes C$, if $A_{n\times m}$, $B_{p\times q}$, $C_{r\times s}$;
6. $(A \otimes B)(C \otimes D) = AC \otimes BD$, if $A_{n\times m}$, $B_{p\times q}$, $C_{n\times r}$, $B_{q\times s}$;
7. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if $A_{n\times m}$ and $B_{p\times q}$ are invertible;
8. $e^{A\oplus B} = e^A \otimes e^B$, for $A_{n\times m}$, $B_{p\times q}$.

Discrete PH-distribution

Analogous to the definition of PH-distribution in continuous-time, a discrete-time PH (DPH) distribution is defined as the distribution of the time until absorption in a discrete-time Markov chain with m transient states, $\{1, 2, ..., m\}$, and one absorbing state $m + 1$,

with transition probability matrix given by

$$P = \begin{pmatrix} \mathbf{T} & T^0 \\ 0 & 1 \end{pmatrix}, \tag{1.15}$$

where $\mathbf{T}$ is a substochastic matrix such that $I - T$ is nonsingular and the exit vector T^0 satisfies $T^0 = \mathbf{1} - T\mathbf{1}$.

Definition 1.4.6. *Let $Z = min\{n \in \mathbb{N} : X_n = m+1\}$ denote a random variable denoting time until absorption in state $m+1$. The probability distribution of Z is said to be the DPH with parameter (α, T), denoted by $DPH(\alpha, T)$, and is given by*

$$p_n = P(Z = n) = \alpha T^{n-1} T^0, \quad n \geq 1 \tag{1.16}$$

$$\text{and} \qquad P(Z \leq n) = 1 - \alpha T^{n-1} \boldsymbol{1}. \tag{1.17}$$

In a DPH distribution $p_{m+1} = P(Z = m+1) = \alpha_{m+1} = 1 - \alpha\mathbf{1}$. We assume $\alpha_{m+1} = 0$ and also that $Q^* = T + T^0\alpha$ is irreducible. The stationary probability vector ξ of Q^* satisfies $\xi Q^* = \xi$, $\xi\mathbf{1} = 1$. The rate of arrival is given by $\lambda = \xi T^0$. Some properties of a DPH distribution are given below.

Lemma 1.4.8. *The probability generating function or z-transform of $DPH(\alpha, T)$ is given by*

$$P(z) = \sum_{n=0}^{\infty} z^n p_n = z\alpha(I - zT)^{-1} T^0, \quad \text{for } |z| < 1. \tag{1.18}$$

Lemma 1.4.9. *If Z follows $DPH(\alpha, T)$, mean and raw moments of Z are respectively given by*

$$E(Z) = \alpha(I - T)^{-1} \boldsymbol{1} \tag{1.19}$$

$$\text{and } E(Z^n) = n!\alpha T^{n-1}(I - T)^{-n} \boldsymbol{1}, \quad \text{for } n = 2, 3, \ldots. \tag{1.20}$$

Another expression for mean, in terms of stationary vector ξ, is given by

$$E(Z) = \left(\xi T^0\right)^{-1}. \tag{1.21}$$

These Lemmas with their detailed proofs can be found in Breuer [32]. The geometric distribution can be written in the form of DPH as follows:

A discrete random variable X is said to geometric distribution with parameter q if its probability mass function (p.m.f.) is given by

$$P(X = k) = (1 - q)q^{k-1}, \quad k = 1, 2, \ldots. \tag{1.22}$$

It has a discrete PH-representation with order $m = 1$, $\alpha = 1$ and $T = q$. The exit vector is given by $T^0 = 1 - q$.

1.5 Markovian arrival process

A renewal process is a recurrent event process with independent and identically distributed interarrival times, which describes an ordered set of points like, arrival instants, service completion epochs and equipment failure instants on $[0, \infty)$. Their main simplifying feature is the independence and equi-distributed interrenewal intervals. A renewal process with a PH-distribution for the interrenewal intervals is known as PH-renewal process. This process offers simplicity and algebraic tractability by describing the underlying Markovian structure in a method of phases. Poisson process is a PH-renewal process with exponentially distributed interarrivals.

A PH-renewal process has its interarrivals times as independent and identically distributed. But in modern communication systems like internet or other computer networks there may be strong correlations between subsequent interarrival times. Thus, to introduce a dependence between the subsequent renewal intervals without changing the structure of the generator matrix, a Markovian arrival process (MAP) is defined [32].

Let, in a PH-renewal process, the interarrival times follow a PH(α, T) distribution. Here, after an arrival (*i.e.*, a renewal event), the process gets instantaneously restarted by selecting a new initial state (independent of the past), using the same probability vector

α each time. We get the infinitesimal generator of such a PH-renewal process as

$$Q = \begin{pmatrix} T & A & & & \\ & T & A & & \\ & & T & A & \\ & & & \ddots & \ddots \end{pmatrix}, \quad \text{with} \quad A = T^0\alpha = \begin{pmatrix} T_1^0\ \alpha \\ T_2^0\ \alpha \\ \vdots \\ T_m^0\ \alpha \end{pmatrix}. \tag{1.23}$$

Relaxing this restriction, if we introduce a new matrix

$$B = \begin{pmatrix} T_1^0\ \alpha_1 \\ T_2^0\ \alpha_2 \\ \vdots \\ T_m^0\ \alpha_m \end{pmatrix} \tag{1.24}$$

such that B is non-negative and $\alpha_i\mathbf{1} = 1 \quad$ for all $i = 1, 2, ..., m$, so that we still have $T^0 = -T\mathbf{1}$. Here the new initial state after an arrival depends on the one immediately before that arrival. Therefore, denoting $D_0 = T$ and $D_1 = B$ we get the new matrix

$$Q = \begin{pmatrix} D_0 & D_1 & & & \\ & D_0 & D_1 & & \\ & & D_0 & D_1 & \\ & & & \ddots & \ddots \end{pmatrix}, \tag{1.25}$$

where D_0 and D_1 are n dimensional substochastic matrices. The matrix D_1 is non-negative *i.e.,* $(D_1)_{ij} \geq 0$ for $1 \leq i, j \leq m$. The matrix D_0 is nonsingular and has negative diagonal entries *i.e.,* $(D_0)_{ij} \geq 0$ for $1 \leq i \neq j \leq m$, $(D_0)_{ii} < 0$ for $1 \leq i \leq m$. Here, $(D_0)_{ij}$ represents the transition rates from phase i to phase j with no arrival and $(D_1)_{ij}$ represents the transition rates from phase i to phase j with only one arrival. A Markov chain with such a generator Q in (1.25) is called a Markovian Arrival Process (MAP) and is denoted by MAP(D_0, D_1). The matrix $D = D_0 + D_1$ is an infinitesimal generator with $D\mathbf{1} = 0$ and is assumed to be irreducible. Let π be the invariant probability vector of the Markov chain described by D, which is assumed to exist, satisfying

$$\pi D = \mathbf{0} \text{ and } \pi\mathbf{1} = 1.$$

The rate of arrival is given by $\lambda = \pi D_1 \mathbf{1}$. The MAP representation of PH-renewal process is $D_0 = T$ and $D_1 = T^0\alpha$.

A discrete-time MAP process (DMAP) is described similarly, by two m-dimensional substochastic matrices D_0 and D_1 and is such that $(I - D_0)$ is nonsingular and is denoted by DMAP(D_0, D_1) [102]. Here $(D_0)_{ij}$, $1 \leq i \neq j \leq m$, represents the transition probability from phase i to phase j with no arrival and $(D_1)_{ij}$ represents the transition probability from phase i to phase j with only one arrival. Therefore, $0 \leq (D_0)_{ij} \leq 1$ and $0 \leq (D_1)_{ij} \leq 1$. The matrix $D = D_0 + D_1$ is an irreducible stochastic matrix. The invariant probability vector π, of the Markov chain described by D, which is assumed to exist, satisfying

$$\pi D = \pi \text{ and } \pi \mathbf{1} = 1.$$

The rate of arrival in MAP is given by $\lambda = \pi D_1 \mathbf{1}$.

The concept of an MAP process is further extended to allow simultaneous batch absorptions in the underlying Markov chain. The resulting process is known as the batch Markovian arrival process (BMAP). Some special cases of MAP and DMAP processes are presented below.

Markov-modulated Poisson process

A Markov-modulated Poisson process (MMPP) is governed by a two-state Markov chain. In this process, when the Markov chain is in state one (state two), it generates an arrival according to a Poisson process of rate $\lambda_1(\lambda_2)$ and may remain in the state in the next time period with rate $\sigma_1(\sigma_2)$, where $\sigma_1, \sigma_2 > 0$ to avoid trivialities. A MMPP is a doubly stochastic Poisson process and can be characterized by a infinitesimal generator matrix $\widehat{Q}$ and an arrival rate matrix Λ, defined as

$$\widehat{Q} = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}. \tag{1.26}$$

The total arrival rate λ of the MMPP process is

$$\lambda = \frac{\sigma_1\lambda_2 + \sigma_2\lambda_1}{\sigma_1 + \sigma_2}. \tag{1.27}$$

The stationary distribution of this Markov chain exists, because of its finite state-space, and is given by

$$\pi = (\pi_1, \pi_2) = \frac{1}{\sigma_1 + \sigma_2}(\sigma_2, \sigma_1).$$

The squared coefficient of variation, c^2, of interarrival times of an MMPP process is an important measure of the degree of traffic burstiness, which is defined as the ratio of the variance to the square of mean of interarrival times. It is given by [45]

$$c^2 = 1 + \frac{2\sigma_1\sigma_2(\lambda_1 - \lambda_2)^2}{(\sigma_1 + \sigma_2)^2(\lambda_1\lambda_2 + \lambda_2\sigma_1 + \lambda_1\sigma_2)}. \tag{1.28}$$

The autocorrelation coefficient, which is the ratio of covariance of interarrival times to the product of their standard deviations (or to the variance in case of an independent and identically distributed interarrivals) is used to measure the amount of correlation in the process. The autocorrelation function of the interarrival time with step-1 is given by

$$\psi_1 = \frac{\lambda_1\lambda_2(\lambda_1 - \lambda_2)^2\sigma_1\sigma_2}{c^2(\sigma_1 + \sigma_2)^2[\lambda_1\lambda_2 + \lambda_2\sigma_1 + \lambda_1\sigma_2]^2}. \tag{1.29}$$

The MAP representation of a MMPP process is $D_0 = \widehat{Q} - \Lambda$, $D_1 = \Lambda$ *i.e.*,

$$D_0 = \begin{bmatrix} -\sigma_1 - \lambda_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 - \lambda_2 \end{bmatrix} \text{ and } D_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Markov-modulated Bernoulli process

Markov-modulated Bernoulli process (MMBP) is the discrete-time analogue of the MMPP process. A two state MMBP arrival process is characterized by the transition probability matrix $\widehat{P}$ and the arrival rate matrix Λ defined as

$$\widehat{P} = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \tag{1.30}$$

where $0 \leq p \leq 1$ to avoid trivialities and the total arrival rate λ is,

$$\lambda = \frac{\lambda_1(1-q) + \lambda_2(1-p)}{2-p-q}. \tag{1.31}$$

The squared coefficient of variation of interarrival times of the MMBP process [110] is

$$\begin{aligned} c^2 &= \frac{2[\lambda_1(1-q)+\lambda_2(1-p)]}{\lambda_1(1-q)+\lambda_2(1-p)+\lambda_1\lambda_2(p+q-1)} \\ &+ \frac{2[\lambda_1(1-p)+\lambda_2(1-q)][\lambda_1(1-q)+\lambda_2(1-p)](p+q-1)}{(2-p-q)^2[\lambda_1(1-q)+\lambda_2(1-p)+\lambda_1\lambda_2(p+q-1)]} \\ &- \frac{\lambda_1(1-q)+\lambda_2(1-p)}{2-p-q} - 1. \end{aligned} \tag{1.32}$$

The autocorrelation function of interarrival time with step-1 is given by

$$\psi_1 = \frac{\lambda_1\lambda_2(\lambda_1-\lambda_2)^2(1-p)(1-q)(p+q-1)^2}{c^2(2-p-q)^2[\lambda_1(1-q)+\lambda_2(1-p)+\lambda_1\lambda_2(p+q-1)]^2}. \tag{1.33}$$

The DMAP representation of a MMBP process is $D_0 = \widehat{P}(I-\Lambda)$, $D_1 = \widehat{P}\Lambda$ and this gives

$$D_0 = \begin{bmatrix} p(1-\lambda_1) & (1-p)(1-\lambda_2) \\ (1-q)(1-\lambda_1) & q(1-\lambda_2) \end{bmatrix} \text{ and } D_1 = \begin{bmatrix} p\lambda_1 & (1-p)\lambda_2 \\ (1-q)\lambda_1 & q\lambda_2 \end{bmatrix}.$$

1.6 Quasi-Birth-Death process

A birth-death process is a continuous-time Markov chain in which only one-step transitions to the neighboring states are allowed. A canonical example of a birth-death process is the M/M/1 queue, as the process of number of customers gives a tridiagonal infinitesimal generator matrix. We can classify the states of a birth-death process into boundary states and repeating states. Generalizing to vector form, if the nearest neighbor transitions are interpreted in terms of vectors of states (called levels) *i.e.*, the state transitions are only possible within a level (called phases) or between adjacent levels, then we will get a repetitive structure of the transition matrix similar to the birth-death process. Only the entries are in vector form instead of scalars. This vector process is called a 'quasi-birth-death process' or simply a QBD process [102].

Let $\{X_t, t \geq 0\}$ be a continuous-time Markov chain on two-dimensional state-space

$$E = \{(n,i),\ n \geq 0, 1 \leq i \leq m\}.$$

The first coordinate of a state (n, i), is called the **level** and the second coordinate i is called the **phase**. The number of phases, m, in each level may be either finite or infinite. The state-space can be partitioned on the basis of the levels as $E = \cup_{n\geq 0} l(n)$, where $l(n) = \{(n,1),(n,2),\ldots,(n,m)\}$ for $n \geq 0$.

Definition 1.6.1. *A Markov chain* $\{X_t, t \geq 0\}$ *with* $l(0) = \{(0,1),(0,2),\ldots,(0,m')\}$ *and* $l(n) = \{(n,1),(n,2),\ldots,(n,m)\}$ *for all* $n \geq 1$, *is called a QBD process if the following properties hold:*

1. *One-step transitions from a state are restricted to states in the same or in the two adjacent levels. In other words, transition from* (n,i) *to* (n',j) *is not possible if* $|n - n'| \geq 2$.

2. *For* $n \geq 1$, *the instantaneous transition rate between two states in the same level* $l(n)$ *or between two states in the levels* $l(n)$ *and* $l(n+1)$ *does not depend on* n *i.e., for* n *and* $n' \geq 1$, *the transition rate from* (n,i) *to* (n',j) *may depend on* i,j *and* $n - n'$, *but not on specific values of* n *and* n'.

The infinitesimal generator matrix of a QBD process has the following structure

$$Q = \begin{pmatrix} B_{00} & B_{01} & & & \\ B_{10} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1.34}$$

where B_{00} *is a square matrix of dimension* m', B_{01} *is of dimension* $m' \times m$, B_{10} *is* $m \times m'$ *and* $A_i, i = 0,1,2$, *are square matrices of dimension* m.

For $n = m = m' = 1$, a QBD process reduces to a simple birth-death process.

Throughout this thesis, we will assume that the generator Q of a QBD process is irreducible. The matrix $A = A_0 + A_1 + A_2$ has negative diagonal and non-negative off diagonal elements. For m finite, A is a finite generator matrix with $A\mathbf{1} = \mathbf{0}$. The matrix

A can be irreducible or reducible. Let $\mathbf{x} = [x_0, x_1, ...]$ be the stationary probability vector of the QBD process Q, if exists, such that

$$\mathbf{x}Q = \mathbf{0}, \quad \mathbf{x}\mathbf{1} = 1, \tag{1.35}$$

from which we get the steady-state equations as

$$\begin{aligned} x_0B_{00} + x_1B_{10} &= \mathbf{0} \\ x_0B_{01} + x_1A_1 + x_2A_2 &= \mathbf{0} \\ \text{and} \quad x_{i-1}A_0 + x_iA_1 + x_{i+1}A_2 &= \mathbf{0}, \quad \forall\, i \geq 2. \end{aligned} \tag{1.36}$$

Lemma 1.6.1. *A continuous-time irreducible QBD process is positive recurrent if and only if the minimal non-negative solution R to the matrix-equation*

$$R^2A_2 + RA_1 + A_0 = \boldsymbol{0}, \tag{1.37}$$

has spectral radius $sp(R) < 1$ and finite system of equations

$$x_0B_{00} + x_1B_{10} = \boldsymbol{0} \tag{1.38}$$

$$x_0\boldsymbol{1} + x_1\left(I - R\right)^{-1}\boldsymbol{1} = 1, \tag{1.39}$$

has a unique positive solution $[x_0, x_1]$. The stationary probability vector is such that

$$x_n = x_1R^{n-1}, \text{ for } n \geq 1. \tag{1.40}$$

R is called the rate matrix which records the expected sojourn times in the states of $l(n+1)$ starting from $l(n)$, avoiding $\cup_{0\leq k\leq n} l(k)$. The matrix R is also interpreted as recording the rate of sojourn times in the states of $l(n+1)$, per unit of the local time of $l(n)$.

Lemma 1.6.2. *Let a continuous-time QBD process be irreducible and with irreducible matrix A. The QBD process is positive recurrent if and only if*

$$\hat{\pi}A_0\boldsymbol{1} < \hat{\pi}A_2\boldsymbol{1}, \tag{1.41}$$

where $\hat{\pi}$ is the unique solution of the system $\hat{\pi}A = \boldsymbol{0}$, $\hat{\pi}\boldsymbol{1} = 1$.

These Lemmas with their detailed proofs can be found in Latouche and Ramaswami [102].

Discrete-time QBD process

A discrete-time QBD process is a Markov chain $\{X_t, t = 0, 1, \ldots\}$ on the state-space $\cup_{n\geq 0} l(n)$. The transition probabilities are assumed to be level independent *i.e.,* for $n, n^{'} \geq 1$, the probability $P\{X_1 = (n^{'}, j)|X_0 = (n, i)\}$ may depend on i, j and $n - n^{'}$ but not on the specific values of n and $n^{'}$. Thus, the transition probability is block diagonal and has the form

$$P = \begin{pmatrix} B_{00} & B_{01} & & & \\ B_{10} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1.42}$$

where B_{00}, B_{01}, B_{10}, and A_i, $i = 0, 1, 2$, are non-negative matrices. The matrix $A = A_0 + A_1 + A_2$ is stochastic. We assume the QBD process is irreducible, aperiodic and positive recurrent. The stationary probability vector π satisfies $\pi P = \pi$ with $\pi \mathbf{1} = 1$.

Lemma 1.6.3. *If a discrete-time irreducible QBD process is positive recurrent, then*

1. *the minimal non-negative solution R to the matrix-equation*

$$R^2 A_2 + R A_1 + A_0 = R, \tag{1.43}$$

has spectral radius $sp(R) < 1$ i.e., $(I - R)^{-1}$ exists,

2. *the stationary probability vector is such that*

$$x_n = x_1 R^{n-1}, \text{ for } n \geq 1, \tag{1.44}$$

3. *the matrix*

$$B[R] = \begin{bmatrix} B_{00} & B_{01} \\ B_{10} & A_1 + RA_2 \end{bmatrix} \tag{1.45}$$

is stochastic, and

4. the finite system of equations

$$x_0 B_{00} + x_1 B_{10} = x_0 \tag{1.46}$$

$$x_0 \boldsymbol{1} + x_1 (I - R)^{-1} \boldsymbol{1} = 1, \tag{1.47}$$

has a unique positive solution $[x_0, x_1]$.

The rate matrix R records the expected number of visits to $l(n+1)$ starting from $l(n)$, avoiding $\cup_{0\le k\le n} l(k)$ *i.e.,* R_{ij} $(1 \le i, j \le m)$ is the expected number of visits to $(n+1, j)$ before return to $l(0) \cup l(1) \cup \cdots \cup l(n)$, given that the process starts in (n, i).

Lemma 1.6.4. *Let a discrete-time QBD process be irreducible and the matrix A be also irreducible. The QBD process is positive recurrent if and only if*

$$\hat{\pi} A_0 \boldsymbol{1} < \hat{\pi} A_2 \boldsymbol{1}, \tag{1.48}$$

where $\hat{\pi}$ is the unique solution of the system $\hat{\pi} A = \hat{\pi}$, $\hat{\pi} \boldsymbol{1} = 1$.

In a discrete-time QBD, if the matrix A is reducible but with finite number of phases, $m < \infty$, then it can be written, possibly after a permutation of its rows and columns, as

$$A = \begin{pmatrix} C^{(1)} & 0 & \dots & 0 & 0 \\ 0 & C^{(2)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & C^{(K)} & 0 \\ D^{(1)} & D^{(1)} & \dots & D^{(K)} & D^{(0)} \end{pmatrix}, \tag{1.49}$$

where the blocks $C^{(k)}$, $1 \le k \le K$, are irreducible and stochastic and $D^{(0)}$ is substochastic [102]. Since the matrices A_i, $i = 0, 1, 2$, all are nonnegative for a discrete-time model and are similarly structured, after the same purmutations as we have for A, we get

$$A_i = \begin{pmatrix} C_i^{(1)} & 0 & \dots & 0 & 0 \\ 0 & C_i^{(2)} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & C_i^{(K)} & 0 \\ D_i^{(1)} & D_i^{(1)} & \dots & D_i^{(K)} & D_i^{(0)} \end{pmatrix}. \tag{1.50}$$

Lemma 1.6.5. *Let a discrete-time QBD be irreducible and the number of phases, m, be finite. If A is partitioned as (1.49), where $K \geq 1$ and the matrices $C^{(k)}$, $1 \leq k \leq K$ are irreducible, the QBD process is recurrent if and only if*

$$\gamma^{(k)} C_0^{(k)} \mathbf{1} \leq \gamma^{(k)} C_2^{(k)} \mathbf{1}, \text{ for all } k,\ 1 \leq k \leq K, \tag{1.51}$$

where $\gamma^{(k)}$ is the stationary probability vector of $C^{(k)}$. The QBD is positive recurrent if and only if all the inequalities are strict i.e.,

$$\gamma^{(k)} C_0^{(k)} \mathbf{1} < \gamma^{(k)} C_2^{(k)} \mathbf{1}, \text{ for all } k,\ 1 \leq k \leq K. \tag{1.52}$$

The QBD is transient if and only if there exists at least one k, $1 \leq k \leq K$, such that

$$\gamma^{(k)} C_0^{(k)} \mathbf{1} > \gamma^{(k)} C_2^{(k)} \mathbf{1}. \tag{1.53}$$

Proofs of these Lemmas can be found in Latouche and Ramaswami [102].

Non-Homogeneous QBD process

A non-homogeneous continuous-time QBD process has the form

$$Q = \begin{pmatrix} A_1^{(0)} & A_0^{(0)} & & & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & & & \\ & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & & \\ & & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1.54}$$

Like the homogeneous case, the state-space of a non-homogeneous QBD is two dimensional and partitioned into levels. The transitions are still allowed between adjacent levels only, but the transition rates out from $l(n)$ depend on the level n. Also different levels may have different number of phases. The stationary distribution of such a process is given in the result below.

Lemma 1.6.6. *Let a continuous-time non-homogeneous QBD be irreducible, aperiodic and positive recurrent. Then the stationary probability distribution of the QBD satisfies*

$$x_n = x_{n-1} R^{(n)}, \text{ for } n \geq 1, \tag{1.55}$$

where $R^{(n)}$ *satisfies the equations*

$$A_0^{(n-1)} + R^{(n)}A_1^{(n)} + R^{(n)}R^{(n+1)}A_2^{(n+1)} = 0, \text{ for } n \geq 1. \tag{1.56}$$

The matrix $R^{(n)}$ *records the expected rate of sojourn time in* $l(n)$ *between two visits to* $l(n-1)$, *[102].*

In case of non-homogeneous QBD processes with infinite state-space, no easily computable analytic expressions are available. In most of the cases, the generator matrix is truncated in some arbitrary but large level and solved numerically the system of equations $\pi Q = 0$ and $\pi \mathbf{1} = 1$. Various algorithms are also introduced to solve such non-homogeneous QBD processes.

1.7 Solution methods of queueing models

There are several techniques to solve a queueing system. The most frequently presented method in queueing literature is the imbedded Markov chain approach in which we look at the queue behavior at the instant of an arrival or of a service completion in order to get a complete information of system state at those observation epochs. But sometimes such observation epochs are not sufficient to describe a system. To get a Markov process out of the non-Markovian one, a method called supplementary variables technique is used in which the joint distribution of queue length and the supplementary variable can be derived in a simple and convenient way. This technique was first proposed by Cox [44]. A Markov chain is analyzed using method of transforms, approximation method, matrix-geometric method and many others. Some of these methods are described below.

1.7.1 Transform methods

In many queueing problems, the form of distribution function or probability mass function of a system performance measure may be so complex that it makes the computations

difficult, if not impossible. A transform can provide a compact description of a distribution, and it is relatively easy to compute the mean, the variance and other moments directly from a transform rather than resorting to a tedious sum or a tedious integral. The transform methods are particularly useful in problems involving sums of independent random variables and in solving differential difference equations related to a queueing process. Some such useful transform methods are z-transform (or probability generating function), Laplace-Stieltjes transform, Fourier transform (or characteristic function) etc.

Probability Generating Function: The probability generating function (PGF) or the z-transform of a discrete random variable is a power series representation of the probability mass function of the random variable. PGFs are often employed for their concise description of the sequence of probabilities with the help of well developed theory of power series with non-negative coefficients. It can be defined as follows:

Definition 1.7.1. *If X is a non-negative integral-valued discrete random variable with $P(X = k) = p_k$. The PGF of X is defined as*

$$G_X(z) = \sum_{k=0}^{\infty} z^k p_k = p_0 + zp_1 + z^2 p_2 + \cdots + z^k p_k + \cdots, \ for\ |z| < 1. \tag{1.57}$$

It can be verified easily that

$$G_X(1) = 1 = \sum_{k=0}^{\infty} p_k.$$

We can determine interesting quantities such as the mean and the variance of the random variable from the PGF itself. A PGF uniquely defines the distribution of a random variable.

Laplace-Stieltjes Transforms: Laplace-Stieltjes Transform (LST) is very effective for solving linear differential equations by reducing it to an algebraic equation. Distinct probability distributions on $[0, \infty)$ have distinct LSTs. The definition of LST of a random variable is given below.

Definition 1.7.2. *If X is a non-negative random variable with distribution function $F(t) = P(X \leq t)$. The LST of probability distribution of X is given as*

$$f^*(s) = \int_{t=0}^{\infty} e^{-st} dF(t), \qquad for\ Re(s) \geq 0. \tag{1.58}$$

We have $f^*(s) = E(e^{-sX})$ *and* $f^*(0) = 1$.

For a discrete random variable $f^*(s) = G_X(e^{-s})$, *i.e.,* the LST of a random variable differs from its PGF only by a replacement of z with e^{-s}. Thus there is a close analogy between the properties of LST and those of PGF.

1.7.2 Matrix-Geometric Method

Matrix-geometric method is a powerful technique to solve and analyze complex queueing models. This method handles the block matrices of system states and transitions within the states instead of dealing with individual states or transition probabilities. This approach of solving makes the expressions in matrix form simpler than the corresponding expressions given in terms of eigenvalues. Also the fundamental matrices have direct probabilistic interpretation, while the eigenvalues do not. In this method, the geometric relation between vectors of the stationary probability distribution allows for simple closed-form formulas for the computation of measures of interest such as the system queue length and busy cycle. One of the practical advantages of this method is that these elementary matrix operations can easily be programmed for a high-speed computer. The use of PH-distributions in the representation of system elements and the matrix-geometric method in their analysis has significantly expanded the scope of queueing systems. Neuts [120] has developed this matrix-geometric method which extends and modifies the earlier transform methods and makes it amenable for an algorithmic solution. Matrix-geometric techniques are applicable to both continuous and discrete-time Markov processes.

If x is the stationary probability vector of a QBD process with generator Q of the form (1.34), from the equations

$$x_0 B_{00} + x_1 B_{10} = 0 \quad \text{and} \quad x_0 \mathbf{1} + x_1 (I - R)^{-1} \mathbf{1} = 1, \tag{1.59}$$

the values of x_0 and x_1 are found, and for $i \geq 2$, x_i's are derived from

$$x_{n+1} = x_n R, \quad n = 1, 2, \ldots, \tag{1.60}$$

where the rate matrix R is the non-negative solution to the matrix-equation

$$R^2 A_2 + RA_1 + A_0 = 0, \text{ with } sp(R) < 1. \tag{1.61}$$

This method of solving a QBD process with the help of the rate matrix R is called the matrix-geometric method. The rate matrix has physical interpretations as given in Section 1.6.

To evaluate the matrix R, several algorithms has been proposed and developed. Neuts [120], Latouche and Ramaswami [102], Diagle and Lucantoni [49] are some of the most important references on such algorithms. Most of the algorithms proceed by successive substitutions and converge linearly. They are easily implementable and are often efficient. Latouche and Ramaswami [101] developed an iterative algorithm 'logarithmic reduction algorithm' which is quadratically convergent and also has good numerical stability having small number of iterations. Based on the theory of matrix factorization, Naoumov further developed the logarithmic algorithm [156].

For non-homogeneous QBD processes, analytic solutions are complicated and sometimes it seems too difficult to derive exactly the stationary probabilities. Neither a closed-form solution nor a direct algorithmic computation of the stationary probabilities can been obtained for models having spatially non-homogeneous generator matrices. So numerical approximation methods are often used to compute the stationary distributions of such processes.

1.8 Outline of the thesis

The research work contained in this thesis is split into chapters as outlined below.

Chapter 2 considers a MAP/PH/1/WV queue which captures both burstiness and correlation in the interarrival times. The model is analyzed in discrete-time as well as in continuous-time to have certain insights on effects of correlation of the arrival traffic on system efficiency.

Chapter 3 deals with a finite-buffer MAP/PH/1/WV queueing model. The emphasis

of the study is laid on the impact of buffer size and loss probabilities on the performance of a system which follows the WV policy.

Chapter 4 includes a PH/M/c/WV model where the server obeys asynchronous multiple WV policy. The customers considered here, become impatient, if upon their arrival all the servers are busy and the arriving customers have to join a queue. This model is studied to observe the influence of number of servers in an WV model having different arrival patterns and impatient customers.

Chapter 5 considers a M/PH/1/WV queue with two priority classes, called higher priority class and lower priority class, governed by different Poisson inputs. The WVs are interrupted if the system finds any customer waiting in queue after completing a service in WV. The significance of this analysis is to see the effect of higher priority class over lower priority class in an WV environment.

Chapter 6 handles the retrial phenomena in a M/G/1/WV model. If an arriving customer finds the system busy upon its arrival it joins an orbit and keep retrying for service. To observe the consequence of repeated attempts of customers in a WV model, we analyze the system for different service time distributions.

Chapter 7 investigates the effect of impatient behavior of customers in a M/M/1/WV model. Two types of WV termination policies are taken, the multiple WV (MWV) policy and the single WV (SWV) policy. Closed-form probabilities are derived for both the models and the models are compared in search of a better one that enhances the system performance.

Chapter 8 gives the conclusion of the research work and some possible extensions of this study.

Chapter 2

A Queue with Correlated Arrivals

In high-speed communication networks, WDM is the most used method of transmitting packets. The packet traffics are highly correlated and bursty. Burstiness describes the tendency of traffic to occur in clusters. Packet burstiness is represented by squared coefficient of variation whereas the autocorrelation function describes the correlation between the packets. To study the effect of bursty and correlated traffic on WDM networks, we analyze in this chapter, the MAP/PH/1 model with multiple WVs. We consider the model in discrete-time as well as in continuous-time. Continuous-time models are often used for network modelling. Though dealing with discrete-time models is more complicated than continuous-time ones, interest is growing for discrete-time models due to the introduction of digital data transmission technology like ATM (Asynchronous Transfer Mode), because these systems are assumed to work synchronously on the basis of a smallest time unit. In this chapter, we give a detailed analysis of the WV model in discrete-time and also outline its continuous-time analogue. As for the traffic model, we have chosen here the MAP process due to the following reasons: the MAP is one of the most powerful traffic models in terms of its ability to mimic complex statistical properties of network traffic *e.g.,* the self-similar nature and the long-range dependent behavior [42]. Also, Servi and Finn [139] have remarked over the necessity of considering correlated arrivals as the traffic sources are highly correlated. A related work in this direction is that of Alfa [2] who, using

the matrix-geometric procedure, has analyzed the discrete-time MAP/PH/1 queue with exhaustive and non-exhaustive services but with classical vacations. Baba [22], Li et al. [107], Tian [148] and Wu and Takagi [163] have analyzed WV queues having independent arrivals of packets.

2.1 The discrete-time MAP/PH/1/WV model

We consider a discrete-time single-server queue where customers arrive according to a DMAP(D_0, D_1). Here, $(D_0)_{ij}$ represents the transition probability from phase i to phase j with no arrival and $(D_1)_{ij}$ represents the transition probability from phase i to phase j with only one arrival. If π is the invariant probability vector of the Markov chain described by $D = D_0 + D_1$ satisfying $\pi D = \pi$ and $\pi\mathbf{1} = 1$, then the customer arrival rate can be given by $\lambda = \pi D_1 \mathbf{1}$, where $\mathbf{1}$ is the column vector of ones of appropriate dimension.

The service times during non-vacation periods are assumed to follow a DPH(β, S) of dimension m and rate μ_b with $\mu_b = \xi S^0$, where ξ is the stationary probability vector of $S + S^0\beta$ satisfying $\xi(S + S^0\beta) = \xi$ and $\xi\mathbf{1} = 1$. Here, $(S)_{ij}$ represents the transition probability from phase i to phase j with no service completion and $(S^0)_i$ represents the probability of service completion from phase i, with the system in non-vacation. The distribution of service times during a WV period is also a DPH$(\beta, \bar{S})$, of dimension m, having the same initial probability vector β. During a WV period, the server serves at a rate μ_v such that $\mu_v < \mu_b$. The rate is $\mu_v = \bar{\xi}\bar{S}^0$, where $\bar{\xi}(\bar{S} + \bar{S}^0\beta) = \bar{\xi}$ and $\bar{\xi}\mathbf{1} = 1$.

The duration of a WV is also DPH(δ, L), of dimension r with rate θ. From a non-vacation period, the server takes a WV as soon as the system becomes empty. If a customer arrives during a WV period, the system serves the customer at the rate μ_v. At the end of a WV, if at least one customer is present in the system, the server switches its service rate from μ_v to μ_b and a non-vacation period starts. Also, if the system has an ongoing service at a vacation termination epoch, the server switches its service rate from μ_v to μ_b and the service is continued at the rate μ_b until completion. On the other hand, if the server finds the system empty at a vacation termination epoch, the system

takes another WV. The customers are served according to the arrival order *i.e.*, FCFS. The interarrival times, service times and vacation times, are all assumed to be mutually independent.

In discrete-time queueing systems, the arrivals, the departures and the beginning/end of vacations may occur at the same time. It is assumed that time is slotted in intervals of fixed-length with length of slot being unity. Without loss of generality, we can assume that if there is an arrival at time t, then its precise occurrence lies somewhere inside the interval (t, t^+), where t^+ is the moment immediately after t. On the other hand, if there is a departure at time t, then its precise occurrence lies somewhere inside the interval (t^-, t), where t^- is the moment immediately before t. This model is referred to as an early arrival system (EAS). The beginning and ending of vacations also occurs at the instant t^+.

In this system, we need to keep track of the phase of arrival process, phase of service time and the phase of vacation duration to have complete information of the system. Given the just defined sign convention, the state of the system under consideration can be given by a discrete-time Markov chain

$$\Delta_d = \left\{ \left(N_t, (Q_t, J_t, (V_t, K_t^v)) \cup (Q_t, J_t, K_t^b)\right), t = 0, 1, \ldots \right\},$$

where N_t is the number of customers in the system at time t^+, Q_t represents the vacation-state of the server at time t^+ *i.e.*, $Q_t = 0$, if the server is in vacation and $Q_t = 1$, if the server is in non-vacation period, J_t is the phase of arrival, V_t is the phase of the vacation duration, K_t^v gives the phase of service during WV and K_t^b gives the phase of the service in non-vacation period. The state-space of the Markov chain Δ_d is given by

$$E = \Big\{ \{\{0\} \times \{(0, i, j)\}\} \cup \{\mathbb{N} \times \{\{(0, i, (j, k))\} \cup \{(1, i, k)\}\}\}, \\ 1 \le i \le n, 1 \le j \le r, 1 \le k \le m \Big\}$$

as, when the the system is empty, there is no need to keep track of the phase of service. The first coordinate n of a state $(n, (0, i, (j, k)))$ (or $(n, (1, i, k))$) is called the level of the process which gives the number of customers in the system.

Using the lexicographical ordering of the states as, $(0,(0,1,1)),\ldots,(0,(0,1,r)),(0,(0,2,1)),\ldots,(0,(0,n,r)),\ (1,(0,1,(1,1))),\ \ldots,\ (1,(0,n,(r,m))),\ (1,(1,1,1)),\ \ldots,\ (1,(1,n,m)),\ (2,(0,1,(1,1))),\ \ldots$, we get the transition probability matrix P of the Markov chain Δ_d as

$$P = \begin{pmatrix} B_{00} & B_{01} & & & & \\ B_{10} & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2.1}$$

where $B_{00} = D_0 \otimes (L + L^0\delta)$, $B_{01} = [D_1 \otimes L \otimes \beta \quad D_1 \otimes L^0\beta]$, $B_{10} = \begin{bmatrix} D_0 \otimes (L + L^0\delta) \otimes \bar{S^0} \\ D_0 \otimes S^0\delta \end{bmatrix}$, $A_i = \begin{bmatrix} A_{i3} & A_{i2} \\ 0 & A_{i0} \end{bmatrix}$, $i = 0, 1, 2$, $A_{13} = D_0 \otimes L \otimes \bar{S} + D_1 \otimes L \otimes \bar{S^0}\beta$, $A_{12} = D_0 \otimes L^0 \otimes \bar{S}(\mathbf{1}\bar{\xi}) + D_1 \otimes L^0 \otimes \bar{S^0}\beta$, $A_{10} = D_0 \otimes S + D_1 \otimes S^0\beta$, $A_{03} = D_1 \otimes L \otimes \bar{S}$, $A_{02} = D_1 \otimes L^0 \otimes \bar{S}(\mathbf{1}\bar{\xi})$, $A_{00} = D_1 \otimes S$, $A_{23} = D_0 \otimes L \otimes \bar{S^0}\beta$, $A_{22} = D_0 \otimes L^0 \otimes \bar{S^0}\beta$, $A_{20} = D_0 \otimes S^0\beta$. The vector $\bar{\xi}$ gives the initial probability of switched service from WV to non-vacation period. $A_0, A_1,$ and A_2 are square matrices of dimension $nm(r+1)$, B_{00} is a square matrix of dimension nr, B_{01} is of dimension $nr \times nm(r+1)$ and B_{10} is of dimension $nm(r+1) \times nr$. The formation of the entries of this transition matrix P is as follows.

$B_{00} = D_0 \otimes (L + L^0\delta) = D_0 \otimes L + D_0 \otimes L^0\delta$, represents the probability of the event that the system remains in level 0. This event can occur if any of the two cases happens. First is, the system has no arrival and no vacation completion, but only internal phase transition of arrival and of vacation. Second is, the system has no arrival and after completing a vacation, the server takes another vacation (since the system is empty) with rate δ.

$B_{01} = [D_1 \otimes L \otimes \beta \quad D_1 \otimes L^0\beta]$, is the probability of the system changing its level from 0 to 1. This can happen only if a customer arrives. The system will remain in vacation if the vacation does not terminate but only internal phase transition happens and the new arrival immediately moves for service with initial probability β. The system will move to

non-vacation from the vacation period, if the vacation terminates when the new customer enters the system.

$B_{10} = \begin{bmatrix} D_0 \otimes (L + L^0\delta) \otimes \bar{S^0} \\ D_0 \otimes S^0\delta \end{bmatrix}$, represents the probability of the system changing its level from 1 to 0. This can happen as follows. If the system was in vacation, either it remains in vacation after completing service of the only customer, during which no arrival happens, or after completing the service during vacation, the system takes another vacation with rate δ. When the system was in non-vacation and the service of the only customer is completed, the system becomes empty and no new arrival happens. So, the system immediately takes a vacation with probability δ.

The matrix A_1 represents the probability of the system remaining in the same level. Its entries are elaborated here.

- A_{13} is the probability of the event that the system remains in WV and has no change in total number of customers. This event occurs if either of the two cases happen. First is, the system has no arrival, no service completion and no vacation completion, but it only has internal phase transitions. The second event is when the system has an arrival and a service completion but not a vacation termination. So, we get

$$A_{13} = D_0 \otimes L \otimes \bar{S} + D_1 \otimes L \otimes \bar{S^0}\beta.$$

- A_{12} gives the probability of the event that the system remains in same level and changes its state from WV to non-vacation. Depending on whether a WV terminates in between an ongoing service or after a service completion, we can have two possibilities. First is, a WV terminates in between an ongoing service and the server switches its service rate with probability $\bar{\xi}$, during which no arrival happens. Second case is when the system has an arrival and a WV terminates with a service completion, to keep the system in the same level. Therefore, we have

$$A_{12} = D_0 \otimes L^0 \otimes \bar{S}(\mathbf{1}\bar{\xi}) + D_1 \otimes L^0 \otimes \bar{S^0}\beta.$$

- A_{10} is the probability of the event that the system remains in non-vacation period and in the same level. It is the probability of the event that either the system has no arrival and no service completion or the system has a service completion in non-vacation with an arrival. We get

$$A_{10} = D_0 \otimes S + D_1 \otimes S^0\beta.$$

A_0 is the probability of the system changing its level from n to $n+1,\ n \geq 1$. Its entries are described below.

- A_{03} gives the probability of the event that the system remains in WV but changes its level from n to $n+1$. This is equal to the event that the system has one arrival with no service completion and no vacation termination. Therefore,

$$A_{03} = D_1 \otimes L \otimes \bar{S}.$$

- A_{02} gives the probability that the server changes its state from WV to non-vacation and also changes its level from n to $n+1$. This is equal to the event that the vacation terminates in between an ongoing service and the service rate switches to the higher one with probability $\bar{\xi}$, during which a new customer arrives. We get

$$A_{02} = D_1 \otimes L^0 \otimes \bar{S}(\mathbf{1}\bar{\xi}).$$

- A_{00} is the probability of the event of the system remaining in non-vacation but changing its level from n to $n+1$. This is equal to the event that a customer arrives and neither a service completes nor a vacation terminates, but only their internal phase transitions happen. This gives

$$A_{00} = D_1 \otimes S.$$

A_2 gives the probability of the system changing its level from n to $n-1,\ n \geq 2$. Its entries are explained below.

- A_{23} is the probability of the system remaining in WV and changing its level from n to $n-1$. This is equal to the event that a service completes and neither an arrival nor a vacation terminates, but only internal transitions of arrival and vacation phases happen. So, we get

$$A_{23} = D_0 \otimes L \otimes \bar{S}^0 \beta.$$

- A_{22} is the probability that the server changes its state from WV to non-vacation and also the system changes its level from n to $n-1$. This probability can happen only when vacation terminates and a service completes but no new customer arrives. Therefore,

$$A_{22} = D_0 \otimes L^0 \otimes \bar{S}^0 \beta.$$

- A_{20} is the probability of the event that the system remains in non-vacation and changes its level from n to $n-1$. This happens if the system has a service completion and no new arrivals.

$$A_{20} = D_0 \otimes S^0 \beta.$$

This Markov chain is a discrete-time QBD process with state-space E. We assume, without loss of generality, that this QBD process is irreducible.

Special case: If the arrival process, service processes and vacation durations, all are taken to be geometric distributions, this transition matrix P exactly resembles the one given by Tian et al. [148].

2.1.1 Stability condition

Theorem 2.1.1. *The irreducible QBD process Δ_d is positive recurrent if and only if*

$$\lambda < \mu_b.$$

Proof. The matrices A_i, $i = 0, 1, 2$, of the transition matrix P, are upper triangular and

the matrix $A = A_0 + A_1 + A_2$ is reducible and stochastic. The matrix A is

$$A = \begin{bmatrix} D \otimes L \otimes (\bar{S} + \bar{S^0}\beta) & D \otimes L^0 \otimes \left(\bar{S}(\mathbf{1}\bar{\xi}) + \bar{S^0}\beta\right) \\ 0 & D \otimes (S + S^0\beta) \end{bmatrix}.$$

Now, Lemma 1.6.5 gives the condition for positive recurrence of the QBD process when A is reducible. The matrix A can be written after permutation of its rows and columns as

$$A = \begin{bmatrix} C^{(1)} & 0 \\ C^{(2)} & C^{(0)} \end{bmatrix},$$

where $C^{(1)}$ is irreducible and stochastic and $C^{(0)}$ is sub-stochastic. Since the matrices A_0, A_1 and A_2 are all non-negative and are similarly structured as A, so after the same permutations of rows and columns, we get for $i = 0, 1, 2$,

$$A_i = \begin{bmatrix} C_i^{(1)} & 0 \\ C_i^{(2)} & C_i^{(0)} \end{bmatrix}$$

with $C_i^{(1)} = A_{i0}, C_i^{(2)} = A_{i2}$ and $C_i^{(0)} = A_{i3}$. Lemma 1.6.5 states that the QBD process is positive recurrent if and only if

$$\hat{\pi} C_2^{(1)} \mathbf{1} > \hat{\pi} C_0^{(1)} \mathbf{1}, \tag{2.2}$$

where $\hat{\pi}$ is the stationary probability vector of $C^{(1)}$. Here $C^{(1)} = D \otimes (S + S^0\beta)$ and $\hat{\pi} = \pi \otimes \xi$ such that $\hat{\pi} C^{(1)} = \hat{\pi}$ and $\hat{\pi}\mathbf{1} = 1$. From equation (2.2), we have

$$(\pi \otimes \xi)(D_0 \otimes S^0\beta)\mathbf{1} \;>\; (\pi \otimes \xi)(D_1 \otimes S)\mathbf{1}$$

which can be simplified to

$$\pi D_0 \mathbf{1} \otimes \xi S^0 \;>\; \pi D_1 \mathbf{1} \otimes \xi S \mathbf{1}.$$

Adding the term $\pi D_1 \mathbf{1} \otimes \xi S^0$ to both sides of the above inequality gives

$$\pi(D_0 + D_1)\mathbf{1} \otimes \xi S^0 \;>\; \pi D_1 \mathbf{1} \otimes \xi(S\mathbf{1} + S^0),$$

which implies

$$\pi D \mathbf{1} \otimes \xi S^0 \;>\; \pi D_1 \mathbf{1} \otimes \xi(S + S^0\beta)\mathbf{1}.$$

Since $\pi D\mathbf{1} = \pi\mathbf{1} = 1$ and $\xi(S + S^0\beta)\mathbf{1} = \xi\mathbf{1} = 1$, the above inequality reduces to $\lambda < \mu_b$. □

2.1.2 The rate matrix R

The Markov chain is a QBD process and can be analyzed by the matrix-geometric method. From Lemma 1.6.3, we can derive the rate matrix R, a square matrix of dimension $nm(r+1)$. The matrix R can be computed using well-known algorithms [102]. From Theorem 2.1.1, we get that, for $\lambda < \mu_b$, the Markov chain Δ_d is positive recurrent and the matrix R has $sp(R) < 1$.

The matrices $A_k,\ k = 0, 1, 2$ are all upper triangular, and so R will also have the same structure, say

$$R = \begin{bmatrix} R_{00} & R_{01} \\ 0 & R_{11} \end{bmatrix}.$$

If the matrices A_i's, $i = 0, 1, 2$, are of higher dimensions, so will be the matrix R. To have computational efficiency, we can break the equation $R = A_0 + RA_1 + R^2A_2$ into three matrix equations as

$$\begin{aligned} R_{00} &= A_{03} + R_{00}A_{13} + R_{00}^2A_{23}, \\ R_{01} &= A_{02} + R_{00}A_{12} + R_{01}A_{10} + R_{00}^2A_{22} + (R_{00}R_{01} + R_{01}R_{11})A_{20}, \\ R_{11} &= A_{00} + R_{11}A_{10} + R_{11}^2A_{20}, \end{aligned} \tag{2.3}$$

from which $R_{00},\ R_{01},\ R_{11}$ can be computed comfortably rather than computing the matrix R. A simple approach for computing $R_{i,j}$ is as follows. Let $R_{i,j}^n$ be the value of $R_{i,j}$ at the n^{th} iteration, where $R_{i,j}^0 = 0$. We insert $R_{i,j}^n$ into the RHS of the equations to obtain $R_{i,j}^{n+1}$ on the LHS. This iteration procedure is continued until $|R_{i,j}^{n+1}(u,v) - R_{i,j}^n(u,v)| < \epsilon,\ \forall(i,j),\ \forall(u,v)$, where $R_{i,j}^n(u,v)$ is the $(u,v)^{th}$ element of the matrix $R_{i,j}^n$ and ϵ is a very small positive number, usually about 10^{-12} [2].

2.1.3 Stationary distribution

The irreducible QBD process is positive recurrent under the condition $\lambda < \mu_b$. So, the stationary distribution of the QBD process exists when $\lambda < \mu_b$. Let x be the stationary

probability vector associated with the transition matrix P such that

$$xP = x,\ x\mathbf{1} = 1.$$

Aggregating terms from the stationary vector x, on the basis of level of the states, gives $x = [x_0, x_1, \ldots]$, where x_0 is a row nr-vector, x_k $(k \geq 1)$ is a row $nm(r+1)$-vector. Further, each vector $x_k = [x_{k0}\ x_{k1}]$, for $k \geq 1$, depending on vacation-state of the server (in vacation or not), where x_{k0} is a nmr-vector and x_{k1} is a nm-vector. Then, we have from the matrix-geometric method that

$$x_{k+1} = x_k R, \quad \text{for } k \geq 1. \tag{2.4}$$

We calculate $[x_0\ x_1]$ from

$$[x_0\ x_1] = [x_0\ x_1]\ B[R], \tag{2.5}$$

where

$$\begin{aligned} B[R] &= \begin{bmatrix} B_{00} & B_{01} \\ B_{10} & A_1 + RA_2 \end{bmatrix} \\ &= \begin{bmatrix} D_0 \otimes (L + L^0\delta) & D_1 \otimes L \otimes \beta & D_1 \otimes L^0\beta \\ D_0 \otimes (L + L^0\delta) \otimes \bar{S^0} & A_{13} + R_0A_{23} & A_{12} + R_0A_{22} + R_{10}A_{20} \\ D_0 \otimes S^0\delta & 0 & A_{10} + R_1A_{20} \end{bmatrix}. \end{aligned} \tag{2.6}$$

The matrix $B[R]$ is stochastic from Lemma 1.6.3 and the matrix $A_1 + RA_2$ is substochastic, since the block B_{10} is not zero. Therefore, Lemma 2.1.1, given below, ensures the existance of the inverse of $I - A_1 - RA_2$, where I is the identity matrix of dimension $nm(r+1)$.

