

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**QUEUEING SYSTEMS WITH PHASE TYPE
SERVICE**

by
Şerife ÖZKAR

July, 2017

İZMİR

QUEUEING SYSTEMS WITH PHASE TYPE SERVICE

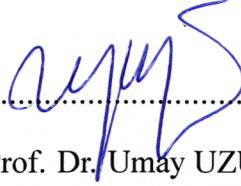
**A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of
Philosophy in Statistics**

**by
Şerife ÖZKAR**

**July, 2017
İZMİR**

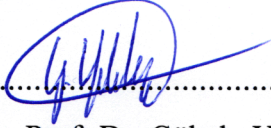
Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**QUEUEING SYSTEMS WITH PHASE TYPE SERVICE**” completed by **ŞERİFE ÖZKAR** under supervision of **ASSOC. PROF. DR. UMay UZUNOĞLU KOÇER** and **PROF. DR. SRINIVAS R. CHAKRAVARTHY** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.



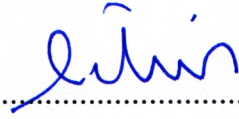
Assoc. Prof. Dr. Umay UZUNOĞLU KOÇER

Supervisor



Assoc. Prof. Dr. Gökarp YILDIZ

Thesis Committee Member



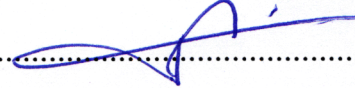
Assoc. Prof. Dr. A. Fırat ÖZDEMİR

Thesis Committee Member



Prof. Dr. Behir Hanalioğlu

Examining Committee Member



Prof. Dr. Mergen Dağr

Examining Committee Member



Prof. Dr. Emine İlknur CÖCEN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Umay UZUNOĞLU KOÇER and to my co-supervisor Prof. Dr. Srinivas R. CHAKRAVARTHY for their continuous support, selfless guidance, generous advice and helpful comments. Especially, I wish to thank my supervisor, for her valuable insights, patience and encouragement during the hard times throughout my research. I am truly grateful to my co-supervisor, for his never ending guidance and support during and after my Ph.D. visit in Kettering University, USA. Their academic perspective created very positive impact on my studies in many aspects.

Besides my advisors, I would like to thank to Assoc. Prof. Dr. Gökalg YILDIZ, for his contributions and invaluable help. I'm also grateful for the insights and efforts put forth by Prof. Dr. Cengiz ÇELİKOĞLU and Assoc. Prof. Dr. A. Fırat ÖZDEMİR.

This thesis is dedicated to my parents Havva ÖZKAR and İsmail ÖZKAR because of their endless love, confidence, encouragement and support in my life. Words can not express my thanks to my lovely family.

Şerife ÖZKAR

QUEUEING SYSTEMS WITH PHASE TYPE SERVICE

ABSTRACT

In this thesis, service times in the queueing systems are modelled by using phase-type distribution. All the systems have the structure of quasi-birth-death process. By means of the advantage of this process, the stationary distributions of the systems are obtained by using matrix geometric method. All studied queueing systems have phase-type service times, however each queueing model presented is designed for a special real-life application, such as healthcare systems, working vacation, Markovian arrival process and crowdsourcing.

Firstly, the service process in a geriatric department is characterized as a queueing system. In this context, three different models are studied. The main difference among these models comes from the capacity of the system. The models have no waiting space, infinite buffer and finite buffer, respectively.

Secondly, queueing model with phase-type service is studied with working vacation policy. The server does not remain idle during the vacation period and the system operates at a lower rate.

Finally, four queueing models are analyzed in the context of crowdsourcing useful in service sectors. In these models, the arrival and service processes are modeled by Markovian arrival process that allows correlated arrivals and phase-type distribution, respectively. The first model has a single server whereas the second model has multi-server. Then, the single-server queueing model with crowdsourcing is developed by working vacation and pure vacation policies. In the working vacation model, the server offers services to customers arriving during a vacation at a lower rate while the server doesn't work during a vacation period in the pure vacation model.

Keywords: Queueing systems, phase-type distributions, matrix-geometric solution, quasi-birth-death process, Markovian arrival process, optimization, working vacation policy, stochastic decomposition, crowdsourcing.

SERVİS SÜRESİ DAĞILIMININ FAZ TİPİ OLDUĞU KUYRUK SİSTEMLERİ

ÖZ

Tez içerisinde çalışılan tüm kuyruk sistemlerinde servis süreleri faz-tipi dağılımlar kullanılarak modellenmiştir. Bu sistemler yarı-doğum-ölüm süreci yapısındadır. Bu sürecin bir avantajı olarak sistemlere ait durağan dağılımlar matris-geometrik metot kullanılarak elde edilmiştir. Faz-tipi dağılımlara sahip kuyruk modellerinin sağlık sistemleri, çalışmalı dinlenme, Markovian geliş süreci ve kitlekaynak kullanımı gibi çeşitli alanlarda kullanımı gösterilmiştir.