Lemma 2.1.1. *For a Markov chain with transition probability matrix* $P = \begin{pmatrix} P_1 & 0 \\ P_2 & P_3 \end{pmatrix}$, *where* P_3 *is substochastic, the inverse* $(I - P_3)^{-1}$ *exists,* I *being the identity matrix of appropriate dimension [28].*

Now, since $I - A_1 - RA_2$ is nonsingular, it is easy to see that $x_1 = x_0B_{01}(I - A_1 - RA_2)^{-1}$. The vector x_0 is then obtained from the normalization condition

$$x_0\mathbf{1} + x_0[B_{01}(I - A_1 - RA_2)^{-1}(I - R)^{-1}]\mathbf{1} = 1. \tag{2.7}$$

Calculating the inverses of $I - A_1 - RA_2$ and $I - R$ can be carried out in a simpler manner because they are both upper triangular blocks, and this fact can be utilized to compute the inverses block by block [3].

The queue length distribution

The vector x_i is the joint probability of the number of customers in system, arrival phase, phase of vacation, service phase during non-vacation period or service phase during WV period, with i customers in the system.

Let y_i be the marginal probability of finding i customers in the system at an arbitrary time t, then

$$y_i = x_i \mathbf{1},$$

and the mean number of customers in the system is

$$\begin{aligned} E(N) &= \sum_{i=0}^{\infty} i x_i \mathbf{1} \\ &= x_1 \sum_{i=1}^{\infty} i R^{i-1} \mathbf{1} \\ &= x_1 (I - R)^{-2} \mathbf{1}. \end{aligned} \tag{2.8}$$

2.2 Waiting time distribution

The distribution of waiting time can be studied by using absorbing Markov chain. The absorbing states are the states when the target customer reaches the head of the queue and at that time epoch, we assume that the customer's wait in queue is over. With the help of the stationary distribution of the number of customers in the system for this absorbing Markov chain, we can finally get the waiting time distribution of the customers [3].

First, let us find the state of the system at the arrival of an arbitrary customer. Let z_i be the stationary probability vector of finding i customers in the system by an arrival

and $z = [z_0, z_1, \ldots]$ with $z\mathbf{1} = 1$. For $i \geq 1$, we get

$$z_i = [z_{i0}\ z_{i1}] \text{ and } z_0 = z_{00},$$

where z_{ij} represents the probability vector that an arriving customer finds i customers in the system with system in vacation-state j ($j = 0$ for the system on WV period and $j = 1$ for the system on non-vacation period). This can happen as a result of three mutually exclusive cases. When the system is in vacation, it may be empty ($j = 0,\ i = 0$) or it may have more than one customer ($j = 0,\ i \geq 1$). But when the system is busy there will be at least one customer ($j = 1,\ i \geq 1$) in the system because if the system becomes empty, it will always be in vacation. So, we don't have the state $j = 1, i = 0$. According to these cases, we get the following probabilities.

Case 1 : The probability that the arriving customer finds an empty system in WV ($j = 0,\ i = 0$) is,

$$z_0 = \frac{1}{\lambda}\left[x_0\ (D_1 \otimes (L + L^0\delta)) + x_{10}\ (D_1 \otimes (L + L^0\delta) \otimes \bar{S^0}) + x_{11}\ (D_1 \otimes S^0\delta)\right].$$

The first term is the probability of the event that the arriving customer finds the system empty while the vacation has its internal phase change or takes another vacation after completing one. The second term gives the probability that the arriving customer finds the system empty because of service completion of the only customer in WV. The third term is the probability of the event that the system completes service of the only customer in non-vacation and takes a WV while the arriving customer enters the system.

Case 2: The probability that the arriving customer finds the system on WV period with more than one customer ($j = 0,\ i \geq 1$) will be,

$$z_{i0} = \frac{1}{\lambda}\left[x_{i0}(D_1 \otimes L \otimes \bar{S}) + x_{i+1,0}\ (D_1 \otimes L \otimes \bar{S^0}\beta)\right].$$

The first term gives the probability of the event that the system is in WV with i customers and without any vacation termination or service completion while the arriving customer enters the system. The second term gives the event of a service

completion, while there are $i+1$ customers in WV, so that the arriving customer finds the system in WV with i customers.

Case 3: The probability that the arriving customer finds the system in non-vacation with more than one customer ($j = 1,\ i \geq 1$) is,

$$z_{i1} = \frac{1}{\lambda}\left[x_{i1}(D_1 \otimes S) + x_{i+1,1}\,(D_1 \otimes S^0\beta) + x_{i0}\left(D_1 \otimes L^0 \otimes \bar{S}(\mathbf{1}\bar{\xi})\right) + x_{i+1,0}\,(D_1 \otimes L^0 \otimes \bar{S}^0\beta)\right].$$

The first term gives the probability of the event that the system is in non-vacation with i customers while the arriving customer enters and the ongoing service continues. The second term gives the probability of the event of a service completion when the system has $i+1$ customers in non-vacation period, so that the arriving customer finds i customers. The last two terms give the chances of the events of terminating the WVs before the arriving customer enters the system; and if system has i customers, it continues its service with the switched rate or if system has $i+1$ customers a service completion happens.

We have, $z_0\mathbf{1} + \sum_{i=1}^{\infty} z_{i0}\mathbf{1} + \sum_{i=1}^{\infty} z_{i1}\mathbf{1} = 1$, because

$$\begin{aligned}
& z_0\mathbf{1} + \sum_{i=1}^{\infty} z_{i0}\mathbf{1} + \sum_{i=1}^{\infty} z_{i1}\mathbf{1} \\
= \ & \frac{1}{\lambda}\Bigg[x_0\left\{D_1\mathbf{1} \otimes (L + L^0\delta)\mathbf{1}\right\} + x_{10}\left\{D_1\mathbf{1} \otimes (L + L^0\delta)\mathbf{1} \otimes \bar{S}^0\right\} + x_{11}\left(D_1\mathbf{1} \otimes S^0\delta\mathbf{1}\right) \\
& + \sum_{i=1}^{\infty} x_{i0}\left\{D_1\mathbf{1} \otimes \left(L\mathbf{1} + L^0\right) \otimes \bar{S}\mathbf{1}\right\} + \sum_{i=1}^{\infty} x_{i1}\,(D_1\mathbf{1} \otimes S\mathbf{1}) \\
& + \sum_{i=1}^{\infty} x_{i+1,0}\left\{D_1\mathbf{1} \otimes (L\mathbf{1} + L^0) \otimes \bar{S}^0\beta\mathbf{1}\right\} + \sum_{i=1}^{\infty} x_{i+1,1}\left(D_1\mathbf{1} \otimes S^0\beta\mathbf{1}\right)\Bigg] \\
= \ & \frac{1}{\lambda}\Bigg[x_0\,(D_1\mathbf{1} \otimes \mathbf{1}) + x_{10}\left\{D_1\mathbf{1} \otimes (L\mathbf{1} + L^0) \otimes (\bar{S}\mathbf{1} + \bar{S}^0)\right\} + x_{11}\ \left\{D_1\mathbf{1} \otimes (S\mathbf{1} + S^0)\right\} \\
& + \sum_{i=1}^{\infty} x_{i+1,0}\left\{D_1\mathbf{1} \otimes (L\mathbf{1} + L^0) \otimes (\bar{S}\mathbf{1} + \bar{S}^0)\right\} + \sum_{i=1}^{\infty} x_{i+1,1}\left\{D_1\mathbf{1} \otimes (S\mathbf{1} + S^0)\right\}\Bigg] \\
= \ & \frac{1}{\lambda}\left[x_0 D_1\mathbf{1} + \sum_{i=1}^{\infty}(x_{i0} + x_{i1})\,D_1\mathbf{1}\right] = 1.
\end{aligned}$$

In order to determine the waiting time distribution of a customer, we define an absorbing Markov chain seen by an arriving customer, where the state of absorption is the one when the customer reaches the head of the queue. The states of this absorbing Markov chain refer to the number of customers that will receive service ahead of the arriving customer and the current service phase (during vacation or during non-vacation) along with the vacation phase. The absorbing state will be $\{(0,(0,i,j)), 0 \leq i \leq n, 1 \leq j \leq r\}$. The distribution of waiting time of a customer is the time till it visits the absorbing state for the first time. We get, $\widetilde{P}$, the transition probability matrix for the absorbing Markov chain as

$$\widetilde{P} = \begin{pmatrix} \widetilde{B}_{00} & 0 & & & \\ \widetilde{B}_{10} & \widetilde{A}_1 & 0 & & \\ & \widetilde{A}_2 & \widetilde{A}_1 & 0 & \\ & & \widetilde{A}_2 & \widetilde{A}_1 & 0 \\ & & & . & . \quad . \end{pmatrix},$$

where $\widetilde{B}_{00} = (L + L^0\delta)$, $\widetilde{B}_{10} = \begin{bmatrix} (L + L^0\delta) \otimes \bar{S}_0 \\ S^0\delta \end{bmatrix}$, $\widetilde{A}_1 = \begin{bmatrix} L \otimes \bar{S} & L^0 \otimes \bar{S}(e \otimes \bar{\xi}) \\ 0 & S \end{bmatrix}$ and $\widetilde{A}_2 = \begin{bmatrix} L \otimes \bar{S^0}\beta & L^0 \otimes \bar{S^0}\beta \\ 0 & S^0\beta \end{bmatrix}$.

Let us define $\widetilde{z}^0 = [\widetilde{z}^0_0, \widetilde{z}^0_1, \widetilde{z}^0_2, ...]$ with $\widetilde{z}^0_i = [\widetilde{z}^0_{i,0}\ \widetilde{z}^0_{i,1}]$ for $i \geq 1$ to be the stationary probability vector of this absorbing Markov chain $\widetilde{P}$. These can be obtained in terms of probability vector z_i, because the waiting time of a customer gets over as it reaches the head of the queue. Therefore, we obtain

$$\begin{aligned} \widetilde{z}^0_0 &= z_0(\mathbf{1} \otimes I_r), \\ \widetilde{z}^0_{i,0} &= z_{i,0}(\mathbf{1} \otimes I_{rm}), \qquad i \geq 1, \\ \widetilde{z}^0_{i,1} &= z_{i,1}(\mathbf{1} \otimes I_m), \qquad i \geq 1, \end{aligned}$$

where I_r is the identity matrix of dimension r. Also, we get

$$\widetilde{z}^{(n+1)} = \widetilde{z}^{(n)}\widetilde{P}, \quad n \geq 0.$$

If W is the random variable giving the waiting time of a customer in queue and s_j is the probability that a customer's waiting time is less than or equal to j units of time, then we have

$$s_j = \widetilde{z}_0^{(j)}\mathbf{1}, \quad j \geq 0.$$

$\widetilde{z}_0^{(j)}$ can be calculated recursively to an appropriate truncation level form as follows:

$$\begin{aligned}
\widetilde{z}_0^{(n+1)} &= \widetilde{z}_0^{(n)}(L + L^0\delta) + \widetilde{z}_{1,0}^{(n)}\left[(L + L^0\delta) \otimes \bar{S}_0\right] + \widetilde{z}_{1,1}^{(n)}S^0\delta, \\
\widetilde{z}_{i,0}^{(n+1)} &= \widetilde{z}_{i,0}^{(n)}(L \otimes \bar{S}) + \widetilde{z}_{i+1,0}^{(n)}(L \otimes \bar{S}^0\beta), \quad i \geq 1, \\
\widetilde{z}_{i,1}^{(n+1)} &= \widetilde{z}_{i,0}^{(n)}\left[L^0 \otimes \bar{S}(e \otimes \bar{\xi})\right] + \widetilde{z}_{i,1}^{(n)}S + \widetilde{z}_{i+1,0}^{(n)}(L^0 \otimes \bar{S}^0\beta) + \widetilde{z}_{i+1,1}^{(n)}S^0\beta, \quad i \geq 1.
\end{aligned}$$

We can obtain the mean waiting time, using Little's law, as

$$E(W) = E(N)/\lambda.$$

2.3 Regular busy period and busy cycle

In a vacation queue, the regular busy period or the non-vacation period is the duration in which the server works at the normal rate μ_b and let it be denoted by D_v. The total WV period, V_g, is the sum of all consecutive WV durations, between two consecutive regular busy periods. According to the WV policy, the system changes to a regular busy period from a WV period, when the system finds at least one customer waiting at the end of vacation, otherwise the server takes another vacation. Therefore, the busy cycle, the time period between ending instants of two consecutive regular busy periods, is given as the sum

$$B_c = D_v + V_g.$$

Let $G^{(j)}(i)$ be a matrix of dimension $m \times m$ with its $(u, v)^{th}$ entry representing the probability that the service time of j customers lasts for i units of time given that the service

of the first customer starts in phase u and that of the j^{th} customer ends in phase v. Then

$$G^{(1)}(i) = S^{i-1}(S^0\beta), \quad i \geq 1, \tag{2.9}$$

$$G^{(i)}(i) = (S^0\beta)^i, \quad i \geq 1, \tag{2.10}$$

$$G^{(j)}(i) = (S^0\beta)\, G^{(j-1)}(i-1) + S\, G^{(j)}(i-1), \quad i \geq j+1, \quad j \geq 2. \tag{2.11}$$

The equation (2.9) represents the probability of the event that the service time of a customer lasts for i units of time, which means that for first $i-1$ units of time, the service continues with internal phase transitions and the service completes only within the last unit of time. Equation (2.10) represents the probability of the event that the service time of i customers lasts i units of time, which can happen only if, exactly one service completes within one time unit, for each of the i time units. The equation (2.11) is the probability of the event that the service of j customers lasts i units of time. This can occur if any one of the following events happen. Either a service completes within the first time unit and the service of rest of the $j-1$ customers lasts $i-1$ units of time; or the service of the first customer is not completed within the first time unit and the service of all the j customers takes $i-1$ time units to complete.

If $d_{i,j}$ is the probability that j customers arrive in i units of time, $i \geq j$, then we have

$$d_{i,j} = \binom{i}{j} D_0^{i-j} D_1^j$$

because, j arrivals will happen in j time units and the rest $i-j$ units of time will have no arrival, where j time units can be chosen from i time units with $\binom{i}{j}$ ways.

Also, we define the matrix $B^{(j)}(i)$ of dimension $mn \times mn$ such that its $(u,v)^{th}$ entry represents the probability that a busy period initiated by j customers lasts i units of time, under the condition that the first customer's service and arrival are in phase u, $1 \leq u \leq mn$, and that the service of the last customer and arrival are in phase v, $1 \leq v \leq mn$, then

$$B^{(j)}(i) = G^{(j)}(i) \otimes D_0^i + \sum_{l=i}^{i-j} \left(G^{(j)}(i-l) \otimes \sum_{k=1}^{l} d_{i-l,k} \right) B^{(k)}(l). \tag{2.12}$$

The first term of the RHS of equation (2.12) gives the conditional probability of the event that service time of j customers lasts i units of time during which no new customer arrives

and the busy period ends after i time unit. The second term is the probability that the service time of j customers lasts $i-l$ units of time during which k new customers join the queue and these k customers initiate a busy period that lasts l units of time, for given arrival and service phases of the first and the last customer.

Unconditioning on the first and the last customer's arrival and service phases, let $b_i(j)$ be the probability that a busy period initiated by j customers lasts i units of time. Let ϕ be the stationary probability vector of having arrival phase as i and service phase as j for $i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m$. As we are dealing with discrete-time systems the service time of j customers must be at least equal to j, thus

$$\begin{aligned} b_i(j) &= 0, && \text{for } i < j; \\ &= \phi B^{(j)}(i)\mathbf{1}, && \text{otherwise.} \end{aligned}$$

Therefore, we get probability that a busy period initiated by j customers lasts i units of time as [63]

$$b_i(j) = \phi\left(G^{(j)}(i) \otimes D_0^i\right)\mathbf{1} + \sum_{l=i}^{i-j}\sum_{k=1}^{l}\binom{i-l}{k}\phi\left(G^{(j)}(i-l) \otimes D_0^{i-l-k}D_1^k\right)\mathbf{1}\, b_l(k),$$

where the first term on the RHS is the unconditional probability that the service time of j customers lasts i units of time and during that period no new customer joins the queue. The second term on the RHS is the probability that the service time of j customers lasts $i-l$ units of time. During the first $i-l$ units of time, k new customers join the queue. These k customers initiate a busy period that lasts l units of time.

Let v_i be the probability that total vacation duration lasts i units of time which is same as the probability that there is no customer in the system for $i-k$ units of time and the system does not return from vacation when customer arrives during last k units of time. Then

$$v_i = \sum_{k=0}^{i}(x_0\mathbf{1})^{i-k}\left[x_1(I-R)^{-1}(e_1 \otimes \mathbf{1})\right]^k,$$

where e_1 is the column vector of zeros and one in the first position.

The mean duration of regular busy period and mean vacation duration are given by

$$E(D_v) = \sum_{i=j}^{\infty} ib_i(j),$$
$$E(V_g) = \sum_{i=0}^{\infty} iv_i$$

and hence, we can have the mean busy cycle as

$$E(B_c) = E(D_v) + E(V_g).$$

2.4 Numerical examples

The model DMAP/DPH/1/WV(DPH) captures both burstiness and correlation of the arrival process which are highly essential features to be considered to measure system performance. To show the impact of burstiness and correlation in queue lengths and waiting times, we compare our model with an equivalent model that does not capture the correlation property. For simplicity, we have taken a special case of DMAP, the 2-state Markov-modulated Bernoulli process (MMBP) and presented some numerical examples. The first part gives the comparison of the performance measures, like mean queue length and mean waiting time of MMBP/Geo/1/WV(Geo) model with that of a DPH/Geo/1/WV(Geo) model. By this comparison, we can see the impact of autocorrelation on the performance measures of a WV model. And, in the next part, we will observe only the MMBP model and will see how the performances change with changes in autocorrelation.

The squared coefficient of variation, c^2, of the interarrival times of an MMBP process is an important measure of the degree of traffic burstiness which is defined as the ratio of the variance to the square of mean of interarrival times and is given by expression (1.32). The autocorrelation coefficient, which is the ratio of covariance of interarrival times to the variance, is used to measure the amount of correlation in arrivals and is given as (1.33). The DMAP representation of a MMBP process is $D_0 = \widehat{P}(I - \Lambda)$, $D_1 = \widehat{P}\Lambda$ and this

gives

$$D_0 = \begin{bmatrix} p(1-\lambda_1) & (1-p)(1-\lambda_2) \\ (1-q)(1-\lambda_1) & q(1-\lambda_2) \end{bmatrix} \text{ and } D_1 = \begin{bmatrix} p\lambda_1 & (1-p)\lambda_2 \\ (1-q)\lambda_1 & q\lambda_2 \end{bmatrix}.$$

Here, we have four variables, $p, q, \lambda_1, \lambda_2$ and two equations, one of total arrival rate λ and the other of squared coefficient of variation, c^2. For given values of λ and c^2, we can obtain the corresponding MMBP parameters $p(\lambda_1)$ and $q(\lambda_1)$ as a function of λ_1 from the formulas by fixing the other arrival rate λ_2. Therefore, by fixing two variables, λ_1 and λ_2 we can get a MMBP process for the given λ and c^2.

For PH-renewal process with interarrivals following DPH(α, T), where $T = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $T^0 = \begin{bmatrix} g \\ h \end{bmatrix}$ and $\alpha = [\alpha_1, \alpha_2]$, with $0 \leq a, b, c, d, g, h \leq 1$, the DMAP representation is

$$\overline{D_0} = T \text{ and } \overline{D_1} = T^0\alpha = \begin{bmatrix} \alpha_1 g & \alpha_2 g \\ \alpha_1 h & \alpha_2 h \end{bmatrix}.$$

Here, we have eight variables and five equations from the conditions $Te + T^0 = e$, $\alpha e = 1$ and the equations of mean and c^2. So, we need to fix three variables.

If we want to have a PH-renewal process which is equivalent to the MMBP process, we can relate the three variables of PH-renewal process to that of MMBP. Different ways of choosing these variables gives different PH-renewal processes. Since we are free to choose three variables, let us pick them in such a way that the transition matrix $\overline{D_0}(= T)$ of PH-renewal process becomes equivalent to D_0 of the MMBP process *i.e.* any three variables of T is taken to be equal to their respective variables in D_0, *e.g.,*

$$a = p(1-\lambda_1),\ c = (1-q)(1-\lambda_1) \text{ and } b = (1-p)(1-\lambda_2). \tag{2.13}$$

Another way of choosing the variables of PH-renewal process is to make the matrix $\overline{D_1}(= T^0\alpha)$ equivalent to the D_1 of the MMBP process for given λ and c^2.

2.4.1 Comparison of MMBP and DPH arrival models

In the examples below, we show how the performances of a MAP arrival model differs from a corresponding model with DPH arrival, by keeping the service time distributions and the WV durations as geometric for all the cases, to see the influence only of the arrival process.

For given arrival rate λ and c^2, we can independently choose two variables for MMBP arrival process and three variables for DPH arrival process. For $\lambda = 0.3$ and $c^2 = 2$, fixing the rates $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$, we get the MMBP process as

$$\widehat{P} = \begin{bmatrix} 0.9815 & 0.0185 \\ 0.0185 & 0.9815 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.1 \end{bmatrix}$$

for which

$$D_0 = \begin{bmatrix} 0.4907 & 0.0167 \\ 0.0093 & 0.8833 \end{bmatrix} \text{ and } D_1 = \begin{bmatrix} 0.4907 & 0.0019 \\ 0.0093 & 0.0981 \end{bmatrix}.$$

The autocorrelation coefficient, ψ, of this MMBP process is 0.1321. In Figure 2.1, we can see the change of autocorrelation of this MMBP process with the change in λ_1.

For the PH-renewal process, taking the parameters $\lambda = 0.3$, $c^2 = 2$, which are same with the MMBP process, and choosing the three variables using relation (2.13), we get

$$T = \begin{bmatrix} 0.4907 & 0.0167 \\ 0.0093 & 0.8833 \end{bmatrix}, \; T^0 = \begin{bmatrix} 0.4926 \\ 0.1074 \end{bmatrix} \text{ and } \alpha = [0.8333, \; 0.1667].$$

The service process and vacation durations for both the models are taken to be geometric. The parameter values are taken as $\mu_b = 0.7$ and $\theta = 0.1$. Computing the matrix R using the Naoumov's algorithm [156] and from equation (2.8), we have calculated the mean queue lengths for both the models with varying parameter values.

In Figure 2.2, we compare both the models for different arrival rates λ. When $\lambda = 0.3$, the mean queue lengths of MMBP model and of corresponding PH model are very close, with a difference only about 1%. For fast vacation-service rates, the mean queue length reduces individually for each model, but the difference in queue lengths between the models remains unaffected even for increased service rates. For $\lambda = 0.2$, the mean queue

length of MMPP model is longer than that of PH model by 35% whereas high vacation-service rates can reduce this difference to 15%. This shows that a MMBP model and a PH model do not have same mean queue lengths however they may be close for some cases.

Figure 2.3 shows the difference between the mean queue lengths of MMBP and PH models for vacation duration rate θ varying from 0.1 to 0.8. The graph shows that for high values of θ, that is models having short durations of vacations are not effected by the vacation-service rates, which is true for both MMBP and PH models. On the other hand, for small values of θ, the MMBP queue lengths reduce faster than that of a PH model as vacation-service rate is increased. Here, the MMBP queues can be longer than the PH queues upto 40%.

In Figure 2.4, the change in queue lengths with variation in θ is shown for different values of λ. This also shows that the queue lengths of a correlated arrival WV model cannot be assumed to be same with that having independent arrivals. The change in the value of c^2 also ensures the difference in queue lengths for MMBP and the corresponding PH model. This is observed in Figure 2.5, where the queue lengths differ progressively as the value of c^2 increases. Finally, if we change the model by changing the value of λ_1, we can see in Figure 2.6 how significantly the queue lengths differ with the change in c^2.

From these observations, we can conclude that there is a significant difference in mean queue lengths when we consider the correlation in the arrival process (MMBP model) from the case where the arrivals are assumed to be independent (PH model). Therefore, assuming the interarrivals to be independent, when it is not, might lead to inaccurate prediction of mean queue lengths. Similar behaviors are also observed for the mean waiting times.

We have seen that for c^2=1 the mean queue lengths of MMBP model and PH model come very close, in Figure 2.5. Here the amount of correlation in the MMBP process is 0.0281. For such a situation, we present in Table 2.1, the distributions of mean queue lengths for both the models. The parameter values taken are $\lambda_1 = 0.7$, $\lambda_2 = 0.1$, $\lambda = 0.4$,

$\theta = 0.7$, $\mu_v = 0.4$ and $c^2 = 1$. It can be seen that, even if the mean queue lengths are close enough, the distributions are not same. For $\mu_b = 0.48$ and $k = 0$, the values of y_0 are the same for both MMBP model (y_k^M) and PH model (y_k^P). Then, from $k = 1$ till $k = 5$, the values of y_k^M remain smaller than y_k^P, and beyond $k = 5$, y_k^M becomes higher than y_k^P. We have tabled the changes for four different values of μ_b. So, same mean does not give the same distribution for those two models.

We can also have situation, like in Figure 2.7, where the queue lengths of MMBP model become less than that of the PH model. The mean waiting times also show similar behaviors in this case.

So, whatever way we choose the parameters of the models, because of the correlation in MMBP process, there is considerable difference in mean queue lengths and also in mean waiting times, which cannot be neglected if we wish to have a good estimation of the performance measures. Thus the correlation in the arrival process is quite an important feature to consider which indicates that this DMAP/DPH/1/WV model can give a better estimation of the performance measures whenever interarrivals are correlated.

2.4.2 Autocorrelation and DMAP/Geo/1/WV(Geo) model

Autocorrelation coefficient for the interarrival times gives the measure of the degree of correlation. This correlation parameter is a function of the arrival rate and the transition probabilities as given in equation (1.33) for the MMBP arrival process.

Figures 2.8 and 2.9 shows the impact of correlation on the mean queue length of MMBP/Geo/1/WV(Geo) model under different traffic burstiness (c^2). The two figures are for two different traffic loads, $\rho = 0.65$ and 0.9 respectively. Other parameter values are taken as $\lambda_2 = 0.9$, $\mu_b = 0.5$, $\mu_v = 0.4$ and $\theta = 0.7$. The graphs show that for $c^2 = 1$, the degree of correlation in traffic does not effect the mean queue lengths even for heavy loaded systems. As the traffic becomes more bursty, that is when c^2 is high, the effect of correlation becomes eminent. This affect can raise the mean queue lengths upto 90% especially when traffic load is too high and traffics are highly correlated.

In Table 2.2, we have presented the queue length distribution for different values of ρ of a DMAP/Geo/1/WV(Geo) model. An example of a DMAP arrival process (taken from [69]) is given by

$$D_0 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.95 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.95 & 0.05 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0.95 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$n = 7, \text{ and } \quad \pi = [0.0018 \;\; 0.0023 \;\; 0.0023 \;\; 0.9011 \;\; 0.0451 \;\; 0.0451 \;\; 0.0023].$$

Here the parameter values are $\lambda = 0.0925$, $\mu_v = 0.03$ and $\theta = 0.08$. The service rates taken are $\mu_b = 0.1, 0.11, 0.12, 0.13, 0.14, 0.15$ and 0.16. For a particular traffic load, ρ, the probability of number of customers in the system increases from $k = 0$ to $k = 1$ and after that it keeps on decreasing as k increases. This behavior is similar for all traffic loads. Comparing the probabilities between different traffic loads, we see that for lighter traffic system ($\rho = 0.5781$), the mass at $k = 0$ is highest compared to the highly loaded traffics. But for $k = 4$, the probability is highest for $\rho = 0.6607$, not for $\rho = 0.5781$. And, when $k = 14$, we see that the probability is highest for $\rho = 0.9250$. This distribution of number of customers says that a heavily loaded system need not always have the highest probability of congestion.

Every discrete-time model has its continuous-time analogue. In the next section, we study the continuous-time analogue of this discrete-time MAP/PH/1/WV model. As the analytical treatment are similar for both the models, we only give the basic differences between the two models and some of the important results.

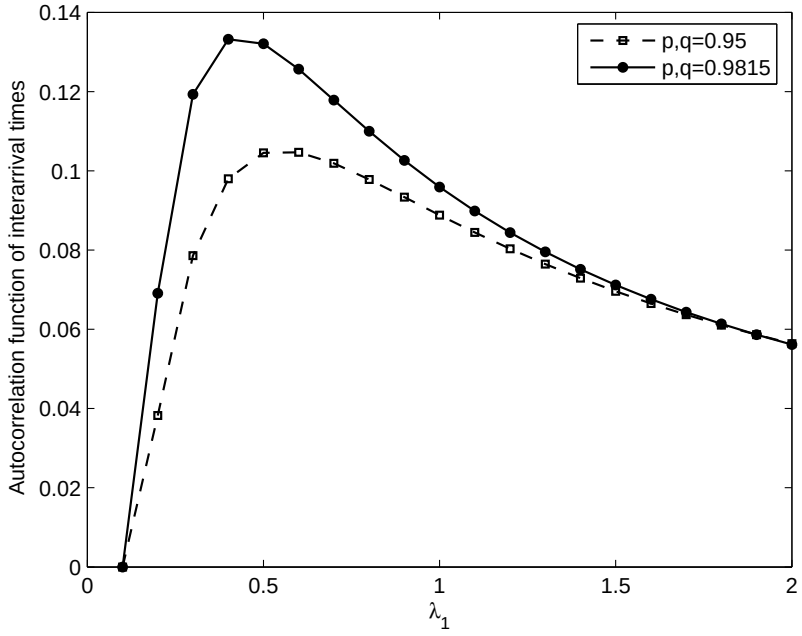

Figure 2.1: Autocorrelation vs λ_1, with $\lambda_2 = 0.1$.

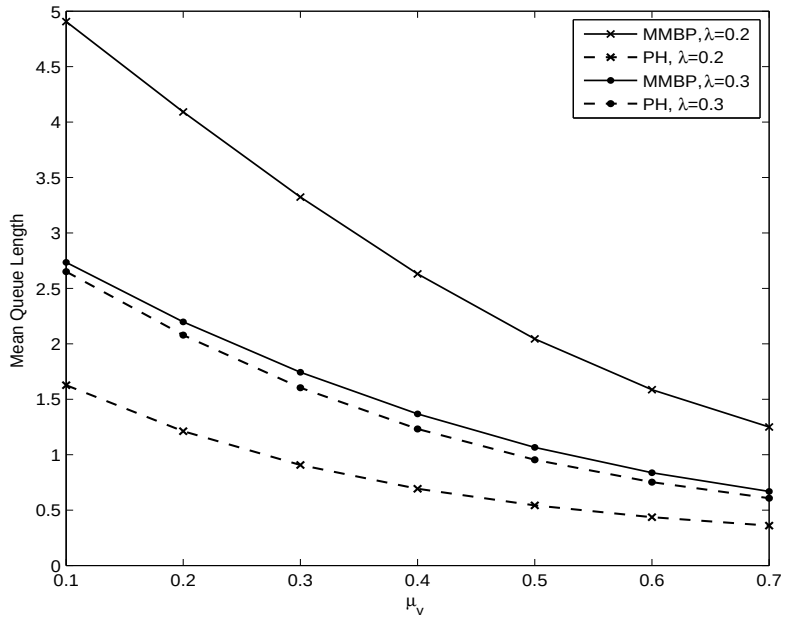

Figure 2.2: Mean queue length vs μ_v, with $\mu_b = 0.7$, $c^2 = 2$, $\theta = 0.1$.

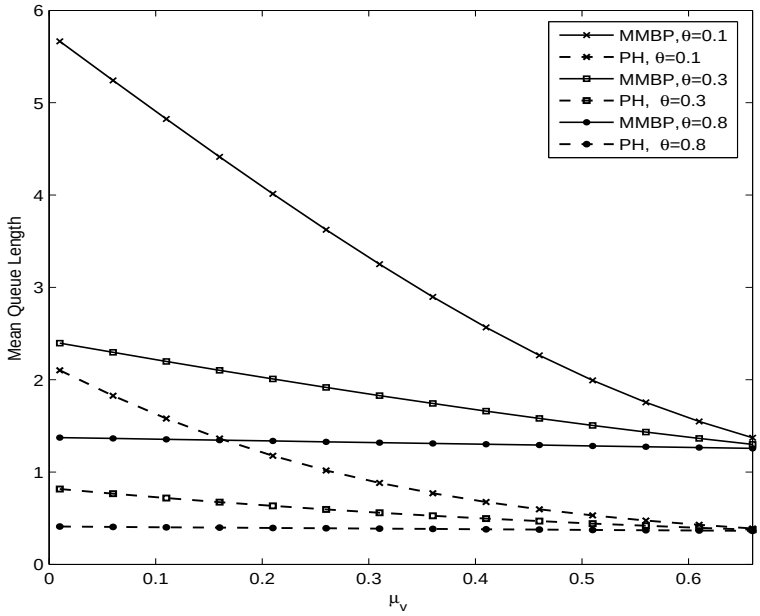

Figure 2.3: Mean queue length vs μ_v, with $\mu_b = 0.7$, $c^2 = 2$, $\lambda = 0.2$.

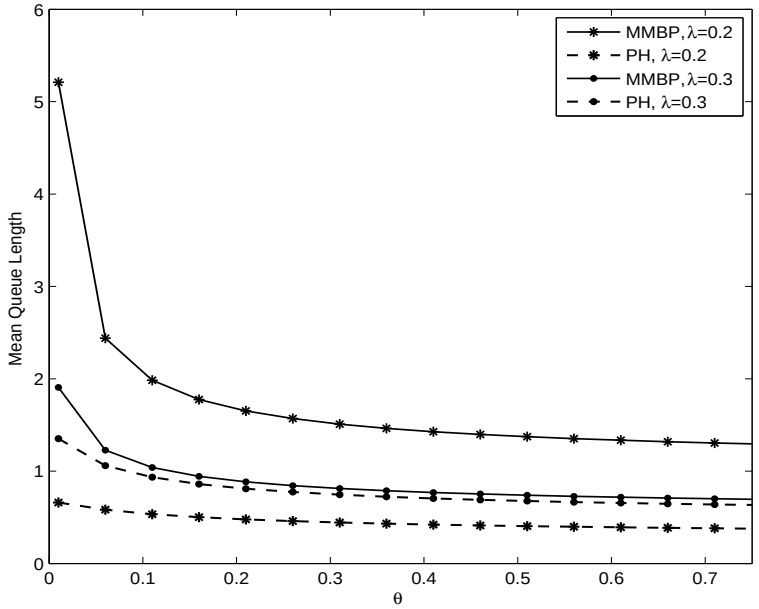

Figure 2.4: Mean queue length vs θ, with $\mu_b = 0.7$, $\mu_v = 0.5$, $c^2 = 2$.

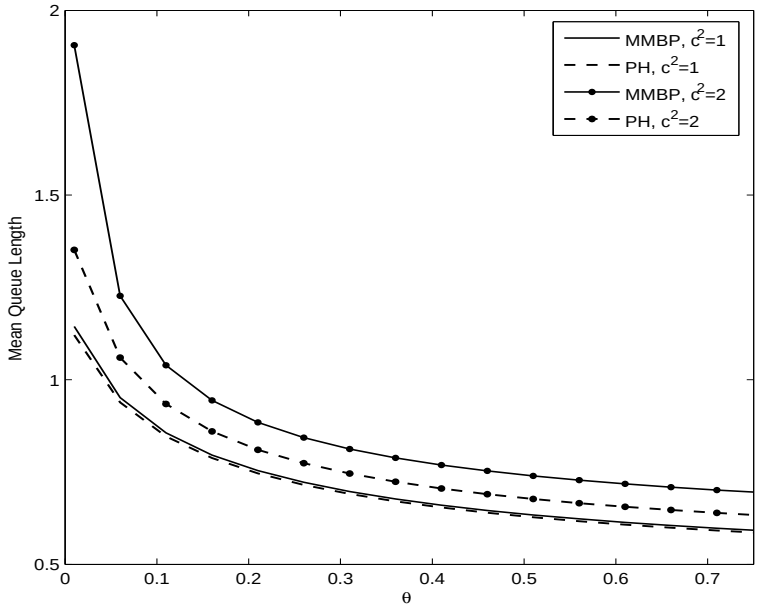

Figure 2.5: Mean queue length vs θ, with $\mu_b = 0.7,\ \mu_v = 0.5,\ \lambda = 0.3$.

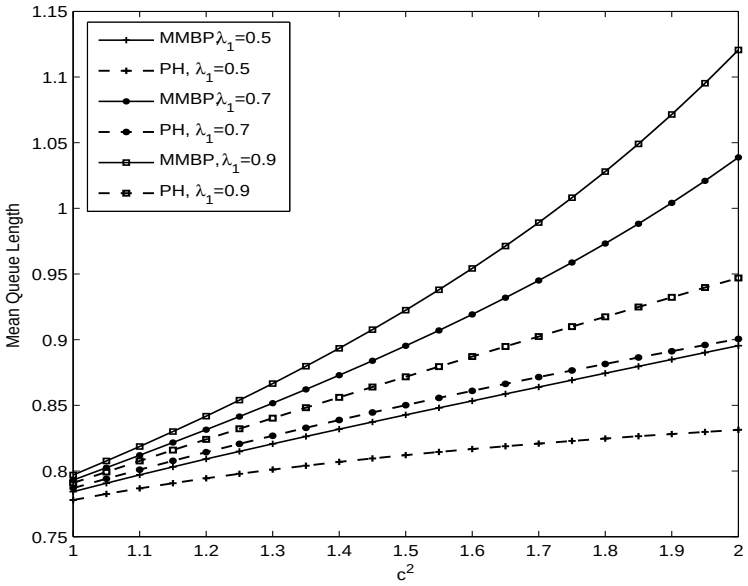

Figure 2.6: Mean queue length vs c^2, with $\mu_b = 0.7,\ \mu_v = 0.3,\ \theta = 0.4,\ \lambda = 0.3$.

Table 2.1: Distribution of number of customers in MMBP and DPH models.

	μ_b=0.48		μ_b=0.5		μ_b=0.52		μ_b=0.54	
k	y_k^M	y_k^P	y_k^M	y_k^P	y_k^M	y_k^P	y_k^M	y_k^P
0	0.1634	0.1634	0.1953	0.1953	0.2246	0.2245	0.2514	0.2514
1	0.1761	0.1783	0.2047	0.2070	0.2293	0.2315	0.2505	0.2525
2	0.1375	0.1420	0.1514	0.1561	0.1607	0.1653	0.1664	0.1708
3	0.1088	0.1116	0.1133	0.1156	0.1135	0.1154	0.1110	0.1124
4	0.0862	0.0875	0.0847	0.0853	0.0801	0.0803	0.0739	0.0736
5	0.0683	0.0686	0.0633	0.0630	0.0565	0.0558	0.0491	0.0482
6	0.0541	0.0537	0.0473	0.0465	0.0399	0.0388	0.0327	0.0315
7	0.0428	0.0421	0.0354	0.0343	0.0281	0.0270	0.0217	0.0206
8	0.0339	0.0330	0.0264	0.0254	0.0198	0.0187	0.0145	0.0135
9	0.0268	0.0259	0.0198	0.0187	0.0140	0.0130	0.0096	0.0088
10	0.0213	0.0203	0.0148	0.0138	0.0099	0.0091	0.0064	0.0058

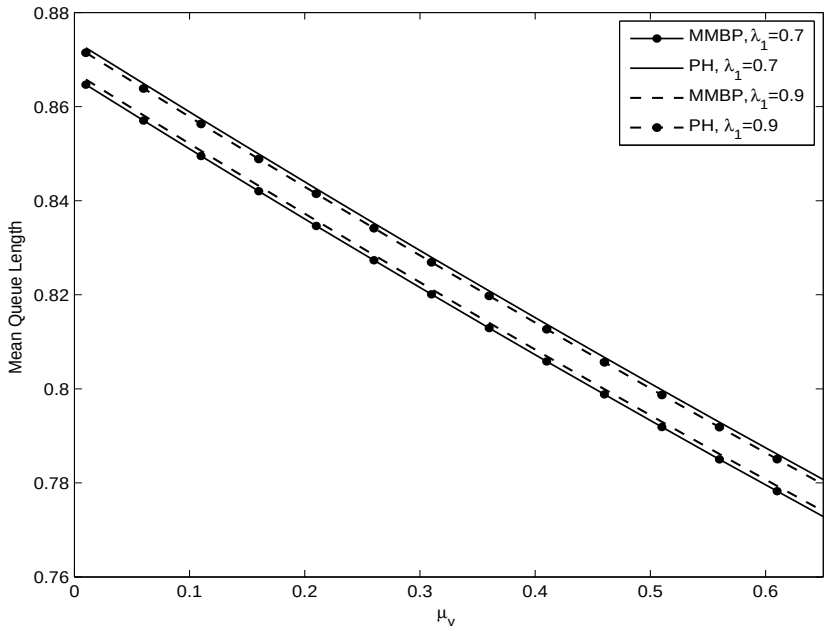

Figure 2.7: Mean queue length vs μ_v, with $\mu_b = 0.7, \theta = 0.8, \lambda = 0.4$, $c^2 = 0.5$.

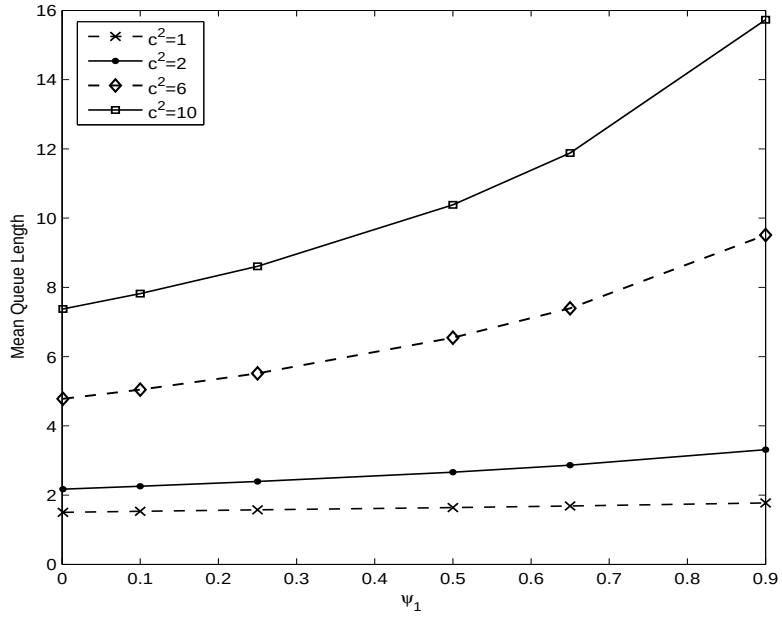

Figure 2.8: Mean queue length vs ψ_1, with $\rho = 0.65$.

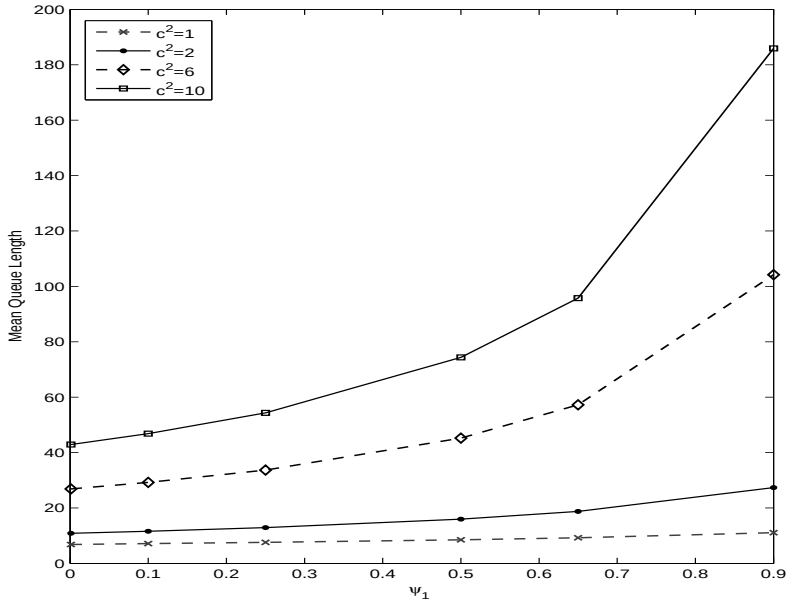

Figure 2.9: Mean queue length vs ψ_1, with $\rho = 0.9$.

Table 2.2: Distribution of number of customers in MAP/Geo/1/WV(Geo) model.

k	ρ=0.9250 y_k	ρ=0.8409 y_k	ρ=0.7708 y_k	ρ=0.7115 y_k	ρ=0.6607 y_k	ρ=0.6167 y_k	ρ=0.5781 y_k
0	0.0549	0.1142	0.1627	0.2030	0.2371	0.2662	0.2915
1	0.0347	0.0669	0.0891	0.1046	0.1155	0.1233	0.1289
2	0.0545	0.1031	0.1354	0.1573	0.1725	0.1832	0.1907
3	0.0441	0.0767	0.0932	0.1008	0.1036	0.1036	0.1021
4	0.0482	0.0804	0.0945	0.0996	0.1002	0.0986	0.0960
5	0.0433	0.0666	0.0726	0.0714	0.0674	0.0625	0.0576
6	0.0426	0.0619	0.0642	0.0606	0.0553	0.0499	0.0450
7	0.0396	0.0532	0.0514	0.0454	0.0390	0.0332	0.0284
8	0.0377	0.0474	0.0432	0.0363	0.0299	0.0245	0.0204
9	0.0354	0.0412	0.0351	0.0277	0.0214	0.0167	0.0131
10	0.0334	0.0363	0.0290	0.0216	0.0159	0.0119	0.0090
11	0.0314	0.0317	0.0236	0.0165	0.0115	0.0081	0.0059
12	0.0296	0.0277	0.0194	0.0128	0.0084	0.0057	0.0039
13	0.0278	0.0242	0.0158	0.0098	0.0061	0.0039	0.0026
14	0.0262	0.0212	0.0129	0.0075	0.0044	0.0027	0.0017
15	0.0246	0.0185	0.0106	0.0058	0.0032	0.0018	0.0011

2.5 The continuous-time MAP/PH/1/WV model

We consider here an analogous WV model in continuous-time, where the state of the system can be given by a Markov chain

$$\Delta = \left\{ \left(N_t, (Q_t, J_t, (V_t, K_t^v)) \cup (Q_t, J_t, K_t^b)\right), t \geq 0 \right\},$$

where N_t is the number of customers in the system at time t, Q_t representing the vacation-state of the server *i.e.*, $Q_t = 0$ if the system in vacation and $Q_t = 1$ if the system in non-vacation period, J_t is the phase of arrival, V_t is the phase of vacation, K_t^v gives the phase of service during WV and K_t^b gives the phase of the service in non-vacation period. Then Δ has a state-space given by

$$E_c = \Big\{ \{\{0\} \times \{(0,i,j)\}\} \cup \{\mathbb{N} \times \{\{(0,i,(j,k))\} \cup \{(1,i,k)\}\}\}, \\ 1 \leq i \leq n, 1 \leq j \leq r, 1 \leq k \leq m \Big\}.$$

Using the lexicographical ordering of the states as, $(0,(0,1,1)), \ldots, (0,(0,1,r)), (0,(0,2,1)), \ldots, (0,(0,n,r)), (1,(0,1,(1,1))), \ldots, (1,(0,n,(r,m))), (1,(1,1,1)), \ldots, (1,(1,n,m)), (2,(0,1,(1,1))), \ldots$, we get the infinitesimal generator matrix Q of the Markov chain Δ as a QBD process, which can be partitioned similar to the transition matrix P in (2.1) and is given by

$$Q = \begin{pmatrix} B_{00} & B_{01} & & & & \\ B_{10} & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

The off-diagonal entries of Q are non-negative and the diagonal entries are strictly negative. The entries of Q are

$B_{00} = D_0 \oplus (L + L^0\delta)$, $B_{01} = [D_1 \otimes I_r \otimes \beta \quad 0]$, $B_{10} = \begin{bmatrix} I_n \otimes I_r \otimes \bar{S^0} \\ I_n \otimes S^0\delta \end{bmatrix}$, $A_i = \begin{bmatrix} A_{i3} & A_{i2} \\ 0 & A_{i1} \end{bmatrix}$, $i = 0,1,2$, $A_{13} = D_0 \oplus L \oplus \bar{S}$, $A_{12} = I_n \otimes L^0 \otimes \mathbf{1}\bar{\xi}$, $A_{11} = D_0 \oplus S$, $A_{03} = D_1 \otimes I_r \otimes I_m$,

$A_{02} = 0$, $A_{01} = D_1 \otimes I_m$, $A_{23} = I_n \otimes I_r \otimes \bar{S^0}\beta$, $A_{22} = 0$, $A_{21} = I_n \otimes S^0\beta$.

Here I_n is the identity matrix of dimension n and the sign '$\oplus$' is the Kronecker sum. The formations of these matrices are given below.

- $B_{00} = D_0 \oplus (L + L^0\delta) = D_0 \otimes I_m + I_n \otimes L + I_n \otimes L^0\delta$, because the system remains in level 0 if no arrival and no vacation terminates. either by internal phase transition without any arrival or, by internal phase transition in vacation without vacation termination or, by starting another vacation after terminating one.

- $B_{01} = [D_1 \otimes I_r \otimes \beta \quad 0]$, because the system can changes its level from 0 to 1 only if there is an arrival; and immediately the customer moves for service with initial probability β.

- $B_{10} = \begin{bmatrix} I_n \otimes I_r \otimes \bar{S^0} \\ I_n \otimes S^0\delta \end{bmatrix}$, because the system changes its level from 1 to 0 by completing service of the only customer in the system. If the system was in vacation, it remains in vacation after the service completion and if the system was in non-vacation, after completing the service, it immediately takes another vacation with probability δ.

The matrices A_i, $i = 0, 1, 2$, can be partitioned into blocks depending on whether the system stays in vacation (A_{i3}), moves from vacation to non-vacation period (A_{i2}) or remains in non-vacation period (A_{i1}). So for $i = 0, 1, 2$, we get

$$A_i = \begin{bmatrix} A_{i3} & A_{i2} \\ 0 & A_{i1} \end{bmatrix}.$$

Here, A_1 gives the probability of the system staying in the same level.

- $A_{13} = D_0 \oplus L \oplus \bar{S}$ because, either the system has only internal phase transition without arrival or, change in internal vacation phases or, internal phase transition of service.

- $A_{12} = I_n \otimes L^0 \otimes I_m(\mathbf{1}\bar{\xi})$ because, after completing the vacation, the system moves to non-vacation period and the ongoing service of the customer, if any, is switched to the service rate of the non-vacation period with probability $\bar{\xi}$.

- $A_{11} = D_0 \oplus S$ because, the system remains in non-vacation without any service completion and without any arrival, but only internal phase transitions.

The other matrices can also be written with similar arguments.

We assume that Q is irreducible. The matrix A becomes reducible and has negative diagonal and non-negative off-diagonal elements. For $\lambda < \mu_b$, this process Q is positive recurrent and its stationary probability vector $x = [x_0\ x_1\ x_2\ ...]$, satisfies the equations

$$x_{k+1} = x_k R, \quad \text{for } k \geq 1, \tag{2.14}$$

$$[x_0\ x_1]\ B[R] = 0 \tag{2.15}$$

$$\text{and} \qquad x_0 e + x_1 (I - R)^{-1} e = 1, \tag{2.16}$$

where

$$B[R] = \begin{bmatrix} B_{00} & B_{01} \\ B_{10} & A_1 + RA_2 \end{bmatrix}$$

and R is the minimum non-negative solution to the matrix equation

$$R^2 A_2 + RA_1 + A_0 = 0, \tag{2.17}$$

with $sp(R) < 1$. The matrix R can also be computed by using algorithms listed in [102].

The stationary vector x_i is the joint probability of the number of customers, arrival phase, phase of vacation duration, service phase during non-vacation period or service phase during WV period, with i customers in the system. Let y_i be the marginal probability of finding i customers in the system at time t, then

$$y_i = x_i \mathbf{1}$$

and the mean number of customers in the system is

$$E(N) = \sum_{i=0}^{\infty} i y_i = x_1 (I - R)^{-2} \mathbf{1}. \tag{2.18}$$

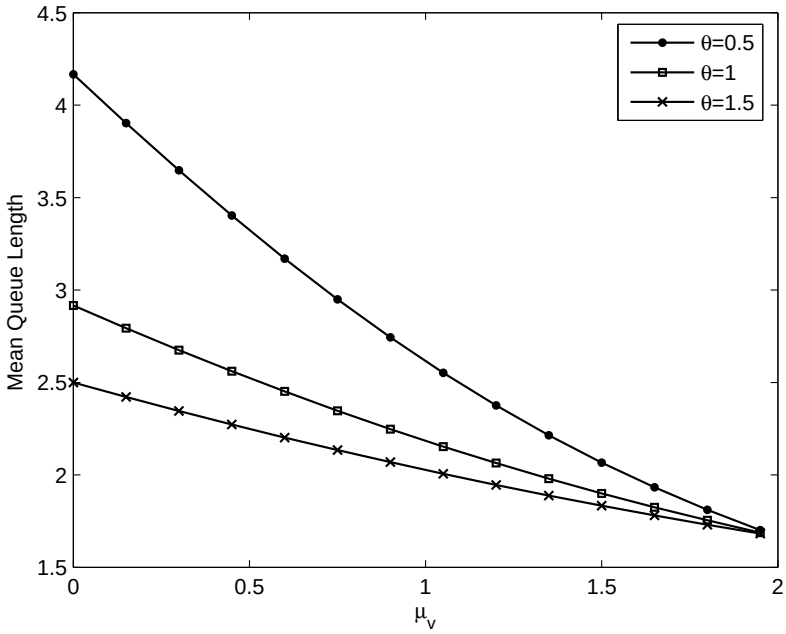

Figure 2.10: Mean queue length vs μ_v of M/M/1/WV(M) model, with $\mu_b = 2,\ \lambda = 1.25$.

The mean waiting time can be obtained by little's law as

$$E(W) = E(N)/\lambda.$$

Special case: When the arrival process, the service distributions as well as the distribution of vacation durations are exponential, we get the generator matrix P which resembles exactly with the one given by Liu et al. [113]. In Figure 2.10 we have plotted the mean queue lengths of a M/M/1/WV model for $\lambda = 1.25$ and $\mu_b = 2$.

Chapter 3

A Finite-buffer Queue with Correlated Arrivals

Analysis of networks of finite capacity queues plays an important role for the effective congestion control of network traffic and quality of service (QoS) protection in communication networks. We discussed about bursty and correlated traffic arising in WDM networks, in the previous chapter, having infinite capacity or infinite buffer space. A buffer is a space in network used to hold network packets temporarily, while they are moved from one place to another. In practice, buffer is limited and heavy traffic affects the buffer occupancy, leading to network congestion and packet loss. Packet loss probabilities also depend upon the type of network. The study of loss probabilities is an important task as it enables better network design in terms of buffer sizing and management, congestion control mechanisms etc. and hence can influence the performance of the network. Thus, it is extremely necessary to study the blocking performance of a finite-buffer network, taking into consideration the burstiness and correlation of the packets. A finite-buffer model GI/M/1/K/WV was presented by Banik et al. [26] considering an independent arrival process. In this chapter, we study a WV model with correlated arrivals having finite-buffer space to lay the emphasis of loss probabilities on system efficiency.

3.1 Model description

We consider here a single-server finite-buffer continuous-time queueing model, where the customers arrive according to a MAP(D_0, D_1) process with rate λ. A customer who could not get service upon arrival will wait in the buffer. The buffer size is K $(K < \infty)$ *i.e.,* the number of customers allowed in the system at any time t, including the one in service, is K, where K is finite. If the buffer is full and a customer arrives, the arriving customer will leave the system without joining the queue. The service times and WV durations are assumed to follow PH-distributions. During a non-vacation period, the server serves the customers according to a PH(β, S) of dimension m and rate $\mu_b = \xi S^0$ with ξ as the stationary probability vector of $S + S^0\beta$. As per the WV policy, the service times during WV follow PH$(\beta, \bar{S})$ and rate μ_v such that $\mu_v < \mu_b$. The rate $\mu_v = \bar{\xi}\bar{S}^0$, where $\bar{\xi}(\bar{S} + \bar{S}^0\beta) = 0$ and $\bar{\xi}\mathbf{1} = 1$.

As soon as the system becomes empty, the server takes a WV whose durations follow a PH(δ, L) distribution, with rate θ. At a vacation termination epoch, if the server finds at least one customer in the system, it switches its service rate from μ_v to μ_b. If the vacation terminates in middle of an ongoing service, the server continues to serve, but with rate μ_b, until completion. Otherwise, if the server finds an empty system, it takes another WV. The interarrival times, service times and vacation times, are all assumed to be mutually independent.

To have complete information of the system, we need to keep track of phase of arrival process, phase of service times and phase of vacation durations. The states of the system then can be described by a continuous-time Markov chain

$$\Delta = \left\{ \left(N_t, (Q_t, J_t, (V_t, U_t^v)) \cup (Q_t, J_t, U_t^b) \right), t \geq 0 \right\},$$

where N_t is the number of customers in the system at time t, Q_t representing the vacation-state of the server *i.e.*, $Q_t = 0$ if the server is in WV period and $Q_t = 1$ if the server is in non-vacation period, J_t is the phase of arrival, U_t^v gives the phase of service during WV and U_t^b gives the phase of service in non-vacation period. Then Δ has a state-space given

by

$$E = \Big\{\{(0,(0,i,j))\} \cup \{\{(u,(0,i,(j,k)))\} \cup \{(u,(1,i,k))\}\},$$
$$1 \le u \le K, 1 \le i \le n, 1 \le j \le r, 1 \le k \le m\Big\}.$$

Using the lexicographical ordering of the states as, $(0,(0,1,1)), \ldots, (0,(0,1,r)), (0,(0,2,1)), \ldots, (0,(0,n,r)), (1,(0,1,(1,1))), \ldots, (1,(0,n,(r,m))), (1,(1,1,1)), \ldots, (1,(1,n,m)), \ldots, (K,(0,1,(1,1))), \ldots, (K,(0,n,(r,m))), (K,(1,1,1)), \ldots, (K,(1,n,m))$, we get the infinitesimal generator matrix Q of the Markov chain Δ given by

$$Q = \begin{pmatrix} B_{00} & B_{01} & & & & \\ B_{10} & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & A_2 & A_1 & A_0 \\ & & & & A_2 & B_{KK} \end{pmatrix}, \tag{3.1}$$

where the matrices are given by

$$B_{00} = D_0 \oplus (L + L^0\delta),\ B_{01} = [D_1 \otimes I_r \otimes \beta \quad 0],\ B_{10} = \begin{bmatrix} I_n \otimes I_r \otimes \bar{S^0} \\ I_n \otimes S^0\delta \end{bmatrix},$$

$$B_{KK} = \begin{bmatrix} L \oplus \bar{S} & I_n \otimes L^0 \otimes I_m \\ 0 & S \end{bmatrix},\ A_i = \begin{bmatrix} A_{i3} & A_{i2} \\ 0 & A_{i1} \end{bmatrix},\ i = 0,1,2,$$

$$A_{13} = D_0 \oplus L \oplus \bar{S},\ A_{12} = I_n \otimes L^0 \otimes \mathbf{1}\bar{\xi},\ A_{11} = D_0 \oplus S,\ A_{03} = D_1 \otimes I_r \otimes I_m,\ A_{02} = 0,$$

$$A_{01} = D_1 \otimes I_m,\ A_{23} = I_n \otimes I_r \otimes \bar{S^0}\beta,\ A_{22} = 0,\ A_{21} = I_n \otimes S^0\beta.$$

Here, I_n is the identity matrix of dimension n. The matrices are similar to the infinite-buffer continuous-time case, as described in Section 2.5 of Chapter 2, except for the matrix B_{KK}. The matrix

$B_{KK} = \begin{bmatrix} L \oplus \bar{S} & I_n \otimes L^0 \otimes I_m \\ 0 & S \end{bmatrix}$ represents the probability that if the system is full, no customers are allowed in the system and it remains in vacation with either internal vacation phase transition or internal service phase transition. If system moves from WV to non-vacation there must be a vacation termination. The system remains in non-vacation with an internal phase change without any service completion.

This Markov chain is a QBD process which we assume to be irreducible. An irreducible Markov chain with finite state-space is always positive recurrent [72], which gives the existence of a unique stationary distribution.

3.1.1 Stationary distribution at arbitrary epochs

Let x be the stationary probability vector associated with the generator matrix Q such that

$$xQ = 0,\ x\mathbf{1} = 1.$$

Grouping terms from the stationary vector x on the basis of level (number of customers) of the states gives $x = [x_0\ x_1\ x_2\ ...]$ and further, depending on the vacation-state, we get $x_k = [x_{k0}\ x_{k1}]$, for $1 \leq k \leq K$. Here, x_0 is a row nr-vector, x_{k0} $(1 \leq k \leq K)$ is a row nmr-vector, and x_{k1} $(1 \leq k \leq K)$ is a row nm-vector. Then, the matrix-geometric solution method gives

$$x_k = x_1 R^{k-1}, 2 \leq k \leq K-1, \tag{3.2}$$

where R is the minimum non-negative solution to the matrix equation $R^2A_2+RA_1+A_0 = 0$ with $sp(R) < 1$ and x_0, x_1, x_k are given by

$$x_0B_{00} + x_1B_{10} = 0,$$
$$x_0B_{01} + x_1A_1 + x_2A_2 = 0,$$
$$x_{K-1}A_0 + x_KB_{KK} = 0,$$

with the normalization condition,

$$\sum_{i=0}^{K} x_i\mathbf{1} = 1.$$

To compute the matrix R, we have used here the 'Linear Level Reduction Algorithm' [102], which is given below. The vector x_i is the joint probability of the number of customers, arrival phase, vacation duration phase, service phase during non-vacation or service phase during WV period, with i customers in the system. Let y_i be the marginal probability of

finding i customers in the system at time t, then we get

$$y_i = x_i \mathbf{1}.$$

Algorithm *Linear Level Reduction Algorithm for Finite QBDs*

1. $C(0) \leftarrow B_{00}$;
2. $C(1) \leftarrow B_{00} + B_{10} * (-C(0))^{-1} * B_{01}$;
3. **for** $i \leftarrow 2$ **to** $K - 1$
4. **do** $C(i) \leftarrow A_1 + A_2 * (-C(i-1))^{-1} * A_0$;
5. $C(K) \leftarrow B_{KK} + A_2 * (-C(K-1))^{-1} * A_0$;
6. **solve** $p(K)C(K) = 0,\ p(K)\mathbf{1} = 1$;
7. **for** $i \leftarrow K - 1$ **to** 0
8. **do** $p(i) \leftarrow p(i+1) * A_2 * (-C(i))^{-1}$;
9. $x \leftarrow (p\mathbf{1})^{-1} * p$;
10. **return** x

3.1.2 Stationary distribution at pre-arrival epochs

Let z_i be the stationary probability vector of finding i customers in the system by an arriving customer and, for $1 \leq i \leq K$, we get

$$z_i = [z_{i0}\ z_{i1}] \text{ with } z\mathbf{1} = 1,$$

where z_{ij} represents the probability that there are i customers in the system seen by an arriving customer with the system in vacation-state j, where $j = 0$ for the system in WV period and $j = 1$ for the system in non-vacation period. These pre-arrival epoch probabilities can be expressed in terms of the arbitrary epoch probabilities as

$$z_i = \frac{1}{\lambda} x_i D_1,$$

because the probability of an arrival at an arbitrary epoch with i customers in the system is same as the probability of finding i customers by an arriving customer.

The vector z_i is the joint probability of the number of customers, phase of arrival, phase of vacation, service phase during non-vacation or service phase during WV period, with i customers in the system, seen by an arriving customer. Let y_i^- be the marginal probability of finding i customers in the system at time t, then

$$y_i^- = z_i \mathbf{1}.$$

3.2 Performance measures

Here, we give various system performances in terms of stationary probabilities at arbitrary epochs. The average number of customers in the system is

$$E(N) = \sum_{i=1}^{K} i(x_{i0}\mathbf{1} + x_{i1}\mathbf{1}). \tag{3.3}$$

The average number of customers in the system when the server is in non-vacation period is

$$E(N_b) = \sum_{i=1}^{K} i x_{i1}\mathbf{1}. \tag{3.4}$$

The average number of customers in the system when the server is in WV is

$$E(N_v) = \sum_{i=1}^{K} i x_{i0}\mathbf{1}. \tag{3.5}$$

The customer loss probability or blocking probability is given by

$$P_B = \frac{1}{\lambda}(x_{K0}\mathbf{1} + x_{K1}\mathbf{1}). \tag{3.6}$$

The average delay or waiting time in the system is given by Little's law

$$E(W) = E(N)/\hat{\lambda}, \tag{3.7}$$

where $\hat{\lambda}$ is the effective arrival rate given by

$$\hat{\lambda} = \lambda(1 - P_B). \tag{3.8}$$

The traffic intensity is $\rho = \lambda/\mu_b$ and the utilization factor or effective traffic load is

$$\hat{\rho} = \lambda(1 - P_B)/\mu_b. \tag{3.9}$$

To have a better insight of this model we have taken the special case, the MMPP/M/1/K model with exponential WV durations and analyzed the system performances with the help of some numerical examples.

3.3 MMPP arrival process model: A special case

We assume the input traffic of the system under study to follow a MMPP process, where the customer (or data packet) arrival process is governed by a two-state Markov chain. When the traffic source is in state one (state two), it generates an arrival according to a Poisson process of rate $\lambda_1(\lambda_2)$ and may remain in the state in the next time period with rate $\sigma_1(\sigma_2)$. The MMPP is a doubly stochastic Poisson process and can be characterized by its generator matrix $\widehat{Q}$ and arrival rate matrix Λ defined in equation (1.26) in Chapter 1. The total arrival rate λ of the MMPP process is

$$\lambda = \frac{\sigma_1\lambda_2 + \sigma_2\lambda_1}{\sigma_1 + \sigma_2}. \tag{3.10}$$

The steady-state vector π of this Markov chain is such that

$$\pi\widehat{Q} = 0, \quad \pi\mathbf{1} = 1.$$

The solution of this system of equations is given by

$$\pi = (\pi_1, \pi_2) = \frac{1}{\sigma_1 + \sigma_2}(\sigma_2, \sigma_1).$$

With the arrival process as MMPP in the system, we consider the service times during WV, service times during non-vacation and WV duration times to follow Exp(μ_v), Exp(μ_b) and Exp(θ) respectively. Then the states of the MMPP/M/1/WV(M) model can be described by a Markov chain

$$\Delta = \{(N_t, Q_t, J_t), t \geq 0\},$$

where

$$\begin{aligned} N_t &= \text{the number of customers in system at arbitrary time } t, \\ Q_t &= \text{the vacation-state of the server } i.e., \\ &= \begin{cases} 0, & \text{if the system in WV period and} \\ 1, & \text{if the system in non-vacation period,} \end{cases} \\ J_t &= \text{the phase of arrival.} \end{aligned}$$

with state-space of Δ given by

$$E = \Big\{\big\{(0,0,k)\big\} \cup \big\{\{(i,0,k)\} \cup \{(i,1,k)\}\big\}\Big\}$$

for $i \in \{1,2,\ldots,K\}$, $k = 1,2$. The first set says that when there is no cells in the system, the system will always be in vacation.

Using the lexicographical ordering of the states as, $(0,0,1)$, $(0,0,2)$, $(1,0,1)$, $(1,0,2)$, $(1,1,1)$, $(1,1,2)$, $(2,0,1)$, $(2,0,2)$, $\ldots$, $(K,1,1)$, $(K,1,2)$, we get the generator matrix Q for this MMPP/M/1/WV model as having a similar structure as Q in (3.1) with the following entries:

$$B_{00} = \begin{bmatrix} -(\lambda_1+\sigma_1) & \sigma_1 \\ \sigma_2 & -(\lambda_2+\sigma_2) \end{bmatrix}, \quad B_{01} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \end{bmatrix}, B_{10} = \begin{bmatrix} \mu_v & 0 \\ 0 & \mu_v \\ \mu_b & 0 \\ 0 & \mu_b \end{bmatrix},$$

$A_0 = \text{diag}(\lambda_1,\lambda_2,\lambda_1,\lambda_2)$, $A_2 = \text{diag}(\mu_v,\mu_v,\mu_b,\mu_b)$, $A_1 = \begin{bmatrix} A_{11} & \theta I \\ 0 & A_{12} \end{bmatrix}$, where

$$A_{11} = \begin{bmatrix} -(\lambda_1+\sigma_1+\mu_v+\theta) & \sigma_1 \\ \sigma_2 & -(\lambda_2+\sigma_2+\mu_v+\theta) \end{bmatrix},$$

$$A_{12} = \begin{bmatrix} -(\lambda_1+\sigma_1+\mu_b) & \sigma_1 \\ \sigma_2 & -(\lambda_2+\sigma_2+\mu_b) \end{bmatrix},$$

and $B_{KK} = \begin{bmatrix} G_{11} & \theta I \\ 0 & G_{12} \end{bmatrix}$, with $G_{11} = \begin{bmatrix} -(\sigma_1+\mu_v+\theta) & \sigma_1 \\ \sigma_2 & -(\sigma_2+\mu_v+\theta) \end{bmatrix}$,

$G_{12} = \begin{bmatrix} -(\sigma_1+\mu_b) & \sigma_1 \\ \sigma_2 & -(\sigma_2+\mu_b) \end{bmatrix}$ and I is the identity matrix of dimension 2.

Let $x = [x_0, x_1, \ldots, x_K]$ be the stationary probability vector at arbitrary epoch, associated with the generator matrix Q of the MMPP/M/1/WV model and, for $1 \leq i \leq K$, $x_i = [x_{i01}\ x_{i02}\ x_{i11}\ x_{i12}]$ with

$$x_{ijk} = \lim_{t\to\infty} P(N_t = i, Q_t = j, J_t = k).$$

x_{ijk} gives the stationary joint probability that there are i customers in the system when it is in vacation-state j ($j = 0$ for system in WV and $j = 1$ for system in non-vacation) with the arrival phase k.

The system queue length distribution at pre-arrival epoch z_i is the probability that an arriving customer (whether he can join the queue or not) finds i customers in the system. We get z_i in terms of x_i as

$$z_i = \frac{1}{\lambda} x_i \Lambda,$$

where $x_{i0} = [x_{i01}\ x_{i02}]$ and $x_{i1} = [x_{i11}\ x_{i12}]$.

In the next section, we will present some numerical examples and show the dependency of system efficiency on system parameters.

3.4 Numerical examples

This section presents numerical results to illustrate the behavior of the MMPP/M/1/K WV model subject to various parameter values. These results focus on evaluation of the system performance in terms of customer loss probability, mean queue length of buffer occupancy and mean waiting time or system delay. The effects of traffic burstiness and of correlation upon various performance measures are also shown. A comparison of this correlated arrival model with an uncorrelated arrival model (Erlang-2 arrival process model) is taken to observe the consequence of correlation.

An MMPP process has four parameters, $\sigma_1, \sigma_2, \lambda_1, \lambda_2$ and two equations, one of total arrival rate λ and the other of squared coefficient of variation, c^2. The squared coefficient of variation, c^2, of the interarrival times of the MMPP process is given by equation (1.28)

and the autocorrelation function of the interarrival time with step 1 is given by equation (1.29) of Chapter 1. Normally, the characteristics of the traffic source–the mean $\frac{1}{\lambda}$ and the squared coefficient of variation c^2 of the interarrival distribution–are given. From the formulae of λ and c^2 given in equations (1.27) and (1.28), by fixing λ_2, we can obtain the corresponding MMPP parameters $\sigma_1(\lambda_1)$ and $\sigma_2(\lambda_1)$ as a function of λ_1. Using these parameters the relation between correlation and λ_2 can be obtained by using the 1-step correlation formula. Therefore by fixing two variables, λ_1 and λ_2, we can get a MMPP process for the given λ and c^2.

We consider an MMPP process with arrival rate $\lambda = 0.7$ and coefficient of variation $c^2 = 2$ with $\lambda_1 = 1, \lambda_2 = 0.1$. The other parameters are taken as $\mu_b = 1, \theta = 0.6$.

3.4.1 Effect of buffer size

In Figure 3.1, we have plotted loss probability against system capacity. When the buffer size is large, it is intuitive that, the loss probability of the system approaches zero. In the graph, we can see that, increase in system size leads to the reduction of loss probability up to 35%. Obviously, increasing system size is not always cost effective. So an alternate is to enhance service rates during vacation periods, if feasible. The service rate enhancement can minimize the customer loss probability by 15%. When $\mu_v = 0$, the system coincides with classical vacation model, which gives maximum loss probability. For $\mu_v = 1 = \mu_b$, this model coincides with a non-vacation model and this model has the minimum loss probability compared to others. This graph shows the variation of loss probability between a classical vacation and a non-vacation model. The effect of μ_v on loss probability vanishes gradually with the increase in system capacity.

Figure 3.2 shows the behavior of loss probabilities for different rates of vacation durations. Here, we can observe that the loss probabilities are comparatively less for systems having large mean of vacation durations $\left(\frac{1}{\theta}\right)$. For systems with long vacations, the loss probabilities can be reduced up to 30% by raising the system capacity whereas for a MMPP model, for systems with small vacations (high θ), this reduction can go up to

50%. Therefore, in a WV model the customer loss probability can be minimized either by raising the system capacity or by increasing service rate during vacation period or, by having longer vacation durations.

With the same parameter values as above, in Figure 3.3 we have plotted the mean queue length against system capacity. The behavior of the queue length is exactly opposite to that of loss probability. The mean queue lengths are long for high capacity systems since they provide more waiting space. It is intuitive that a WV model has lower queue length than a model with complete vacation. The graph gives that when $\mu_v = 0.6$ and $K = 20$, the mean queue length can be shortened by 25%, if we impose WV into the model. Also, the graph indicates a higher percentage of decrease in queue lengths if the capacity is further increased. This means that, the difference in queue lengths between these two models becomes more prominent as the system capacity increases.

To see the effect on mean queue length for different MMPP arrival processes, we have plotted in Figure 3.4 the mean queue length against the service rate during vacation μ_v for three different arrival rates. It shows that, an increase of 20% in arrival rate can raise the queue lengths by 40% for low vacation-service rates.

Table 3.1 compares the distribution of number of customers in MMPP/M/1/10/WV(M) for three different service rates $\mu_b = 0.5, 0.9$ and 1.5. The distribution of customers is different for different traffic intensities. When $\rho > 1$, the probabilities first decrease at the beginning and then increase with the increase in number of customers in the system, that is, the system has a point of inflexion at the beginning of the column whereas for $\rho < 1$, it is towards the end of the column *i.e.,* when i=9, 10.

In Table 3.2, we have compared the distribution of number of customers for the MMPP/M/1/10/WV model considering the system at two different time epochs, arbitrary epoch (y_i) and pre-arrival epoch (y_i^-), for $i = 1, 2, \ldots, K$. Here y_{i0} gives the distribution of number of customers at arbitrary epoch when the system is in WV and y_{i1} gives the distribution when the system in in non-vacation period. Similarly, y_{i0}^- and y_{i1}^- gives the same in pre-arrival epochs when the system is in WV and non-vacation respectively.

3.4.2 Effect of burstiness

To demonstrate the impact of traffic burstiness upon system performance for various values of squared coefficient of variation c^2, we have in Figure 3.5, the plot of customer loss probability against the system size K. The loss probabilities are almost same for all values of c^2 when the buffer sizes are too small, that is, bursty traffic does not effect loss probabilities when the system capacity is very small. For $c^2 = 1$, when the buffer size increases, the loss probabilities drop to zero at a faster rate compared to the models with high c^2. Its behavior is more concave in nature whereas the concavity is less for high bursty traffic. Because of this nature, the variation in loss probability among all the MMPP models, with different c^2, is maximum around $K = 7$.

3.4.3 Effect of correlation

The impact of correlation can be viewed by comparing the correlated arrival model to one with uncorrelated arrivals. We have taken a WV model with the arrivals as a Erlang-2 process where the interarrival times are independent, keeping the parameters same with the MMPP model. Figure 3.6 is the plot of autocorrelation coefficient of the interarrival times against the arrival rate λ_1 for an MMPP arrival process. The autocorrelation is highest as λ_1 reaches value 0.2 when λ_2 is fixed to 1.

In Figure 3.7, comparisons are given for the loss probabilities of the correlated model and the Erlang-2 model, with the change in vacation service rates. The figure shows that the loss probabilities of a correlated arrival model and an uncorrelated arrival model vary a lot from each other. The mean queue lengths with system capacity for both Erlang-2 and MMPP arrival models are plotted in Figure 3.8. The difference in queue lengths of MMPP models with the corresponding Erlang-2 model is about 60%, which shows that correlation in arrivals affect the system performance which is not negligible. Hence, we can conclude that the negligence in correlated arrivals can be a cause of system drawback.

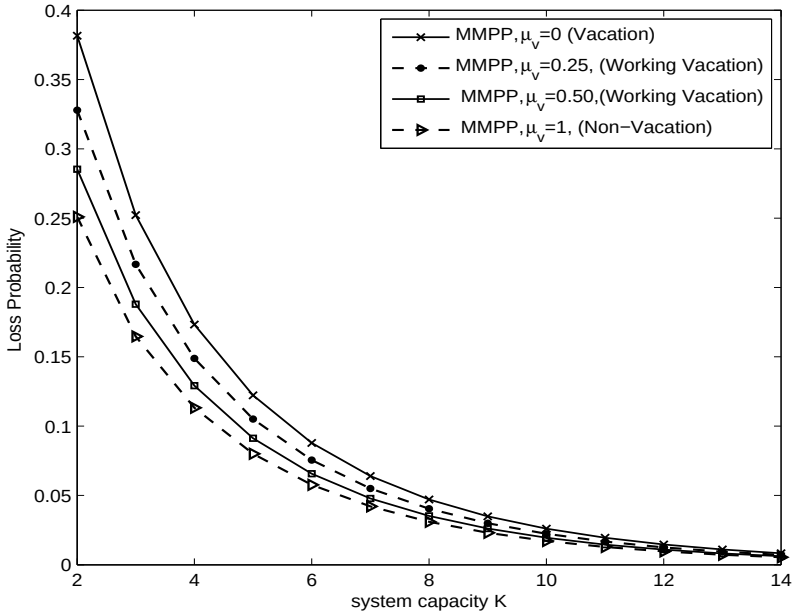

Figure 3.1: Loss probability vs system capacity with $c^2 = 2, \theta = 0.6$.

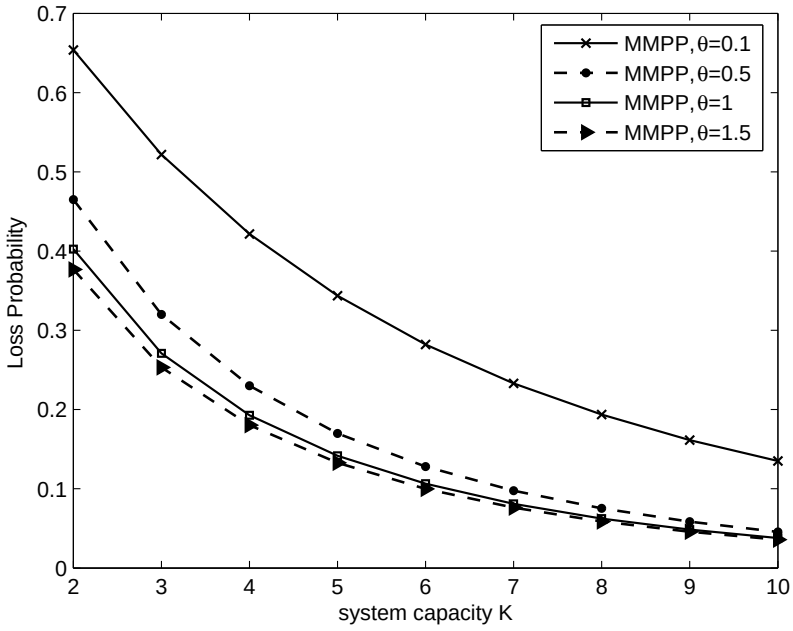

Figure 3.2: Loss probability vs system capacity with $c^2 = 2, \mu_v = 0.3$.

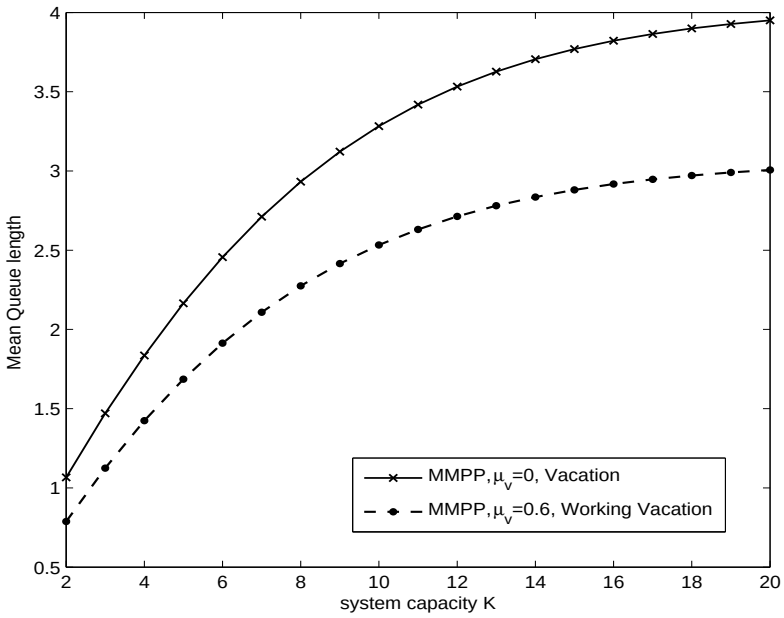

Figure 3.3: Mean queue length vs system capacity with $c^2 = 2, \theta = 0.6$.

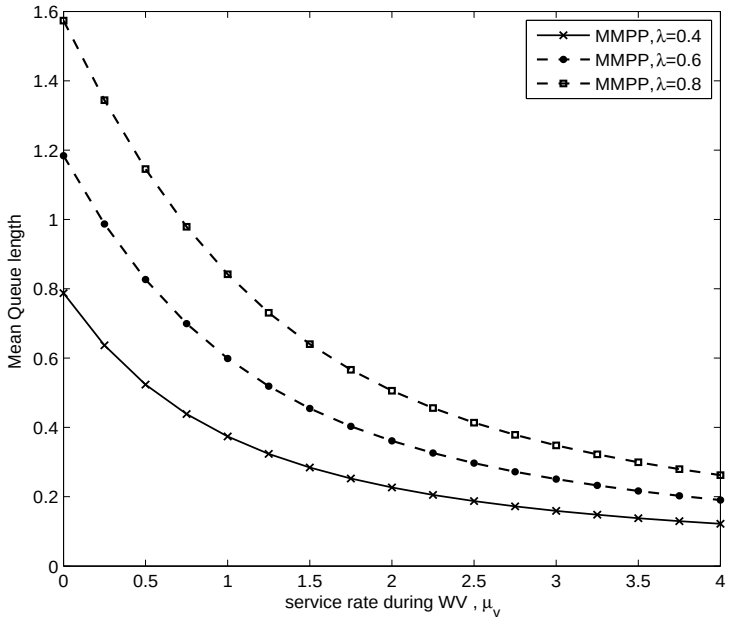

Figure 3.4: Mean queue length vs vacation-service rate with $c^2 = 1.5$, $K = 5$, $\theta = 0.6$.

Table 3.1: Distribution of number of customers in finite model for $\lambda_1 = 1,\ \lambda_2 = 0.1, \lambda = 0.7,\ \theta = 0.1,\ \mu_v = 0.4$.

i	ρ=1.40 y_i	ρ=0.77 y_i	ρ=0.47 y_i
0	0.0337	0.1354	0.2027
1	0.0291	0.0972	0.1331
2	0.0344	0.0936	0.1144
3	0.0423	0.0928	0.1008
4	0.0521	0.0905	0.0875
5	0.0641	0.0870	0.0751
6	0.0793	0.0829	0.0643
7	0.0993	0.0790	0.0557
8	0.1281	0.0765	0.0504
9	0.1749	0.0775	0.0508
10	0.2627	0.0876	0.0650
sum	1.0000	1.0000	1.0000

Table 3.2: Distribution of number of customers at various epochs for $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda = 0.7$, $\theta = 0.5$, $\mu_v = 0.6$, $\mu_b = 1$ and $P_B = 0.1045$.

i	y_{i0}	y_{i1}	y_i	y_{i0}^-	y_{i1}^-	y_i^-
0	0.1709	–	0.1709	0.1075	–	0.1075
1	0.0850	0.0242	0.1092	0.0851	0.0192	0.1043
2	0.0660	0.0334	0.0994	0.0717	0.0306	0.1023
3	0.0558	0.0381	0.0939	0.0611	0.0374	0.0985
4	0.0482	0.0400	0.0882	0.0525	0.0407	0.0932
5	0.0421	0.0400	0.0821	0.0457	0.0418	0.0875
6	0.0377	0.0386	0.0763	0.0406	0.0414	0.0820
7	0.0348	0.0365	0.0713	0.0374	0.0402	0.0776
8	0.0341	0.0338	0.0679	0.0371	0.0386	0.0757
9	0.0365	0.0311	0.0676	0.0417	0.0373	0.0790
10	0.0449	0.0283	0.0732	0.0561	0.0363	0.0924
sum	0.6561	0.3439	1.0000	0.6365	0.3635	1.0000

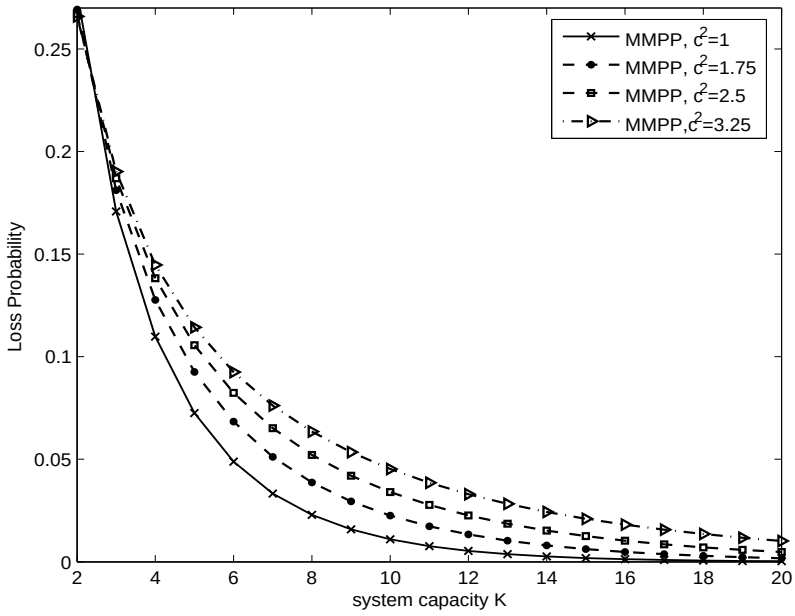

Figure 3.5: Loss probability vs system capacity with $\theta = 0.6, \mu_v = 0.3$.

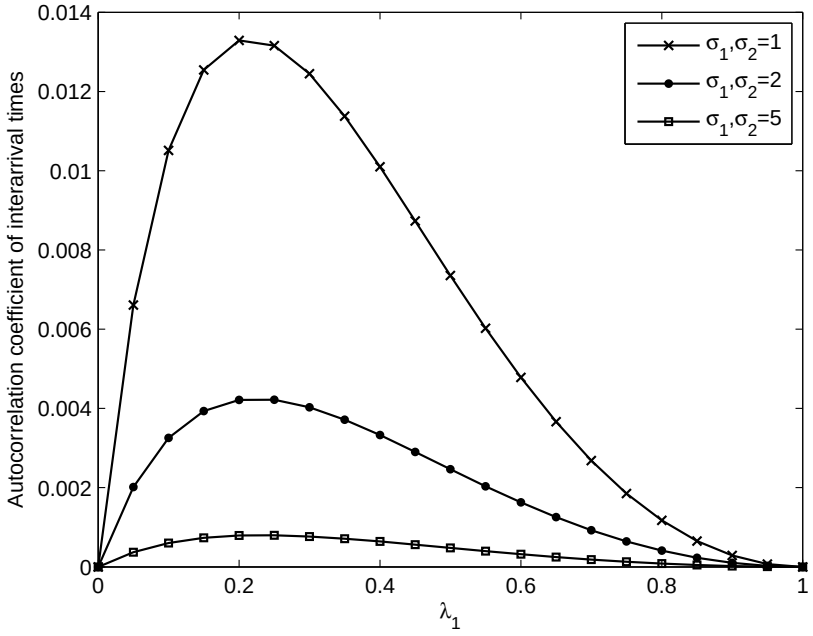

Figure 3.6: Autocorrelation function vs arrival rate λ_1.

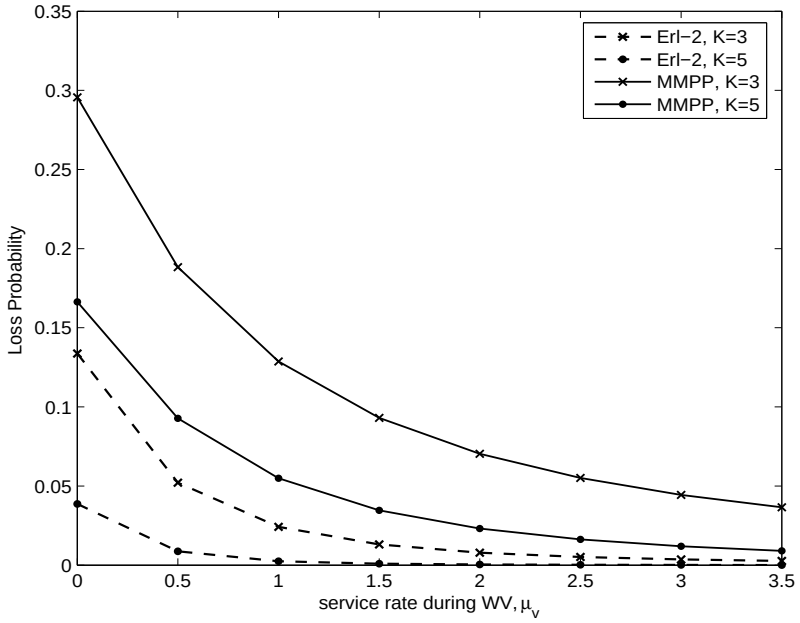

Figure 3.7: Loss probability vs vacation-service rate with $c^2 = 2$, $\theta = 0.6$.

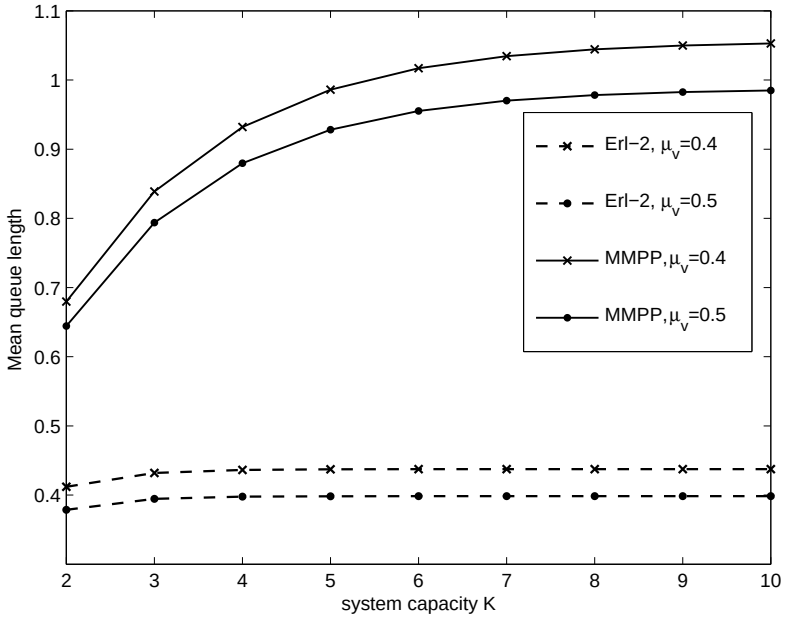

Figure 3.8: Mean queue length vs system capacity with $c^2 = 2$, $\theta = 0.6$.

Chapter 4

A Multi-server Queue with Impatience

In communication networks, multiple servers are used to reduce the traffic congestion and improve server performance. Also, multiple services are used in highly efficient bandwidth-intensive applications. Different services may require different channel capacities and capacity of a channel depends upon the number of resources allocated to it. Short distance networks, like local area networks (LANs) mostly use multi-mode WDM links. Multi-mode link is a single fiber link that supports many propagation paths or transverse modes through it. By using WDM in multi-mode fiber, the bandwidth of the fiber is multiplied by the number of paths used [10]. In a WDM network, the available fiber bandwidth is divided into WDM channels and this division of bandwidth or channel allocation is based on the capacities required for various services provided by the network. Multiple channels allocated for a particular service can be viewed as a multi-server system. For a high performance system, WDM channel allocation should lead to optimized resource utilization in a given network which is physically feasible and cost-effective. The multi-mode fiber links are predominant in LANs due to their low-cost installation and maintenance. LAN over Internet Protocol (IP) allows the forwarding of LAN packets over the internet or an intranet network. One of the most critical performance measures in

LAN over IP is the percentage of packets that are transmitted within hard delay bound or time constraint. If quality of service requirements is not met within the time bound, end users may terminate the internet connections. A connection is terminated by pressing the stop button, refreshing the connection or following a different link. This behavior can be termed as the impatience of a user in LANs. To study the effect of multiple servers in a WDM network performance, we consider in this chapter, a multi-server model with asynchronous multiple working vacation (AMWV) policy and impatient customers. In a AMWV policy, the servers take vacations individually and continue taking vacations till they do not find any customer in the system. Lin and Ke [111] presented a multi-server WV queue with exponential interarrivals but this model does not represent systems with non-exponential arrivals or state-dependent systems. To study the role of arrival processes on a multi-server model having impatient customers, we consider here the PH-renewal arrival process.

4.1 Model description

We consider a PH/M/c queue with multiple WVs and impatient customers. The interarrival times of customers follow a PH(α, T) distribution of dimension n and with arrival rate λ. The customers are served according to a FCFS basis. An arriving customer who finds all the c servers busy has to wait in queue i.e., when the number of customers in the system is more than c, a queue begins to form. Each server works independent of each other. The service times of each server during the non-vacation period follow an Exp(μ_b) distribution. For each server, the duration of WVs follows an Exp(θ) distribution. During a WV period of a server, the customers are served with Exp(μ_v) distribution, where $\mu_v < \mu_b$. When a server returns from its vacation, the server switches its service rate from μ_v to μ_b if it finds atleast one customer in queue waiting for service or finds an ongoing service in that server. Otherwise, if the server finds an empty queue it immediately leaves for another WV.

Whenever an arriving customer finds all the servers in WV and all are occupied, the customer joins the queue and activates an impatience timer X. This impatient timer X follows an Exp(ξ) distribution and is independent of the number of customers in the queue at that moment. If no server returns to its non-vacation period by the time X expires, the customer leaves the system and never returns. Otherwise, if any of the servers return from its vacation before the time X expires, the customer stops the timer and stays in the system until its service is completed. The interarrival times, service times, vacation duration times and the impatient times, all are taken to be mutually independent.

To model this system, we define a continuous-time Markov chain

$$\Delta = \{(N_t, Q_t, J_t),\ t \geq 0\},$$

where N_t denotes the total number of customers in the system, Q_t denotes the number of busy servers in non-vacation state and J_t gives the phase of the arrival process. The state space of this Markov chain is

$$E = \{(i,j,k);\ 0 \leq i < c,\ j \leq i,\ 0 \leq k \leq n\} \cup \{(i,j,k);\ i \geq c,\ 0 \leq j \leq c,\ 0 \leq k \leq n\} \tag{4.1}$$

The lexicographical order of the states *i.e.,* $(0,0,0)$, $(0,0,1)$, ..., $(0,0,n)$, $(1,0,0)$, ..., $(1,0,n)$, $(1,1,1)$, ..., $(1,1,n)$, $(2,0,1)$, ..., $(2,0,n)$, $(2,1,1)$, ..., $(2,1,n)$, $(2,2,1)$, ..., $(2,2,n)$, ..., $(c,0,1)$, ..., (c,c,n), $(c+1,0,1)$, ..., $(c+1,c,n)$, ..., gives the infinitesimal generator matrix of the Markov chain as with

$$Q = \begin{pmatrix} A_1^{(0)} & A_0^{(0)} & & & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & & & \\ & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & & \\ & & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.2}$$

where, for $0 < i < c$,

$$A_2^{(i)} = \begin{bmatrix} i\mu_v I & & & & & \\ & (\mu_b + (i-1)\mu_v)I & & & & \\ & & (2\mu_b + (i-2)\mu_v)I & & & \\ & & & \ddots & & \\ & & & & ((i-1)\mu_b + \mu_v)I & \\ & & & & & i\mu_b I \end{bmatrix}_{(i+2)n \times (i+1)n,}$$

for $0 \leq i < c$

$$A_1^{(i)} = \begin{bmatrix} T - (c\theta + i\mu_v)I & c\theta I & & & \\ & T - ((c-1)\theta + \mu_b + (i-1)\mu_v)I & (c-1)\theta I & & \\ & & \ddots & \ddots & \\ & & & & T - i\mu_b I \end{bmatrix}_{(i+1)n \times (i+1)n}$$

and $A_0^{(i)} = \begin{bmatrix} T^0\alpha & & & 0 \\ & T^0\alpha & & 0 \\ & & \ddots & \\ & & & T^0\alpha \quad 0 \end{bmatrix}_{(i+1)n \times (i+2)n.}$

For $i \geq c$, *i.e.*, for $i = c + l,\ l = 0, 1, 2, \ldots,$

$$A_2^{(c+l)} = \begin{bmatrix} (c\mu_v + l\xi)I & & & & \\ & (\mu_b + (c-1)\mu_v)I & & & \\ & & (2\mu_b + (c-2)\mu_v)I & & \\ & & & \ddots & \\ & & & & c\mu_b I \end{bmatrix}_{(c+1)n \times (c+1)n,}$$

$$A_1^{(c+l)} = \begin{bmatrix} T - (c(\theta + \mu_v) + l\xi)I & c\theta I & & \\ & \ddots & \ddots & \\ & & T - (\theta + \mu_v + (c-1)\mu_b)I & \theta I \\ & & & T - (c\mu_b)I \end{bmatrix}_{(c+1)n \times (c+1)n}$$

$$\text{and } A_0^{(c+l)} = \begin{bmatrix} T^0\alpha & & & \\ & T^0\alpha & & \\ & & \ddots & \\ & & & T^0\alpha \end{bmatrix}_{(c+1)n\times(c+1)n.}$$

The matrix I is a identity matrix of dimension n. Here, for $0 \leq i \leq c-1$, dimension of the matrices $A_0^{(i)}, A_1^{(i)}$ and $A_2^{(i)}$ increases with the levels; and for $i \geq c$ the matrices are of dimensions $(c+1)n \times (c+1)n$ each. This is a non-homogeneous QBD process which we assume to be irreducible.

4.2 Stationary distribution

The queueing system under study is stable for $\rho = \lambda/c\mu_b < 1$ [50]. Let x be the stationary probability vector associated with Q satisfying

$$xQ = 0,\ x\mathbf{1} = 1. \tag{4.3}$$

Aggregating terms depending on levels we get $x = [x_0\ x_1\ x_2\ ...]$. Further depending on number of busy servers in non-vacation, we get for $0 \leq i \leq c-1$, $x_i = [x_{1j}, \ldots, x_{ij}, \ldots]$, which are row $(i+1)n$-vectors and for $i \geq c$, x_i are row $(c+1)n$-vectors. Each x_{ij} vector is a n-dimensional row vector, $x_{ij} = [x_{ij1}, \ldots, x_{ijn}]$ for $j \leq i$, depending on the phases of arrivals.

In this model, the generator matrix Q is spatially non-homogeneous and a closed-form analytical solution or a direct algorithmic computation of the stationary probability vector x is quite difficult, if not impossible. For such level-dependent QBDs (LDQBDs) the stationary vectors are usually approximated by using various numerical approximation methods. Those approximation methods can be classified into the following categories.

1. **Finite truncation methods:** A finite truncation method or a direct truncation method consists of replacing the original infinite state-space model by a finite one, so that the resulting model is solvable. That is, the system size is restricted to a

sufficiently large value, say K_f, such that the arriving customer finding the system full is considered lost. A number of approaches are available for determining the cut-off point K_f depending on the required system performance measures. For approximating system performance, the value of K_f can be increased until the largest individual change in the elements of stationary probabilities x, for two successive values, is less than ϵ, a predetermined infinitesimal value [13, 37].

2. **Generalized truncation methods:** Falin [57] introduced the generalized truncation method where the original infinite state space is replaced by another infinite but simplified and solvable state space. Based on this, Diamond and Alfa [50] and later Artalejo and Poze [18] modified the truncation method for computing the stationary distribution for single-server and multi-server models respectively. Artalejo [13] compared the finite truncation method with generalized truncation by studying the stationary distribution of a multi-server model.

3. **Truncation method using LDQBD processes:** Bright and Taylor [33] gave a procedure for obtaining a dominating process to arrive at a truncation level of a QBD process. The dominating process is constructed in the same state-space of the process under study. This algorithm requires the computation of several matrices obtained as solutions to quadratic equations. A considerable amount of quantities need to be evaluated and the truncation process is based on tail probabilities. By using this truncated method, Krishnamoorthy et al. [97] studied mean waiting time of a tagged customer in a priority queue model.

4. **Matrix-geometric approximations:** For high level congestion systems, Neuts and Rao [122] developed an efficient algorithmic solution by making a simplifying approximation which yields an efficiently computable infinitesimal generator with modified matrix-geometric steady-state vector. Artalejo and Chakravarthy [15] studied a retrial queue using this method. Chakravarthy [37] analyzed multi-server retrial queue using this method and gave efficient algorithms to compute various steady-state performances.

Different methods have different levels of computable efficiency but it is expected that whichever method is used, the general behavior of the performance measures of a system with a change in system parameters, is not affected by the method used. Since the finite truncation method is comparatively tractable than the others, we choose this method to derive the stationary distributions of the non-homogeneous QBD (4.2).

4.3 The finite truncation method

In the finite truncation method, the infinite generator matrix is truncated at a finite level K. That is, the system of equations given by $xQ = 0$ and $x\mathbf{1} = 1$, is truncated at a sufficiently large value, say K, and the resulting finite system is solved for the equilibrium probability vector. The level K is arbitrary but fixed and it is chosen such that customer loss probability due to truncation is small. As, for higher dimension generator matrices, the level K is difficult to find analytically, a trial and error approach needs to be adopted. An appropriate level, say K_f, is determined by starting with a reasonable initial value for K and increasing it progressively until an appropriately chosen cut-off criterion is met. Stationary probability vector x can then be evaluated by an iterative method, such as Gauss-Seidel, which takes advantage of the sparsity and structure of Q. For each new value of K, the previously computed vector x is used as the initial solution to reduce the number of iterations required [122]. Thus, the numerical implementation of the approximation based on finite truncation implies the determination of an appropriate cut-off level K_f. Here, we use the algorithm given by Artalejo [13], the steps of which are described below.

For K as the cut-off point, the modified generator will be

$$\hat{Q}(K) = \begin{pmatrix} A_1^{(0)} & A_0^{(0)} & & & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & & & \\ & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & & \\ & & \ddots & \ddots & \ddots & \\ & & & A_2^{(K-1)} & A_1^{(K-1)} & A_0^{(K-1)} \\ & & & & A_2^{(K)} & \phi_1^{(K)} \end{pmatrix}, \tag{4.4}$$

where $\phi_1^{(K)} = A_1^{(K)} + A_0^{(K)}$. Let π be the stationary distribution of $\hat{Q}(K)$ which satisfies

$$\pi\hat{Q}(K) = 0,\ \pi\mathbf{1} = 1, \tag{4.5}$$

where $\pi = [\pi(0), \pi(1), \ldots, \pi(K)]$, by aggregating terms of the QBD $\hat{Q}(K)$, depending on levels. Define $z = [z_0(K), z_1(K)]$ with

$$z_0(K) = [\pi(0), \pi(1), \ldots, \pi(K-1)], \tag{4.6}$$

$$z_1(K) = \pi(K). \tag{4.7}$$

and $z(K,i) = \pi(i),\ 0 \le i \le K$. Here $z_0(K)$ is a row vector of dimension $m = \frac{n(c+1)(c+2)}{2} + n(c+1)(K-c)$ and $z_1(K)$ is a row vector with dimension $n(c+1)$. By partitioning $\hat{Q}(K)$ according to $z_0(K)$ and $z_1(K)$, we have

$$(z_0(K), z_1(K)) \begin{pmatrix} B_{00}(K) & B_{01}(K) \\ B_{10}(K) & B_{11}(K) \end{pmatrix} = \left(0_m, 0_{n(c+1)}\right), \tag{4.8}$$

where $B_{00}(K)$ is the matrix obtained by deleting last column matrices and last row matrices from $\hat{Q}(K)$, $B_{01}(K) = \text{trans}\left[0, 0, \ldots, 0, A_0^{(K-1)}\right]$, $B_{10}(K) = \left[0, 0, \ldots, 0, A_2^{(K)}\right]$ and $B_{11}(K) = \phi_1^{(K)}$. These are block structured matrices with $(K \times K)$, $(K \times 1)$, $(1 \times K)$ and (1×1) blocks respectively. From (4.8), we find that

$$z_1(K)B_{10}(K)B_{00}^{-1}(K) = -z_0(K), \tag{4.9}$$

$$z_1(K)\left[(B_{11}(K) - B_{10}(K)B_{00}^{-1}(K)B_{01}(K)\right] = 0_{n(c+1)}. \tag{4.10}$$

Further, we can partition $B_{00}(K)$ as

$$B_{00}(K) = \begin{pmatrix} B_{00}(K-1) & B_{01}(K-1) \\ C_0(K-1) & C_1(K-1) \end{pmatrix}, \tag{4.11}$$

where

$$C_0(K-1) = \left[0_n, 0_{2n}, \ldots, 0, A_2^{(K-1)}\right], \tag{4.12}$$

$$C_1(K-1) = A_1^{(K-1)}. \tag{4.13}$$

The inverse of matrix $B_{00}(K)$ can be determined, using methods given in [74], as

$$B_{00}^{-1}(K) = \begin{pmatrix} D_{00}(K) & D_{01}(K) \\ D_{10}(K) & D_{11}(K) \end{pmatrix}, \tag{4.14}$$

where

$$\begin{aligned} D_{00}(K) &= \left[B_{00}(K-1) - B_{10}(K-1)C_1^{-1}(K-1)C_0(K-1)\right]^{-1}, \\ D_{10}(K) &= -C_1^{-1}(K-1)C_0(K-1)D_{00}(K), \\ D_{11}(K) &= \left[C_1(K-1) - C_0(K-1)B_{00}^{-1}(K-1)B_{01}(K-1)\right]^{-1}, \\ D_{01}(K) &= -B_{00}^{-1}(K-1)B_{01}(K-1)D_{11}(K). \end{aligned}$$

As the dimensions of the matrices increase in each iteration, the calculations to compute the above matrices involve multiplications and inverse of increasingly large matrices. If we exploit the structure of matrices in the above equations, we notice that the sparse blocks of $B_{01}(K-1)$ and $C_0(K-1)$ simplifies the calculations. $B_{01}(K-1)$ has only one non-zero square matrix $A_0^{(K-1)}$ of dimension $n(c+1)$ in the last rows and $C_0(K-1)$ has one $A_2^{(K-1)}$ in the last columns. So $B_{00}^{-1}(K-1)B_{01}(K-1)$ can be written in simplified form as $[D_{10}(K-1), D_{11}(K-1)]A_0^{(K-1)}$. Further $C_0(K-1)B_{00}^{-1}(K-1)B_{01}(K-1)$ becomes $A_2^{(K-1)}D_{11}(K-1)A_0^{(K-1)}$. These substitutions make the remaining operations in $D_{01}(K)$ and $D_{11}(K)$ simple, as they involve multiplications and inverse of only known simple matrices of size $n(c+1)$. The key step is to compute the matrix $D_{00}(K)$. The inverse in the definition of $D_{00}(K)$ can be computed by using small-rank adjustment, i.e., if we have the inverse of a matrix A and we want the inverse of its adjustment $B = A + XWY$, where W is a matrix of smaller order than A, then we have

$$B^{-1} = \left[I - A^{-1}X(W^{-1} + YA^{-1}X)^{-1}Y\right]A^{-1}.$$

Here, we have $A = B_{00}(K-1)$, $X = -B_{01}(K-1)$, $W^{-1} = C_1^{-1}(K-1)$ and $Y = C_0(K-1)$. Thus, we obtain that

$$D_{00}(K) = B^{-1} = \left[I - D_{01}C_0(K-1)\right]B_{00}^{-1}(K-1),$$

so D_{00} is obtained by multiplications and additions of already computed matrices. Finally, we have

$$D_{11}(K) = \left[C_1(K-1) - A_2^{(K-1)}D_{11}(K-1)A_0^{(K-1)}\right]^{-1}, \tag{4.15}$$

$$D_{01}(K) = -\left[D_{10}(K-1), D_{11}(K-1)\right]A_0^{(K-1)}D_{11}(K), \tag{4.16}$$

$$D_{00}(K) = [I - D_{01}C_0(K-1)]\, B_{00}^{-1}(K-1), \tag{4.17}$$

$$D_{10}(K) = -C_1^{-1}(K-1)C_0(K-1)D_{00}(K). \tag{4.18}$$

So, the computation of vector $z_1(K)$ is reduced to solving the system (4.10) subject to the normalization condition

$$\pi(K)\left[e_{n(c+1)} - B_{10}(K)B_{00}^{-1}(K)e_{nK(c+1)}\right] = 1. \tag{4.19}$$

Finally, the vector $z_0(K)$ can be solved substituting $z_1(K)$ in (4.9). To get the cut-off value, successive increments of K are made, starting from $K = c+1$ and we stop at the point $K = K_f$ when

$$\max_{0\le i\le K_f} \parallel z(K_f, i) - z(K_f - 1, i) \parallel_\infty < \epsilon, \tag{4.20}$$

where ϵ is an infinitesimal quantity and $\parallel . \parallel_\infty$ is the infinity norm. The whole method of computing the stationary distribution using Finite Truncation Method is summarized in the following algorithm.

Algorithm *Finite Truncation Method*

1. $K := c+1$;
2. **compute** $B_{00}^{-1}(K)$
3. **compute** $z_1(K)$ by (4.10) and (4.19)
4. **compute** $z_0(K)$ by (4.9)
5. **store** $B_{00}(K)$, $B_{00}^{-1}(K)$, $B_{01}(K)$ and $B_{10}(K)$
6. $K := K+1$
7. **while** $K \ge c$
8. compute $B_{00}^{-1}(K)$ by (4.14)
9. compute $z_1(K)$ by (4.10) and (4.19)
10. compute $z_0(K)$ by (4.9)
11. **if** $\max_{0\le i\le K_f} \parallel z(K_f, i) - z(K_f - 1, i) \parallel_\infty < \epsilon$
12. $K = K_f$
13. break;

14. update $B_{00}(K)$, $B_{00}^{-1}(K)$, $B_{01}(K)$ and $B_{10}(K)$

15. $K := K + 1$

16. **do**

4.4 Performance measures

The performance measures are the qualitative behavior of the model under study. In a multi-server queueing model the efficiency of the model depends upon the mean number of busy servers, the mean queue length, the blocking probability and the mean number of customers lost due to impatience.

In our model, the server serves even during its vacation. Therefore, the number of busy servers will be i, $0 \leq i \leq c$, if there are i customers in the system and when the system has more than c customers, all the servers will be busy serving customers either in WV or in non-vacation with rate μ_v and μ_b respectively. The mean number of servers busy in non-vacation is

$$B_s = \sum_{i=0}^{c-1}\sum_{j=1}^{i} jx_{ij}\mathbf{1} + \sum_{i=c}^{\infty}\sum_{j=1}^{c} jx_{ij}\mathbf{1}. \tag{4.21}$$

The mean queue length of the system under study is

$$N = \sum_{i=0}^{\infty} ix_i\mathbf{1}. \tag{4.22}$$

Availability of the server, R, is the probability that an arrival finds a server free and is given by

$$R = P(N < c) = \sum_{i=1}^{c-1} x_i\mathbf{1}. \tag{4.23}$$

The blocking probability of a multi-server queue is the probability of refraining a customer from service. In our model, a customer is kept waiting in the queue for service when all the servers are in busy state, either in WV or in non-vacation.

$$B_p = P(N \geq c) = 1 - \sum_{i=1}^{c-1} x_i\mathbf{1} = 1 - R. \tag{4.24}$$

The mean number of customers lost by the system will be

$$N_c = \sum_{i=0}^{\infty} i x_{i0} \mathbf{1}. \tag{4.25}$$

4.5 Numerical examples

Let us illustrate the behavior of our PH/M/c/WV queue with the help of some numerical examples. The algorithm '*Finite Truncation Method* ' is coded in MATLAB©. The algorithm computes the stationary distribution and its main objective is to find the termination criteria of the level K_f. Starting with an initial value $K \geq c+1$ and progressively increasing the value of K until a change in the stationary probability z is sufficiently small due to the increased K. We choose the smallest value of K_f such that $\max_{0 \leq i \leq K_f} \parallel z(K_f, i) - z(K_f - 1, i) \parallel_\infty < \epsilon$, for $\epsilon = 10^{-6}$. With this selection criteria, we find the values of K_f, the mean queue lengths and the blocking probabilities for various sets of parameter values and for different arrival processes. Here we take some examples (from [37]) of well known distributions and give their PH-representations below:

1. Exponential (Exp)

$$T = -1, \quad T^0 = 1 \quad \text{and} \quad \alpha = 1.$$

2. Erlang-2 (Erl)

$$T = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}, \quad T^0 = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \quad \text{and} \quad \alpha = [1 \;\; 0].$$

3. Hyperexponential-2 (Hyp)

$$T = \begin{pmatrix} -1.9 & 0 \\ 0 & -0.19 \end{pmatrix}, \quad T^0 = \begin{pmatrix} 1.9 \\ 0.19 \end{pmatrix} \quad \text{and} \quad \alpha = [0.9 \;\; 0.1].$$

All these PH distributions have the same mean arrival rate $\lambda = 1$. The standard deviations of the three distributions are, 1.0, 0.70711 and 2.24472 respectively. The service rate during non-vacation period, μ_b, is calculated for specific values of ρ using the formula

$\rho = \frac{\lambda}{c\mu_b}$. We have chosen $\rho = 0.1,\ 0.5$ and 0.9 for given values of c ($c = 1, 3, 6$). The effect of parameters on system performance are illustrated here. We will mention the models having interarrivals as hyperexponential, Erlang and exponentially distributed as hyperexponential model, Erlang model and exponential model respectively.

4.5.1 Effect on cut-off value K_f

We have illustrated here the effect of the parameters, namely traffic intensity (ρ), rate of vacation duration (θ), service rate during WV (μ_v) and the type of arrival process on the truncation cut-off value K_f. For three different values of c ($c = 1, 3, 6$) different tables are presented. The impatience rate is fixed at $\xi = 0.1$ for the tables. We have the following observations from the Tables 4.1, 4.2 and 4.3 (given at the end of the chapter) :

1. The cut-off value increases with the increase in the variance of the distribution of the interarrival times. For Erlang model the termination is fastest whereas for hyperexponential it is the slowest. This behavior seems to be same for all sets of parameter values and for all c.

2. For a particular arrival process when the traffic load is small, the value of K_f decreases with increase in μ_v and also with the increase in θ. But for high ρ (> 0.5) and high θ it shows the reverse property for all arrival processes and for any number of servers c *i.e.,* when the system load is heavy and the system has small vacation durations, the cut-off value seems to be high for all types of arrival processes and any number of servers.

3. When the vacation duration rate θ is too high ($= 100$), the K_f value remains unaffected by vacation-service rate μ_v for any number of servers.

These observations show that the cut-off value depends on the system parameters and also on the arrival process, but becomes independent of vacation-service rates when we have systems with small vacation durations.

4.5.2 Effect on mean queue length

1. The mean queue length of the system depends upon the arrival process. The Tables 4.1, 4.2 and 4.3, show that systems with interarrival distributions of high variance have higher number of customers in the queue for any number of servers. We have fixed the impatient rate at $\xi = 0.1$. For $c = 1, 3, 6$ with $\rho = 0.5$, we have Figures 4.1,4.2 and 4.3 respectively, and with $\rho = 0.9$, we have Figures 4.4, 4.5 and 4.6, where the change in mean queue lengths are given for increasing vacation-service rates. A hyperexponential model always has the highest mean queue length compared to the corresponding Erlang and exponential models, irrespective of the number of servers.

2. When the traffic load is heavy, $\rho = 0.9$, increase in vacation-service rate does not affect the mean queue lengths, whatever be the arrival process or the number of servers (Figures 4.4, 4.5 and 4.6).

3. For $\rho = 0.1$ and $c = 1, 3, 6$, we plot the Figures 4.7, 4.8 and 4.9 respectively. Here hyperexponential model has the least queue length compared to the corresponding Erlang and exponential models. For $c = 6$, the queue lengths are same for all of them are shown in Figure 4.9. Also, it can be seen that for increased vacation-service rates, arrival processes do not have much influence on mean queue lengths.

4. The impatience rate ξ affects the queue lengths significantly, especially when $\rho = 0.1$. In Figures 4.10, 4.11 and 4.12, the change in queue lengths with the increase in impatient rate is shown. When the impatient rate is small, the mean queue length for Erlang model becomes minimum. As the impatient rate increases it shows the reverse behavior. The point of inflection depends upon the service rate μ_v. But the impatient rate does not have much effect on queue lengths when the arrival process is Erlang, whereas for hyperexponential model, the mean queue length decreases significantly with the increase in impatient rate.

Therefore, systems with hyperexponential arrivals have longest queues compared to corresponding Erlang or exponential arrivals. For light loaded systems ($\rho = 0.1$) and

highly impatient customers $\left(\frac{1}{\xi} < 10\right)$, hyperexponential arrivals give the minimum queue lengths. When the customer impatient rates are small, the system behavior depends on the vacation-service rates.

4.5.3 Effect on blocking probability

From the tables and the graphs plotted for blocking probability we have seen the following properties for the models under study.

1. Figure 4.13 gives that for a single-server system the blocking probability of a hyperexponential model is minimum and that for Erlang is maximum while $\theta = 1, \xi = 1$ and $\rho = 0.9$. This behavior is also shown in multiserver models when θ is too small. For higher values of θ multiserver models follow the reverse nature *i.e.,* Erlang model gives the mimimum queue length and hyperexponential gives the maximum of those three different arrival models. That is, the chance of blocking a customer with Poisson arrival in a single-server as well as in a multi-server queue is always sandwiched between those with Erlang and hyperexponential arrivals.

2. When we have a single-server Erlang model, the blocking probabilities seem to reduce up to 6% with an increased rate of service during vacation.

3. Figures 4.14 and 4.15 show that the hyperexponential model is not much effected by the vacation-service rate, whereas the Erlang model can reduce the blocking probability up to 4% for increased vacation-service rates.

A model with c=1 and Erlang-2 arrival process has the maximum chance to make a customer wait in queue compared to exponential and hyperexponential arrivals, but as the number of servers are increased hyperexponential arrival model has the highest blocking probability. The exponential arrival model always remains in between these two.

4.5.4 Average number of servers busy in non-vacation

The mean number of servers that is in working status during non-vacation period are shown in subsequent figures.

1. Figure 4.16 is a plot of blocking probabilities with changing θ. Here, for an increase in vacation duration, the number of servers that remain busy is high, because the servers serve at a low rate but for longer time, and a new arrival will be served by an idle server if any. Consequently, it increases the number of busy servers in the system.

2. Figure 4.17 shows that if the service rate is fast, customers are served at a faster rate which results in less number of busy servers in non-vacation period. This is true for all the three types of arrival models.

3. The impatience makes a customer leave the system unserved and for high impatience rates more servers remain idle (Figure 4.18). But if the impatient rate is increased beyond a certain value ($\xi > 6$) the mean number of busy servers remain unaffected.

4.5.5 Average customer loss

We plot the mean number of customers who abandon the queue without getting served in Figures 4.19 and 4.20. The values of θ for these plots are $\theta = 0.1$ and $\theta = 1$ respectively, keeping the other parameters fixed for both cases.

When $\theta = 0.1$, *i.e.,* the system has longer vacations, the number of lost customers is less compared to the corresponding model for $\theta = 1$. In both the cases, the hyperexponential models have the maximum customer loss, which is up to 60% more than the Erlang model. Also, the effect of impatient rates on customer loss is negligible for a system having small vacation durations.

Thus, we have seen the role of various parameters on system performances and we are now in a position to handle them to enhance the system efficiency.

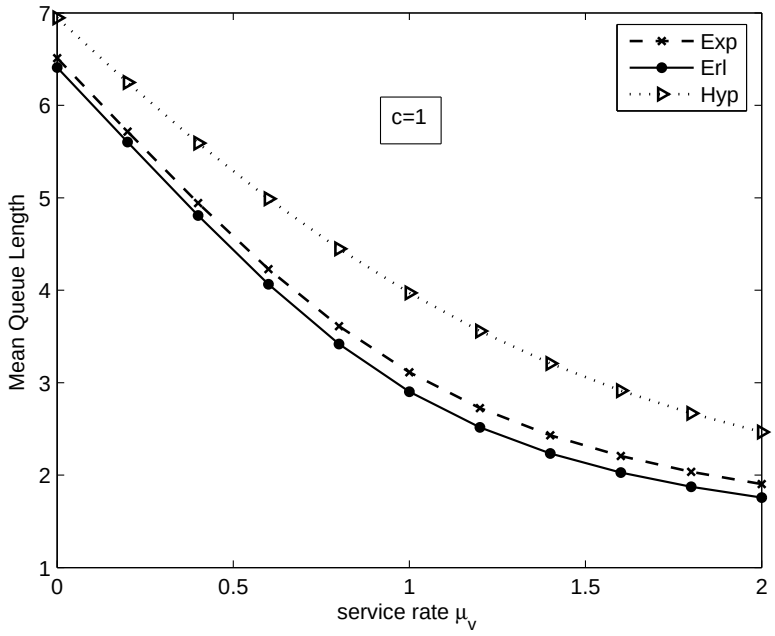

Figure 4.1: Mean queue length vs vacation-service rate with $\rho = 0.5$, $\xi = 0.1$, $\theta = 0.1$, $c = 1$.

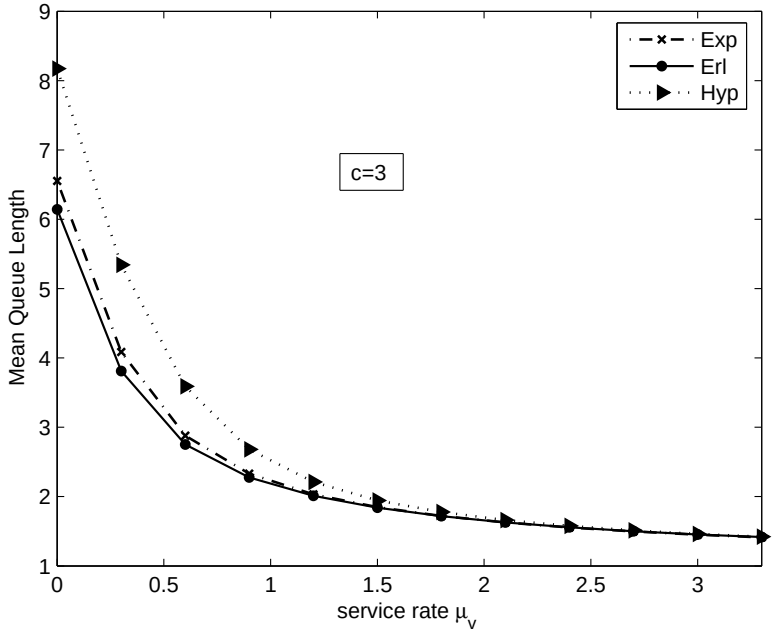

Figure 4.2: Mean queue length vs vacation-service rate with $\rho = 0.5$, $\xi = 0.1$, $\theta = 0.1$, $c = 3$.

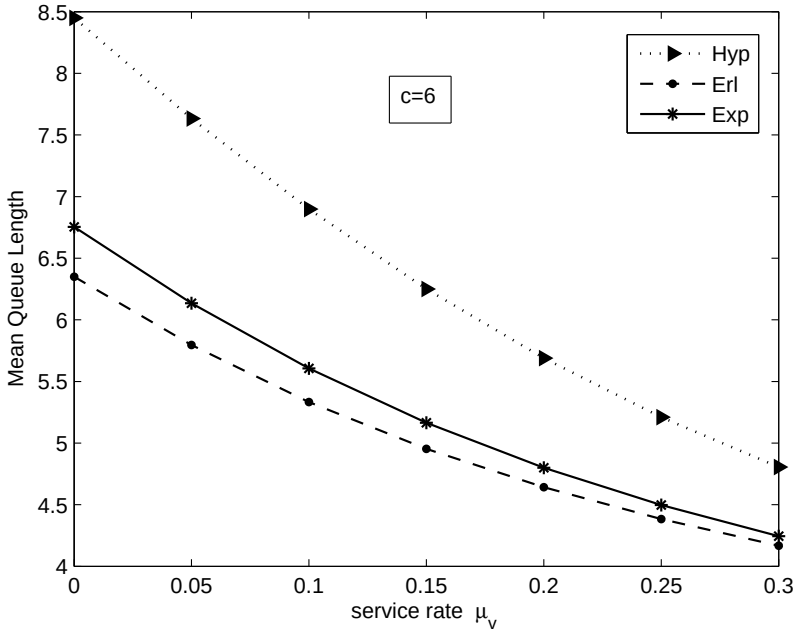

Figure 4.3: Mean queue length vs vacation-service rate with $\rho = 0.5$, $\xi = 0.1$, $\theta = 0.1$, $c = 6$.

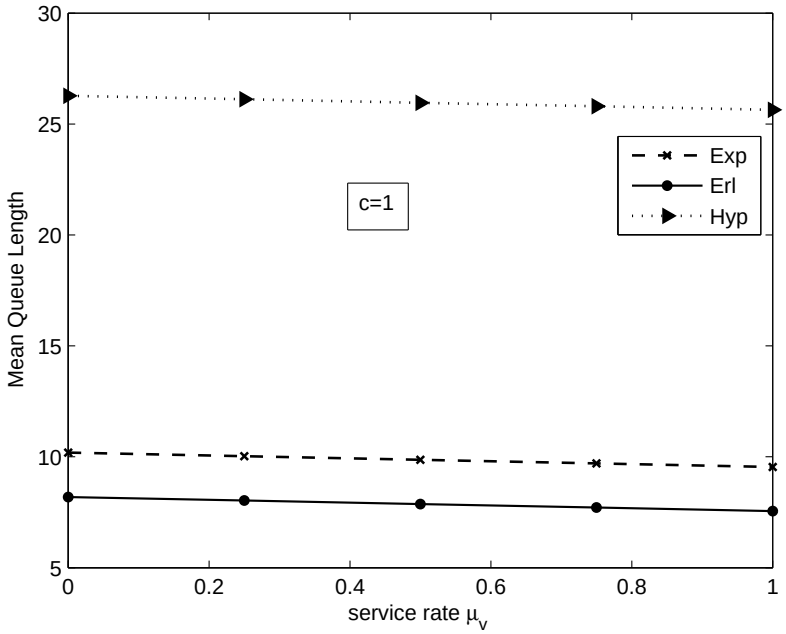

Figure 4.4: Mean queue length vs vacation-service rate with $\rho = 0.9$, $\xi = 1$, $\theta = 1$, $c = 1$.

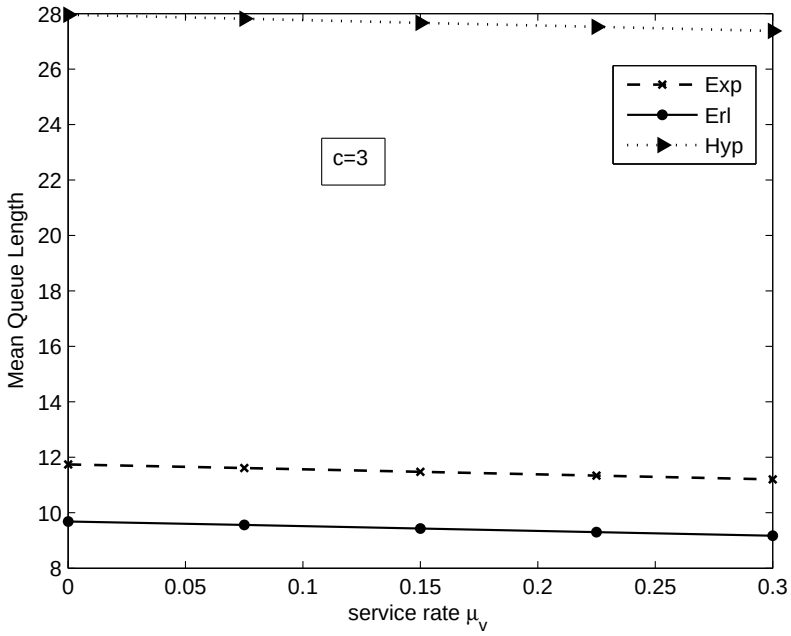

Figure 4.5: Mean queue length vs vacation-service rate with $\rho = 0.9$, $\xi = 1$, $\theta = 1$, $c = 3$.

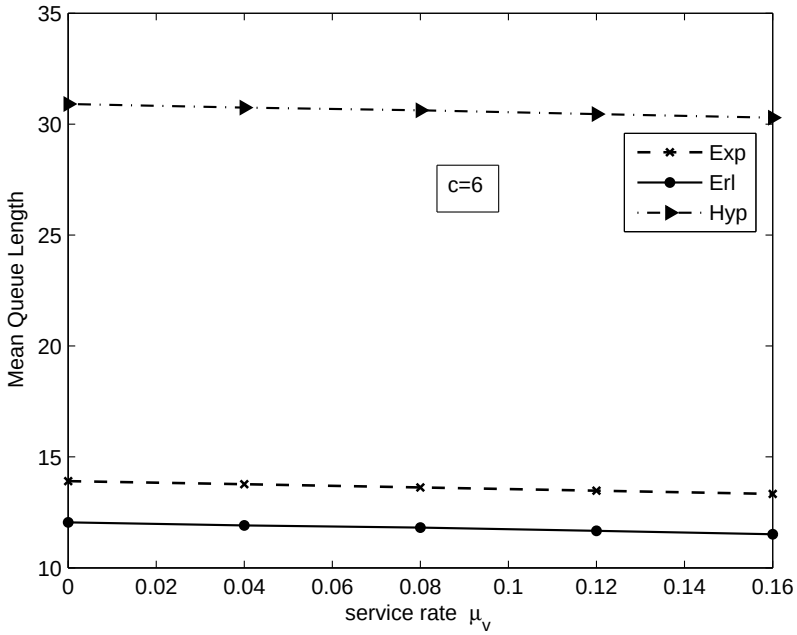

Figure 4.6: Mean queue length vs vacation-service rate with $\rho = 0.9$, $\xi = 1$, $\theta = 1$, $c = 6$.

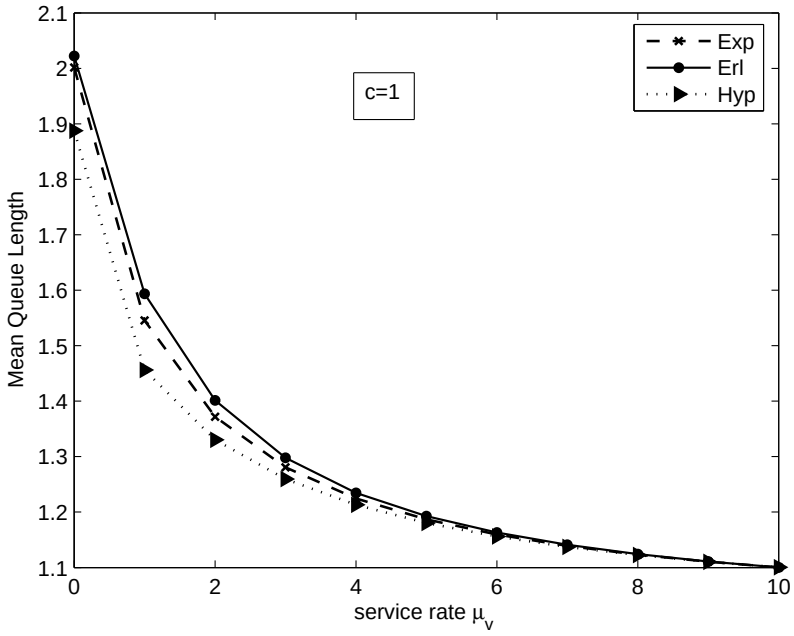

Figure 4.7: Mean queue length vs vacation-service rate with $\rho = 0.1$, $\xi = 10$, $\theta = 0.1$, $c = 1$.

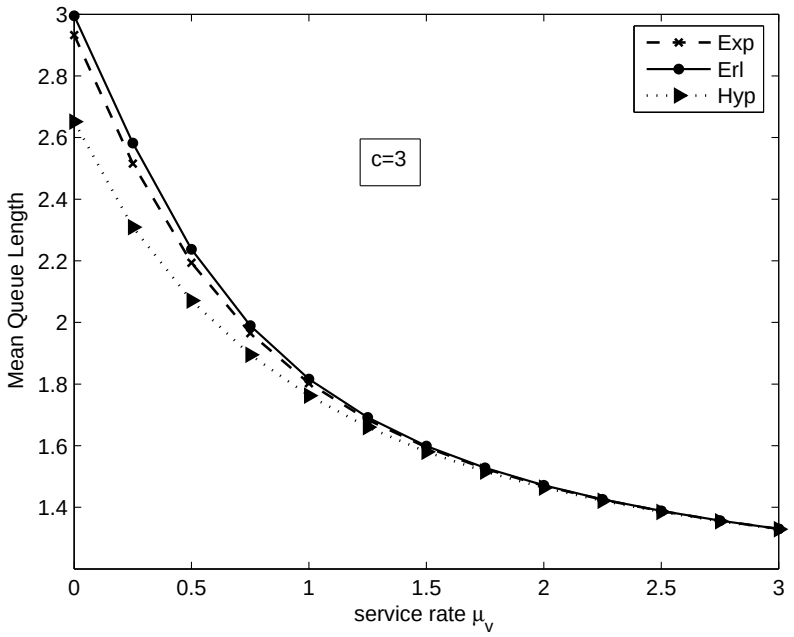

Figure 4.8: Mean queue length vs vacation-service rate with $\rho = 0.1$, $\xi = 10$, $\theta = 0.1$, $c = 3$.

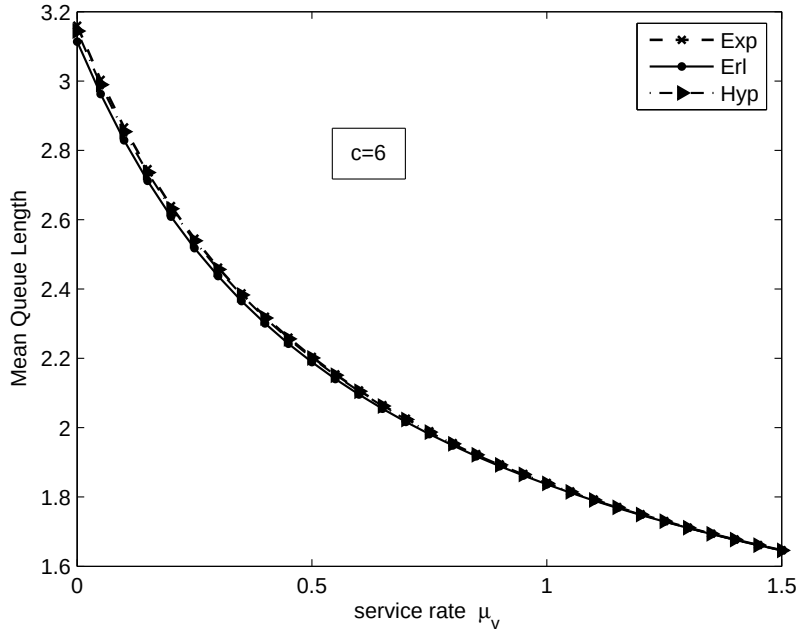

Figure 4.9: Mean queue length vs vacation-service rate with $\rho = 0.1$, $\xi = 10$, $\theta = 0.1$, $c = 6$.

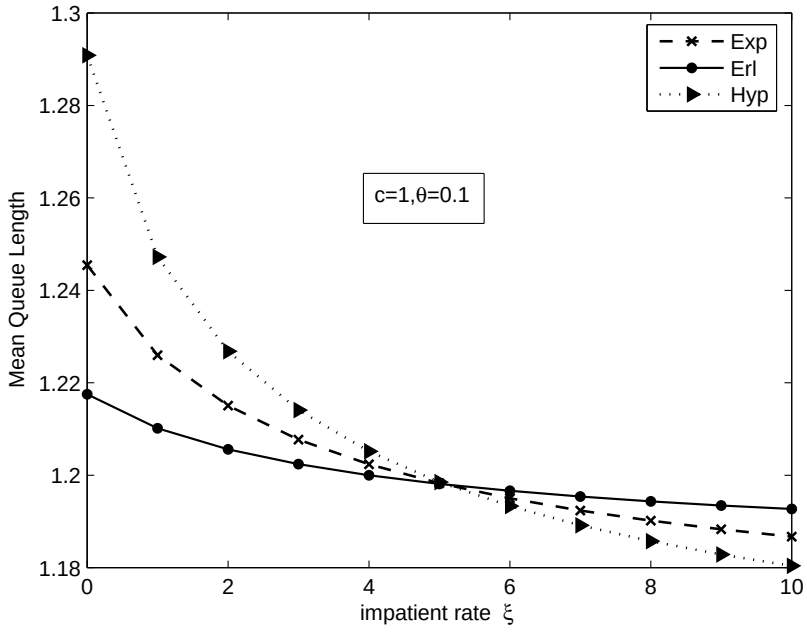

Figure 4.10: Mean queue length vs impatient rate with $\rho = 0.1$, $\mu_v = 0.5$, $c = 1$.

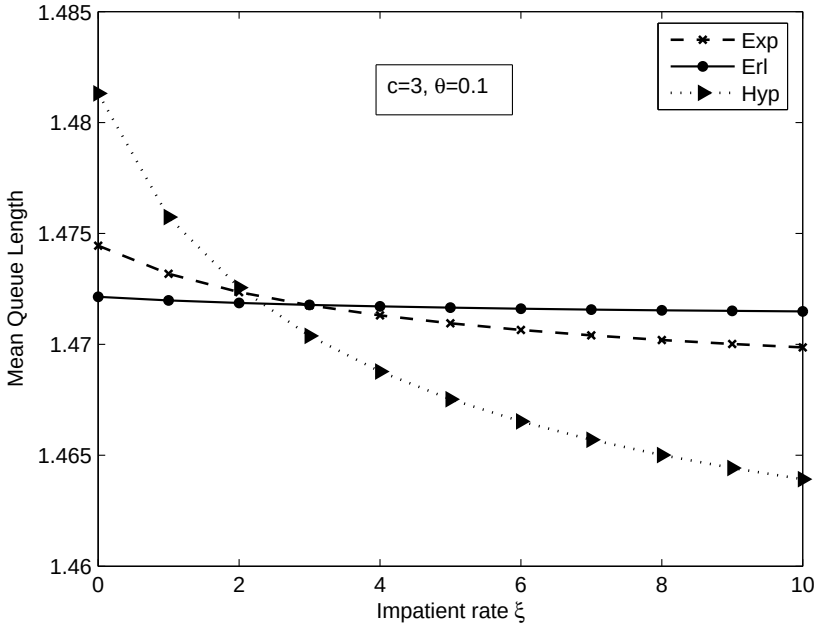

Figure 4.11: Mean queue length vs impatient rate with $\rho = 0.1$, $\mu_v = 0.2$, $c = 3$.

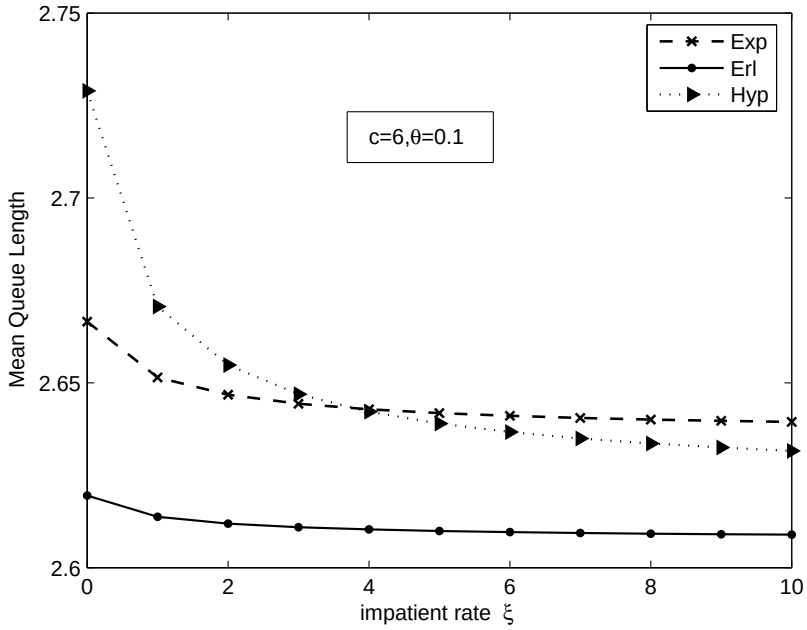

Figure 4.12: Mean queue length vs impatient rate with $\rho = 0.1$, $\mu_v = 0.2$, $c = 6$.

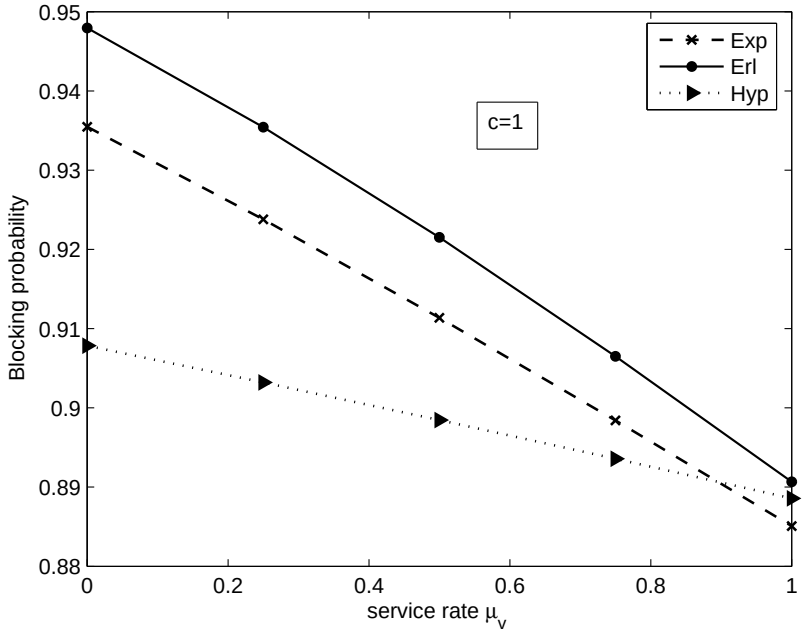

Figure 4.13: Blocking probability vs vacation-service rate with $\rho = 0.9$, $\theta = 1$, $\xi = 1$, $c = 1$.

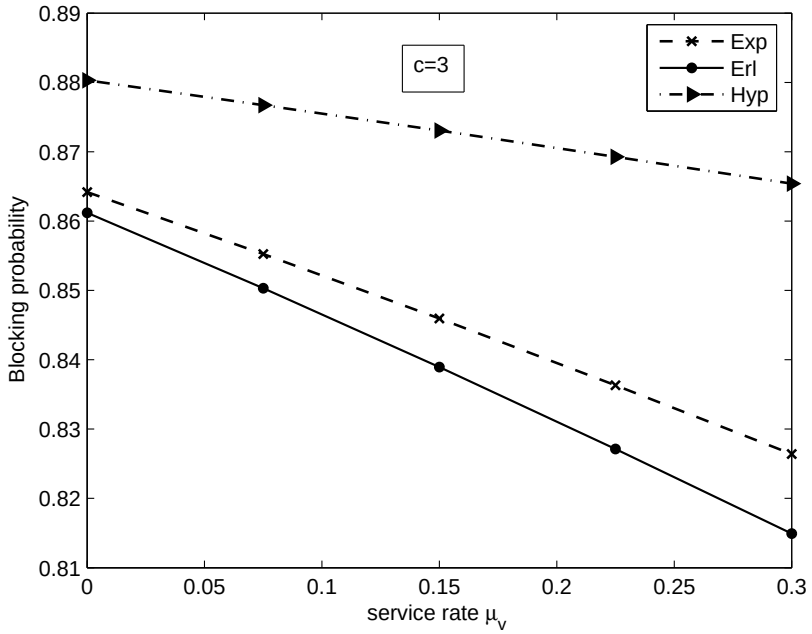

Figure 4.14: Blocking probability vs vacation-service with $\rho = 0.9$, $\theta = 1$, $\xi = 1$, $c = 3$.

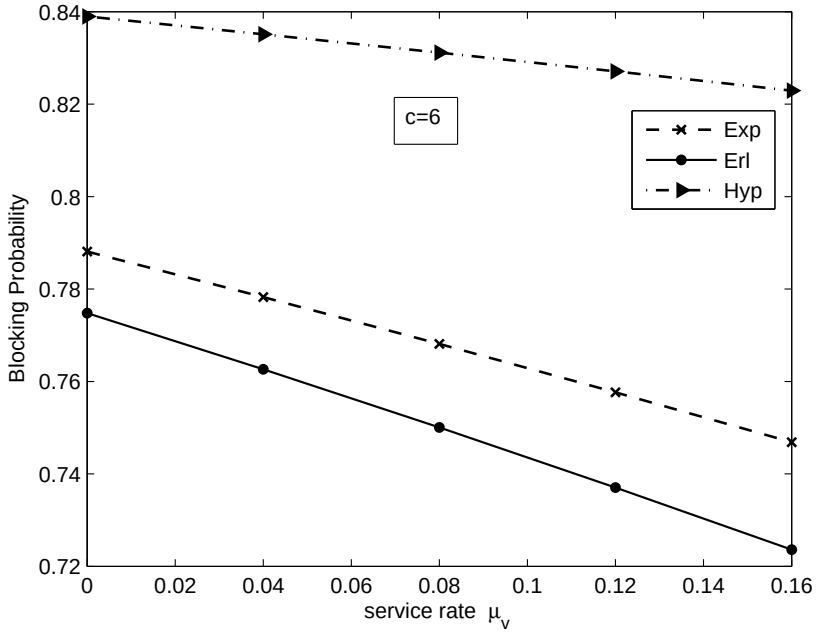

Figure 4.15: Blocking probability vs vacation-service with $\rho = 0.9$, $\theta = 1$, $\xi = 1$, $c = 6$.

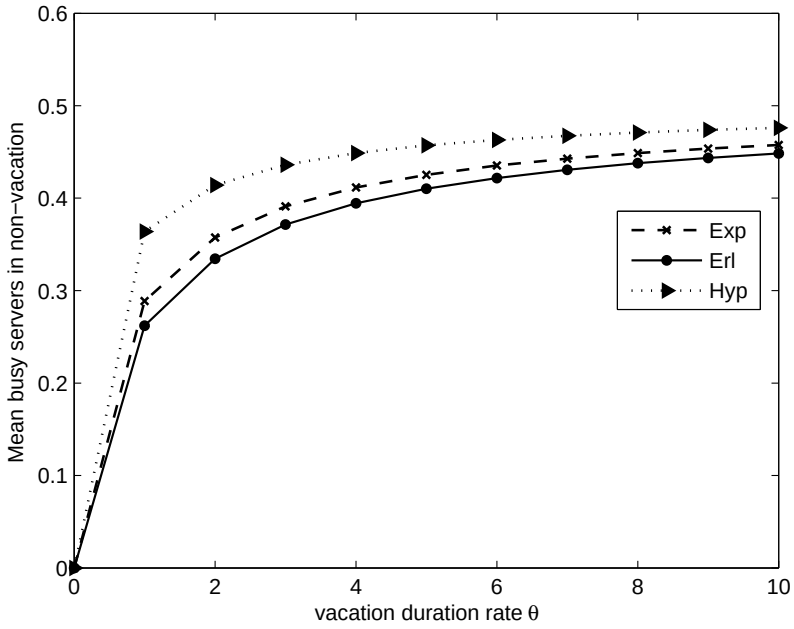

Figure 4.16: Mean number of busy servers vs vacation duration rate with $c = 1$, $\rho = 0.5$, $\xi = 0.1$, $\mu_b = 10$, $\mu_v = 0.4$.

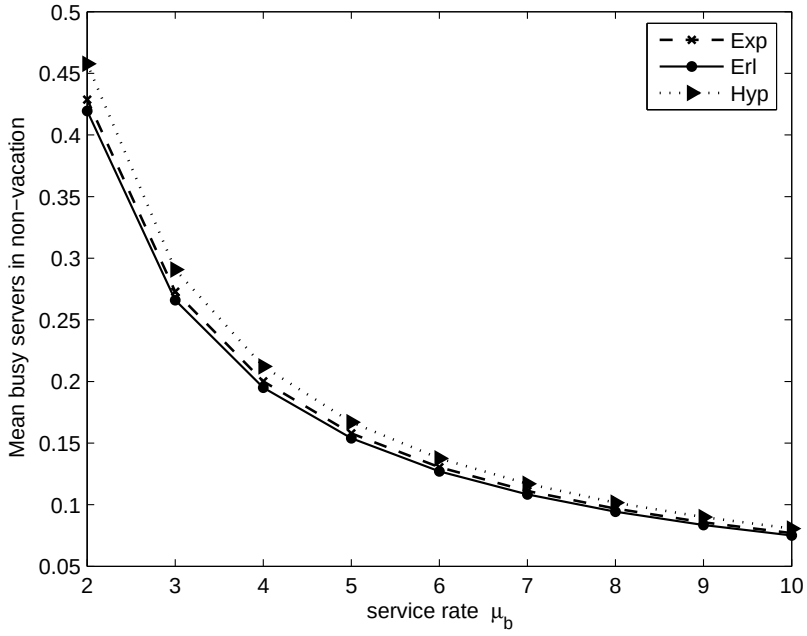

Figure 4.17: Mean number of busy servers vs service rate with $c = 1$, $\rho = 0.5$, $\xi = 0.1$, $\mu_v = 0.4$.

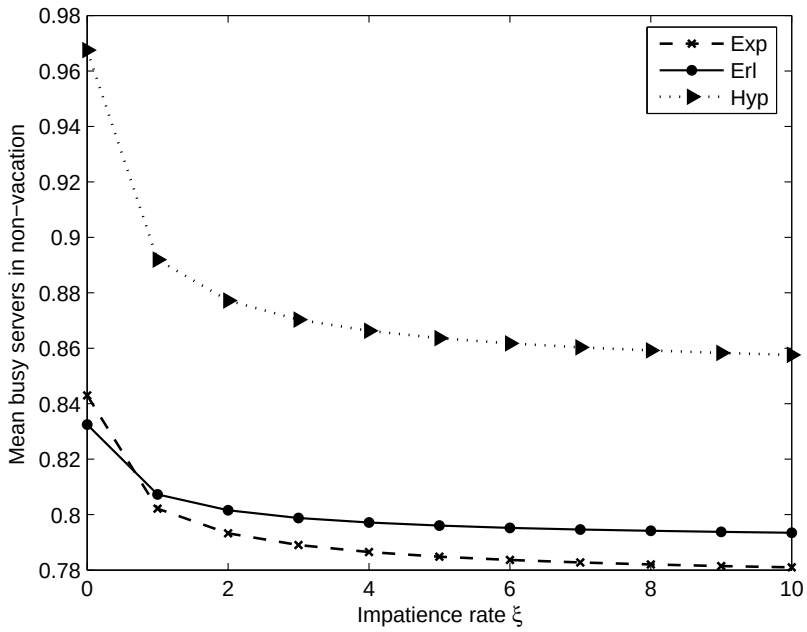

Figure 4.18: Mean number of busy servers vs impatient rate with $c = 3$, $\rho = 0.5$, $\mu_v = 0.4$, $\theta = 0.1$.

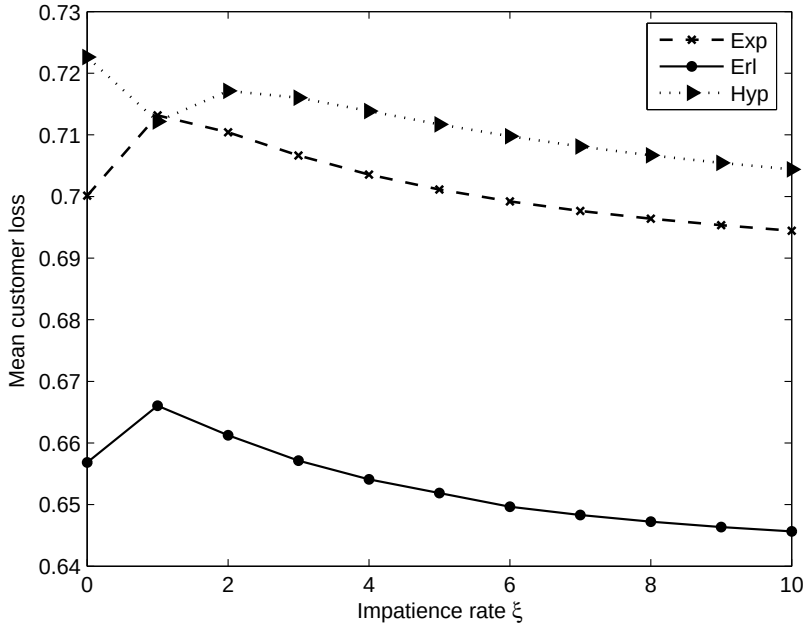

Figure 4.19: Mean customer loss vs impatient rate with $c = 3$, $\rho = 0.5$, $\theta = 0.1$, $\mu_v = 0.4$.

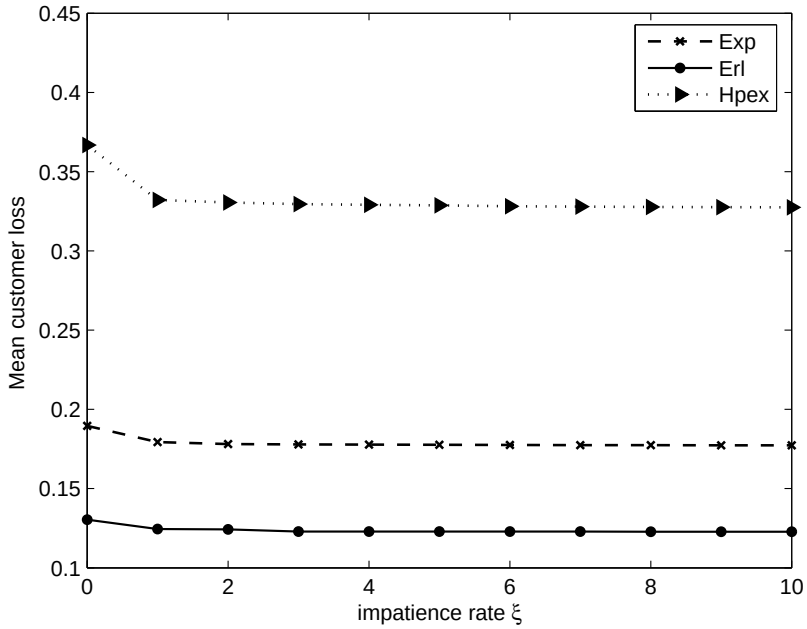

Figure 4.20: Mean customer loss vs impatient rate with $c = 3$, $\rho = 0.5$, $\mu_v = 0.4$, $\theta = 1$.

Table 4.1: Multi-server model with $c = 1$.

			K_f			L			B_p		
ρ	θ	μ_v	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*
0.1	0.1	0.0	25	28	36	6.4229	6.4127	6.3354	0.9354	0.9143	0.7955
		3.0	10	13	18	1.3951	1.4544	1.5802	0.3211	0.3173	0.3088
		6.0	7	9	13	1.1771	1.1963	1.2238	0.1652	0.1647	0.1639
		9.0	6	8	11	1.1149	1.1244	1.1358	0.1109	0.1108	0.1107
	1	0.0	14	16	22	2.0525	2.0533	2.0481	0.6062	0.5479	0.4325
		3.0	8	11	15	1.3040	1.3342	1.3793	0.2650	0.2562	0.2421
		6.0	6	8	12	1.1656	1.1811	1.2011	0.1561	0.1545	0.1524
		9.0	6	7	11	1.1137	1.1229	1.1337	0.1099	0.1097	0.1094
	100	0.0	5	6	10	1.1131	1.1215	1.1305	0.1097	0.1089	0.1082
		3.0	5	6	10	1.1099	1.1184	1.1273	0.1066	0.1061	0.1056
		6.0	5	6	10	1.1068	1.1153	1.1243	0.1037	0.1034	0.1031
		9.0	5	6	10	1.1040	1.1124	1.1214	0.1009	0.1008	0.1008
0.5	0.1	0.0	25	28	42	6.4099	6.5109	6.9481	0.9579	0.9412	0.8236
		0.5	21	25	41	4.4283	4.5761	5.2829	0.9062	0.8788	0.7457
		1.0	17	21	40	2.9028	3.1115	3.9708	0.7716	0.7451	0.6520
		1.5	15	20	38	2.1227	2.3112	3.0542	0.6121	0.5987	0.5579
	1	0.0	17	19	38	2.7191	2.8952	3.6542	0.7863	0.7442	0.6270
		0.5	16	21	38	2.3956	2.5839	3.3851	0.7113	0.6784	0.5949
		1.0	16	21	38	2.1386	2.3326	3.1440	0.6331	0.6121	0.5616
		1.5	15	20	38	1.9441	2.1339	2.9316	0.5605	0.5505	0.5280
	100	0.0	14	18	37	1.8203	2.0117	2.8092	0.5061	0.5049	0.5027
		0.5	14	20	37	1.8178	2.0080	2.8066	0.5046	0.5037	0.5021
		1.0	15	20	37	1.8146	2.0055	2.8041	0.5030	0.5025	0.5014
		1.5	15	20	37	1.8121	2.0030	2.8016	0.5015	0.5012	0.5007
0.9	0.1	0.0	74	99	246	11.2890	13.2254	28.5672	0.9893	0.9846	0.9509
		0.3	73	98	253	10.0678	11.9600	27.1965	0.9805	0.9735	0.9350
		0.6	71	97	260	8.7709	10.6206	25.7831	0.9593	0.9495	0.9139
		0.9	75	102	265	7.4716	9.2882	24.3478	0.9142	0.9057	0.8880
	1	0.0	73	96	231	8.7160	10.8465	27.2102	0.9574	0.9479	0.9187
		0.3	72	96	233	8.4580	10.5880	26.9545	0.9445	0.9361	0.9135
		0.6	72	96	234	8.2028	10.3341	26.7009	0.9289	0.9228	0.9080
		0.9	71	95	235	7.9573	10.0868	26.4490	0.9111	0.9080	0.9022
	100	0.0	81	110	265	7.8679	10.0268	26.8945	0.9013	0.9010	0.9004
		0.3	81	110	265	7.8651	10.0240	26.8905	0.9010	0.9007	0.9003
		0.6	81	110	265	7.8623	10.0213	26.8865	0.9006	0.9005	0.9002
		0.9	81	110	265	7.8596	10.0185	26.8825	0.9003	0.9002	0.9001

Table 4.2: Multiserver model with $c = 3$.

			K_f			L			B_p		
ρ	θ	μ_v	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*
0.1	0.1	0.0	22	25	33	4.1401	4.1552	4.1155	0.4953	0.4909	0.4452
		1.0	10	13	18	1.8312	1.8449	1.8849	0.0376	0.0606	0.1064
		2.0	8	10	13	1.4721	1.4743	1.4805	0.0050	0.0133	0.0294
		3.0	7	8	11	1.3301	1.3307	1.3326	0.0013	0.0049	0.0117
	1	0.0	9	12	16	1.6336	1.6535	1.6721	0.0203	0.0421	0.0730
		1.0	7	10	13	1.4723	1.4792	1.4875	0.0056	0.0163	0.0343
		2.0	7	9	12	1.3787	1.3811	1.3847	0.0022	0.0078	0.0181
		3.0	6	8	11	1.3165	1.3172	1.3189	0.0011	0.0044	0.0107
	100	0.0	6	7	11	1.3037	1.3051	1.3062	0.0009	0.0039	0.0096
		1.0	6	7	11	1.3026	1.3039	1.3049	0.0009	0.0039	0.0095
		2.0	6	7	11	1.3015	1.3027	1.3036	0.0009	0.0038	0.0093
		3.0	6	7	11	1.3005	1.3016	1.3023	0.0009	0.0037	0.0092
0.5	0.1	0.0	36	41	58	6.1463	6.5540	8.1834	0.7544	0.7535	0.7094
		0.2	26	30	48	4.4310	4.7592	6.1809	0.5820	0.5952	0.6062
		0.4	20	24	43	3.3472	3.5687	4.6449	0.3792	0.4116	0.4878
		0.6	16	20	39	2.7521	2.8807	3.5933	0.2308	0.2694	0.3736
	1	0.0	17	19	37	3.1041	3.2843	4.0155	0.3311	0.3663	0.4414
		0.2	16	19	37	2.9373	3.0968	3.8141	0.2832	0.3217	0.4116
		0.4	17	21	37	2.7875	2.9234	3.6269	0.2421	0.2821	0.3825
		0.6	16	21	37	2.6620	2.7814	3.4534	0.2074	0.2473	0.3544
	100	0.0	16	19	37	2.6304	2.7513	3.4302	0.1989	0.2385	0.3468
		0.2	16	21	37	2.6286	2.7437	3.4280	0.1984	0.2380	0.3463
		0.4	16	21	37	2.6267	2.7418	3.4257	0.1978	0.2375	0.3459
		0.6	16	21	37	2.6249	2.7399	3.4234	0.1973	0.2370	0.3455
0.9	0.1	0.0	78	100	224	14.0668	16.3843	33.6740	0.9618	0.9582	0.9403
		0.1	75	98	224	12.7466	14.9913	31.9614	0.9424	0.9390	0.9251
		0.2	73	95	229	11.3801	13.5544	30.1618	0.9089	0.9077	0.9052
		0.3	71	93	234	9.9871	12.0651	28.2970	0.8541	0.8593	0.8796
	1	0.0	76	101	239	9.7276	11.7783	28.3667	0.8612	0.8643	0.8806
		0.1	76	100	238	9.5526	11.6045	28.1630	0.8466	0.8523	0.8758
		0.2	75	100	237	9.3873	11.4217	27.9598	0.8312	0.8397	0.8708
		0.3	75	99	237	9.2075	11.2421	27.7421	0.8149	0.8265	0.8656
	100	0.0	82	111	269	9.1219	11.1422	29.3701	0.8041	0.8176	0.8622
		0.1	82	111	269	9.1194	11.1398	29.3614	0.8038	0.8175	0.8621
		0.2	82	111	269	9.1169	11.1375	29.3527	0.8036	0.8173	0.8621
		0.3	82	111	269	9.1144	11.1351	29.3440	0.8033	0.8171	0.8620

Table 4.3: Multiserver model with $c = 6$.

			K_f			L			B_p		
ρ	θ	μ_v	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*	*Erl*	*Exp*	*Hyp*
0.1	0.1	0.0	20	23	32	3.2118	3.2976	3.4094	0.0679	0.0887	0.1405
		0.5	11	13	18	2.1893	2.2065	2.2152	0.0008	0.0034	0.0159
		1.0	9	11	14	1.8365	1.8407	1.8401	0.0000	0.0003	0.0024
		1.5	8	9	12	1.6458	1.6464	1.6462	0.0000	0.0001	0.0005
	1	0.0	8	11	15	1.7766	1.7958	1.8184	0.0000	0.0005	0.0035
		0.5	8	10	14	1.7120	1.7219	1.7334	0.0000	0.0002	0.0016
		1.0	8	9	13	1.6587	1.6631	1.6680	0.0000	0.0001	0.0008
		1.5	8	9	12	1.6136	1.6145	1.6155	0.0000	0.0000	0.0004
	100	0.0	8	8	12	1.6022	1.6050	1.6039	0.0000	0.0000	0.0004
		0.5	8	8	12	1.6016	1.6041	1.6028	0.0000	0.0000	0.0004
		1.0	8	8	12	1.6009	1.6032	1.6016	0.0000	0.0000	0.0004
		1.5	8	8	12	1.6002	1.6024	1.6005	0.0000	0.0000	0.0004
0.5	0.1	0.0	30	34	52	6.3770	6.8052	8.4877	0.4072	0.4526	0.5456
		0.1	24	30	46	5.3485	5.6091	6.9287	0.2546	0.3073	0.4411
		0.2	21	25	41	4.6431	4.8028	5.7143	0.1445	0.1924	0.3376
		0.3	18	22	39	4.1697	4.2487	4.8192	0.0797	0.1170	0.2476
	1	0.0	18	20	39	4.5180	4.6593	5.1256	0.1208	0.1617	0.2889
		0.1	18	20	39	4.3618	4.4869	4.9560	0.1007	0.1399	0.2679
		0.2	19	23	38	4.2119	4.2803	4.8001	0.0837	0.1208	0.2477
		0.3	18	22	38	4.0862	4.1494	4.6449	0.0696	0.1042	0.2284
	100	0.0	17	20	38	4.0659	4.1500	4.6956	0.0662	0.1001	0.2233
		0.1	17	20	38	4.0633	4.1474	4.6923	0.0659	0.0998	0.2230
		0.2	18	23	38	4.0514	4.1099	4.6889	0.0657	0.0995	0.2226
		0.3	18	23	38	4.0489	4.1074	4.6856	0.0655	0.0993	0.2223
0.9	0.1	0.00	78	101	229	15.1572	17.3703	35.4516	0.9045	0.9054	0.9082
		0.05	76	99	227	14.2268	16.3514	34.0044	0.8719	0.8759	0.8903
		0.10	75	98	225	13.1999	15.2452	32.4807	0.8265	0.8360	0.8684
		0.15	73	96	222	12.1491	14.0835	30.9017	0.7663	0.7840	0.8418
	1	0.00	79	105	246	12.0928	13.9233	31.2097	0.7748	0.7881	0.8391
		0.05	79	104	245	11.9039	13.7614	30.9647	0.7595	0.7758	0.8343
		0.10	78	103	245	11.7512	13.5946	30.6805	0.7436	0.7629	0.8292
		0.15	78	103	244	11.5535	13.4025	30.4553	0.7270	0.7496	0.8241
	100	0.00	84	114	272	11.5312	13.3305	35.6288	0.7160	0.7407	0.8207
		0.05	84	114	272	11.5273	13.3271	35.6049	0.7157	0.7405	0.8206
		0.10	84	114	272	11.5235	13.3237	35.5810	0.7155	0.7403	0.8205
		0.15	84	113	272	11.5196	13.3446	35.5572	0.7152	0.7401	0.8204

Chapter 5

A Priority Queue with Vacation Interruptions

As discussed in the last chapter, multiple services are provided by an efficient network for better quality of service (QoS) support. However with a prominent increase in number of users and diversity of applications, users require different QoS, with different priority levels. The classes of jobs that need priority over others are kept separate to allocate different sets of wavelengths on priority basis. At a packet switched computer network over WDM technology, priority based wavelength assignment is a necessary task. It incorporates priority in the selection of light path for optical traffic management. A light path is a semi-permanent optical pipe connecting nodes through which packets are transmitted. Priority based channel assignment ensures the transmission of a high priority packet prior to a low priority one. Lack of priorities in wavelength assignment techniques can severely limit the viability of WDM networks as the next generation networks. For example, an online trading or video-conferential connection should normally be considered more important than an ordinary file transfer application [55]. Thus a study of priority queues needs to account for better understanding of a system like WDM which can only be modelled as a WV queue. Dutta and Chaubey [55] studied priority assignment to incoming call connection requests for optical WDM networks by modelling the system

as a M/M/K/K queue. This model does not consider systems having non-exponential service times. In this chapter, we analyze a priority based WDM network with a light path between nodes and a traffic with two priority classes. We model the system, where service times follow PH-distributions and WVs undergo Vacation Interruptions (VIs). VIs can enhance a system performance in a sense that when the system has waiting customers, it gets back to non-vacation period to give fast services and reduce the system congestion. Here the service times are taken as PH-distributed to study the role of different service time distributions in the system performance.

5.1 Model description

We consider a M/PH/1/WV priority system with two types of customers, denoted as class-1 and class-2. A class-i, $i = 1, 2$, customer arrives with Poisson arrival rate λ_i. Class-1 (or priority) customers have priority over the class-2 (or ordinary) customers. Customers are served with nonpreemptive discipline and customers of the same class are served according to the FCFS basis.

The service times for class-i customers, during a non-vacation period, are $PH(\beta_i, S_i)$ with rate μ_{bi}. After completing a service, if the server finds no customer to serve, it takes a WV. During a WV, class-i customers are served with $PH(\bar{\beta}_i, \bar{S}_i)$ and rate μ_{vi}, and according to their priority. Here we assume that $S_i\bar{S}_i = \bar{S}_iS_i$. The duration of a WV is Exp(θ). After serving a customer in vacation, if the server finds any customer waiting in queue, the vacation is interrupted and the server switches its service rate from μ_{vi} to μ_{bi} and starts a non-vacation period; otherwise, the server takes another WV. When a server starts a non-vacation period, it starts serving the class-1 customers first. In case, a vacation terminates in between an ongoing service, the server switches its service rate from μ_{vi} to μ_{bi} and the service is continued for that customer with rate μ_{bi} till it is completed. The interarrival times, service times and vacation duration times are all mutually independent.

We define

$$\lambda = \lambda_1 + \lambda_2 \quad \text{and} \quad \rho = \rho_1 + \rho_2, \tag{5.1}$$

$$\text{where } \rho_i = \frac{\lambda_i}{\mu_{bi}}, \quad i = 1, 2. \tag{5.2}$$

Let $B_i(x)$ be the distribution function of service times of a class-i customer in non-vacation period and $f_i(x)$ be the corresponding density function. Since service times during non-vacation follow PH(β_i, S_i), we get

$$B_i(x) = 1 - \beta_i \exp(S_i x)\mathbf{1} \tag{5.3}$$

$$f_i(x) = \beta_i \exp(S_i x) S_i^0. \tag{5.4}$$

If $\bar{B}_i(x)$ is the distribution of service times during WV period with its density function $\bar{f}_i(x)$, then

$$\bar{B}_i(x) = 1 - \bar{\beta}_i \exp(\bar{S}_i x)\mathbf{1}, \tag{5.5}$$

$$\bar{f}_i(x) = \bar{\beta} i \exp(\bar{S}_i x) \bar{S}_i{}^0. \tag{5.6}$$

For a class-i customer, $i = 1, 2$, the LST and the mean of the service time distribution during non-vacation period respectively are

$$B_i^*(u) = \int_0^\infty e^{-ux} dB_i(x) = \beta_i (uI - S_i)^{-1} S_i^0, \text{for } u \geq 0 \tag{5.7}$$

$$b_i = \int_0^\infty x dB_i(x) = -B_i^{*(1)}(0) = \beta_i S_i^{-2} S_i^0 = -\beta_i S_i^{-1}\mathbf{1}. \tag{5.8}$$

Using the identity

$$\frac{d^n}{du^n}(uI - S)^{-1} = (-1)^n n!(uI - S)^{-(n+1)},$$

which exists for any nonsingular matrix S, the nth raw moment, $n = 2, 3 \ldots$, of the service time distribution during non-vacation period can be derived as

$$b_i^{(n)} = \int_0^\infty x^n dB_i(x) = (-1)^n B_i^{*(n)}(0) = (-1)^{n+1} n! \beta_i S_i^{-(n+1)} S_i^0. \tag{5.9}$$

Let $a_i(k)$ be the probability of arrival of k class-1 customers during service of a class-i customer in non-vacation period. Then

$$a_i(k) = \int_0^\infty e^{-\lambda_1 x} \frac{(\lambda_1 x)^k}{k!} dB_i(x). \tag{5.10}$$

We obtain its generating function as

$$A_i(z) = \sum_{k=0}^{\infty} a_i(k) z^k \tag{5.11}$$

$$= \int_0^{\infty} e^{-\lambda_1 x} \sum_{k=0}^{\infty} \frac{(\lambda_1 x)^k}{k!} z^k dB_i(x) \tag{5.12}$$

$$= \beta_i \left[(\lambda_1 - \lambda_1 z)I - S_i\right]^{-1} S_i^0, \tag{5.13}$$

where

$$A_i(1) = -\beta_i S_i^{-1} S_i^0 = -\beta_i S_i^{-1}(-S_i \mathbf{1}) = 1, \quad i = 1, 2 \tag{5.14}$$

$$\text{and} \quad A_i^{(n)}(1) = \frac{d^n}{dz^n} A_i(z)|_{z=1} = \lambda_1^n (-1)^{n+1} n! \beta_i S_i^{-(n+1)} S_i^0 = \lambda_1^n b_i^{(n)}, \quad i = 1, 2 \tag{5.15}$$

Let the server be idle in a WV when a class-i customer arrives. For convenience, let us denote V as the residual vacation time and U as service time of that customer respectively. We can have two scenarios here. Either the duration of residual vacation, from the arrival epoch, is longer than the service time of that customer i.e. $V > U$, or the residual vacation is not longer than the service time in WV period, $V \leq U$. In the first case, after a service completion during WV, vacation interruption happens and the server comes back to the non-vacation period. In the case $V \leq U$, when the vacation terminates during an ongoing service with rate μ_{vi}, the service rate is switched from μ_{vi} to μ_{bi} at the vacation termination epoch and a non-vacation period starts.

For the case $V > U$, let $c_i(k)$, $i = 1, 2$, be the probability of arrival of k customers of class-1 during the service of a class-i customer in WV period. This is given by

$$c_i(k) = \int_0^{\infty} e^{-\theta x} e^{-\lambda_1 x} \frac{(\lambda_1 x)^k}{k!} d\bar{B}_i(x)$$

$$= \int_0^{\infty} e^{-(\theta+\lambda_1)x} \frac{(\lambda_1 x)^k}{k!} \bar{\beta}_i \exp(\bar{S_i} x) \bar{S}_i^0 dx.$$

$$= \bar{\beta}_i \int_0^{\infty} \frac{(\lambda_1 x)^k}{k!} \exp\left\{-\left((\theta + \lambda_1)I - \bar{S}_i\right) x\right\} \bar{S}_i^0 dx,$$

with its generating function given by

$$C_i(z) = \sum_0^{\infty} c_i(k) z^k = \bar{\beta}_i \left[(\theta + \lambda_1 - \lambda_1 z)I - \bar{S}_i\right]^{-1} \bar{S}_i^0. \tag{5.16}$$

For the case $V \leq U$, the probability of arrival of k customers of class-1 during the switched service is

$$\begin{aligned} d_i(k) &= \int_{t=0}^{\infty}\int_{x=0}^{t} \theta e^{-\theta x} e^{-\lambda_1 x}\frac{(\lambda_1 x)^k}{k!}\bar{\beta}_i \exp\left\{\bar{S}_i x + S_i(t-x)\right\} S_i^0 dx dt \\ &= \theta\bar{\beta}_i \int_{t=0}^{\infty} e^{-\lambda_1 t}\frac{(\lambda_1 t)^k}{k!}\left[\int_{x=0}^{t}\exp\left\{-(\theta I - \bar{S}_i + S_i)x\right\}dx\right]\exp(S_i t)S_i^0 dt \\ &= \theta\bar{\beta}_i\left[\theta I - \bar{S}_i + S_i\right]^{-1}\int_{t=0}^{\infty}\frac{(\lambda_1 t)^k}{k!}\Big[\exp\left\{-(\lambda_1 I - S_i)t\right\} \\ &\qquad - \exp\left\{-((\lambda_1+\theta)I - \bar{S}_i)t\right\}\Big]S_i^0 dt. \quad (5.17) \end{aligned}$$

with its generating function

$$D_i(z) = \theta\bar{\beta}_i\left[\theta I - \bar{S}_i + S_i\right]^{-1}\left[\{(\lambda_1 - \lambda_1 z)I - S_i\}^{-1} - \{(\theta + \lambda_1 - \lambda_1 z)I - \bar{S}_i\}^{-1}\right]S_i^0. \quad (5.18)$$

The nth derivatives of $C_i(z)$ and $D_i(z)$, at $z = 1$, are

$$\begin{aligned} C_i^{(n)}(1) &= \lambda^n n!\bar{\beta}_i\left(\theta I - \bar{S}_i\right)^{-(n+1)}\bar{S}_i^0 = \lambda^n \bar{b_i}^{(n)}, \\ D_i^{(n)}(1) &= \theta\lambda^n n!(-1)^{n+1}\bar{\beta}_i\left(\theta I - \bar{S}_i + S_i\right)^{-1}\left\{S_i^{-(n+1)} + (\theta I - \bar{S}_i)^{-(n+1)}\right\}S_i^0 = \lambda^n \hat{b_i}^{(n)}, \end{aligned}$$

where

$$\bar{b_i}^{(n)} = n!\bar{\beta}_i\left(\theta I - \bar{S}_i\right)^{-(n+1)}\bar{S}_i^0, \quad (5.19)$$

$$\hat{b_i}^{(n)} = \theta n!(-1)^{n+1}\bar{\beta}_i\left(\theta I - \bar{S}_i + S_i\right)^{-1}\left\{S_i^{-(n+1)} + (\theta I - \bar{S}_i)^{-(n+1)}\right\}S_i^0. \quad (5.20)$$

5.2 Length of busy period

Let Θ be the length of the busy period generated by a customer in the system during non-vacation with LST $\Theta^*(s)$. If K is the number of customers arriving during the service time of the first customer in non-vacation period, then we can write the busy period in a recursive way as, the sum of the first customer's service time and all the busy periods generated by the customers that arrive during the first service time. Thus, we have

$$\Theta = X + \Theta^{(1)} + \Theta^{(2)} + \ldots + \Theta^{(K)}, \quad (5.21)$$

where $\{\Theta^{(k)}; k = 1, 2, \ldots, K\}$ are mutually independent random variables with the same distribution as Θ. We get

$$\begin{aligned} E[e^{-s\Theta} \mid X = x, K = k] &= E[e^{-s(x+\Theta^{(1)}+\Theta^{(2)}+\ldots+\Theta^{(k)})}] \\ &= e^{-sx} E[e^{-s\Theta^{(1)}}] \ldots E[e^{-s\Theta^{(k)}}] \\ &= e^{-sx} [\Theta^*(s)]^k. \end{aligned} \tag{5.22}$$

First unconditioning on K, we find

$$\begin{aligned} E[e^{-s\Theta} \mid X = x] &= \sum_{k=0}^{\infty} E[e^{-s\Theta} \mid X = x, K = k] \, P[K = k \mid X = x] \\ &= \sum_{k=0}^{\infty} e^{-sx} [\Theta^*(s)]^k e^{-\lambda x} \frac{(\lambda x)^k}{k!} \\ &= e^{-(s+\lambda)x} \sum_{k=0}^{\infty} \frac{[\lambda \Theta^*(s)]^k}{k!} x^k \\ &= e^{-[s+\lambda-\lambda\Theta^*(s)]x}. \end{aligned} \tag{5.23}$$

Unconditioning again on X, the service time during non-vacation period, we get

$$\begin{aligned} \Theta^*(s) = E[e^{-s\Theta}] &= \int_0^{\infty} E[e^{-s\Theta} \mid X = x] \, f_X(x) dx \\ &= \int_0^{\infty} e^{-[s+\lambda-\lambda\Theta^*(s)]x} dB(x) \\ &= B^*(s + \lambda - \lambda\Theta^*(s)), \end{aligned} \tag{5.24}$$

where

$$B(x) = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} [1 - \beta_i \exp(S_i x)\mathbf{1}] \quad \text{and} \quad B^*(s) = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \beta_i (sI - S_i)^{-1} S_i^0. \tag{5.25}$$

The mean length of the busy period will be

$$E(\Theta) = -\Theta^{*(1)}(0) = \frac{\rho}{(1-\rho)\lambda}. \tag{5.26}$$

5.3 Busy cycle

The busy cycle consists of an idle period and a busy period of the server. In our system, the server can be idle only during WV period and it can be busy (or active) during WV as

well as during non-vacation period. A busy period is initiated by an arrival of a customer during WV and after getting served (completely or partially) at a lower service rate, the server switches to the non-vacation period. Therefore the busy cycle can be seen as a busy period initiated by an exceptional service for the first customer.

Theorem 5.3.1. *The LST of the busy cycle of M/PH/1/WV model with exceptional service time for the first customer is given by*

$$\Theta_c^*(s) = B_0^*(s + \lambda - \lambda\Theta^*(s)), \tag{5.27}$$

with mean busy period

$$E(\Theta_c) = \frac{1}{1-\rho} b_0, \tag{5.28}$$

where $B_0(x)$ is the service distribution of the first customer with mean b_0 given by

$$\begin{aligned} B_0(x) &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \Bigg[\left(\theta I - \bar{S}_i\right)^{-1} \left\{1 - e^{-(\theta I - \bar{S}_i)x}\right\} \bar{S}_i^0 \\ &\quad + \theta \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left\{ S_i^{-1} \left(e^{S_i x} - 1\right) - \left(\theta I - \bar{S}_i\right)^{-1} \left(1 - e^{-(\theta I - \bar{S}_i)x}\right) \right\} S_i^0 \Bigg], \end{aligned}$$

and

$$b_0 = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left[\left(\theta I - \bar{S}_i\right)^{-2} \bar{S}_i^0 + \theta \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left\{ S_i^{-2} + \left(\theta I - \bar{S}_i\right)^{-2} \right\} S_i^0 \right].$$

Proof. Let X_0 be the actual service time of the first customer who finds the system idle upon arrival. As discussed in Section 5.1, either the service of the first customer ends before the vacation terminates; or if the vacation terminates before the service completion, the service rate is switched from lower to higher and a non-vacation period starts. For the case $V > U$, the distribution of service time is given by

$$\begin{aligned} P(X_0 \le x, V > U) &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \int_0^x e^{-\theta t} \bar{\beta}_i \exp(\bar{S_i} t) \bar{S}_i^0 dt \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \int_0^x e^{-(\theta I - \bar{S}_i)x} dt \bar{S}_i^0 \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left(\theta I - \bar{S}_i\right)^{-1} \left[1 - e^{-(\theta I - \bar{S}_i)x}\right] \bar{S}_i^0. \end{aligned} \tag{5.29}$$

But if $V \leq U$, the distribution service times of the switched service is

$$\begin{aligned} P(X_0 \leq x, V \leq U) &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \int_{t=0}^{x} \int_{t_1=0}^{t} \theta e^{-\theta t_1} \bar{\beta}_i \exp\left\{\bar{S}_i t_1 + S_i(t-t_1)\right\} S_i^0 dt_1 dt \\ &= \theta \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \int_{t=0}^{x} \bar{\beta}_i \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left\{1 - e^{-(\theta I - \bar{S}_i + S_i)t}\right\} e^{S_i t} S_i^0 dt \\ &= \theta \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left[S_i^{-1}\left(e^{S_i x} - 1\right) - \left(\theta I - \bar{S}_i\right)^{-1} \right. \\ &\qquad \left. \times \left\{1 - e^{-(\theta I - \bar{S}_i)x}\right\}\right] S_i^0. \quad (5.30) \end{aligned}$$

Therefore, the distribution of service times X_0 is

$$\begin{aligned} B_0(x) &= P(X_0 \leq x, V > U) + P(X_0 \leq x, V \leq U) \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left[\left(\theta I - \bar{S}_i\right)^{-1} \left\{1 - e^{-(\theta I - \bar{S}_i)x}\right\} \bar{S}_i^0 \right. \\ &\quad \left. + \theta \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left\{S_i^{-1}\left(e^{S_i x} - 1\right) - \left(\theta I - \bar{S}_i\right)^{-1} \left(1 - e^{-(\theta I - \bar{S}_i)x}\right)\right\} S_i^0 \right]. \\ & \quad (5.31) \end{aligned}$$

This is a distribution function as (i)$0 \leq B_0(x) \leq 1$, for $-\infty < x < \infty$, (ii) B_0 is a non-decreasing function, (iii) $B_0(x+) = B_0(x)$ and (iv)$\lim_{x \to -\infty} B_0(x) = 0$ and $\lim_{x \to +\infty} B_0(x) = 1$, because we have

$$\begin{aligned} \lim_{x \to +\infty} B_0(x) &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left[\left(\theta I - \bar{S}_i\right)^{-1} \bar{S}_i^0 - \theta \left(\theta I - \bar{S}_i + S_i\right)^{-1} \left\{S_i^{-1} + \left(\theta I - \bar{S}_i\right)^{-1}\right\} S_i^0 \right] \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \bar{\beta}_i \left[\left(\theta I - \bar{S}_i\right)^{-1} \bar{S}_i^0 - \theta \left\{\left(\theta I - \bar{S}_i\right)\left(\theta I - \bar{S}_i + S_i\right)\right\}^{-1} \right. \\ &\qquad \left. \times \left(\theta I - \bar{S}_i + S_i\right) S_i^{-1} S_i^0 \right]. \quad (5.32) \end{aligned}$$

For $S_i \bar{S}_i = \bar{S}_i S_i$, we get

$$\left(\theta I - \bar{S}_i\right)\left(\theta I - \bar{S}_i + S_i\right) = \left(\theta I - \bar{S}_i + S_i\right)\left(\theta I - \bar{S}_i\right).$$

Therefore, equation (5.32) reduces to

$$\begin{aligned}\lim_{x\to+\infty} B_0(x) &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\bar{\beta}_i\Big[(\theta I - \bar{S}_i)^{-1}\bar{S}_i^0 - \theta(\theta I - \bar{S}_i)^{-1}S_i^{-1}S_i^0\Big] \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\bar{\beta}_i(\theta I - \bar{S}_i)^{-1}\left(\bar{S}_i^0 + \theta\mathbf{1}\right) \\ &= \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\bar{\beta}_i(\theta I - \bar{S}_i)^{-1}\left(-\bar{S}_i + \theta I\right)\mathbf{1} = 1. \end{aligned} \tag{5.33}$$

The mean and the moments of this distribution of service times are

$$b_0 = -B_0^{*(1)}(0) = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\left(\bar{b}_i^{(1)} + \hat{b}_i^{(1)}\right) \tag{5.34}$$

$$\text{and} \quad b_0^{(n)} = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\left(\bar{b}_i^{(n)} + \hat{b}_i^{(n)}\right), \quad \text{for } n = 1, 2, \ldots. \tag{5.35}$$

Let Θ_c be the length of a busy period with exceptional service time X_0 for the first customer then, using equation (5.24), we have

$$\Theta_c^*(s) = B_0^*(s + \lambda - \lambda\Theta^*(s)) \tag{5.36}$$

with mean length of a busy period

$$\begin{aligned} E(\Theta_c) &= -\frac{d}{ds}B_0^*(s + \lambda - \lambda\Theta^*(s))|_{s=0} \\ &= -\left\{1 - \lambda\Theta^{*(1)}(0)\right\}B_0^{*(1)}(0) \\ &= \frac{1}{1-\rho}b_0. \end{aligned}$$

□

The idle period is a time interval of a system being empty. Since the arrival process is Poisson with rate λ, the interarrival time is exponential with mean $1/\lambda$. Thus the mean length of the idle period is

$$E(I) = \frac{1}{\lambda}. \tag{5.37}$$

The system state alternates between idle period and busy period. Therefore, a customer arrives during the idle period with probability

$$P_0 = \frac{E(I)}{E(\Theta_c) + E(I)} = \frac{\lambda b_0}{1 + \lambda b_0 - \rho}. \tag{5.38}$$

and its waiting time is zero. A customer arrives during the busy period with probability

$$1 - P_0 = \frac{E(\Theta_c)}{E(\Theta_c) + E(I)} = \frac{1-\rho}{1+\lambda b_0 - \rho}. \tag{5.39}$$

5.4 Queue length

Let X_n be the number of class-1 customers at the vacation starting epoch or at a service starting epoch in non-vacation period. Then $\{X_n, n = 1, 2, \ldots\}$ is a Markov chain. Here the service starting epochs during WV periods are not Markov points. Also, we need not keep track of the phases of service. Let the stationary distribution of this Markov chain be

$$\pi_k = \lim_{n\to\infty} P(X_n = k), \quad k = 0, 1, \ldots, \quad \text{with } \sum_{k=0}^{\infty} \pi_k = 1. \tag{5.40}$$

Also,

$$\pi_k = \sum_{j=0}^{\infty} p_{jk}, \quad \text{where } p_{jk} = \lim_{n\to\infty} P(X_{n+1} = k \mid X_n = j). \tag{5.41}$$

Here, π_0 is the probability of the system having no class-1 customers. The PGF of queue length of class-1 customers is given in the theorem below.

Theorem 5.4.1. *The PGF of queue length of class-1 customers at the service starting epochs is given by*

$$N_p(z) = P_0 \{C_1(z) + D_1(z)\} + \frac{\lambda}{\lambda_1} A_1(z) \left[P_0 \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \{C_i(z) + D_i(z)\} + \{\pi_0 - P_0\} A_2(z) - \pi_0 \right], \tag{5.42}$$

where

$$\begin{aligned}
A_i(z) &= \beta_i \left[(\lambda_1 - \lambda_1 z)I - S_i\right]^{-1} S_i^0, \\
C_i(z) &= \bar{\beta}_i \left[(\theta + \lambda_1 - \lambda_1 z)I - \bar{S}_i\right]^{-1} \bar{S}_i^0, \\
D_i(z) &= \theta\bar{\beta}_i \left[\theta I - \bar{S}_i + S_i\right]^{-1} \left[\{(\lambda_1 - \lambda_1 z)I - S_i\}^{-1} - \{(\theta + \lambda_1 - \lambda_1 z)I - \bar{S}_i\}^{-1}\right] S_i^0 \\
\textit{and } \pi_0 &= 1 - \frac{\lambda_1}{\lambda}(1 - P_0).
\end{aligned}$$

Proof. The transition probabilities of the Markov Chain $\{X_n, n = 1, 2, \ldots\}$, with the stationary distribution $\{\pi_k, k = 0, 1, \ldots\}$, can be as follows:

1. For $X_n = j,\ j \geq 1$ and $X_{n+1} = k$, we have that the system has j customers of class-1 in non-vacation period gives that during the service of a class-1 customer, $k - j + 1$ customers of class-1 arrive. Thus,

$$p_{jk} = a_1(k - j + 1). \tag{5.43}$$

2. For $X_n = 0$ with class-2 customers in the system and $X_{n+1} = k \geq 0$, we have that during the service of a class-2 customer in non-vacation, k class-1 customers arrive. Thus,

$$p_{0k} = a_2(k) \qquad \text{with probability } \pi_0 - P_0. \tag{5.44}$$

3. If $X_n = 0$ and the system is idle, the first customer will be a class-i customer with probability λ_i/λ and during its service k class-1 customers arrive. So, we have

$$p_{0k} = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \left[c_i(k) + d_i(k)\right] \qquad \text{with probability } P_0. \tag{5.45}$$

The generating function of number of class-1 customers in the system is

$$\begin{aligned} N(z) &= \sum_{k=0}^{\infty} \pi_k z^k \\ &= \sum_{k=0}^{\infty} (\pi_0 p_{0k} + \sum_{j=1}^{\infty} \pi_j p_{jk}) z^k \\ &= P_0 \sum_{k=0}^{\infty} \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \left\{c_i(k) + d_i(k)\right\} z^k + (\pi_0 - P_0) \sum_{k=0}^{\infty} a_2(k) z^k + \sum_{k=0}^{\infty} \sum_{j=1}^{k+1} \pi_j a_1(k - j + 1) z^k \\ &= \frac{N(z) - \pi_0}{z} A_1(z) + (\pi_0 - P_0) A_2(z) + P_0 \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \left\{C_i(z) + D_i(z)\right\}, \end{aligned} \tag{5.46}$$

which gives,

$$N(z) = \frac{1}{z - A_1(z)} \left[P_0 \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \left\{C_i(z) + D_i(z)\right\} + (\pi_0 - P_0) A_2(z) - \pi_0 A_1(z) \right] \tag{5.47}$$

From the condition $N(1) = 1$ and using L'Hospital's rule, we have

$$\begin{aligned}\pi_0 &= \frac{(1-\rho_1) + \lambda_1 b_2 P_0 - \lambda_1 P_0 \left[\sum_{i=1}^{2} \frac{\lambda_i}{\lambda}\left(\bar{b_i} + \hat{b_i}\right)\right]}{1 - \rho_1 + \lambda_1 b_2} \\ &= 1 - \frac{\lambda_1 b_2 (1-P_0) + \lambda_1 b_0 P_0}{1 - \rho_1 + \lambda_1 b_2}. \end{aligned} \tag{5.48}$$

Using the relations

$$b_2 = \frac{\rho_2}{\lambda_2} = \frac{\rho - \rho_1}{\lambda - \lambda_1} \quad \text{and} \quad \lambda b_0 P_0 = (1-\rho)(1-P_0), \tag{5.49}$$

in equation (5.48), we obtain

$$\pi_0 = 1 - \frac{\lambda_1}{\lambda}(1 - P_0). \tag{5.50}$$

Let $N_p(z)$ be the PGF of number of class-1 customers in the system just before the service of a class-1 customer begins. An arbitrary service starting epoch is an instant of starting service of a class-1 customer, with probability $\frac{\lambda_1}{\lambda}$, and the PGF of the number of class-1 customers, seen only at the service starting epoch of a class-1 customer, in the Markov chain $\{X_n, n = 1, 2, \ldots\}$ is given by the R.H.S of (5.46). Thus we get,

$$\begin{aligned} N_p(z) &= \left(\frac{\lambda_1}{\lambda}\right)^{-1} \left[\frac{N(z) - \pi_0}{z} A_1(z) + \frac{\lambda_1}{\lambda} P_0 \left\{C_1(z) + D_1(z)\right\}\right] \\ &= P_0 \left\{C_1(z) + D_1(z)\right\} + \frac{\lambda}{\lambda_1} A_1(z) \left[P_0 \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} \left\{C_i(z) + D_i(z)\right\} + \left\{\pi_0 - P_0\right\} A_2(z) - \pi_0\right] \end{aligned} \tag{5.51}$$

The mean number of class-1 customers present in the system at the service starting epoch of a class-1 customer is

$$\begin{aligned} E(N_p) &= \frac{d}{dz} N_p(z)|_{z=1} \\ &= \frac{\lambda P_0}{2(1-\rho_1)^2}\left[(b_0 - b_2)\left\{\lambda_1^2 b_1^{(2)} + 2\rho_1(1-\rho_1)\right\} + \lambda_1(1-\rho_1)\left(b_0^{(2)} - b_2^{(2)}\right)\right] \\ &\quad + \frac{\lambda \pi_0}{2(1-\rho_1)^2}\left[(1-\rho_1)\left(2\rho_1 b_2 + \lambda_1 b_2^{(2)}\right) + \lambda_1^2 b_2 b_2^{(2)}\right] + \lambda_1 b_0 P_0 \end{aligned} \tag{5.52}$$

□

5.5 Waiting time

The class-1 customers are served according to the FCFS order in the system. An arriving customer initiates a busy period with probability P_0 and its waiting time in queue is zero. A customer arrives during a busy period with probability $1 - P_0$ and we get the LST of the distribution function for the waiting time in queue of a class-1 customer, W_{q1}, from the following relation.

$$N_p(z) = W^*_{q1}(\lambda_1(1-z)|\text{busy})B^*_1(\lambda_1 - z\lambda_1)$$

which gives,

$$W^*_{q1}(s|\text{busy}) = \frac{N_p\left(1-\frac{s}{\lambda_1}\right)}{B^*_1(s)}. \tag{5.53}$$

and $W_{q1}(s|\text{idle}) = 1$. The unconditioned waiting time is given by

$$\begin{aligned} W^*_{q1}(s) &= P_0 W^*_{q1}(s|\text{idle}) + (1-P_0)W^*_{q1}(s|\text{busy}) \\ &= P_0 + (1-P_0)\beta_1 \left(sI - S_1\right)^{-1} S_1^0 N_p\left(1-\frac{s}{\lambda_1}\right). \end{aligned} \tag{5.54}$$

The mean waiting time of a class-1 customer

$$E(W_{q1}) = -\frac{d}{ds}W^*_{q1}(0) = \frac{(1-P_0)}{\lambda_1}E(N_p). \tag{5.55}$$

The waiting time of a class-2 customer in queue, W_{q2}, is equal to W_{q1} plus sum of service times of customers of class-1 that arrive during the delay period generated by W_{q1}. Therefore, we have

$$W^*_{q2}(s) = W^*_{q1}(s + \lambda_1 - \lambda_1\Theta^*_1(s)), \tag{5.56}$$

where $\Theta^*_1(s)$ is the LST of the length of busy period generated by customers of class-1 and is given by

$$\Theta^*_1(s) = B^*_1(s + \lambda_1 - \lambda_1\Theta^*_1(s)).$$

Let us denote $\sigma_1(s) = s + \lambda_1 - \lambda_1\Theta^*_1(s)$. Then

$$\begin{aligned} W^*_{q2}(s) &= W^*_{q1}(\sigma_1(s)) \\ &= P_0 + (1-P_0)\beta_1 \left\{\sigma_1(s)I - S_1\right\}^{-1} S_1^0 N_p\left(1-\frac{\sigma_1(s)}{\lambda_1}\right) \end{aligned} \tag{5.57}$$

and

$$\sigma_1^{(1)}(0) = 1 - \lambda_1 \Theta_1^{*(1)}(0) = \frac{1}{1-\rho_1}. \tag{5.58}$$

The mean waiting time of a class-2 customer is

$$\begin{aligned} E(W_{q2}) &= -\frac{d}{ds} W_{q2}^*(\sigma_1(s))|_{s=0} \\ &= -\sigma_1^{(1)}(0) W_{q2}^{*(1)}(0) \\ &= \frac{(1-P_0)}{\lambda_1(1-\rho_1)} E(N_p). \end{aligned} \tag{5.59}$$

The generating function of the number of class-2 customers in a queue can be derived from the LST of waiting time distribution,

$$\begin{aligned} N_0(z) &= W_{q2}^*(\lambda_1 - \lambda_1 z) B_2^*(\lambda_1 - \lambda_1 z) \\ &= \left[P_0 + (1-P_0)\beta_1 \left\{ \sigma_1(\lambda_1 - \lambda_1 z) I - S_1 \right\}^{-1} S_1^0 N_p \left(1 - \frac{\sigma_1(\lambda_1 - \lambda_1 z)}{\lambda_1} \right) \right] \\ &\quad \times \beta_2 \left\{ (\lambda_1 - \lambda_1 z) I - S_2 \right\}^{-1} S_2^0. \end{aligned} \tag{5.60}$$

and the mean number of class-2 customers will be

$$\begin{aligned} E(N_0) &= \frac{d}{dz} \left[W_{q2}^*(\lambda_1 - \lambda_1 z) B_2^*(\lambda_1 - \lambda_1 z) \right] |_{z=1} \\ &= -\lambda_1 \left[W_{q2}^{*(1)}(0) B_2^*(0) + W_{q2}^*(0) B_2^{*(1)}(0) \right] \\ &= \lambda_1 \left[E(W_{q2}) + b_2 \right]. \end{aligned} \tag{5.61}$$

5.6 Response time

The customer response time is defined as the time interval from the arrival time of an arbitrary customer to the time when it leaves the system after service completion. The mean response time is said to be the single most important performance measure for systems without blocking. Let W be the customer response time or waiting time in system with LST $W^*(s)$. Since the waiting time of a customer in queue and its service time are independent, we have waiting time of the customer in the system as

$$W^*(s) = W_q^*(s) B^*(s). \tag{5.62}$$

In our model, the server remains idle with probability P_0, where waiting time of a customer is zero. The LST $W_i^*(s)$, $i = 1, 2$, of the distribution function for the response time W_i of a customer of class-i is

$$W_i^*(s) = \frac{\lambda_i}{\lambda} P_0 \left[C_i \left(1 - \frac{s}{\lambda_1} \right) + D_i \left(1 - \frac{s}{\lambda_1} \right) \right] + (1 - P_0) W_{qi}^*(s) B_i^*(s) \quad (5.63)$$

The mean response time for a class-i customer, $i = 1, 2$, is

$$\begin{aligned} E(W_i) &= -\frac{d}{ds} W_i^*(s)|_{s=0} \\ &= \frac{\lambda_i}{\lambda} P_0 \left(\bar{b}_i^{(1)} + \hat{b}_i^{(1)} \right) + (1 - P_0) \left[W_{qi}^{*(1)}(0) B_i^*(0) + W_{qi}^*(0) B_i^{*(1)}(0) \right] \\ &= \frac{\lambda_i}{\lambda} P_0 \left(\bar{b}_i^{(1)} + \hat{b}_i^{(1)} \right) + (1 - P_0) \left[E(W_{qi}) - b_i \right]. \end{aligned} \quad (5.64)$$

5.7 Numerical examples

In a queueing model with priority classes, important measures of system performance are the queue lengths of the different classes of customers. First, we assume the service times for both classes to be Exp(μ_v) during a WV and Exp(μ_b) during non-vacation. Later, we take different service distributions and observe their effect on system performance. The service distributions taken are given here in PH-representation.

1. Exponential (Exp)

$$S_i = -1 \quad S_i^0 = 1 \quad \text{and} \quad \beta_i = 1, \qquad \text{for } i = 1, 2.$$

2. Erlang-2 (Erl)

$$S_i = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix} \quad S_i^0 = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \quad \text{and} \quad \beta_i = [1 \;\; 0], \qquad \text{for } i = 1, 2.$$

3. Hyperexponential-2 (Hyp)

$$S_i = \begin{pmatrix} -1.9 & 0 \\ 0 & -0.19 \end{pmatrix} \quad S_i^0 = \begin{pmatrix} 1.9 \\ 0.19 \end{pmatrix} \quad \text{and} \quad \beta_i = [0.9 \;\; 0.1], \qquad \text{for } i = 1, 2.$$

These distributions have mean $\mu_{bi} = 1,\ i = 1, 2$. We have taken $\bar{S}_i = aS_i, \bar{S}_i^0 = aS_i^0$ and $\bar{\beta}_i = \beta_i$, where a is a positive scalar. We assume $a = 0.3$ and the arrival rates $\lambda_1 = 0.85$ and $\lambda_2 = 0.06$. First, we assume the service times are exponential and study the role of system parameters on system efficiency. Later, we compare the performance measures of the system for different service time distributions.

In Figure 5.1, we plot the queue length of class-i, $i = 1, 2$, customers against the traffic load for various values of vacation duration rate θ. The service time distribution is taken to be exponential. For a particular value of θ, we see that $E(N_0)$ monotonously increases, as λ grows, because of queueing of customers and the effects from class-1 customers. The waiting times also show similar behavior (Figure 5.2). We get, for any vacation duration rate

$$E(N_p) < E(N_0) \text{ and } E(W_{q1}) < E(W_{q2}). \tag{5.65}$$

This can also be verified from the equations (5.52) and (5.61). But, from the graph we can quantify the amount by which the queue lengths of both types of customers differ. Comparison of queue lengths of class-1 customers to that of class-2 customers shows that for high traffic intensity (> 0.7), increase in queue length of class-2 customers is much faster than class-1 customers. As the traffic load increases the difference between the queue lengths of the two classes increases up to 35%. But for small vacations ($\theta > 0.1$), this difference is about 20%. Therefore, a system having small vacations can significantly reduce the queue lengths for both types of customers, when the traffic load is high. However, systems having light traffic loads ($\rho < 0.3$) are not much affected by the vacation duration rates.

Now, we will concentrate on the effect of parameters on only class-1 customers. The impact of vacation-service rates of class-1 customers on their mean waiting times in queue is seen in Figure 5.3. As expected, $E(W_{q1})$ monotonically decreases with increased service rate during WV. When the vacation duration rate is small *i.e.,* when we have longer vacations, the effect of μ_v is more significant. Systems having small vacations ($\theta \geq 10$) are not much affected by the increased vacation service rate. It is because, if we have

longer vacation, the vacation gets interrupted after completing one service during the vacation period and a non-vacation period begins. This graph also shows the span of queue lengths between a classical vacation model ($\mu_{v1} = 0$) and a non-vacation model ($\mu_{v1} = 1 = \mu_{b1}$). Therefore, the waiting times of class-1 customers is reduced up to 40% by having WV model instead of a classical vacation one. Consequently, the waiting times of class-2 customers will also be reduced by having systems with WVs in place of a classical vacations.

In Figure 5.4, the response times of class-1 customers are plotted and compared for increasing intensity of traffic. The response time of a customer is the total of its waiting time in queue and its service time. This response time shows a drastic increase when the system is heavily loaded. For system with long vacations ($\theta = 0$) the response time increases up to 90%, whereas for system with small vacations ($\theta = 10$) it increases up to 70% as the offered system load is heavy ($> 60\%$). So for a heavy loaded system, small vacation durations can reduce the wait of a customer in the system by 20%.

The system performance is affected by the service time distribution. In Figure 5.5, we compare the waiting times of class-1 customers for three different service distributions– hyperexponential, Erlang-2 and exponential, for $\theta = 1$. Hyperexponential service time distribution seems to raise the waiting times of class-1 customers by a significant degree. When system load is heavy ($\rho > 0.7$) a hyperexponential service model enhanced the waiting time of a class-1 customer up to 55% compared to a system having exponential service time distribution. The Erlang-2 service model always gives the minimum waiting time between the three. A system with Erlang-2 service can cut down the waiting time in a system with exponential service by 10%, when the system load is heavy (> 0.8). However, for $\rho < 0.4$, exponential service time distributions and Erlang-2 service time distributions give the same waiting times. In this case, the waiting times in hyperexponential service time distribution model is higher by 20%. So, we may conclude that, for any amount of system load, we have

$$E(W_{q1})_{Erl} < E(W_{q1})_{Exp} < E(W_{q1})_{Hyp}.$$

We can conclude from this study that in a priority system with WV the waiting times of customers, which is an important performance measure of a priority system, are effected by the vacation durations and vacation-service rates. By choosing a proper service time distribution, the waiting times of customers can be reduced.

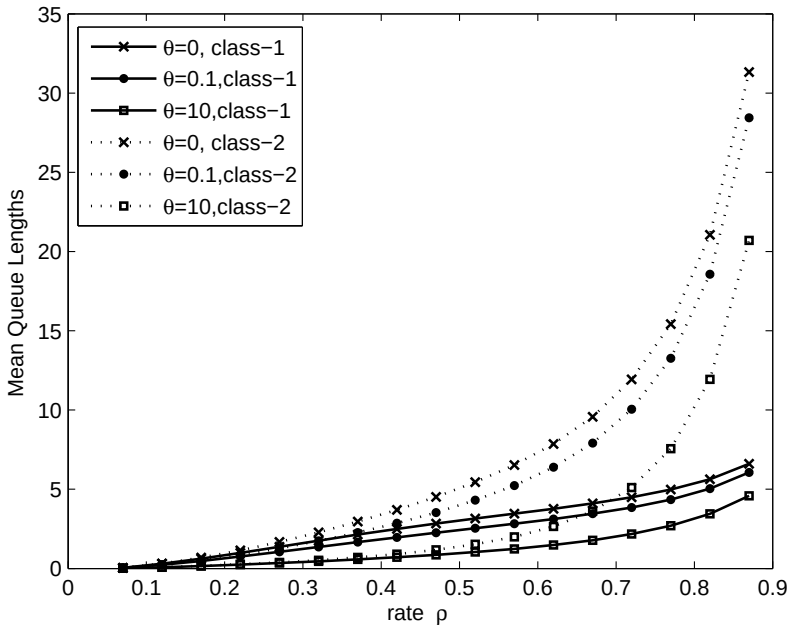

Figure 5.1: Mean queue length vs traffic intensity.

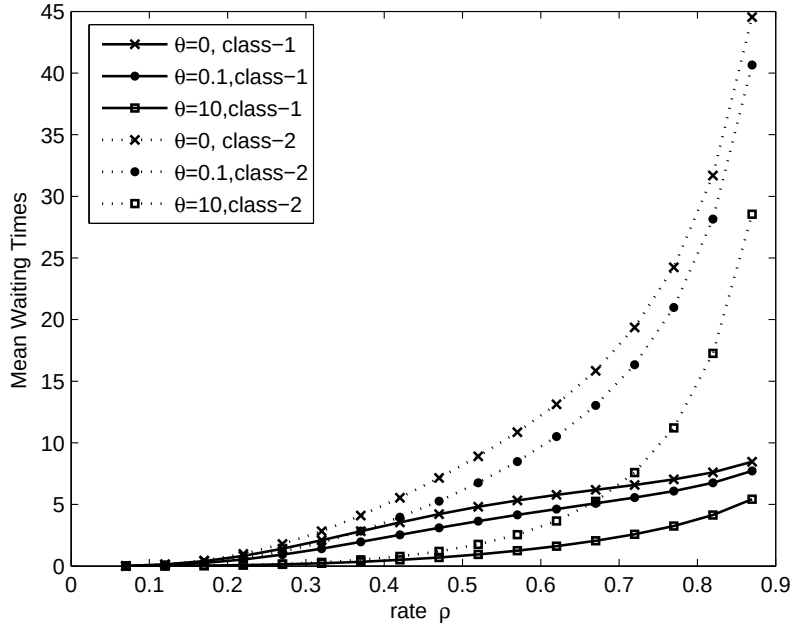

Figure 5.2: Mean waiting times vs traffic intensity.

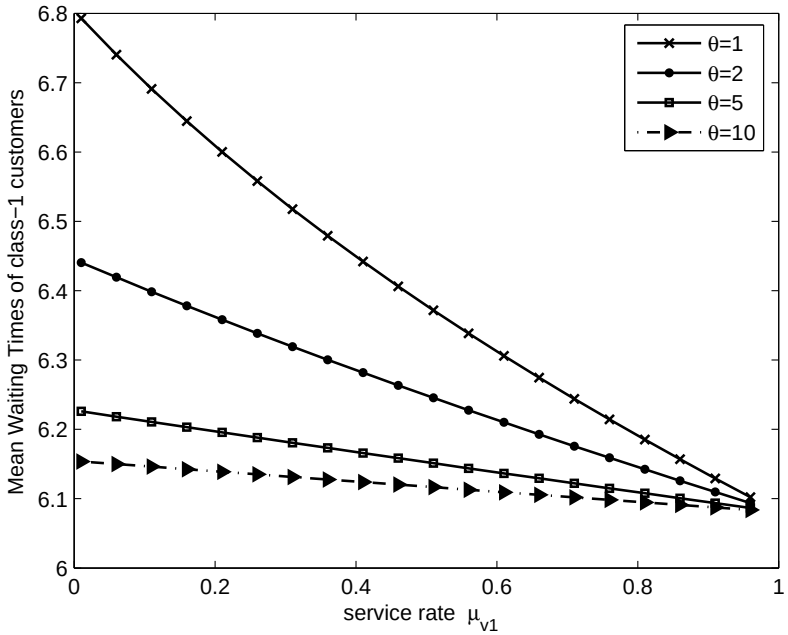

Figure 5.3: Mean waiting times of class-1 customers vs vacation-service rate.

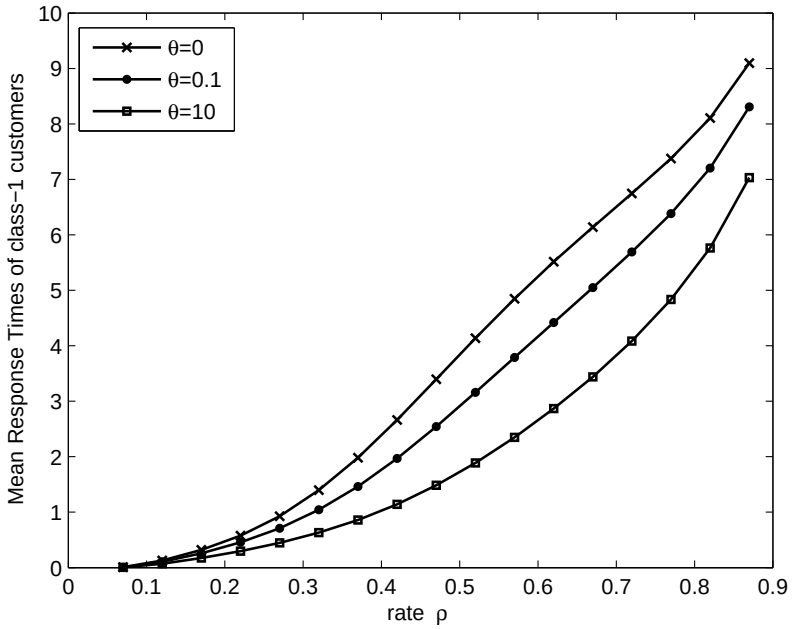

Figure 5.4: Mean response time of class-1 customers vs traffic intensity.

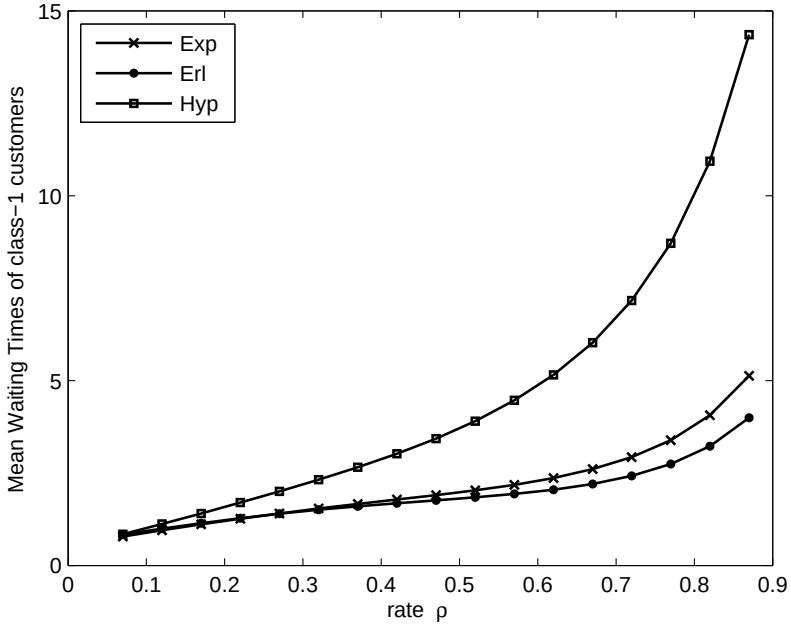

Figure 5.5: Mean waiting times of class-1 customers vs traffic intensity.

Chapter 6

A Retrial Queue

Priority assignment of jobs in WDM networks is required when the jobs are of varying importance and wavelengths have to be assigned according to their priority, as discussed in the previous chapter. Another important property, seen in WDM networks, is the retrial or repeated attempts of jobs for service, where the jobs have to be kept in a buffer space (a miniature conduit contained within the optical fiber) until they receive the requested service. Such situations mostly arise in mobile ad hoc networks (MANET). MANET is a self-configuring network of mobile devices in which mobile subscribers are connected to a base station by wireless links. When a call request comes to a base station, it assigns the mobile subscriber a link to the destination. Due to traffic congestion, if no links are available, the call either keeps retrying till a link is allocated successfully or it balks the system. To dynamically shift available capacity towards the actual traffic concentrations, radio in the fiber arrangements is employed in the access networks of Universal Mobile Telecommunication Systems (UMTS) and Mobile Broadband Systems (MBS). Both UMTS and MBS are benefited, when WDM is employed to add a further level of network management flexibility by allowing several mobile service providers to share the same access infrastructure operated by a single network operator. That is, to dynamically shift available capacity towards the actual traffic concentrations, suitable arrangements are employed in the access networks of UMTS and MBS with the help of

WDM technology. Hence it becomes crucial to study the retrial effect on WDM networks. This chapter models such a WDM network having retrial phenomena and analyzes the system behavior.

In literature, Ke and Chang [88] have considered a M/G/1 retrial queue with modified vacation policy where the server takes a classical vacation when the orbit becomes empty. At a vacation completion epoch, if the orbit is nonempty, the server waits for customers from the orbit or for new customers to arrive. If no customer appears in the orbit, the server leaves for another vacation. The server takes a maximum of J vacations. If the orbit is empty by the end of the Jth vacation, the server remains idle for customers in the system. Do [51] studied a M/M/1 retrial queue with WVs, which was motivated by the performance analysis of a Media Access Control (MAC) function in wireless systems, and derived closed-form expressions for steady-state probabilities. We consider a M/G/1 queue with WV and retrials, where the interretrial times follow general distribution.

6.1 Model description

We consider a single-server queueing system in which new customers arrive from outside the system according to a Poisson process with rate λ. Our server can be in one of the four states: idle during a WV period, idle during a non-vacation period, busy during a WV period and busy during a non-vacation period. We assume that there is no waiting space and therefore if an arriving customer finds the server idle, either during a WV period or during a non-vacation period, the customer obtains service immediately and leaves the system after service completion. During WV the server serves the customers at a lower rate than the non-vacation period. If the arriving customer finds the server busy in a WV period, it either joins a orbit with probability a or balks the system without being served with probability $1 - a$. When an arriving customer finds the system busy in non-vacation period it always joins the orbit. The orbit is a queue of infinite capacity and the customers in the orbit retry for service on a FCFS basis *i.e.*, only the customer at the head of the queue is allowed to access the server. Successive interretrial times

follow a general distribution with distribution function $R(.)$, p.d.f. $r(.)$ and retrial rate $\alpha(.)$ defined by

$$r(x) = \alpha(x)\exp\left\{-\int_0^x \alpha(t)dt\right\} \quad \text{and} \quad \alpha(x) = \frac{r(x)}{\overline{R}(x)}, \tag{6.1}$$

where $\overline{R}(x) = 1 - R(x)$. These definitions are taken from [159]. The service times of this single server model during the non-vacation period is generally distributed with distribution function $B(.)$, p.d.f. $b(.)$ and service completion rate μ_b. The server takes a WV as soon as the system (both the orbit and the server) becomes empty, *i.e.*, if all the customers have been served, no new customer arrives and the orbit is empty. During WV, the customers are served at rate μ_v, which is less than μ_b, and this service time is generally distributed with distribution function $V(.)$ and p.d.f. $v(.)$. The duration of such a WV is Exp(θ). So, we have

$$b(x) = \mu_b(x)\exp\left\{-\int_0^x \mu_b(t)dt\right\}, \qquad \mu_b(x) = \frac{b(x)}{\overline{B}(x)}, \tag{6.2}$$

$$v(x) = \mu_v(x)\exp\left\{-\int_0^x \mu_v(t)dt\right\} \quad \text{and} \quad \mu_v(x) = \frac{v(x)}{\overline{V}(x)}, \tag{6.3}$$

where $\overline{B}(x) = 1 - B(x)$, $\overline{V}(x) = 1 - V(x)$. If no customer appears in the system at the vacation completion epoch, the server again leaves for another WV. This continues until the server returns from a vacation to find at least one customer in the system. When a vacation ends in the middle of an ongoing service, the server switches its service rate from μ_v to μ_b and continues to serve at the rate μ_b until the service completes. If a vacation ends when the server is free but the orbit is not empty, the server will stay idle in non-vacation period and will wait for a customer to retry from the orbit or for a new customer to arrive. The interarrival times, service times, interretrial times and vacation duration times all are mutually independent.

To analyze this system, let us define a continuous-time process $\{C(t), N(t)\}$, where $C(t)$ denotes the server state $0, 1, 2$, or 3 depending on whether the server is idle during WV, idle during non-vacation period, busy during WV and busy during non-vacation period respectively, and $N(t)$ is the number of customers in the orbit at time t. This process is not Markovian as the service times are not exponentially distributed. So we

introduce a supplementary variable $\xi(t)$ which denotes the elapsed service time of the customer under service at time t. If the system is in idle state, $\xi(t)$ will define the elapsed retrial time. This expanded process $\{C(t), N(t), \xi(t)\}$ is a continuous-time Markov process with state space

$$E = \left\{\{0,1,2,3\} \cup \{0,1,2,\ldots\} \cup R^+\right\},$$

where R^+ denotes the set of non-negative real numbers.

6.2 Stationary distribution

In this section, we first develop the stationary differential difference equations for joint distributions of server state, orbit length and elapsed service time (or the elapsed retrial time). We define the partial PGFs on the basis of the server state $C(t)$. The probability that the orbit is empty and the system is in WV at time t is

$$P_0(t) = P\{C(t) = 0, N(t) = 0\}, \quad t \geq 0.$$

For $n \geq 1$,

$$P_n(t,x)dx = P\{C(t) = 0, N(t) = n, x < \xi(t) \leq x + dx\}, \quad t \geq 0,$$

which is the probability that at time t the system is idle in WV with n customers in the orbit and the elapsed retrial time of the customer at the head of the orbit lies between x and $x + dx$. Similarly, for the idle system in non-vacation period, we get

$$H_0(t) = P\{C(t) = 1, N(t) = 0\}, \quad t \geq 0,$$
$$H_n(t,x)dx = P\{C(t) = 1, N(t) = n, x < \xi(t) \leq x + dx\}, \quad t \geq 0, n \geq 1.$$

Also,

$$S_n(t,x)dx = P\{C(t) = 2, N(t) = n, x < \xi(t) \leq x + dx\}, \quad t \geq 0, n \geq 0, x > 0,$$

which is the probability that at time t the server is busy in WV with n customers in the orbit and the elapsed service time of the customer under service lies between x and $x+dx$.

For the busy system in non-vacation period, we define similarly,

$$W_n(t,x)dx = P\{C(t)=3, N(t)=n, x<\xi(t)\leq x+dx\}, \quad t\geq 0, n\geq 0, x>0.$$

By considering transitions of the process between t and $t+\Delta t$ and letting $\Delta t\to 0$, we derive the following system of Kolmogorov's forward equations

$$\left[\frac{d}{dt}+\lambda\right]P_0(t) = \int_0^\infty \mu_v(x)S_0(t,x)dx + \int_0^\infty \mu_b(x)W_0(t,x)dx, \tag{6.4}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\theta+\alpha(x)\right]P_n(t,x) = 0, \quad n\geq 1, \tag{6.5}$$

$$H_0(t) = 0. \tag{6.6}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\alpha(x)\right]H_n(t,x) = 0, \quad n\geq 1, \tag{6.7}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\theta+\mu_v(x)\right]S_0(t,x) = \lambda(1-a)S_0(t,x), \tag{6.8}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\theta+\mu_v(x)\right]S_n(t,x) = \lambda a S_{n-1}(t,x)+\lambda(1-a)S_n(t,x), \quad n\geq 1, \tag{6.9}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\mu_b(x)\right]W_0(t,x) = \theta S_0(t,x), \tag{6.10}$$

$$\left[\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\lambda+\mu_b(x)\right]W_n(t,x) = \lambda W_{n-1}(t,x)+\theta S_n(t,x), \quad n\geq 1. \tag{6.11}$$

These equations are to be solved under the boundary conditions at $x=0$ which includes $P_n(t,0)$, $H_n(t,0)$, $S_0(t,0)$, $S_n(t,0)$, $W_0(t,0)$ and $W_n(t,0)$. We elaborate the expressions for these boundary conditions as follows:

1. $P_n(t,0)$ is the probability of the event that at time t, the server is idle in WV, with exactly n customers in the orbit and the customer at the head of the orbit has just started its retrial. This event occurs when, at a service completion epoch, the customer leaves the system idle in WV with n customers in the orbit. Thus, we get

$$P_n(t,0) = \int_0^\infty \mu_v(x)S_n(t,x)dx, \quad n\geq 1.$$

2. $H_n(t,0)$ is the probability of the event that at time t, the system is idle in non-vacation, with n customers in orbit and the customer at the head of the orbit has

just started its retrial. Such an event occurs when, at the service completion epoch, the customer leaves the system idle in non-vacation with n customers in the orbit. Therefore,

$$H_n(t,0) = \int_0^\infty \mu_b(x) W_n(t,x)dx, \quad n \geq 1.$$

3. $S_0(t,0)$ is the probability of the event that the system is in WV with an empty orbit at service completion epoch t. Such an event can occur if one of the two following cases happen. First is that the system is in WV with an empty orbit while one customer arrives and gets service immediately. The other case is that the only customer in the orbit retries successfully, leaving behind an empty orbit. Hence,

$$S_0(t,0) = \lambda P_0(t) + \int_0^\infty \alpha(x) P_1(t,x)dx.$$

4. $S_n(t,0)$ is the probability of the event that the system is in WV and there are exactly n customers in the orbit at the service completion epoch t. Such an event can occur if either of the following two events happen. First is, when the system is in WV with n customers in the orbit and the arriving customer gets service immediately upon his arrival. The second event is when there are $(n+1)$ customers in the orbit, the system is in WV and the retrial of a customer to get service is successful. Thus,

$$S_n(t,0) = \lambda \int_0^\infty P_n(t,x)dx + \int_0^\infty \alpha(x) P_{n+1}(t,x)dx, \quad n \geq 1.$$

5. $W_0(t,0)$ is the probability of the event that the system is in non-vacation with an empty orbit at the service completion epoch t. This is equivalent to the event that when the system is in non-vacation and the only customer in the orbit retries successfully, leaving behind an empty orbit. We obtain

$$W_0(t,0) = \int_0^\infty \alpha(x) H_1(t,x)dx.$$

6. $W_n(t,0)$ is the probability of the event that there are exactly n customers in the orbit at the service completion epoch t, with the system in non-vacation. Such an event occurs if either of the following two events happen. The first event is that

there are n customers in the orbit with the system in non-vacation and the arriving customer gets service immediately. The other event is that there are $n+1$ customers in the orbit and the retrial of a customer to get service is successful. So,

$$W_n(t,0) = \lambda \int_0^\infty H_n(t,x)dx + \int_0^\infty \alpha(x) H_{n+1}(t,x)dx, \quad n \geq 1.$$

Thus, under these boundary conditions we have to solve the Kolmogorov's forward equations (6.4) to (6.11). We also have the normalization condition for the system given by

$$P_0(t) + H_0(t) + \sum_{n=1}^{\infty} \int_0^\infty [P_n(t,x) + H_n(t,x)]\, dx + \sum_{n=0}^{\infty} \int_0^\infty [S_n(t,x) + W_n(t,x)]\, dx = 1.$$

We assume that the stability condition, $\lambda a < \mu_b$, holds and that the stationary behavior of the system can be analyzed by defining

$$\begin{aligned} P_0 &= \lim_{t\to\infty} P_0(t), & H_0 &= \lim_{t\to\infty} H_0(t), \\ P_n(x) &= \lim_{t\to\infty} P_n(t,x), & H_n(x) &= \lim_{t\to\infty} H_n(t,x), \\ S_n(x) &= \lim_{t\to\infty} S_n(t,x), & W_n(x) &= \lim_{t\to\infty} W_n(t,x). \end{aligned}$$

The Kolmogorov's forward equations for $t \to \infty$ become

$$\lambda P_0 = \int_0^\infty \mu_v(x) S_0(x)dx + \int_0^\infty \mu_b(x) W_0(x)dx. \quad (6.12)$$

$$\left[\frac{d}{dx} + \lambda + \theta + \alpha(x)\right] P_n(x) = 0, \quad n \geq 1, \quad (6.13)$$

$$H_0 = 0, \quad (6.14)$$

$$\left[\frac{d}{dx} + \lambda + \alpha(x)\right] H_n(x) = 0, \quad n \geq 1, \quad (6.15)$$

$$\left[\frac{d}{dx} + \lambda + \theta + \mu_v(x)\right] S_0(x) = \lambda(1-a) S_0(x), \quad (6.16)$$

$$\left[\frac{d}{dx} + \lambda + \theta + \mu_v(x)\right] S_n(x) = \lambda a S_{n-1}(x) + \lambda(1-a) S_n(x), \quad n \geq 1, \quad (6.17)$$

$$\left[\frac{d}{dx} + \lambda + \mu_b(x)\right] W_0(x) = \theta S_0(x), \quad (6.18)$$

$$\left[\frac{d}{dx} + \lambda + \mu_b(x)\right] W_n(x) = \lambda W_{n-1}(x) + \theta S_n(x), \quad n \geq 1, \quad (6.19)$$

with the boundary conditions

$$P_n(0) = \int_0^\infty \mu_v(x) S_n(x) dx, \quad n \geq 1 \tag{6.20}$$

$$H_n(0) = \int_0^\infty \mu_b(x) W_n(x) dx, \quad n \geq 1 \tag{6.21}$$

$$S_0(0) = \lambda P_0 + \int_0^\infty \alpha(x) P_1(x) dx, \tag{6.22}$$

$$S_n(0) = \lambda \int_0^\infty P_n(x) dx + \int_0^\infty \alpha(x) P_{n+1}(x) dx, \quad n \geq 1 \tag{6.23}$$

$$W_0(0) = \int_0^\infty \alpha(x) H_1(x) dx, \tag{6.24}$$

$$W_n(0) = \lambda \int_0^\infty H_n(x) dx + \int_0^\infty \alpha(x) H_{n+1}(x) dx, n \geq 1, \tag{6.25}$$

and the normalizing equation is

$$P_0 + H_0 + \sum_{n=1}^{\infty} \int_0^\infty [P_n(x) + H_n(x)] \, dx + \sum_{n=0}^{\infty} \int_0^\infty [S_n(x) + W_n(x)] \, dx = 1.$$

We use the method of PGFs to solve these equations. First, we define the following partial PGFs

$$Q_P(z,x) = \sum_{n=1}^{\infty} z^n P_n(x), \qquad Q_H(z,x) = H_0 + \sum_{n=1}^{\infty} z^n H_n(x),$$

$$Q_S(z,x) = \sum_{n=0}^{\infty} z^n S_n(x), \qquad Q_W(z) = \sum_{n=0}^{\infty} z^n W_n(x), \quad |z| \leq 1.$$

and integrating with respect to x, we get

$$Q_P(z) = \int_0^\infty Q_P(z,x) dx,$$

$$Q_H(z) = \int_0^\infty Q_H(z,x) dx,$$

$$Q_S(z) = \int_0^\infty Q_S(z,x) dx,$$

$$Q_W(z) = \int_{n=0}^\infty Q_W(z,x) dx.$$

Theorem 6.2.1. *The joint stationary distribution of the server state and the queue length has the partial PGFs given by*

$$Q_P(z) = \lambda P_0 \overline{R}^*(\lambda+\theta)\left[\frac{v^*(\lambda a(1-z)+\theta) - K_1 v^*(\lambda a+\theta)}{1-K_2 v^*(\lambda a(1-z)+\theta)}\right], \tag{6.26}$$

$$Q_H(z) = \lambda P_0 \theta \overline{R}^*(\lambda)\left[\frac{g_z^*(\lambda(1-z))\{1-K_1K_2 v^*(\lambda a+\theta)\}}{1-K_2 v^*(\lambda a(1-z)+\theta)} - K_1 g_0^*(\lambda)\right], \tag{6.27}$$

$$Q_S(z) = \lambda P_0 \overline{V}^*(\lambda a(1-z)+\theta)\left[\frac{1-K_1K_2 v^*(\lambda a+\theta)}{1-K_2 v^*(\lambda a(1-z)+\theta)}\right], \tag{6.28}$$

$$\text{and } Q_W(z) = \lambda\theta P_0 g_z^*(\lambda(1-z))\left[\frac{1-K_1K_2 v^*(\lambda a+\theta)}{1-K_2 v^*(\lambda a(1-z)+\theta)}\right], \tag{6.29}$$

where

$$K_1 = \frac{1}{v^*(\lambda a+\theta)+\theta g_0^*(\lambda)}, \qquad K_2 = \lambda\overline{R}^*(\lambda+\theta) + r^*(\lambda+\theta), \quad \text{and}$$

$$g_z^*(\lambda(1-z)) = \int_0^\infty e^{-\lambda(1-z)x} b(x)\left[\int_0^x e^{-[\theta-\lambda(1-a)(1-z)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt\right]dx.$$

Proof. Multiplying equation (6.13) with z^n and summing over $n = 1, 2, \ldots$, we obtain

$$\frac{\partial}{\partial x}Q_P(z,x) + (\lambda+\theta+\alpha(x))Q_P(z,x) = 0,$$

which implies that

$$Q_P(z,x) = Q_P(z,0)e^{-(\lambda+\theta)x}\overline{R}(x). \tag{6.30}$$

Similarly, from equation (6.15) we have

$$Q_H(z,x) = Q_H(z,0)e^{-\lambda x}\overline{R}(x). \tag{6.31}$$

Equations (6.16) and (6.17) yield

$$Q_S(z,x) = Q_S(z,0)e^{-[\lambda a(1-z)+\theta]x}\overline{V}(x). \tag{6.32}$$

Equations (6.18) and (6.19) give rise to a non-homogeneous equation

$$\frac{\partial}{\partial x}Q_W(z,x) + [\lambda(1-z)+\mu_b(x)]Q_W(z,x) = \theta Q_S(z,x)$$

with the integrating factor $e^{\lambda(1-z)x}\frac{1}{\overline{B}(x)}$, which after solving reduces to

$$Q_W(z,x) = \theta Q_S(z,0)e^{-\lambda(1-z)x}\overline{B}(x)\int_0^x e^{-[\theta-\lambda(1-a)(1-z)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt. \tag{6.33}$$

The initial conditions (6.20)–(6.25) give rise to the PGFs

$$Q_P(z,0) = \int_0^\infty \mu_v(x)Q_S(z,x)dx - \int_0^\infty \mu_v(x)S_0(x)dx, \tag{6.34}$$

$$Q_H(z,0) = \int_0^\infty \mu_b(x)Q_W(z,x)dx - \int_0^\infty \mu_b(x)W_0(x)dx, \tag{6.35}$$

$$Q_S(z,0) = \lambda P_0 + \lambda\int_0^\infty Q_P(z,x)dx + \int_0^\infty \alpha(x)Q_P(z,x)dx, \tag{6.36}$$

$$Q_W(z,0) = \theta Q_S(z,0) + \lambda\int_0^\infty Q_H(z,x)dx + \int_0^\infty \alpha(x)Q_H(z,x)dx. \tag{6.37}$$

From (6.16), we have

$$S_0(x) = S_0(0)e^{-(\lambda a+\theta)x}\overline{V}(x). \tag{6.38}$$

And, from (6.18) and (6.38)

$$W_0(x) = \theta S_0(0)e^{-\lambda x}\overline{B}(x)\int_0^x e^{-[\theta-\lambda(1-a)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt. \tag{6.39}$$

Inserting (6.38) and (6.39) in (6.12), we obtain

$$S_0(0) = \frac{\lambda P_0}{v^*(\lambda a+\theta)+\theta g_0^*(\lambda)} = \lambda P_0 K_1, \tag{6.40}$$

where

$$K_1 = \frac{1}{v^*(\lambda a+\theta)+\theta g_0^*(\lambda)}$$

and

$$g_0^*(\lambda) = \int_0^\infty e^{-\lambda x}b(x)\left[\int_0^x e^{-[\theta-\lambda(1-a)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt\right]dx.$$

Using (6.32) and (6.40) in (6.34), we get

$$Q_P(z,0) = Q_S(z,0)v^*(\lambda a(1-z)+\theta) - \lambda P_0 K_1 v^*(\lambda a+\theta). \tag{6.41}$$

Again, using the above equation in (6.36), we get

$$Q_S(z,0) = \frac{\lambda P_0[1-K_1K_2v^*(\lambda a+\theta)]}{1-K_2v^*(\lambda a(1-z)+\theta)}, \tag{6.42}$$

where

$$K_2 = \lambda \overline{R}^*(\lambda+\theta) + r^*(\lambda+\theta).$$

Therefore, from (6.32), we find that

$$Q_S(z,x) = \frac{\lambda P_0[1 - K_1 K_2 v^*(\lambda a+\theta)]e^{-[\lambda a(1-z)+\theta]x}\overline{V}(x)}{1 - K_2 v^*(\lambda a(1-z)+\theta)} \tag{6.43}$$

and, from (6.30), we find that

$$\begin{aligned} Q_P(z,x) &= \left[\frac{\lambda P_0 v^*(\lambda a(1-z)+\theta)\{1 - K_1 K_2 v^*(\lambda a+\theta)\}}{1 - K_2 v^*(\lambda a(1-z)+\theta)} \right. \\ &\qquad \left. -\lambda P_0 K_1 v^*(\lambda a+\theta)\right] e^{-(\lambda+\theta)x}\overline{R}(x) \\ &= \lambda P_0 \left[\frac{v^*(\lambda a(1-z)+\theta) - K_1 v^*(\lambda a+\theta)}{1 - K_2 v^*(\lambda a(1-z)+\theta)}\right] e^{-(\lambda+\theta)x}\overline{R}(x). \end{aligned} \tag{6.44}$$

In a similar fashion, the other generating functions can be derived as

$$Q_H(z,x) = \lambda\theta P_0 \left[\frac{g_z^*(\lambda(1-z))\{1 - K_1 K_2 v^*(\lambda a+\theta)\}}{1 - K_2 v^*(\lambda a(1-z)+\theta)} - K_1 g_0^*(\lambda)\right] e^{-\lambda x}\overline{R}(x), \tag{6.45}$$

$$Q_W(z,x) = \lambda\theta P_0 \left[\frac{1 - K_1 K_2 v^*(\lambda a+\theta)}{1 - K_2 v^*(\lambda a(1-z)+\theta)}\right] e^{-\lambda(1-z)x}\overline{B}(x)\int_0^x e^{-[\theta-\lambda(1-a)(1-z)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt, \tag{6.46}$$

where

$$g_z^*(\lambda(1-z)) = \int_0^\infty e^{-\lambda(1-z)x} b(x)\left[\int_0^x e^{-[\theta-\lambda(1-a)(1-z)]t}\frac{\overline{V}(t)}{\overline{B}(t)}dt\right]dx.$$

Integrating equations (6.43)-(6.46) with respect to x, we get the desired partial PGFs (6.26)-(6.29). □

All these partial PGFs are in terms of the unknown quantity P_0. The derivation of P_0 using the normalization equation is given in the next theorem.

Theorem 6.2.2. *The probability P_0 is given by the following expression*

$$\begin{aligned} P_0 &= [1 - K_2 v^*(\theta)]\Big/\Big[\lambda\overline{R}^*(\lambda+\theta)\{v^*(\theta) - K_1 v^*(\lambda a+\theta)\} + \lambda\theta\overline{R}^*(\lambda)\{g_1^*(0) - K_1 g_0^*(\lambda)\} \\ &\quad +\lambda\theta\overline{R}^*(\lambda)K_1K_2\{g_0^*(\lambda)v^*(\theta) - g_1^*(0)v^*(\lambda a+\theta)\} + [1 - K_2 v^*(\theta)] \\ &\quad +\lambda\left\{\overline{V}^*(\theta) + \theta g_1^*(0)\right\}\{1 - K_1 K_2 v^*(\lambda a+\theta)\}\Big]. \end{aligned} \tag{6.47}$$

Proof. Taking the limit as z approaches 1, the partial PGFs from Theorem (6.2.1) give

$$Q_P(1) = \lambda P_0 \overline{R}^*(\lambda+\theta)\left[\frac{v^*(\theta)-K_1 v^*(\lambda a+\theta)}{1-K_2 v^*(\theta)}\right], \tag{6.48}$$

$$Q_H(1) = \lambda P_0 \theta \overline{R}^*(\lambda)\left[\frac{g_1^*(0)\{1-K_1K_2 v^*(\lambda a+\theta)\}}{1-K_2 v^*(\theta)} - K_1 g_0^*(\lambda)\right], \tag{6.49}$$

$$Q_S(1) = \lambda P_0 \overline{V}^*(\theta)\left[\frac{1-K_1K_2 v^*(\lambda a+\theta)}{1-K_2 v^*(\theta)}\right], \tag{6.50}$$

$$Q_W(1) = \lambda\theta P_0 g_1^*(0)\left[\frac{1-K_1K_2 v^*(\lambda a+\theta)}{1-K_2 v^*(\theta)}\right]. \tag{6.51}$$

The normalization condition gives

$$P_0 + Q_P(1) + Q_H(1) + Q_S(1) + Q_W(1) = 1. \tag{6.52}$$

Substituting (6.48)-(6.51) in (6.52) and rearranging it, we get the expression (6.47) for P_0. □

6.3 System efficiency

For WDM networks with retrials, the system efficiency mainly depends upon the network availability (probability that the server is not busy) and the probability of blocking. Blocking probability is the ratio of the number of unsuccessful connection requests to the total number of connection requests in the network. The blocked request will stay in the candidate queue and the request can re-enter the queue unlimited number of times until the connection is established successfully. The average request holding time (or waiting time in orbit) is another important QoS metric. We have listed some of the important performance measures below.

6.3.1 The availability of the server

Let the availability of the server be defined as

$$P_A(t) = P\{\text{the server is available at time } t\}.$$

It is the probability that the server is idle (free), either in WV period or in non-vacation period, at time t. Defining the stationary server availability as the limiting case of $P_A(t)$, we get

$$P_A = P_0 + Q_P(1) + Q_H(1).$$

6.3.2 The blocking probability

The blocking probability is the probability that the server is not available *i.e.*, $P_B = 1 - P_A$. It can also be written as the probability that the server is busy either during WV period or during non-vacation period and is given by

$$P_B = Q_S(1) + Q_W(1).$$

Blocking probability is a very important measure of the system performance; higher the blocking probability lower the system efficiency.

6.3.3 The distribution of number of customers in the orbit

If $N_0(t)$ is the random variable of the number of customers in the orbit at time t, with stationary probability distribution given as $u_n = P\{N_0 = n\}$. The PGF of N_0 is given by

$$N_0(z) = P_0 + Q_P(z) + Q_H(z) + Q_S(z) + Q_W(z).$$

Using Theorem (6.2.1) and (6.2.2), we get

$$\begin{aligned} N_0(z) = P_0 + \frac{\lambda P_0}{1 - K_2 v^*(\lambda a(1-z)+\theta)} \Big[& v^*(\lambda(1-z)+\theta) \left\{ \overline{R}^*(\lambda+\theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \right\} \\ & + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \left\{ \overline{V}^*(\lambda a(1-z)+\theta) + \theta \left(1 + \overline{R}^*(\lambda)\right) g_z^*(\lambda(1-z)) \right\} \\ & - K_1 \overline{R}^*(\lambda+\theta) v^*(\lambda a + \theta) - \theta K_1 \overline{R}^*(\lambda) g_0^*(\lambda) \Big]. \end{aligned} \tag{6.53}$$

Differentiating the above equation with respect to z and evaluating it at $z = 1$ gives the average number of customers in orbit as

$$\begin{aligned} E(N_0) = & \frac{\lambda P_0}{1 - K_2 v^*(\theta)} \Big[\beta_1(\theta) \left\{ \overline{R}^*(\lambda + \theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \right\} + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \\ & \times \left\{ \gamma_1(\theta) + \theta \left(1 + \overline{R}^*(\lambda) \right) \delta_1(0) \right\} \Big] + \frac{\lambda P_0 K_2 \beta_1(\theta)}{[1 - K_2 v^*(\theta)]^2} \Big[v^*(\theta) \{ \overline{R}^*(\lambda + \theta) \\ & + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \} + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \left\{ \overline{V}^*(\theta) + \theta \left(1 + \overline{R}^*(\lambda) \right) g_1^*(0) \right\} \\ & - K_1 \overline{R}^*(\lambda + \theta) v^*(\lambda a + \theta) - \theta K_1 \overline{R}^*(\lambda) g_0^*(\lambda) \Big], \end{aligned} \tag{6.54}$$

where we define

$$\beta_1(z) = \frac{d}{dz} v^*(z), \quad \gamma_1(z) = \frac{d}{dz} \overline{V}^*(z) \quad \text{and} \quad \delta_1(z) = \frac{d}{dz} g_z^*(z).$$

6.3.4 The distribution of number of customers in the system

Let $N(t)$ be a random variable of the number of customers in the orbit including the one in service at time t with stationary probability distribution $q_n = P\{N = n\}$. The PGF of N is given by

$$N(z) = P_0 + Q_P(z) + Q_H(z) + z Q_S(z) + z Q_W(z),$$

which is derived to be

$$\begin{aligned} N(z) = & P_0 + \frac{\lambda P_0}{1 - K_2 v^*(\lambda a(1 - z) + \theta)} \Big[v^*(\lambda(1 - z) + \theta) \left\{ \overline{R}^*(\lambda + \theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \right\} \\ & + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \left\{ z \overline{V}^*(\lambda a(1 - z) + \theta) + \theta \left(z + \overline{R}^*(\lambda) \right) g_z^*(\lambda(1 - z)) \right\} \\ & - K_1 \overline{R}^*(\lambda + \theta) v^*(\lambda a + \theta) - \theta K_1 \overline{R}^*(\lambda) g_0^*(\lambda) \Big]. \end{aligned} \tag{6.55}$$

The mean number of customers in the system is

$$\begin{aligned} E(N) = & \frac{\lambda P_0}{1 - K_2 v^*(\theta)} \Big[\beta_1(\theta) \left\{ \overline{R}^*(\lambda + \theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \right\} \\ & + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \left\{ \gamma_1(\theta) + \overline{V}^*(\theta) + \theta g_1^*(0) + \theta \left(1 + \overline{R}^*(\lambda) \right) \delta_1(0) \right\} \Big] \\ & + \frac{\lambda P_0 K_2 \beta_1(\theta)}{[1 - K_2 v^*(\theta)]^2} \Big[v^*(\theta) \left\{ \overline{R}^*(\lambda + \theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda) \right\} \\ & + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \times \left\{ \overline{V}^*(\theta) + \theta \left(1 + \overline{R}^*(\lambda) \right) g_1^*(0) \right\} \\ & - K_1 \overline{R}^*(\lambda + \theta) v^*(\lambda a + \theta) - \theta K_1 \overline{R}^*(\lambda) g_0^*(\lambda) \Big]. \end{aligned} \tag{6.56}$$

6.4 Stochastic decomposition

The stochastic decomposition property indicates the effects of vacations on system performance measures like queue length and waiting times, and plays an important role in vacation models. We have proved in the following result that our retrial model with a WV policy also satisfies such a decomposition for the stationary queue length and stationary waiting time.

Theorem 6.4.1. *For $\rho < 1$, the stationary queue length of customers, N, in the M/G/1/WV retrial queue can be decomposed into sum of two independent random variables as*

$$N = N_c + N_d,$$

where N_c is the queue length of a classical M/G/1 retrial queue without vacations and N_d is the additional queue length due to the effect of WV with its PGF given by

$$\begin{aligned} N_d(z) &= \frac{P_0\left[b^*(\lambda a(1-z))-z\right]}{(1-\rho)(1-z)b^*(\lambda a(1-z))[1-K_2v^*(\lambda a(1-z)+\theta)]}\Big[v^*(\lambda(1-z)+\theta) \\ &\quad \left\{\overline{R}^*(\lambda+\theta)+\theta\overline{R}^*(\lambda)K_1K_2g_0^*(\lambda)\right\}+\left\{1-K_1K_2v^*(\lambda a+\theta)\right\}\left\{z\overline{V}^*(\lambda a(1-z)+\theta)\right. \\ &\quad \left.+\theta\left(z+\overline{R}^*(\lambda)\right)g_z^*(\lambda(1-z))\right\}-K_1\overline{R}^*(\lambda+\theta)v^*(\lambda a+\theta)-\theta K_1\overline{R}^*(\lambda)g_0^*(\lambda)\Big] \end{aligned} \tag{6.57}$$

$$\text{and } N_c(z) = (1-\rho)\frac{(1-z)b^*(\lambda a(1-z))}{b^*(\lambda a(1-z))-z}.$$

Proof. The stationary queue length of customers in our system is given by (6.55). For simplicity, the PGF of queue length can be written as $N(z) = P_0M(z)$, where

$$\begin{aligned} M(z) &= 1+\frac{\lambda}{1-K_2v^*(\lambda a(1-z)+\theta)}\Big[v^*(\lambda(1-z)+\theta)\left\{\overline{R}^*(\lambda+\theta)+\theta\overline{R}^*(\lambda)K_1K_2g_0^*(\lambda)\right\} \\ &\quad +\left\{1-K_1K_2v^*(\lambda a+\theta)\right\}\left\{z\overline{V}^*(\lambda a(1-z)+\theta)+\theta\left(z+\overline{R}^*(\lambda)\right)g_z^*(\lambda(1-z))\right\} \\ &\quad -K_1\overline{R}^*(\lambda+\theta)v^*(\lambda a+\theta)-\theta K_1\overline{R}^*(\lambda)g_0^*(\lambda)\Big]. \end{aligned}$$

Then

$$\begin{aligned} N(z) &= \frac{(1-\rho)(1-z)b^*(\lambda a(1-z))}{b^*(\lambda a(1-z))-z}\times\frac{P_0\{b^*(\lambda a(1-z))-z\}M(z)}{(1-\rho)(1-z)b^*(\lambda a(1-z))} \\ &= N_c(z)\times N_d(z). \end{aligned}$$

$N_d(z)$ gives the number of customers in the system when the system is in WV given that the system is idle in WV or busy in WV. So this term can be also be obtained through $[\{P_0 + Q_P(z) + zQ_S(z)\} / \{P_0 + Q_P(1) + Q_S(1)\}]$ and $N_d(0) = 0, N_d(1) = 1$ which ensures that $N_d(z)$ is a PGF.

□

Theorem 6.4.2. *If $\rho < 1$, the stationary waiting time of customers W in the M/G/1/WV with retrial system can be decomposed into the sum of two independent random variables as*

$$W = W_c + W_d,$$

where W_c is the waiting time of a customer corresponding to classical M/G/1 retrial queue and has exponential distribution with parameter $\mu_b(1-\rho)$ and W_d is the additional delay due to the effect of WV with its LST given by

$$\begin{aligned} W_d^*(s) &= \frac{\lambda P_0 \left[\lambda b^*(as) - (\lambda - s)\right]}{(1-\rho)sb^*(as)[1 - K_2 v^*(as+\theta)]} \Bigg[v^*(s+\theta)\left\{\overline{R}^*(\lambda+\theta) + \theta \overline{R}^*(\lambda) K_1 K_2 g_0^*(\lambda)\right\} \\ &\quad + \{1 - K_1 K_2 v^*(\lambda a + \theta)\} \left\{ \left(1 - \frac{s}{\lambda}\right) \overline{V}^*(as+\theta) + \theta \left(1 - \frac{s}{\lambda} + \overline{R}^*(\lambda)\right) g_z^*(s) \right\} \\ &\quad - K_1 \overline{R}^*(\lambda+\theta) v^*(\lambda a + \theta) - \theta K_1 \overline{R}^*(\lambda) g_0^*(\lambda) \Bigg]. \end{aligned} \tag{6.58}$$

Proof. We have from the distributional form of Little's law that $N(z) = W^*(\lambda(1-z))$ [91]. Let $s = \lambda(1-z)$ which gives

$$z = \left(1 - \frac{s}{\lambda}\right) \text{ and } 1 - z = \frac{s}{\lambda}.$$

Putting these relations in (6.57) we get the expression (6.58). □

6.5 Numerical results

To observe the efficiency of the proposed model, we perform some numerical experiments in this section. The performance measures observed are the availability of the server and the average queue length. Mainly, we want to depict the influence of retrial on the WV

system. We plot the change in the performance measures for different retrial rates along with the change in the other parameter values. For simplicity, we first assume that the service time distribution is exponential. Later, we compare the effect of different service distributions on the model.

Effect on system availability

In Figure 6.1, the plot of system availability against vacation service rate is shown. We have taken $a = 0.7, \lambda = 1.85, \mu_b = 2$ and $\theta = 0.3$. When $\mu_v = 0$, the system does not serve during the vacation period and our model boils down to a pure vacation model. For $\mu_v = \mu_b = 2$, the server serves the customer at the same rate throughout and so this becomes equivalent to a system where the server never takes vacations. The graph gives the bridge between a pure vacation retrial model and a non-vacation one. Within this range of μ_v, the system availability can increase by upto 60%. Another observation is that if there are no retrials ($\alpha = 0$) the availability of the system increases by upto 50% with the increase in vacation service rate. The number of retrials (per unit time) increases the system availability, but the system almost becomes unaffected by retrial rate beyond $\alpha = 10$. This is because, when the mean retrial time $\frac{1}{\alpha}$ is too small, this is similar to a system without retrials.

Figure 6.2 and Figure 6.3 are plots of system availability against the retrial rate for different vacation duration rates θ. The retrial rate increases the availability of the server. This increase is monotone for smaller rates of θ (in Figure 6.2). But this behavior changes as the vacation duration rate increases beyond $\theta = 1.8$. For $\theta = 8$, the system availability remains almost stable with respect to the retrial rates as we can see in the graph (Figure 6.3). So the retrial rate affects the system availability if the vacation duration rate is below 8, *i.e.,* for smaller durations of vacations.

The blocking probability is the probability that the system is not available *i.e.*, $P_B = 1 - P_A$. Therefore the retrial effect on the blocking probability will be the reverse of the effect on system availability.

Effect on average queue length

We compare the mean queue lengths of this system for three different values of vacation duration rates in Figure 6.4. The retrial rate can decrease the mean queue length. Here the vacation duration rate plays a role to minimize the average queue length even further. Figure 6.5 shows that the increase in retrial rate can enhance the efficiency of the model by decreasing the mean queue length. The value of queue length, or equivalently the value of waiting time in queue, can be reduced if the retrial rates as well as the vacation duration rates are increased.

The mean waiting time is the ratio of mean queue length over the arrival rate λ. Figure 6.6 gives the mean queue length of our system without retrial. When mean retrial tends to ∞, our model gives almost the same result as in Lui [113].

Effect of service distribution

To see the effect of service distribution on the model, we have plotted in Figure 6.7 the system availability for an exponentially distributed service time distribution and for a Erlang-2 distributed service times having the same mean. The parameter values are given in the graph. We have taken the same mean values for both the distributions. The graph shows a vast difference in server availability if the service distribution is changed. Server availability for the exponential distributed service model is about 40% more than the one with Erlang-2. Therefore the choice of service distribution is extremely important in order to enhance the efficiency of the model.

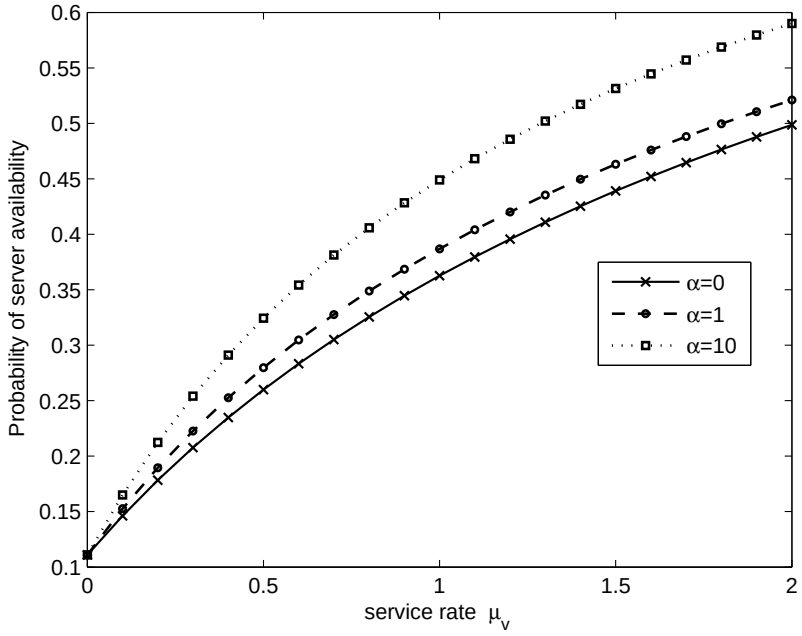

Figure 6.1: Probability of server availability vs vacation-service rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\theta = 0.3$.

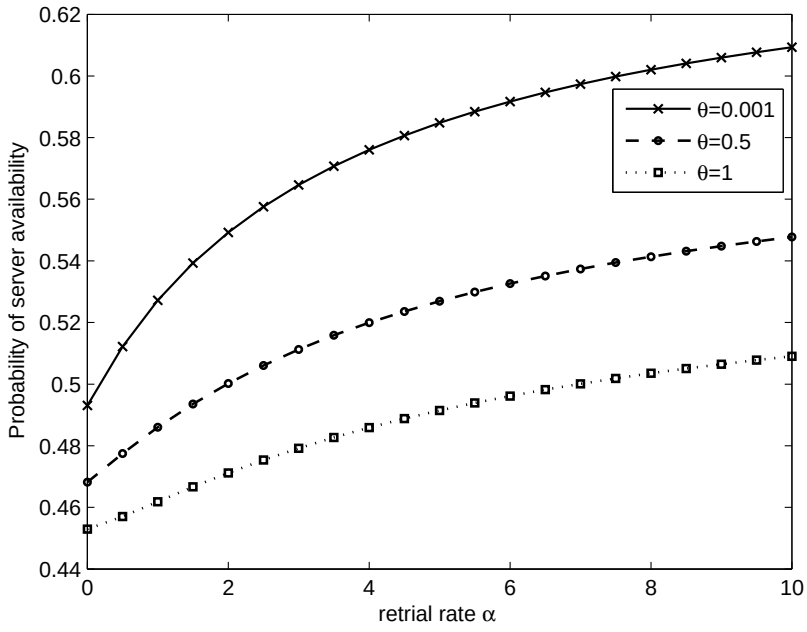

Figure 6.2: Probability of server availability vs retrial rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\mu_v = 1.8$.

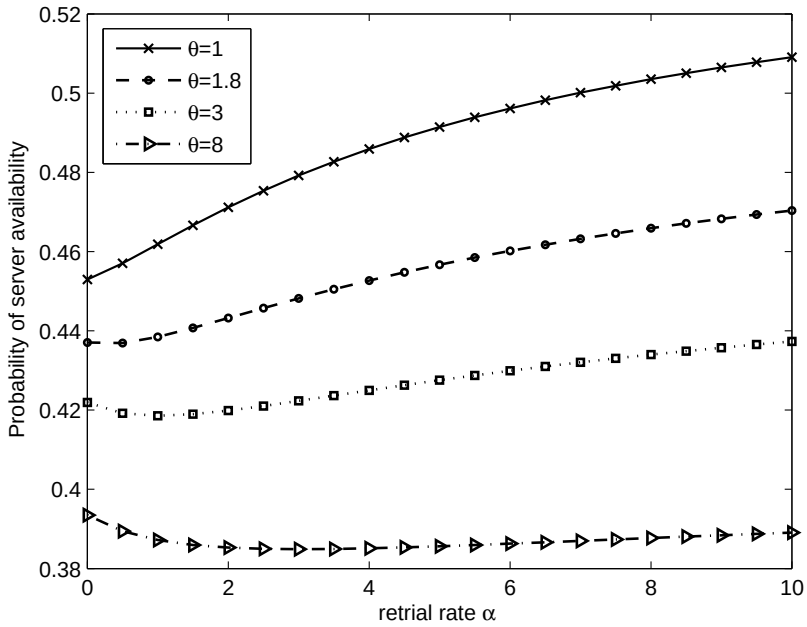

Figure 6.3: Probability of server availability vs retrial rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\mu_v = 1.8$.

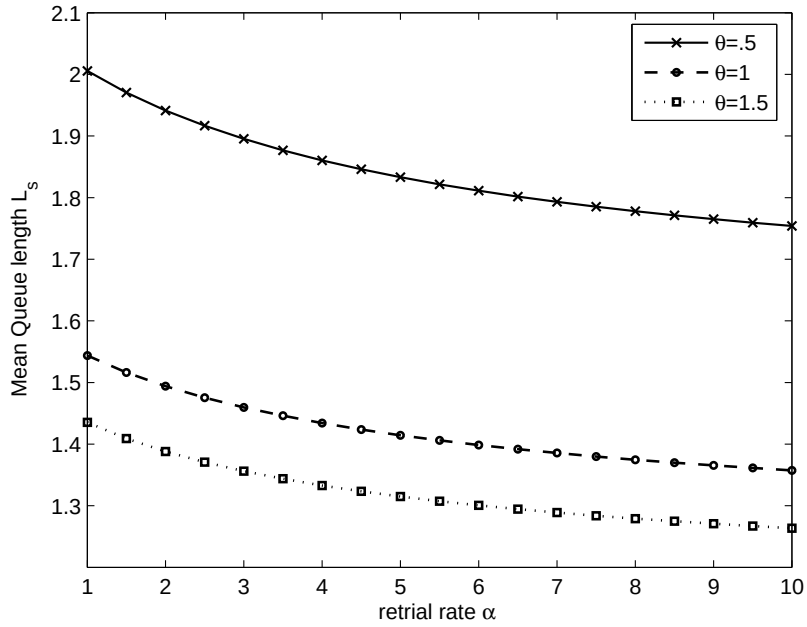

Figure 6.4: Mean queue length vs retrial rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\mu_v = 1.8$.

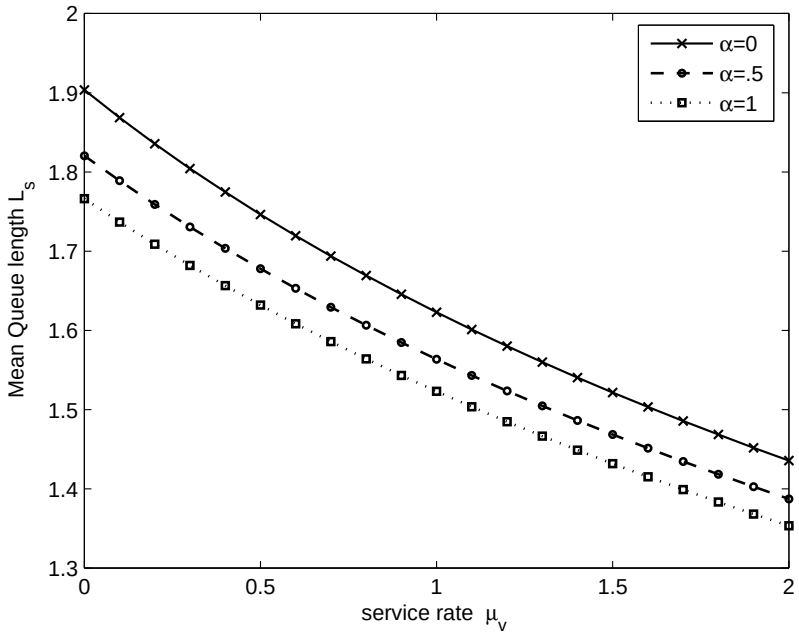

Figure 6.5: Mean queue length vs vacation-service rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\theta = 1.8$.

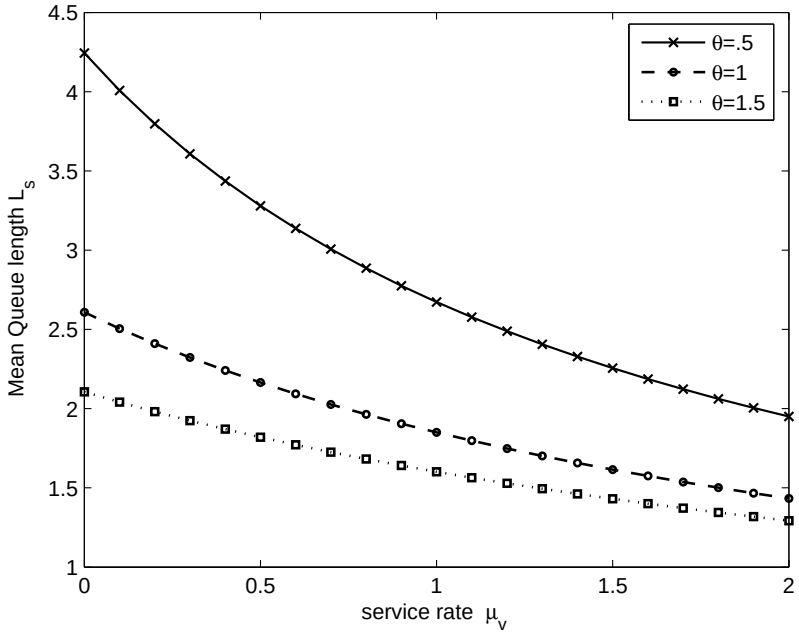

Figure 6.6: Mean queue length vs vacation-service rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\alpha = 100$.

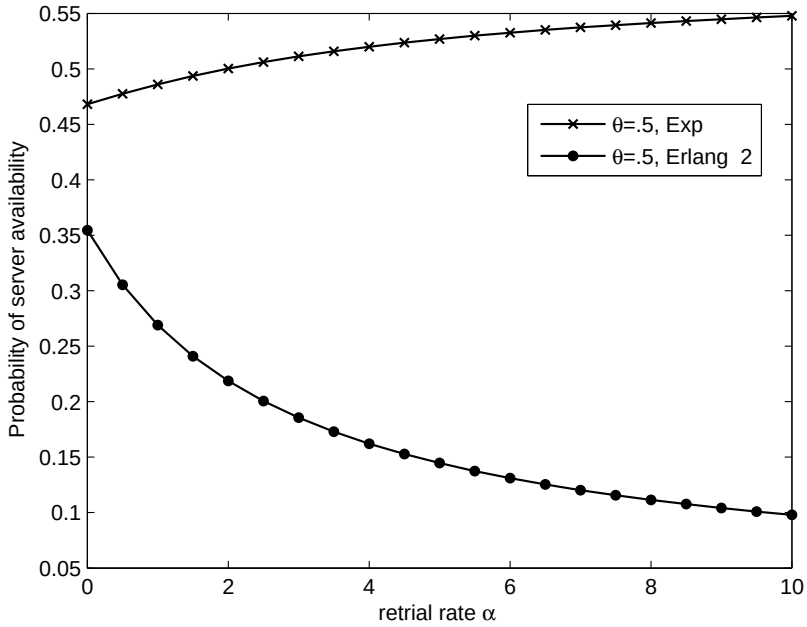

Figure 6.7: Probability of server availability vs retrial rate with $a = 0.7$, $\lambda = 1.85$, $\mu_b = 2$, $\mu_v = 1.8$.

Chapter 7

A Queue with Impatient Customers

In a communication network the jobs which do not get service upon arrival wait in a queue. The jobs can become impatient due to high waiting times or due to uncertainty of receiving services and may leave the system unserved. This scenario often occurs in Optical Burst Switching (OBS) networks. OBS is a technology for reducing the gap between transmissions and switching speeds. In OBS, incoming traffic from clients at the edge of the network is aggregated at the ingress of the network according to a particular parameter. This is then assembled and is further transmitted through WDM links. The operation of an OBS controller can be seen as a queue with reneging or impatience [30]. In OBS networks, a control packet is sent first, on a separate signaling channel, to set up a connection. It is followed by a data burst without waiting for an acknowledgement for path establishment [131]. In particular, when a path is not assigned, the burst control packet is accepted to the queue and is kept waiting for a path. If its delay budget is lower than the effective processing delay, the packet becomes impatient and leaves the system unserved. To have more efficient use of the network, the loss of packets has to be reduced for better performance of the network. Therefore, study on impatience of a request in WDM links is essential for the benefit of better system performance. This chapter presents a detailed study of stationary distribution of packet waiting times, the queue length distribution and the mean number of packets served (which do not abandon

the system), in such a WDM network.

Altman and Yechiali [7] studied a M/M/1 queue with impatient customers in a classical vacation setup, where each arriving customer who finds no server on duty, activates an independent random impatient timer. If a server does not return from vacation by the time the timer expires, the customer abandons the queue and never returns. If the server returns from its vacation before the timer expires, the customer stays in the system until his service is completed. We consider, in this chapter, a M/M/1 queue under the WV policy with impatient customers. We study the model with both MWV and SWV policies and derive the explicit expressions for the queue length distribution and the waiting time distribution. The motivation of this work is to have answers to queries like, how impatient customers change the system performance in WV models; which WV model (MWV or SWV) works better if the customers are impatient and how the system parameters affect the system behavior.

7.1 Multiple working vacation model

We consider a system where the arrivals happen according to a Poisson process with rate λ. The service time during a non-vacation period follows an Exp(μ_b) distribution, where $\lambda < \mu_b$. The server takes a WV as soon as the system becomes empty. During a WV period, the server serves customers according to an Exp(μ_v) distribution, where $\mu_v < \mu_b$. The duration of a vacation is also Exp(θ). A MWV policy requires the server to keep taking vacations until it finds at least one customer waiting in the system at a vacation completion instant. When the server returns from its vacation and finds atleast one customer in the system, the server switches its service rate from μ_v to μ_b and a non-vacation period starts; otherwise it immediately leaves for another WV. If the vacation terminates in between an ongoing service, the server switches its service rate and continues the service at the higher rate until completion. The customers in this system are impatient. Whenever a customer arrives and finds the system in WV, the customer activates an impatient timer 'T', which follows an Exp(ξ) distribution and is independent of the number of customers

in the system at that moment. If the server ends the vacation before the time T expires, the customer stays in the system till his service is completed; otherwise, the customer leaves the system and never returns.

Here we assume that the service rate during vacation, μ_v, is strictly positive. In WV policy, the customer who arrives during a WV period starts getting service immediately upon its arrival if it finds the server empty and therefore it can never become impatient. This type of impatience of customers under WV policy is different from the impatient policy studied in Altman and Yechiali [7], in which all customers arriving during vacation period activate impatient timers. The interarrival times, service times, vacation duration times and the impatient times, are all taken to be mutually independent. The server serves the customers as per FCFS basis.

The interarrival times, service times and vacation durations all are exponential for this model. To have a full information of the system at an arbitrary time t, the number of customers in the system at that time and the vacation-state of the server are sufficient. So, to describe the state of the system, we define a continuous-time Markov chain, $\Delta = \{(N_t, J_t),\ t \geq 0\}$, where N_t denotes the total number of customers in the system and J_t denotes the vacation-state of the server, *i.e.*,

$$J_t = \begin{cases} 1, & \text{if the server is in non-vacation period at time t;} \\ 0, & \text{if the server is in WV period at time t.} \end{cases}$$

The state-space of this Markov chain is

$$E = \Big\{ \{(0,0)\} \cup \{(i,j)\},\ i = 1, 2, \ldots,\ j = 0, 1 \Big\}.$$

and the state transition diagram can be constructed from the model description, as shown in Figure 7.1.

7.1.1 Stationary distribution

Let us define the stationary probabilities for the Markov chain Δ as

$$p_{i,j} = P\{N_t = i, J_t = j\},\ i = 0, 1, 2, \ldots\ j = 0, 1.$$

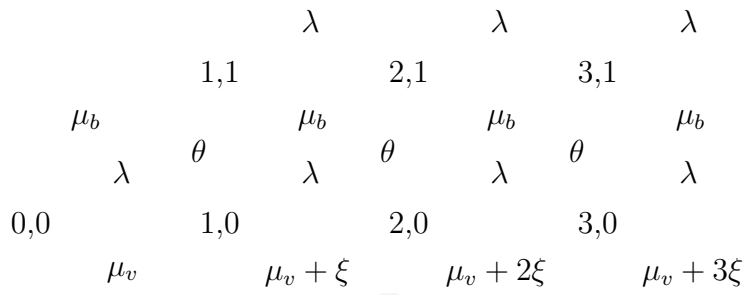

Figure 7.1: State transition diagram for a M/M/1 queue with MWV

The balance equations for this model are given by

$$\lambda p_{0,0} = \mu_v p_{1,0} + \mu_b p_{1,1}, \qquad \text{if } n = 0, \tag{7.1}$$

$$[\lambda + \mu_v + \theta + (n-1)\xi]\, p_{n,0} = \lambda p_{n-1,0} + (\mu_v + n\xi) p_{n+1,0}, \quad \text{if } n \geq 1 \tag{7.2}$$

$$(\lambda + \mu_b) p_{1,1} = \theta p_{1,0} + \mu_b p_{2,1}, \qquad \text{if } n = 1, \tag{7.3}$$

$$(\lambda + \mu_b) p_{n,1} = \lambda p_{(n-1),1} + \theta p_{n,0} + \mu_b p_{n+1,1}, \qquad \text{if } n \geq 2. \tag{7.4}$$

Let us define the PGFs

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{n,0},\ 0 < z \leq 1,$$

$$P_1(z) = \sum_{n=1}^{\infty} z^n p_{n,1},\ 0 < z \leq 1,$$

with $P_0(1) + P_1(1) = 1$ and $P_0'(z) = n \sum_{n=1}^{\infty} z^{n-1} p_{n,0}$.

Theorem 7.1.1. *For $\rho = \frac{\lambda}{\mu_b} < 1$ and $\xi < \mu_v$, the PGFs can be expressed in terms of $p_{0,0}$ as*

$$P_0(z) = p_{0,0}\left(\frac{\mu_v}{\xi} - 1\right)\left[\sum_{n=0}^{\infty}\sum_{r=0}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{J(r)}{(n-r)!}z^n - \sum_{n=1}^{\infty}\sum_{r=1}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{1}{(n-r)!}\left\{J(r-1) + \frac{C(1)}{A(1)}H(r-1)\right\}z^n\right] \tag{7.5}$$

and

$$P_1(z) = \left[\frac{\theta}{\mu_b} z P_0(z) - \frac{(\mu_v - \xi)C(1)}{\mu_b A(1)} z p_{0,0}\right](\rho z - 1)^{-1}(1 - z)^{-1}, \tag{7.6}$$

with

$$J(n)=\sum_{k=0}^{n}\frac{(-1)^{n-k}}{\left(\frac{\mu_v}{\xi}+n-k-1\right)(n-k)!}\frac{\left(n-k+\frac{\mu_v}{\xi}-\frac{\theta}{\xi}\right)_k}{(n+\frac{\mu_v}{\xi})_k}\left(\frac{\lambda}{\xi}\right)^{n-k},$$

$$H(n)=\sum_{k=0}^{n}\frac{(-1)^{n-k}}{\left(\frac{\mu_v}{\xi}+n-k\right)(n-k)!}\frac{\left(n-k+\frac{\mu_v}{\xi}+\frac{\theta}{\xi}\right)_k}{(n+\frac{\mu_v}{\xi}+1)_k}\left(\frac{\lambda}{\xi}\right)^{n-k},$$

$$A(1)=K\left(\frac{\lambda}{\xi},\frac{\mu_v}{\xi},\frac{\theta}{\xi}\right)\quad and\quad C(1)=K\left(\frac{\lambda}{\xi},\frac{\mu_v}{\xi}-1,\frac{\theta}{\xi}+1\right),$$

where $K(a,b,c)=B(b,c)_1F_1(c;b+c;a)$ *with the Beta function* $B(b,c)=\int_0^1 t^{(b-1)}(1-t)^{c-1}dt,\ b>0,c>0$; *the degenerate hypergeometric function* ${}_1F_1(\alpha;\beta;z)=\sum_{k=0}^{\infty}\frac{(\alpha)_k}{(\beta)_k}\frac{z^k}{k!}$ *and the Pochhammer symbol* $(\alpha)_k=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$.

Proof. Multiplying (7.2) with z^n and summing over n gives

$$[\lambda+\mu_v+\theta+(n-1)\xi]\sum_{n=1}^{\infty}z^np_{n,0}=\lambda\sum_{n=1}^{\infty}z^np_{n-1,0}+\sum_{n=1}^{\infty}(\mu_v+n\xi)z^np_{n+1,0}$$

and rearranging the terms gives rise to the differential equation, for $\xi>0$,

$$P_0'(z)-\left[\frac{\lambda z-\mu_v+\xi}{z\xi}+\frac{\theta}{\xi(1-z)}\right]P_0(z)+\left[\frac{\theta}{\xi(1-z)}-\frac{(\mu_v-\xi)}{z\xi}\right]p_{0,0}+\frac{\mu_b}{\xi(1-z)}p_{1,1}=0. \tag{7.7}$$

To solve the above first order linear differential equation, the integrating factor is found as

$$I.F.=\exp\left\{-\int\left[\left(\frac{\lambda}{\xi}-\frac{\mu_v}{z\xi}\right)+\frac{1}{z}+\frac{\theta}{\xi(1-z)}\right]dz\right\}=e^{-\frac{\lambda}{\xi}z}z^{\left(\frac{\mu_v}{\xi}-1\right)}(1-z)^{\frac{\theta}{\xi}}. \tag{7.8}$$

Using (7.8), and integrating from 0 to z, the general solution of (7.7) becomes

$$P_0(z)\ =z^{-\left(\frac{\mu_v}{\xi}-1\right)}(1-z)^{-\frac{\theta}{\xi}}\left[\left(\frac{\mu_v}{\xi}-1\right)p_{0,0}C(z)-\left(\frac{\theta}{\xi}p_{0,0}+\frac{\mu_b}{\xi}p_{1,1}\right)A(z)\right], \tag{7.9}$$

with

$$A(z)\ =\int_0^z e^{\frac{\lambda}{\xi}(z-x)}x^{\frac{\mu_v}{\xi}-1}(1-x)^{\frac{\theta}{\xi}-1}dx \tag{7.10}$$

$$\text{and}\ \ C(z)\ =\int_0^z e^{\frac{\lambda}{\xi}(z-x)}x^{\frac{\mu_v}{\xi}-2}(1-x)^{\frac{\theta}{\xi}}dx. \tag{7.11}$$

To get $p_{1,1}$ in terms of $p_{0,0}$, let us determine (7.10) and (7.11) for limit z tending to 1. We have the identity [1],

$$\int_0^w x^{v-1}(w-x)^{u-1}e^{\beta x}dx = B(u,v)w^{u+v-1}\ {}_1F_1(v;u+v;\beta w), \quad Re(u)>0, Re(v)>0, \tag{7.12}$$

with the Beta function $B(b,c) = \int_0^1 t^{(b-1)}(1-t)^{c-1}dt,\ b>0, c>0$; the degenerate hypergeometric function ${}_1F_1(\alpha;\beta;z) = \sum_{k=0}^{\infty}\frac{(\alpha)_k}{(\beta)_k}\frac{z^k}{k!}$ and the Pochhammer symbol $(\alpha)_k = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$. Substituting $z=1$, $1-x=t$ and using identity (7.12), in equations (7.10) and (7.11), we get

$$A(1) = B\left(\frac{\mu_v}{\xi},\frac{\theta}{\xi}\right){}_1F_1\left(\frac{\theta}{\xi};\frac{\mu_v}{\xi}+\frac{\theta}{\xi};\frac{\lambda}{\xi}\right) \tag{7.13}$$

$$\text{and } C(1) = B\left(\frac{\mu_v}{\xi}-1,\frac{\theta}{\xi}+1\right){}_1F_1\left(\frac{\theta}{\xi}+1;\frac{\mu_v}{\xi}+\frac{\theta}{\xi};\frac{\lambda}{\xi}\right), \tag{7.14}$$

which can be written as

$$A(1) = K\left(\frac{\lambda}{\xi},\frac{\mu_v}{\xi},\frac{\theta}{\xi}\right) \text{ and } C(1) = K\left(\frac{\lambda}{\xi},\frac{\mu_v}{\xi}-1,\frac{\theta}{\xi}+1\right), \tag{7.15}$$

where $K(a,b,c) = B(b,c){}_1F_1(c;b+c;a)$. $C(1)$ is valid only if $\left(\frac{\mu_v}{\xi}-1\right)>0$, which leads to the assumption $\xi<\mu_v$. Now, determining $P_0(z)$ for limit z tending to 1 gives

$$\lim_{z\to1}P_0(z) = P_0(1) = \left[\left(\frac{\mu_v}{\xi}-1\right)p_{0,0}C(1) - \left(\frac{\theta}{\xi}p_{0,0}+\frac{\mu_b}{\xi}p_{1,1}\right)A(1)\right]\times\lim_{z\to1}(1-z)^{-\frac{\theta}{\xi}}.$$

Since $1>P_0(1)=\sum_{n=0}^{\infty}p_{n,0}\geq 0$ and $\lim_{z\to1}(1-z)^{-\frac{\theta}{\xi}}=\infty$, we must have the term

$$\left(\frac{\mu_v}{\xi}-1\right)p_{0,0}C(1) - \left(\frac{\theta}{\xi}p_{0,0}+\frac{\mu_b}{\xi}p_{1,1}\right)A(1) = 0.$$

Rearranging it, we get $p_{1,1}$ in terms of $p_{0,0}$ as

$$p_{1,1} = \left[\frac{(\mu_v-\xi)C(1)}{\mu_b A(1)} - \frac{\theta}{\mu_b}\right]p_{0,0}. \tag{7.16}$$

Applying the above relation in (7.9), we get

$$P_0(z) = p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[C(z)-\frac{C(1)}{A(1)}A(z)\right]z^{-\left(\frac{\mu_v}{\xi}-1\right)}(1-z)^{-\frac{\theta}{\xi}}.$$

To have $P_0(z)$ as a series summation, we now expand the integrals $A(z)$ and $C(z)$ in series. The function $A(z)$ can be written in terms of incomplete beta function $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ as

$$A(z) = e^{\frac{\lambda}{\xi}z}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\left(\frac{\lambda}{\xi}\right)^n B\left(z; n+\frac{\mu_v}{\xi}, \frac{\theta}{\xi}\right). \tag{7.17}$$

The incomplete Beta function can again be expressed in terms of Gauss hypergeometric function as

$$B(x;a,b) = \frac{x^a}{a}\,{}_2F_1(a, 1-b; a+1; x) \tag{7.18}$$

and further we have the relation

$${}_2F_1(\alpha,\beta;\gamma;z) = (1-z)^{\gamma-\alpha-\beta}\,{}_2F_1(\gamma-\alpha,\gamma-\beta;\gamma;z). \tag{7.19}$$

Using these identities in (7.17), we have

$$A(z) = e^{\frac{\lambda}{\xi}z}(1-z)^{\frac{\theta}{\xi}}z^{\frac{\mu_v}{\xi}-1}\sum_{n=0}^{\infty}\frac{(-1)^n z^{n+\frac{\mu_v}{\xi}}}{n!(n+\frac{\mu_v}{\xi})}\left(\frac{\lambda}{\xi}\right)^n {}_2F_1\left(1, n+\frac{\mu_v}{\xi}+\frac{\theta}{\xi}; n+\frac{\mu_v}{\xi}+1; z\right).$$

Further, the series expansion of ${}_2F_1(\alpha,\beta;\gamma;z) = \sum_{k=0}^{\infty}\frac{(\alpha)_k(\beta)_k}{(\gamma)_k}\frac{z^k}{k!}$ reduces the above expression to

$$A(z) = e^{\frac{\lambda}{\xi}z}(1-z)^{\frac{\theta}{\xi}}z^{\frac{\mu_v}{\xi}-1}\sum_{n=0}^{\infty}H(n)z^{n+1}, \tag{7.20}$$

where

$$H(n) = \sum_{k=0}^{n}\frac{(-1)^{n-k}}{\left(\frac{\mu_v}{\xi}+n-k\right)(n-k)!}\frac{\left(n-k+\frac{\mu_v}{\xi}+\frac{\theta}{\xi}\right)_k}{(n-k+\frac{\mu_v}{\xi}+1)_k}\left(\frac{\lambda}{\xi}\right)^{n-k}. \tag{7.21}$$

Similarly, we get

$$C(z) = e^{\frac{\lambda}{\xi}z}(1-z)^{1+\frac{\theta}{\xi}}z^{\frac{\mu_v}{\xi}-1}\sum_{n=0}^{\infty}J(n)z^n, \tag{7.22}$$

where

$$J(n) = \sum_{k=0}^{n}\frac{(-1)^{n-k}}{\left(\frac{\mu_v}{\xi}+n-k-1\right)(n-k)!}\frac{\left(n-k+\frac{\mu_v}{\xi}+\frac{\theta}{\xi}\right)_k}{(n-k+\frac{\mu_v}{\xi})_k}\left(\frac{\lambda}{\xi}\right)^{n-k}. \tag{7.23}$$

Using (7.20) and (7.22), and expanding the exponential terms, e^{az}, in series, we get

$$\begin{aligned}P_0(z) &= p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[\sum_{n=0}^{\infty}\sum_{r=0}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{J(r)}{(n-r)!}\right.\\ &\quad\left.-\sum_{n=1}^{\infty}\sum_{r=1}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{1}{(n-r)!}\left\{J(r-1)+\frac{C(1)}{A(1)}H(r-1)\right\}\right]z^n.\end{aligned}$$

This $P_0(z)$ is the solution to the differential equation (7.7).

Next we will find the generating function $P_1(z)$. Multiplying (7.4) with z^n and summing over n gives,

$$(\lambda+\mu_b)\sum_{n=2}^{\infty} z^n p_{n,1} = \lambda\sum_{n=2}^{\infty} z^n p_{n-1,1} + \theta\sum_{n=2}^{\infty} z^n p_{n,0} + \mu_b\sum_{n=2}^{\infty} z^n p_{n+1,1},$$

which after some simple algebra reduces to

$$(\lambda z-\mu_b)(1-z)P_1(z) = \theta z P_0(z) - z\left[\theta p_{0,0} + \mu_b p_{1,1}\right]$$

and, for $0 < z < 1$, we have

$$P_1(z) = \frac{\theta z P_0(z) - z\left[\theta p_{0,0} + \mu_b p_{1,1}\right]}{(\lambda z-\mu_b)(1-z)} \tag{7.24}$$

which simplifies to

$$P_1(z) = \left[\frac{\theta}{\mu_b} z P_0(z) - \frac{(\mu_v-\xi)C(1)}{\mu_b A(1)} z p_{0,0}\right]\left[(1+\rho)z - \rho z^2 - 1\right]^{-1}.$$

□

The probabilities $\{p_{n,0},\ n \geq 1\}$ and $\{p_{n,1},\ n \geq 1\}$ can be found from the generating functions $P_0(z)$ and $P_1(z)$ respectively.

The PGF $P_0(z)$ gives the probabilities $p_{n,0}$, for $n \geq 1$ in terms of $p_{0,0}$ as

$$\begin{aligned} p_{n,0} &= p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[\sum_{r=0}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{J(r)}{(n-r)!}\right. \\ &\quad \left. -\sum_{r=1}^{n}\left(\frac{\lambda}{\xi}\right)^{n-r}\frac{1}{(n-r)!}\left\{J(r-1)+\frac{C(1)}{A(1)}H(r-1)\right\}\right]. \end{aligned} \tag{7.25}$$

For example, the probability

$$\begin{aligned} p_{1,0} &= p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[\left(\frac{\lambda}{\xi}-1\right)J(0) + J(1) - \frac{C(1)}{A(1)}H(0)\right] \\ &= p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[\left(\frac{\lambda}{\xi}-1\right)\frac{\xi}{\mu_v-\xi} + \frac{\xi(\mu_v+\theta)-\lambda(\mu_v-\xi)}{\mu_v(\mu_v-\xi)} - \frac{C(1)}{A(1)}\frac{\xi}{\mu_v}\right] \\ &= p_{0,0}\left(\frac{\mu_v}{\xi}-1\right)\left[\frac{\xi(\lambda+\theta)}{\mu_v(\mu_v-\xi)} - \frac{C(1)}{A(1)}\frac{\xi}{\mu_v}\right]. \end{aligned}$$

reduces to

$$p_{1,0} = p_{0,0}\left[\frac{\lambda+\theta}{\mu_v} - \frac{(\mu_v-\xi)C(1)}{\mu_v A(1)}\right].$$

The same expression can be derived from (7.1), using (7.16). Similarly, the probabilities $p_{n,1}$, $n \geq 1$, can be found from $P_1(z)$ in terms of $p_{0,0}$.

We have seen that all the probabilities can be derived in terms of $p_{0,0}$, which can be found from the result below.

Theorem 7.1.2. *If $\rho < 1$ and $\xi < \mu_v$, then the probability $p_{0,0}$ is given as*

$$p_{0,0} = \frac{(\xi+\theta)(\mu_b-\lambda)}{\theta(\mu_v-\xi) + \frac{(\mu_v-\xi)C(1)}{\theta A(1)}\left[\theta(\xi+\mu_b-\mu_v)+\xi(\mu_b-\lambda)\right]}, \tag{7.26}$$

where $A(1)$ and $C(1)$ are products of the Beta function and the degenerate hypergeometric series as in (7.15) above.

Proof. Applying L'Hopital's rule in (7.9),

$$\lim_{z\to 1} P_0(z) = \lim_{z\to 1} \frac{\left(\frac{\mu_v}{\xi}-1\right)p_{0,0}(1-z) - \left(\frac{\theta}{\xi}p_{0,0} + \frac{\mu_b}{\xi}p_{1,1}\right)z}{\left(\frac{\mu_v}{\xi}-1\right)(1-z) - \frac{\theta}{\xi}z}, \tag{7.27}$$

which gives

$$\theta P_0(1) = \theta p_{0,0} + \mu_b p_{1,1}. \tag{7.28}$$

Applying L'Hopital's rule in (7.24) gives

$$\lim_{z\to 1} P_1(z) = \lim_{z\to 1} \frac{\theta P_0(z) + \theta z P_0'(z) - [\theta p_{0,0} + \mu_b p_{1,1}]}{\lambda(1-z) - (\lambda z - \mu_b)},$$

which reduces to

$$P_1(1) = \frac{\theta P_0(1) + \theta P_0'(1) - [\theta p_{0,0} + \mu_b p_{1,1}]}{\mu_b - \lambda} \tag{7.29}$$

and applying result (7.28) in (7.29), we obtain

$$P_0'(1) = \frac{\mu_b - \lambda}{\theta} P_1(1). \tag{7.30}$$

The expression of $P_0'(z)$ can also be found from (7.7) as

$$P_0'(z) = \frac{\left[(\lambda z - \mu_v + \xi)(1-z) + \theta z\right]P_0(z) - \left[(\theta z - (\mu_v-\xi)(1-z))p_{0,0} + \mu_b z p_{1,1}\right]}{\xi z(1-z)}. \tag{7.31}$$

Applying L'Hopital's rule

$$\lim_{z\to 1} P_0'(z) = \lim_{z\to 1} \frac{[\lambda(1-2z)+\mu_v-\xi+\theta]\,P_0(z)-[\{\theta+(\mu_v-\xi)\}p_{0,0}+\mu_b p_{1,1}]}{\xi(1-2z)-[(\lambda z-\mu_v+\xi)(1-z)+\theta z]}.$$

Therefore, using (7.28)

$$P_0'(1) = \frac{(\lambda-\mu_v+\xi)\,P_0(1)+(\mu_v-\xi)p_{0,0}}{\xi+\theta}. \tag{7.32}$$

Equations (7.30) and (7.32) imply that

$$\frac{\mu_b-\lambda}{\theta}P_1(1) = \frac{(\lambda-\mu_v+\xi)\,P_0(1)+(\mu_v-\xi)p_{0,0}}{\xi+\theta},$$

which simplifies to

$$[\theta\,(\lambda-\mu_v+\xi)+(\xi+\theta)(\mu_b-\lambda)]\left[p_{0,0}+\frac{\mu_b}{\theta}p_{1,1}\right] = (\xi+\theta)(\mu_b-\lambda)+\theta(\mu_v-\xi)p_{0,0}.$$

Putting $p_{1,1}$ in terms of $p_{0,0}$, we have the expression for $p_{0,0}$ as

$$p_{0,0} = \frac{(\xi+\theta)(\mu_b-\lambda)}{\theta(\mu_v-\xi)+\frac{(\mu_v-\xi)C(1)}{\theta A(1)}\,[\theta(\xi+\mu_b-\mu_v)+\xi(\mu_b-\lambda)]}.$$

□

7.1.2 Performance measures

From (7.28), the probability of the system being in WV becomes

$$P_0(1) = \frac{(\mu_v-\xi)C(1)}{\theta A(1)}p_{0,0} \tag{7.33}$$

and the probability that the system is in non-vacation period is

$$P_1(1) = 1-P_0(1) = 1-\frac{(\mu_v-\xi)C(1)}{\theta A(1)}p_{0,0}. \tag{7.34}$$

The mean number of customers when the system is in WV period is

$$E(N_0) = P_0'(1) = \lim_{z\to 1} P_0'(z) = \frac{\mu_b-\lambda}{\theta}\left[1-\frac{(\mu_v-\xi)C(1)}{\theta A(1)}p_{0,0}\right] \tag{7.35}$$

and when the server is in non-vacation period is

$$E(N_1) = \lim_{z\to 1} P_1'(z) = \frac{\theta}{\mu_b}E(N_0) = \frac{\mu_b-\lambda}{\mu_b}\left[1-\frac{(\mu_v-\xi)C(1)}{\theta A(1)}p_{0,0}\right]. \tag{7.36}$$

Hence, the mean number of customers in the system is

$$E(N) = E(N_0) + E(N_1) = (\mu_b - \lambda)\left(\frac{1}{\theta} + \frac{1}{\mu_b}\right)\left[1 - \frac{(\mu_v - \xi)C(1)}{\theta A(1)}p_{0,0}\right]. \tag{7.37}$$

Using Little's law, the mean waiting time in the system is

$$E(W) = E(N)/\lambda.$$

Another important measure of performance is the total waiting time of a customer who completes its service before leaving the system. Let this be denoted by W_{served}. Let W_{nj} denote the conditional waiting time of a customer who does not abandon the system, given that the state upon arrival is (n, j). The waiting time of a customer is measured from the moment of arrival until departure, either after completion of service or due to abandonment. Then

$$E(W_{n1}) = \frac{n+1}{\mu_b},\ n = 1, 2, 3, \ldots. \tag{7.38}$$

For $j = 0$ and $n \geq 1$,

$$\begin{aligned} E(W_{n0}) &= \frac{\theta}{\theta + \lambda + \mu_v + (n+1)\xi}\left[\frac{1}{\theta + \lambda + \mu_v + (n+1)\xi} + E(W_{n1})\right] \\ &+ \frac{\lambda}{\theta + \lambda + \mu_v + (n+1)\xi}\left[\frac{1}{\theta + \lambda + \mu_v + (n+1)\xi} + E(W_{n0})\right] \\ &+ \frac{(n+1)\xi}{\theta + \lambda + \mu_v + (n+1)\xi}\left(\frac{n}{n+1}\right)\left[\frac{1}{\theta + \lambda + \mu_v + (n+1)\xi} + E(W_{n-1,0})\right] \\ &+ \frac{\mu_v}{\theta + \lambda + \mu_v + (n+1)\xi}\left[\frac{1}{\theta + \lambda + \mu_v + (n+1)\xi} + E(W_{n-1,0})\right]. \end{aligned}$$

In the second term of the above expression, the arrival of a customer does not change the waiting time of a customer present in the system and in the third term only n customers can abandon the system as our customer is not impatient. This expression can be further rewritten as

$$E(W_{n0}) = \frac{1}{\theta + \mu_v + (n+1)\xi}\left[\frac{\theta + \lambda + \mu_v + n\xi}{\theta + \lambda + \mu_v + (n+1)\xi} + \frac{\theta(n+1)}{\mu_b} + (\mu_v + n\xi)E(W_{n-1,0})\right]. \tag{7.39}$$

For $j = 0$ and $n = 0$,

$$\begin{aligned} E(W_{00}) &= \frac{\theta}{\theta + \lambda + \mu_v + \xi}\left(\frac{1}{\theta + \lambda + \mu_v + \xi} + \frac{1}{\mu_b}\right) + \frac{\lambda}{\theta + \lambda + \mu_v + \xi} \\ &\times \left(\frac{1}{\theta + \lambda + \mu_v + \xi} + E(W_{00})\right) + \frac{\mu_v}{\theta + \lambda + \mu_v + \xi}\left(\frac{1}{\theta + \lambda + \mu_v + \xi}\right), \end{aligned}$$

which can be simplified to

$$E(W_{00}) = \frac{1}{\theta + \mu_v + \xi}\left[\frac{\theta + \lambda + \mu_v}{\theta + \lambda + \mu_v + \xi} + \frac{\theta}{\mu_b}\right]. \tag{7.40}$$

Using (7.40) and iterating (7.39), we obtain for $n \geq 0$,

$$\begin{aligned} E(W_{n0}) &= \frac{1}{\theta + \mu_v + (n+1)\xi}\left[\frac{\theta + \lambda + \mu_v + n\xi}{\theta + \lambda + \mu_v + (n+1)\xi} + \frac{(n+1)\theta}{\mu_b}\right. \\ &\quad \left. + \sum_{k=1}^{n}\left(\frac{\theta + \lambda + \mu_v + (k-1)\xi}{\theta + \lambda + \mu_v + k\xi} + \frac{k\theta}{\mu_v}\right)\prod_{i=k}^{n}\left(\frac{\mu_v + i\xi}{\theta + \mu_v + i\xi}\right)\right]. \end{aligned} \tag{7.41}$$

Finally, we get the mean waiting time of customers served by the system as

$$E(W_{served}) = \sum_{n=0}^{\infty} p_{n,0} E(W_{n0}) + \sum_{n=1}^{\infty} p_{n,1} E(W_{n1}),$$

which after using (7.38) becomes

$$E(W_{served}) = \sum_{n=0}^{\infty} p_{n,0} E(W_{n0}) + \frac{E(N_1) + P_1(1)}{\mu_b}. \tag{7.42}$$

7.1.3 Stochastic decomposition in MWV model

Theorem 7.1.3. *For $\rho < 1$ and $\xi < \mu_v$, the stationary queue length N can be decomposed into a sum of two independent random variables*

$$N = N_c + N_d,$$

where N_c is the queue length of a classical M/M/1 queue with impatient customers and without vacations; and N_d is the additional queue length due to the effect of multiple WV with its PGF given by

$$N_d(z) = \frac{p_{0,0}}{(\mu_b - \lambda)(1 - z)}\left[\frac{(\mu_b - \lambda z)(1 - z) - \theta z}{p_{0,0}} P_0(z) + \frac{(\mu_v - \xi)C(1)}{A(1)} z\right]. \tag{7.43}$$

Proof.

$$\begin{aligned}
N(z) &= P_0(z) + P_1(z) \\
&= P_0(z) + \frac{\theta z P_0(z) - z\left[\theta p_{0,0} + \mu_b p_{1,1}\right]}{(\lambda z - \mu_b)(1-z)} \\
&= \left[1 + \frac{\theta z}{(\lambda z - \mu_b)(1-z)}\right] P_0(z) - \frac{(\mu_v - \xi)C(1)z}{(\lambda z - \mu_b)(1-z)A(1)} p_{0,0} \\
&= \left(\frac{\mu_b - \lambda}{\mu_b - \lambda z}\right) \times \left[\left\{\frac{\mu_b - \lambda z}{\mu_b - \lambda} - \frac{\theta z}{(\mu_b - \lambda)(1-z)}\right\} P_0(z) + \frac{(\mu_v - \xi)C(1)z}{(\mu_b - \lambda)(1-z)A(1)} p_{0,0}\right] \\
&= \left(\frac{1-\rho}{1-\rho z}\right) \times \frac{p_{0,0}}{(\mu_b - \lambda)(1-z)} \left[\frac{(\mu_b - \lambda z)(1-z) - \theta z}{p_{0,0}} P_0(z) + \frac{(\mu_v - \xi)C(1)}{A(1)} z\right] \\
&= \left(\frac{1-\rho}{1-\rho z}\right) \times N_d(z).
\end{aligned}$$

$P_0(z)$ and $P_1(z)$, being PGFs, are positive and so $P_0(z) + P_1(z) > 0$ and $\left(\frac{1-\rho}{1-\rho z}\right) > 0$, for $0 < z < 1$ and $\rho < 1$. Therefore, $N_d(z)$ is positive. Also, for $z = 1$, $N_d(1) = 1$. Hence, $N_d(z)$ is a PGF. □

Theorem 7.1.4. *If $\rho < 1$ and $\xi < \mu_v$, the stationary waiting time W can be decomposed into a sum of two independent random variables*

$$W = W_c + W_d,$$

where W_c is the waiting time of a customer corresponding to a classical M/M/1 queue with impatient customers and without vacations, has an exponential distribution with parameter $\mu_b(1-\rho)$; and W_d is the additional delay due to the effect of multiple WV with its LST given by

$$W_d^*(s) = \frac{p_{0,0}}{(\mu_b - \lambda)s}\left[\frac{(\mu_b - s - \lambda)s - \theta(\lambda - s)}{p_{0,0}} P_0\left(1 - \frac{s}{\lambda}\right) + (\lambda - s)\frac{(\mu_v - \xi)C(1)}{A(1)}\right] \tag{7.44}$$

Proof. From the distributional form of Little's law [91], we have the relation

$$N(z) = W^*(\lambda(1-z)).$$

Let $s = \lambda(1-z)$ which gives $z = \left(1 - \frac{s}{\lambda}\right)$ and $1 - z = \frac{s}{\lambda}$. Putting these relations in (7.43), we get the desired expression. □

7.2 Single working vacation model

The $M/M/1$ queue with impatience and SWV is different from the MWV model in a way that when the server returns from its WV period and finds no customer in the system, it does not go for another vacation but remains idle until the next arrival. So in a SWV model the server may stay idle, for some period, whereas in the MWV it does not. In a SWV model the customer becomes impatient and activates the impatient timer 'T', if upon its arrival, it finds a non-empty server in WV servicing at a lower rate. For the

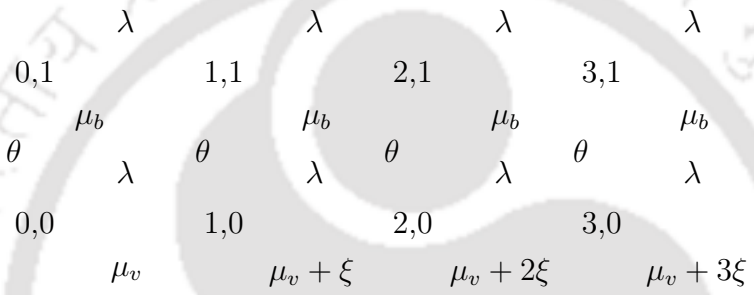

Figure 7.2: State transition diagram for a M/M/1 queue with SWV

SWV model, the Markov chain $\Delta = \{(N_t, Q_t), t \geq 0\}$ can be defined as in MWV case but with the state space $E = \{(i, j), i = 0, 1, \ldots, \ j = 0, 1\}$ and the state transition diagram can be constructed as given in Figure 7.2. Here $j = 1$ if the server is active (idle or busy) and $j = 0$ for the server in WV. We get the balance equations as follows;

$$(\lambda + \theta)p_{0,0} = \mu_v p_{1,0} + \mu_b p_{1,1}, \text{ if } n = 0, \tag{7.45}$$

$$(\lambda + \mu_v + \theta + (n-1)\xi)p_{n,0} = \lambda p_{n-1,0} + (\mu_v + n\xi)p_{n+1,0}, \text{ if } n \geq 1, \tag{7.46}$$

$$\lambda p_{0,1} = \theta p_{0,0}, \text{ if } n = 0, \tag{7.47}$$

$$(\lambda + \mu_b)p_{n,1} = \lambda p_{(n-1),1} + \theta p_{n,0} + \mu_b p_{n+1,1}, \text{ if } n \geq 1. \tag{7.48}$$

Let us define the PGFs

$$G_0(z) = \sum_{n=0}^{\infty} z^n p_{n,0}, \ 0 < z < 1 \tag{7.49}$$

$$G_1(z) = \sum_{n=0}^{\infty} z^n p_{n,1}, \ 0 < z < 1. \tag{7.50}$$

We have $\sum_{n=0}^{\infty} p_{0,n} + \sum_{n=0}^{\infty} p_{1,n} = 1$ *i.e.,* $G_0(1) + G_1(1) = 1$. Multiplying equation (7.46) with z^n and summing over n implies,

$$\xi z(1-z)G_0'(z) - \left[(\lambda z - \mu_v + \xi)(1-z) + \theta z\right] G_0(z) + \left[(\xi - \mu_v)(1-z)\right] p_{0,0} + \mu_b z p_{1,1} = 0. \tag{7.51}$$

For limit z tending to 1, we get

$$\theta G_0(1) = \mu_b p_{1,1}.$$

For $0 < z < 1$,

$$G_0'(z) - \left[\frac{\lambda z - \mu_v + \xi}{z\xi} + \frac{\theta}{\xi(1-z)}\right] G_0(z) + \frac{\xi - \mu_v}{z\xi} p_{0,0} + \frac{\mu_b}{\xi(1-z)} p_{1,1} = 0. \tag{7.52}$$

Solving this differential equation, as in the MWV case, we get

$$G_0(z) = z^{-\left(\frac{\mu_v}{\xi} - 1\right)} (1-z)^{-\frac{\theta}{\xi}} \left[\left(\frac{\mu_v}{\xi} - 1\right) p_{0,0} C(z) - \frac{\mu_b}{\xi} p_{1,1} A(z)\right]. \tag{7.53}$$

Here, we have

$$G_0(0) = p_{0,0} = \frac{\mu_b A(1)}{(\mu_v - \xi)\, C(1)} p_{1,1}$$

which gives,

$$p_{1,1} = \frac{(\mu_v - \xi) C(1)}{\mu_b A(1)} p_{0,0}.$$

Therefore, we get an expression similar as to the case of MWV model

$$G_0(z) = p_{0,0} \left(\frac{\mu_v}{\xi} - 1\right) \left[C(z) - \frac{C(1)}{A(1)} A(z)\right] z^{-\left(\frac{\mu_v}{\xi} - 1\right)} (1-z)^{-\frac{\theta}{\xi}}. \tag{7.54}$$

Multiplying (7.48) with z^n and summing over n gives

$$(\lambda z - \mu_b)\,(1-z) G_1(z) = \theta z G_0(z) - \mu_b (1-z) p_{0,1} - \mu_b z p_{1,1}.$$

For $0 < z < 1$,

$$G_1(z) = \left[\frac{\theta}{\mu_b} z G_0(z) - \left\{(1-z)\frac{\theta}{\lambda} + z\frac{(\mu_v - \xi) C(1)}{\mu_b A(1)}\right\} p_{0,0}\right] \left[(1+\rho)z - \rho z^2 - 1\right]^{-1}. \tag{7.55}$$

To find $G_0(z)$ and $G_1(z)$ explicitly, we have to find the explicit expression of $p_{0,0}$. Applying L'Hopital's rule in (7.55),

$$G_0^{'}(1) = \left(\frac{\mu_b - \lambda}{\theta}\right) G_1(1) - \left(\frac{\mu_b}{\lambda}\right) p_{0,0}. \tag{7.56}$$

Applying L'Hopital's rule in (7.52) gives

$$\lim_{z\to 1} G_0^{'}(z) = \lim_{z\to 1} \frac{[\lambda(1-2z) + \mu_v - \xi + \theta]\, G_0(z) - (\mu_v - \xi)p_{0,0} - \mu_b p_{1,1}}{\xi(1-2z) - (\lambda z - \mu_v + \xi)(1-z) - \theta z}, \tag{7.57}$$

which becomes

$$G_0^{'}(1) = \frac{(\lambda - \mu_v + \xi)G_0(1) + (\mu_v - \xi)p_{0,0}}{\xi + \theta}. \tag{7.58}$$

Equating $G_0^{'}(1)$ from (7.56) and (7.58)

$$[(\lambda - \mu_v + \xi) + (\xi + \theta)(\mu_b - \lambda)]\, G_0(1) = (\xi + \theta)(\mu_b - \lambda) - \left(\frac{\mu_b}{\lambda}\right)\theta(\xi + \theta)p_{0,0}.$$

We finally get

$$p_{0,0} = \frac{(\xi + \theta)(\mu_b - \lambda)}{\theta(\mu_v - \xi) + \frac{\mu_b}{\lambda}\theta(\xi + \theta) + \frac{(\mu_v - \theta)C(1)}{\theta A(1)}\left[\theta(\xi + \mu_b - \mu_v) + \xi(\mu_b - \lambda)\right]}. \tag{7.59}$$

As we have given in the MWV case, we give below the performance measures of this model in a similar manner.

7.2.1 Performance measures

The system performances of this SWV impatient model are given below. The probability of the system being in WV is

$$G_0(1) = \frac{(\mu_v - \theta)C(1)}{\theta A(1)} p_{0,0} \tag{7.60}$$

and that if the system being in non-vacation period is

$$G_1(1) = 1 - G_0(1) = 1 - \frac{(\mu_v - \theta)C(1)}{\theta A(1)} p_{0,0}. \tag{7.61}$$

The mean number of customers during WV period can be given as

$$E(N_0) \;=\; \lim_{z\to 1} G_0^{'}(z) = \frac{\mu_b - \lambda}{\theta}\left[1 - \frac{(\mu_v - \theta)C(1)}{\theta A(1)} p_{0,0}\right] - \frac{\mu_b}{\lambda} p_{0,0} \tag{7.62}$$

and the mean number of customers during non-vacation period as

$$E(N_1) = \lim_{z\to 1} G_1'(z) = \frac{\mu_b - \lambda}{\mu_b}\left[1 - \frac{(\mu_v - \theta)C(1)}{\theta A(1)}p_{0,0}\right]. \tag{7.63}$$

The average number of customers in the system will be

$$E(N) = E(N_0) + E(N_1) = (\mu_b - \lambda)\left(\frac{1}{\theta} + \frac{1}{\mu_b}\right)\left[1 - \frac{(\mu_v - \theta)C(1)}{\theta A(1)}p_{0,0}\right] - \frac{\mu_b}{\lambda}p_{0,0}. \tag{7.64}$$

The average waiting time of customers in system who are served before leaving, W_{served}, can be derived similarly as in the MWV case with $E(W_{n1}) = \frac{n+1}{\mu_b}$ for $n = 0, 1, 2, \ldots$ (rather than $n \geq 1$) and $E(W_{n0})$ given by (7.39). Finally, $E(W_{served})$ is given by (7.42) but with the second sum starting from $n = 0$.

7.2.2 Stochastic decomposition in SWV model

Theorem 7.2.1. *For $\rho < 1$ and $\xi < \mu_v$, the stationary queue length N can be decomposed into a sum of two independent random variables*

$$N = N_c + N_d,$$

where N_c is the queue length of a classical M/M/1 queue with impatient customers and without vacations, and N_d is the additional queue length due to the effect of single WV with its PGF given by

$$N_d(z) = \frac{p_{0,0}}{(\mu_b - \lambda)(1-z)}\left[\frac{(\mu_b - \lambda z)(1-z) - \theta z}{p_{0,0}}P_0(z) + \mu_b\left\{(1-z)\frac{\theta}{\lambda} + z\frac{(\mu_v - \theta)C(1)}{A(1)}\right\}\right]$$

and $P(N_d = 0) = \frac{\mu_b}{\mu_b - \lambda}p_{0,0}$.

Theorem 7.2.2. *If $\rho < 1$ and $\xi < \mu_v$, the stationary waiting time W can be decomposed into a sum of two independent random variables*

$$W = W_c + W_d,$$

where W_c is the waiting time of a customer corresponding to a classical M/M/1 queue with impatient customers and without vacations, has exponential distribution with parameter

$\mu_b(1-\rho)$; *and* W_d *is the additional delay due to the effect of single WV with its LST given by*

$$W_d^*(s) = \frac{p_{0,0}}{(\mu_b-\lambda)s}\left[\frac{(\mu_b-s-\lambda)s-\theta(\lambda-s)}{p_{0,0}}P_0\left(1-\frac{s}{\lambda}\right)+\mu_b\left\{\frac{s\theta}{\lambda}+(\lambda-s)\frac{(\mu_v-\theta)C(1)}{A(1)}\right\}\right]. \tag{7.65}$$

These results can be proved similar to the MWV case in the previous section.

7.3 Comparison of the models

We provide here an analytic comparison of the MWV and SWV models. Since customer waiting time is an important measure of system efficiency, a model can be said to be more efficient than another, if its mean waiting time of the customers is less compared to the other. For a M/M/1/WV queue without impatience, a MWV model is always better than a SWV in the sense that the mean queue length in MWV is always less than that in the SWV model. This result can be verified from the expressions for mean queue length of MWV model and SWV model in [113] and [154] respectively. From the definitions of MWV and SWV policy, it seems that, with impatient customers the SWV policy is more efficient compared to the MWV one. But because of impatience the behavior differs.

To differentiate the terms of both the models, we will use superscript M and S for the MWV and SWV models respectively. From the expressions of $p_{0,0}$ in (7.26) and (7.59), it can be seen that the probability of SWV is less than that of MWV.

$$\frac{1}{p_{0,0}^S} = \frac{1}{p_{0,0}^M} + \frac{\mu_b\theta}{\lambda(\mu_b-\lambda)}, \tag{7.66}$$

which gives $p_{0,0}^S < p_{0,0}^M$. The mean number of customers in these systems, when in non-vacation, are

$$\begin{aligned} E^M(N_1) &= \frac{\mu_b-\lambda}{\mu_b}\left[1-\frac{(\mu_v-\theta)C(1)}{\theta A(1)}p_{0,0}^M\right] \\ \text{and } E^S(N_1) &= \frac{\mu_b-\lambda}{\mu_b}\left[1-\frac{(\mu_v-\theta)C(1)}{\theta A(1)}p_{0,0}^S\right], \end{aligned}$$

which implies

$$E^M(N_1) < E^S(N_1). \tag{7.67}$$

For the system on vacation,

$$\begin{aligned} E^M(N_0) &= \frac{\mu_b - \lambda}{\theta} - \left[\frac{(\mu_b - \lambda)(\mu_v - \theta)C(1)}{\theta^2 A(1)}\right] p^M_{0,0} \\ \text{and} \quad E^S(N_0) &= \frac{\mu_b - \lambda}{\theta} - \left[\frac{(\mu_b - \lambda)(\mu_v - \theta)C(1)}{\theta^2 A(1)} + \frac{\mu_b}{\lambda}\right] p^S_{0,0}; \end{aligned}$$

and the total queue lengths are

$$\begin{aligned} E^M(N) &= (\mu_b - \lambda)\left(\frac{1}{\theta} + \frac{1}{\mu_b}\right)\left[1 - \frac{(\mu_v - \xi)C(1)}{\theta A(1)} p_{0,0}\right] \\ \text{and} \quad E^S(N) &= (\mu_b - \lambda)\left(\frac{1}{\theta} + \frac{1}{\mu_b}\right)\left[1 - \frac{(\mu_v - \theta)C(1)}{\theta A(1)} p_{0,0}\right] - \frac{\mu_b}{\lambda} p_{0,0}. \end{aligned}$$

From the expressions of $E^M(N_0)$ and $E^S(N_0)$, we cannot conclude any relation between them. Therefore the sum, $E^M(N)\big(= E^M(N_0) + E^M(N_1)\big)$, may be greater than, equal or less than $E^S(N)$. For example, let us take an impatience model in which $\lambda = 0.7, \mu_b = 1$ and $\mu_v = 0.5$. For three different values of θ the mean queue lengths of the system are shown in Figure 7.3. When $\theta = 1.7$ the queue length of the SWV model is less than the MWV model and for $\theta = 2.1$, the MWV model gives lesser queue length. If $\theta = 1.9$, it is seen that at point $\xi = 0.2$ both the models give the same value. So, for $\xi < 0.2$, the SWV model becomes better compared to the MWV one and, for $\xi > 0.2$, the MWV model works better. This means that, when the mean of vacation duration time is less, the MWV model gives better performance than the SWV model.

In Figure 7.4, we have compared the mean queue lengths during the WV periods of both the models. The parameter values are $\lambda = 1.8, \mu_b = 2, \mu_v = 1.3$ and $\theta = 0.09$. The difference between the queue lengths increases with the increase in impatient rate ξ. So, if the impatient rate is small, the MWV and SWV models give the same queue lengths during WV periods but for higher impatient rates SWV model will give better performance. This study reveals that not only the vacation duration rate but the impatience rate also plays a major role in the performance of the model. Thus, for a given set of parameter values, we can choose a model to get better performance of a single-server impatient queue with WVs.

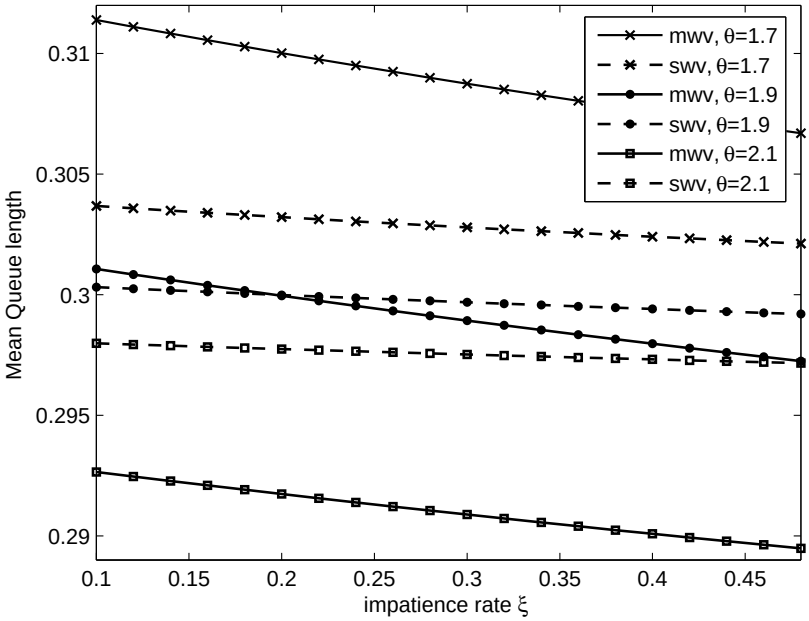

Figure 7.3: Mean queue length of system vs impatient rate.

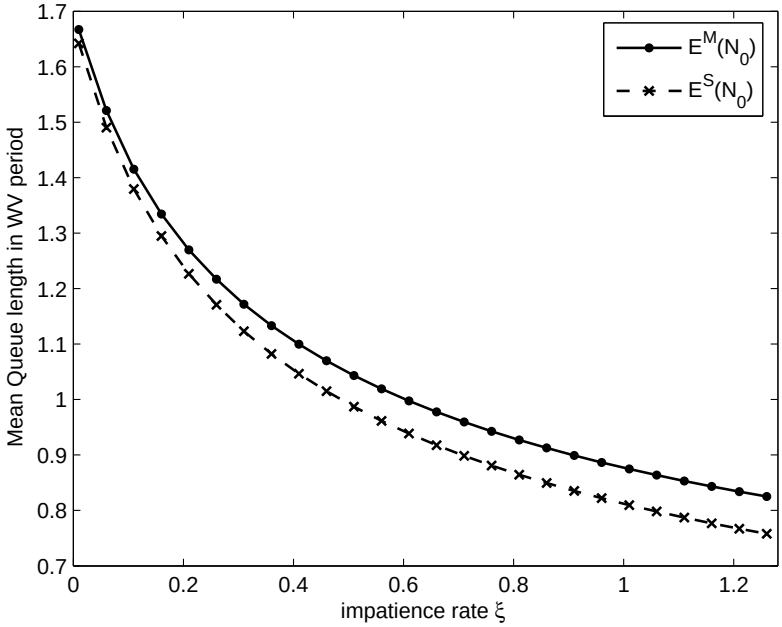

Figure 7.4: Mean queue length during WV vs impatient rate.

Chapter 8

Conclusions

Performance analysis of communication networks deals with the evaluation of the level of efficiency that a network achieves. Queueing theory, with its latest methodologies, has become an effective tool to model and analyze such complex systems. The ultimate objective of the queueing theoretic analysis of systems is to understand the behavior of the system and to make intelligent decisions in their management. WDM is an effective and emerging method of transmitting packets in communication networks to satisfy the demand of fast transmission over links of high capacity. A WDM system can be modelled as a queue with working vacations, a new type of vacation policy different from the classical vacation policy. In a WV policy, the service does not cease completely during vacations unless the system is empty. We study, in this thesis, some WV queues incorporating some special features of networks which leave a prominent impact on system performances.

A MAP/PH/1/WV queue is considered which captures both burstiness and correlation in the interarrival times. The model is analyzed in discrete-time as well as in continuous-time. The constructed QBD is solved using the matrix-geometric solution method to obtain the stationary probabilities. The stability condition and stationary queue length distribution have been derived. The waiting time distribution, the distribution of regular busy period and expressions for vacation durations and the busy cycle are also obtained. Certain insights to the models are presented through some numerical illustrations which

include, comparisons of the considered models with equivalent models having independent arrivals and the effects of correlation parameters on the performance measures of the models. We have seen from the analysis that if the interarrivals are not independent, impelling them to be independent just to acquire a simple model, might lead to erroneous prediction of performance measures.

Next, we have dealt with a finite-buffer MAP/PH/1/WV queueing model. We have formulated the system as a QBD process and obtained the stationary distribution. The stationary queue length distribution and customer loss probabilities were also presented. In this finite-buffer model, numerical illustrations focused on evaluation of the system performance in terms of customer loss probability, mean queue length of buffer occupancy and mean waiting time or system delay. The effects of buffer size, traffic burstiness and of correlation upon various performance measures are also shown.

We have considered a PH/M/c/WV model where the server obeys asynchronous multiple working vacation policy. The customers considered here become impatient, if upon their arrival all the servers are busy and the arriving customers have to join a queue. We have formulated the system as a three-dimensional Markov chain whose generator matrix is a level-dependent (non-homogeneous) QBD. Using the finite truncation method, the stationary distribution is computed. The cut-off value has been observed to depend on the number of servers, the interarrival distribution and also on vacation parameters. The mean queue length of the model is compared for various distributions for interarrivals. Effect of system parameters on blocking probability and the average customer loss due to impatience are also studied.

Further, we have taken a M/PH/1/WV queue with two priority classes, a higher priority class and a lower priority class, governed by different Poisson inputs. The working vacations are interrupted if the system finds a customer waiting in queue. We have used the method of embedded Markov chains to analyze the non-preemptive priority model which follows a FCFS discipline within each priority class. We modelled the system as a queue where an exceptional service time was needed for the first customer in each non-idle period. The probability generating function, for the number of higher priority customers

at service starting epochs, is derived first. Laplace-Stieltjes transform of the probability distribution function for the waiting time of the higher priority customers and the delay busy periods are used to find the waiting times of lower priority customers. The average response time of a customer is also given for each priority class. Numerical experiments are carried out to quantify the difference in waiting times for both the priority classes. The increase in blocking probability with the increased system traffic intensity is shown for higher priority customers and the increased rate is compared for different distributions of interarrival times.

Another model we studied is the M/G/1/WV retrial model. An embedded Markov process is defined by applying the supplementary variable technique and system reliability and probability of a blocked server are derived. The distributions of number of customers in the orbit and in the system are given. We have proved that the retrial model with the working vacation scheme also satisfies the stochastic decomposition property for stationary queue lengths and stationary waiting times. We have observed that increase in retrial rate can increase the system availability when the durations of vacation are large. The average waiting time in the queue can be reduced by higher vacation-service rates and higher retrial rates. Our studies quantify the system efficiency when the blocked requests are allowed to retry and shows the amount of network availability that can be achieved by repeated attempts. The choice of service distribution also plays an important role in such models.

Finally, we have investigated the effect of impatient behavior of customers in a M/M/1/WV model. Two types of working vacation termination policies are taken, the multiple working vacation (MWV) policy and the single working vacation (SWV) policy. Closed-form probabilities are derived using the identities involving beta functions and degenerate hypergeometric functions. The system performances are measured for an average number of customers in WV period as well as in non-vacation period. Stochastic decomposition properties are verified for both MWV and SWV cases. This work underlines the fact that if the system is not in WV, average number of customers in the MWV model is always less than the SWV model. We carried out numerical experiments and found that when the

average vacation duration time is less, the MWV model gives better performance than the SWV model, whereas for higher impatient rates the SWV model becomes more efficient as it reduces the waiting time of the customers in the system.

Future Directions

We have studied working vacation queueing models, mostly seen in WDM networks incorporating certain characteristics. There is ample scope for future extension of this study, a few examples of which are mentioned below.

- Throughout the thesis, we had considered the arrival pattern of customers to be single arrivals. We can further extend the analysis for customers with batch arrivals, since communication networks mostly deal with bulk arrivals of packets. Not only the arrivals but the service patterns in batches are also found in networks.
- The vacation starting policy considered here is the exhaustive one, where the system goes to vacation only if the system becomes empty. Analysis of WV systems with other policies, like the N-policy and gated vacation policies is another direction for research.
- Approximation method used in Chapter 4 is the finite truncation method. There are other approximation methods of finding the stationary distribution and comparisons of various methods and error analysis of those algorithms are still to be done for WV models.

References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover publication, 1972.

[2] A.S. Alfa. A discrete MAP/PH/1 queue with vacations and exhaustive time-limited service. *Operations Research Letters*, 18:31–40, 1995.

[3] A.S. Alfa. Matrix-geometric solution of discrete time MAP/PH/1 priority queue. *Naval Research Logistics*, 45:23–50, 1998.

[4] A.S. Alfa. Discrete time queues and matrix-analytic methods. *TOP*, 10(2):147–210, 2002.

[5] A.S. Alfa. Vacation models in discrete time. *Queueing Systems*, 44:5–30, 2003.

[6] A.S. Alfa and I. Frigui. Discrete NT-policy single server queue with Markovian arrival process and phase type service. *European Journal of Operational Research*, 88:599–613, 1996.

[7] E. Altman and U. Yechiali. Analysis of customer's impatience in queue with server vacations. *Queueing Systems*, 52:261–279, 2006.

[8] E. Altman and U. Yechiali. Infinite-server queues with system's additional tasks and impatient customers. *Probability in the Engineering and Informational Sciences*, 22:477–493, 2008.

[9] H. Anand, C. Reardon, R. Subramaniyan, and A.D. George. Ethernet adaptive link rate (ALR): Analysis of a MAC handshake protocol. *Proceedings of the IEEE Conference on Local Computer Networks*, pages 533–534, 2006.

[10] L.B. Aronson, B.E. Lemoff, and L.A. Buckman. Low-cost multimode WDM for local area networks up to 10 gb/s. *IEEE Photonics Technology Letters,*, 10(10):1489–1491, 1998.

[11] J.R. Artalejo. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers and Operations Research*, 24:493–504, 1997.

[12] J.R. Artalejo. Accessible bibliography on retrial queues: Progress in 1990–1999. *Mathematical and Computer Modelling*, 30:1–6, 1999.

[13] J.R. Artalejo. Algorithmic analysis of the Geo/Geo/c retrial queue. *European Journal of Operational Research*, 189:1042–1056, 2008.

[14] J.R. Artalejo. Accessible bibliography on retrial queues: next term progress in 2000-2009. *Mathematical and Computer Modelling*, In press, 2010.

[15] J.R. Artalejo and S.R. Chakravarthy. Algorithmic analysis of the MAP/PH/1 retrial queue. *TOP*, 14(2):293–332, 2006.

[16] J.R. Artalejo and G.I. Falin. Stochastic decomposition for retrial queues. *TOP*, 2:329–342, 1994.

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

[18] J.R. Artalejo and M. Pozo. Numerical calculation of the stationary distribution of the main multi-server retrial queue. *Annals of Operations Research*, 116:41–56, 2002.

[19] I. Atencia, G. Bouza, and P. Moreno. An $M^{[X]}/G/1$ retrial queue with server breakdowns and constant rate of repeated attempts. *Annals of Operations Research*, 157:225–243, 2008.

[20] I. Atencia, I. Fortes, P. Moreno, and S. Sanchez. An M/G/1 retrial queue with active breakdowns and Bernoulli schedule in the server. *International Journal of Information and Management Sciences*, 17(1):1–17, 2006.

[21] M.C. Ausina, R.E. Lillob, and M.P. Wiper. Bayesian control of the number of servers in a GI/M/c queueing system. *Journal of Statistical Planning and Inference*, 137(10):3043–3057, 2007.

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

[23] F. Baccelli, P. Boyer, and G. Hebuterne. Single server queues with impatient customers. *Advances in Applied Probability*, 16:887–905, 1984.

[24] F. Baccelli and G. Hebuterne. On queues with impatient customers. *Performance '81, Amsterdam*, pages 159–179, 1981.

[25] A.D. Banik, U.C. Gupta, and S.S. Pathak. Finite buffer vacation models under e-limited with limit variation service and markovian arrival process. *Operations Research Letters*, 34(5):539–547, 2006.

[26] A.D. Banik, U.C. Gupta, and S.S. Pathak. On the GI/M/1/N queue with multiple working vacations-analytic analysis and computation. *Applied Mathematical Modelling*, 31(9):1701–1710, 2007.

[27] D.Y. Barrer. Queuing with impatient customers and ordered service. *Operations Research*, 5:650–656, 1957.

[28] U.N. Bhat and G.K. Miller. *Elements of Applied Stochastic Processes*. Wiley, Third edition, 2002.

[29] C. Blondia. Finite capacity vacation model with non-renewal input. *Journal of Applied Probability*, 28:174–197, 1991.

[30] S. Bocquet. Queueing theory with reneging. *Scientific and Technical Report of DSTO*, 2005.

[31] O.J. Boxma and R. Syski. *Queueing Theory and Its Applications: Liber Amicorum for J.W. Cohen*. North Holland, Amsterdam, 1988.

[32] L. Breuer and D. Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, Netherlands, 2006.

[33] L.W. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death process. *Stochastic Models*, 11:497–525, 1995.

[34] F. Cali and M. Conti. Dynamic tuning of the IEEE 802.11 protocol to achieve a theoritical throughput limit. *IEEE Transactions on Networking*, 8(6):785–799, 2000.

[35] K.C. Chae, H.W. Lee, and C.W. Ahn. An arrival time approach to M/G/1-type queues with generalized vacations. *Queueing Systems*, 38:91–100, 2001.

[36] S. R. Chakravarthy. Analysis of a multi-server queue with Markovian arrivals and synchronous phase type vacations. *Asia-Pacific Journal of Operational Research*, 26(1):85–113, 2009.

[37] S. R. Chakravarthy, A. Krishnamoorthy, and V. C. Joshua. Analysis of multi-server retrial queues with search of customers from the orbit. *Performance Evaluation*, 63:776–798, 2006.

[38] M.L. Chaudhry, U.C. Gupta, and M. Agarwal. On exact computational analysis of distributions of numbers in systems for M/G/1/N + 1 and GI/M/1/N + 1 queues using roots. *Computers and Operations Research*, 18(8):679–694, 1991.

[39] H. Chen, F. Wang, N. Tian, and D. Lu. Study on N-policy working vacation polling system for WDM. *Proceedings of IEEE International Conference on Communication Technology*, pages 394–397, 2008.

[40] H. Chen, F. Wang, N. Tian, and D. Lu. Study on working vacation polling system for WDM with PH distribution service time. *International Symposium on Computer Science and Computational Technology*, pages 426–429, 2008.

[41] D.I. Choi, T.S. Kim, and S. Lee. Analysis of an MMPP/G/1/K queue with queue length dependent arrival rates, and its application to preventive congestion control in telecommunication networks. *European Journal of Operational Research*, 187(2):652–659, 2008.

[42] A. Chydzinski. Packet loss process in queues with Markovian arrivals. *IEEE Proceedings of International Conference on Networking*, pages 524–529, 2008.

[43] J.W. Cohen. *The Single Server Queue*. North Holland, London, 1969.

[44] D.R. Cox. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society*, 51(3):433–441, 1955.

[45] Z. Cui and A.A. Nilsson. The impact of correlation on delay performance of high speed networks. *Southeastern Symposium on System Theory, Athens, Ohio*, 1994.

[46] J.H. Daigle. *Queueing Theory with Applications to Packet Telecommunication*. Springer, Boston, 2005.

[47] D.J. Daley. General customer impatience in the queue GI/G/1. *Journal of Applied Probability*, 2:186–205, 1965.

[48] G. R. Dattatreya. *Performance Analysis of Queuing and Computer Networks*. Chapman and Hall, 2008.

[49] J.N. Diagle and D.M. Lucantoni. *Queueing systems having phase-dependent arrival and service rates: In Numerical Solution of Markov Chains*. Marcel Dekker, New York, 1991.

[50] J.E. Diamond and A.S. Alfa. The MAP/PH/1 retrial queue. *Stochastic Models*, 14:1151–1177, 1998.

[51] T.V. Do. M/M/1 retrial queue with working vacations. *Acta Informatica*, 47(1):67–75, 2010.

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

[53] B.T. Doshi and D.L. Jagerman. An M/G/1 queue with class dependent balking (reneging). *Proceedings of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation*, pages 225–243, 1986.

[54] J.H. Dshalalow. *Frontiers in Queueing Models and Applications in Science and Engineering*. CRC Press, Boca Raton, Florida, 1997.

[55] M.K. Dutta and V.K. Chaubey. Priority based wavelength routed WDM networks: A queueing theory approach. *International Journal of Recent Trends in Engineering*, 1(3):253–256, 2009.

[56] A. Economou and S. Kapodistria. Synchronized abandonments in a single server unreliable queue. *European Journal of Operational Research*, 203(1):143–155, 2010.

[57] G.I. Falin. Calculation of probability characteristics of a multiline system with repeated calls. *Moscow University Computational Mathematics and Cybernetics*, 1:43–49, 1983.

[58] G.I. Falin. The M/M/1 retrial queue with retrials due to server failures. *Queueing Systems*, 58:155–160, 2008.

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

[60] D. Fiems, T. Maertens, and H. Bruneel. Queueing systems with different types of server interruptions. *European Journal of Operational Research*, 188(3):835–845, 2008.

[61] D. Fiems, B. Steyaert, and H. Bruneel. Discrete-time queues with generally distributed service times and renewal-type server interruptions. *Performance Evaluation*, 55(3-4):277–298, 2004.

[62] P.D. Finch. Deterministic customer impatience in the queueing system GI/M/1. *Biometrika*, 47(1,2):45–52, 1960.

[63] I. Frigui and A.S. Alfa. Analysis of a time limited polling system. *Computer Communications*, 21:558–571, 1998.

[64] T.C. Fry. *Probability and Its Engineering Uses*. Van Nostrand, Princeton, NJ, 1928.

[65] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5:79–141, 2003.

[66] F. Gebali. *Analysis of Computer and Communication Networks*. Springer, 2008.

[67] D. Gross, J. Shortle, J. Thompson, and C. Harris. *Fundamentals of Queueing Theory*. Wiley, Fourth edition, 2008.

[68] U.C. Gupta, A.D. Banik, and S.S. Pathak. Complete analysis of MAP/G/1/N queue with single (multiple) vacation(s) under limited service discipline. *Journal of Applied Mathematics and Stochastic Analysis*, 3:353–373, 2005.

[69] U.C. Gupta, S.K. Samanta, R.K. Sarma, and M.L. Chaudhry. Discrete-time single-server finite-buffer queues under discrete Markovian arrival process with vacations. *Performance Evaluation*, 64:1–19, 2007.

[70] U.C. Gupta and K. Sikdar. Computing queue length distributions in MAP/G/1/N queue under single and multiple vacation. *Applied Mathematics and Computation*, 174(2):1498–1525, 2006.

[71] C.S. Ho and C. Woei. Study of re-provisioning mechanism for dynamic traffic in WDM optical networks. *International Conference on Advanced Communication Technology*, pages 288–291, 2008.

[72] P.G. Hoel, S.C. Port, and C.J. Stone. *Introduction to Stochastic Processes*. Thomson Brooks/Cole, 1975.

[73] B.V. Houdt and C. Blondia. Analyzing previous termpriority queues with 3 classes using tree-like processes. *Queueing Systems*, 54(2):99–109, 2006.

[74] J.J. Hunter. *Mathematical Techniques of Applied Probability. Discrete Time Models: Basic Theory*, volume 1. Academic press, New York, 1983.

[75] N. Igaki. Exponential two server queue with N-policy and multiple vacations. *Queueing Systems*, 10:279–294, 1992.

[76] M. Jain and P.K. Agrawal. $M/E_k/1$ queueing system with working vacation. *Quality Technology and Quantitative Management*, 4(4):455–470, 2007.

[77] M. Jain and A. Jain. Working vacations queueing model with multiple types of server breakdowns. *Applied Mathematical Modelling*, 24:1–13, 2010.

[78] N.K. Jaiswal. *Priority Queues*, volume 2. Academic Press, New York, 1968.

[79] X. Jin and G. Min. Comprehensive analytical model for priority queuing with heterogeneous traffic. *Electronics Letters*, 43(24):1395–1396, 2007.

[80] O. Jouini, A. Pot, G. Koole, and Y. Dallery. Online scheduling policies for multiclass call centers with impatient customers. *European Journal of Operational Research*, 2010, In press.

[81] E.S. Jung and N.H. Vaidya. A power control MAC protocol for ad hoc networks. *Wireless Networks*, 11(1-2):55–66, 2005.

[82] O.M. Jurkevic. On the investigation of many-server queueing systems with bounded waiting time. *Izv. Akad. Nauk SSSR Techniceskaja Kibernetika*, 5:50–58, 1970. (in Russian).

[83] O.M. Jurkevic. On many-server systems with stochastic bounds for the waiting time. *Izv. Akad. Nauk SSSR Techniceskaja Kibernetika*, 4:39–46, 1971. (in Russian).

[84] F. Karaesmen and S.M. Gupta. The finite capacity GI/M/1 queue with server vacations. *Journal of the Operational Research Society*, 47:817–828, 1996.

[85] S. Kasahara, T. Takine, Y. Takahashi, and T. Hasegawa. MAP/G/1 queues under N-policy with and without vacations. *Journal of the Operational Research Society of Japan*, 39:188–212, 1996.

[86] T. Katayama. Priority queues with semiexhaustive service. *Queueing Systems*, 21:161–181, 1995.

[87] T. Katayama and K. Kobayashi. Analysis of a nonpreemptive priority queue with exponential timer and server vacations. *Performance Evaluation*, 64(6):495–506, 2007.

[88] J.C. Ke and F. M. Chang. Modified vacation policy for M/G/1 retrial queue with balking and feedback. *Computers and Industrial Engineering*, 57:433–443, 2009.

[89] J.C. Ke and F. M. Chang. Analysis of a batch retrial queue with bernoulli vacation and starting failures. *International Journal of Services Operations and Informatics*, 5(2):95–114, 2010.

[90] J.C. Ke and K.H. Wang. A recursive method for the N-policy G/M/1 queueing system with finite capacity. *European Journal of Operational Research*, 142(3):577–594, 2002.

[91] J. Keilson and L.D. Servi. A distributional form of Littles law. *Operations Research Letters*, 7(5):223–227, 1988.

[92] M. Kijima. *Markov Processes for Stochastic Modeling.* Chapman and Hall, New York, 1997.

[93] L. Kleinrock. *Queueing Systems, Volume I, Theory.* Wiley, New York, 1975.

[94] L. Kleinrock. *Queueing Systems, Volume II, Computer Applications.* Wiley, New York, 1976.

[95] A.G. Kok and H.C. Tijms. A two-moment approximation for a buffer design problem requiring a small rejection probability. *Performance Evaluation*, 5:77–84, 1985.

[96] G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113:41–59, 2002.

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

[98] P. J. Kuehn. Reminder on queueing theory for ATM networks. *Telecommunication Systems*, 5:1–24, 1996.

[99] B.K. Kumar and D. Arivudainambi. The M/G/1 retrial queue with bernoulli schedules and general retrial times. *Computers and Mathematics with Applications*, 43:15–30, 2002.

[100] M.S. Kumar and R.Arumuganathan. Performance analysis of an M/G/1 retrial queue with non-persistent calls, two phases of heterogeneous service and different vacation policies. *International Journal of Open Problems in Computer Science and Mathematics*, 2(2):196–214, 2009.

[101] G. Latouche and V. Ramaswami. A logarithimic reduction algorithm for quasi-birth-death process. *Journal of Applied Probability*, 30:650–674, 1993.

[102] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Applied Probability. 1999.

[103] H.W. Lee, B.Y. Ahn, and N.I. Park. Decomposition of the queue length distributions in the MAP/G/1 queue under multiple and single vacations with N-policy. *Stochastic Models*, 17:157–190, 2001.

[104] Y. Levy and U. Yechiali. An M/M/c queue with servers vacations. *INFOR*, 14:153–163, 1976.

[105] H. Li and T. Yang. A single-server retrial queue with server vacations and a finite number of input sources. *European Journal of Operational Research*, 85:149–160, 2005.

[106] J. Li and N. Tian. The discrete-time GI/Geo/1 queue with working vacations and vacation interruption. *Applied Mathematics and Computation*, 185:1–10, 2007.

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

[108] J. Li, N. Tian, and Z. Ma. Performance analysis of GI/M/1 queue with working vacations and vacation interruptions. *Applied Mathematical Modelling*, 32:2715–2730, 2008.

[109] X.X. Li, Z.J. Zhang, and N. Tian. Analysis for the $M^{[X]}$/M/1 working vacation queue. *International Journal of Information and Management Sciences*, 20:379–394, 2009.

[110] J. Liao, L. Lemin, and S. Hairong. The influence of burstiness and correlation of traffic on an ATM multiplexer. *International Conference on Communication Technology, Beijing, China*, pages 20–23, 1996.

[111] C.H. Lin and J. Ke. Multi-server system with single working vacation. *Applied Mathematical Modelling*, 33(7):2967–2977, 2009.

[112] J.D.C. Little. A proof for the queueing formula L=λ W. *Operations Research*, 9:383–387, 1961.

[113] W. Liu, X. Xu, and N. Tian. Stochastic decompositions in the M/M/1 queue with working vacations. *Operations Research Letters*, 35:595–600, 2007.

[114] D.M. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts. A single-server queue with server vacations and a class of non-renewal process. *Advances in Applied Probability*, 22:676–705, 1990.

[115] Z. Lui and J. Wu. An MAP/G/1 G-queues with preemptive resume and multiple vacations. *Applied Mathematical Modelling*, 33(3):1739–1748, 2009.

[116] M. Mandjes. *Large Deviations for Gaussian Queues, Modelling Communication Networks*. Wiley, England, 2007.

[117] J. Medhi. *Stochastic Models in Queueing Theory*. Academic Press, Second edition, 2003.

[118] E.C. Molina. Application of the theory of probability to telephone trunking problems. *Bell Systems Technical Journal*, 6:461–494, 1953.

[119] P.M. Morse. *Queues, Inventories and Maintenance : The Analysis of Operational Systems with Variable Demand and Supply*. Dover, 2004.

[120] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. John Hopkins University Press, Baltimore, 1981.

[121] M.F. Neuts. *Structured Stochastic Matrices of M/G/1 type and their Applications*. Marcel Dekker, New York, 1989.

[122] M.F. Neuts and B.M. Rao. Numerical investigation of a multiserver retrial model. *Queueing Systems*, 7:169–190, 1990.

[123] G.F. Newell. Approximations for superposition arrival processes in queues. *Management Science*, 7:623–632, 1984.

[124] Z. Niu, T. Shu, and Y. Takahashi. A vacation queue with set up and close-down times and batch Markovian arrival processes. *Performance Evaluation*, 54:225–248, 2003.

[125] Z. Niu and Y. Takahashi. A finite-capacity queue with exhaustive vacation/close-down/setup times and Markovian arrival processes. *Queueing Systems*, 31:1–23, 1999.

[126] C. Palm. Methods of judging the annoyance caused by congestion. *Telecommunications*, 4:153–163, 1953.

[127] H.T. Papadopoulos, C. Heavey, and J. Browne. *Queueing Theory in Manufacturing Systems- Analysis and Design*. Springer, 1993.

[128] P.R. Parthasarathy and R.B. Lenin. Applied birth and death models - A time-dependent perspective. *American Journal of Mathematical and Management Sciences*, 24:1–216, 2004.

[129] N. Perel and U. Yechial. Queues with slow servers and impatient customers. *European Journal of Operational Research*, 201(1):247–258, 2010.

[130] A. Pilehvar. *New Queueing Network Approximations for Vaccination Clinics - Studying the Batch Arrival, Batch Service Processes and Stations with no Real Servers*. Vdm Verlag, 2007.

[131] C. Qiao and M. Yoo. Optical burst switching (OBS) - A new paradigm for an optical internet. *Journal of High Speed Networks*, 8(1):69–84, 1999.

[132] X. Qin and Y. Yang. Blocking probability in WDM multicast switching networks with limited wavelength conversion. *Proceedings of IEEE International Symposium on Network Computing and Applications*, pages 322–332, 2003.

[133] A. Rajabi, A. Khonsari, and A. Dadlani. On modeling optical burst switching networks with fiber delay lines: A novel approach. *Computer Communications*, 33(2):240–249, 2010.

[134] R. Ramaswami and K.N. Sivarajan. *Optical Networks: A Practical Perspective*. Second. San Mateo, CA: Morgan Kaufmann, 2002.

[135] V. Ramaswami and D.V. Lucantoni. On the merits of an approximation to the busy period of GI/G/1 queue. *Management Sciences*, 25:285–289, 1979.

[136] V. Ramaswami and M.F. Neuts. Some explicit formulas and computation methods for infinite server queues with ohase type arrivals. *Journal of Applied Probability*, 17:498–514, 1980.

[137] T.L. Saaty. *Elements of Queueing Theory and Applications*. McGraw Hill, New York, 1961.

[138] L.D. Servi. Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules. *IEEE Journal on Selected Areas in Communications*, 4(6):813–822, 1986.

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

[140] R.E. Stanford. Reneging phenomena in single server queues. *Mathematics of Operations Research*, 4:162–178, 1979.

[141] R.E. Stanford. On queues with impatience. *Advances in Applied Probability*, 22:768–769, 1990.

[142] R. Syski. *Introduction to Congestion Theory in Telephone Systems*. Oliver and Boyd, Edinburgh, 1960.

[143] H. Takagi. *Queueing Analysis; A Foundation of Performance Evaluation, Volume 1, Vacation and Priority Systems*. North Holland, New York, 1991.

[144] H. Takagi. *Queueing Analysis; A Foundation of Performance Evaluation, Volume 2, Finite Systems*. North Holland, New York, 1993.

[145] A.N. Tam, S.G. Finn, and M. Mdard. Analysis of reconfiguration in IP over WDM access networks. *Proceedings of the Optical Fiber Communication Conference*, 54:MN4.1 – MN4.3, 2001.

[146] N. Tian and Q. Li. The M/M/c queue with PH synchronous vacations. *System in Mathematical Science*, 13(1):7–16, 2000.

[147] N. Tian, Q. Li, and J. Cao. Conditional stochastic decomposition in M/M/c queue with server vacations. *Stochastic Models*, 15(2):367–377, 1999.

[148] N. Tian, Z. Ma, and M. Liu. The discrete time Geom/Geom/1 queue with multiple working vacations. *Applied Mathematical Modelling*, 32(12):2941–2953, 2008.

[149] N. Tian and Z.G. Zhang. M/M/c queue with synchronous vacations of some servers and its application to electronic commerce operations. page Working Paper, 2000.

[150] N. Tian and Z.G. Zhang. The discrete time GI/Geo/1 queue with multiple vacations. *Queueing Systems*, 40:283–294, 2002.

[151] N. Tian and Z.G. Zhang. Stationary distributions of GI/M/c queue with PH type vacations. *Queueing Systems*, 44(2):183–202, 2003.

[152] N. Tian and Z.G. Zhang. A two threshold vacation policy in multiserver queueing systems. *European Journal of Operational Research*, 168:153–163, 2006.

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

[154] N. Tian, X. Zhao, and K. Wang. The M/M/1 queue with single working vacation. *International Journal of Information and Management Sciences*, 19(4):621–634, 2008.

[155] O. Tickoo and B. Sikdar. Queueing analysis and delay mitigation in IEEE 802.11 random acess MAC based wireless networks. *IEEE INFOCOM*, 2004.

[156] H.T. Tran and T.V. Do. Computational aspects of steady state analysis of QBD processes. *Periodica Polytechnica*, 44(2):179–200, 2000.

[157] K.S. Trivedi. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*. Wiley, New York, Second edition, 2002.

[158] J. Wang. Queueing analysis of WDM-based acess networks with reconfiguration delay. *Proceedings of the 4th International Conference on Queueing Theory and Network Applications, Singapore*, 13, 2009.

[159] J. Wang, J. Cao, and Q. Li. Reliability analysis of the retrial queue with server breakdowns and repairs. *Queueing Systems*, 38:363–380, 2001.

[160] J. Wang, L. Sahasrabuddhe, and B. Mukherjee. Path vs. subpath vs. link restoration for fault management in IP-over-WDM Networks: performance comparisons using GMPLS control signaling. *IEEE Communications Magazine*, 40(11):80–87, 2002.

[161] K. Wang, W. Chen, and D. Yang. Optimal management of the machine repair problem with working vacation: Newton's method. *Journal of Computational and Applied Mathematics*, 233:449–458, 2009.

[162] W. Whitt. Approximating a point process by a renewal process, I-Two basic methods. *Operations Research*, 1(30):125–147, 1982.

[163] D. Wu and H. Takagi. M/G/1 queue with multiple working vacations. *Performance Evaluation*, 63(7):654–681, 2006.

[164] Wei Xiong, David Jagerman, and Tayfur Altiok. M/G/1 queue with deterministic reneging times. *Performance Evaluation*, 65(3-4):308–316, 2008.

[165] C. Xiu, N. Tian, and Y. Liu. The M/M/1 queue with single working vacation serving at a slower rate during the start-up period. *Journal of Mathematics Research*, 2(1):98–102, 2010.

[166] X. Xu, C. Liu, G. Lu, and X. Zhao. Stationary analysis for the bulk input Geom$^{[X]}$/Geom/1 queue with working vacation. *International Journal of Management Science and Engineering Management*, 4(2):118–128, 2008.

[167] X. Xu, Z. Zhang, and N. Tian. The M/M/1 queue with single working vacation queue with setup times. *International Journal of Operational Research*, 6(9):420–434, 2009.

[168] X. Yang and A.S. Alfa. A class of multi-server queueing system with server failures. *Computers and Industrial Engineering*, 56(1):33–43, 2009.

[169] U. Yechiali. Queues with system disasters and impatient customers when system is down. *Queueing Systems*, 56(3-4):195–202, 2007.

[170] X.W. Yi, J.D. Kim, D.W. Choi, and K.C. Chae. The Geo/G/1 queue with disasters and multiple working vacations. *Stochastic Models*, 23(4):537–549, 2007.

[171] M. Yoo, C. Qiao, and S. Dixit. QoS performance of optical burst switching in IP-over-WDM networks. *IEEE Journal on Selected Areas in Communications*, 18(10):2062–2071, 2000.

[172] Z.G. Zhang and N. Tian. Geo/G/1 queue with multiple adaptive vacations. *Queueing Systems*, 38:419–429, 2001.

[173] G. Zhao, X. Du, and N. Tian. GI/M/1 queue with set-up period and working vacation and vacation interruption. *International Journal of Information and Management Sciences*, 20:351–363, 2009.

[174] Y.Q. Zhao and A.S. Alfa. Performance analysis of a telephone system with both patient and impatient customers. *Telecommunication Systems*, 4:201–215, 1995.

[175] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583, 2002.

List of Publications based on the Thesis

- Cosmika Goswami and N. Selvaraju, *The discrete-time $MAP/PH/1$ queue with multiple working vacations.* Applied Mathematical Modelling, 34(4): 931-946, 2010.

- Cosmika Goswami and N. Selvaraju, *Performance measures of a finite-buffer queue with working vacations and correlated arrivals.* IEEE proceedings of Next Generation Internetworking Workshop (NGIntW) during COMSNETS2009. (To appear).

- Cosmika Goswami and N. Selvaraju, *Analysis of M/G/1 retrial queue with multiple working vacations.* (Communicated).

- Cosmika Goswami and N. Selvaraju.*Phase-type arrivals and impatient customers in multiserver working vacation queues.* (Communicated).

- N. Selvaraju and Cosmika Goswami. *Impatient Customers in an M/M/1 queue with single and multiple working vacations.* (Communicated)

- Cosmika Goswami and N. Selvaraju.*Working vacation queues and priority customers.* (Communicated).

- *Correlated Markovian arrival queue with multiple working vacations and limited buffer space.* (To be Communicated).