İlk olarak, geriatri departmanındaki hizmet süreci üç farklı kuyruk sistemi düşünülerek modellenmiştir. Bu modeller arasındaki temel fark, çalışılan sistemlerin kapasiteli olup olmamasından ileri gelmektedir. Şöyle ki, sistemler kapasitesiz, sonlu ve sonsuz kapasitelidir.

İkinci olarak, çalışmalı dinlenme politikasına sahip evre-tipi servisli kuyruk modeli çalışılmıştır. Bu politikada, hizmet birimi dinlenme periyodunda iken tamamen durmak yerine gelen müşteriye düşük hızla hizmet vermeye devam eder.

Son olarak, hizmet sektöründe kitle kaynaklı çalışmanın kullanıldığı dört kuyruk modeli analiz edilmiştir. Bu modellerde, geliş ve servis süreçleri için sırasıyla korelasyonlu gelişlere izin veren Markovian geliş süreci ve evre-tipi dağılımlar kullanılmıştır. Birinci model tek hizmet birimine ve ikinci model birden fazla hizmet birimine sahiptir. Tek hizmet birimine sahip kuyruk modeli, dinlenme ve çalışmalı dinlenme politikaları kullanılarak iki modele daha geliştirilmiştir. Çalışmalı dinlenme politikasına sahip modelde, hizmet birimi dinlenme esnasında gelen müşteriye düşük hızla hizmet verirken, sadece dinlenmenin olduğu modelde hizmet birimi çalışmaz.

Anahtar kelimeler: Kuyruk sistemleri, faz-tipi dağılımlar, matris-geometrik çözüm, yarı-doğum-ölüm süreci, Markovian geliş süreci, optimizasyon, çalışmalı dinlenme politikası, stokastik parçalama, kitle kaynaklı çalışma.

CONTENTS

	Page
THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZ.....	v
LIST OF FIGURES	ix
LIST OF TABLES.....	xi
 CHAPTER ONE – INTRODUCTION.....	 1
1.1 Scope of the Study.....	1
1.2 Literature	5
1.2.1 Queueing Models with Finite Capacity	5
1.2.2 Phase-type Distributions in Healthcare Systems	7
1.2.3 Queueing Models with Working Vacation.....	10
1.2.4 Queueing Models with Crowdsourcing.....	12
1.3 Main Contributions.....	14
 CHAPTER TWO – BACKGROUND MATERIAL	 16
2.1 Phase-type Distributions	17
2.1.1 Special Phase-type Distributions	20
2.1.2 Fitting Phase-type Distributions	28
2.2 From Common Arrival Processes to Markovian Arrival Process.....	34
2.2.1 Common Arrival Processes.....	35
2.2.2 Markovian Arrival Process	38
2.3 Quasi-birth-death Process and Matrix Geometric Solution	42
2.3.1 $M/PH/1$ Queueing Model.....	44
2.3.2 $MAP/PH/1$ Queueing Model	47

CHAPTER THREE – COX SERVICE QUEUEING MODELS AND OPTIMIZATION FOR A GERIATRIC CARE CENTER..... 51

3.1 $M/C_2/c/c$ Queueing Model	52
3.1.1 Model Description	52
3.1.2 Performance Measures and Optimization.....	53
3.1.3 Numerical Study	55
3.2 $M/C_2/c$ Queueing Model.....	57
3.2.1 Model Description	57
3.2.2 QBD Process	58
3.2.3 Steady-state Probabilities.....	61
3.2.4 Performance Measures and Optimization.....	64
3.2.5 Numerical Study	66
3.3 $M/C_2/c/K$ Queueing Model.....	67
3.3.1 Model Description	68
3.3.2 QBD Process	68
3.3.3 Steady-state Probabilities.....	71
3.3.4 Performance Measures and Optimization.....	73
3.3.5 Case Study- The Planning of A Geriatric Care Center	74
3.3.5.1 Data	75
3.3.5.2 Verification.....	79
3.3.5.3 Analysis of Model for Geriatric Data.....	80
3.3.5.4 The Optimization	84
3.3.5.5 Compare the Studied Model with the models in Literature	86

CHAPTER FOUR – QUEUEING MODELS WITH COX SERVICE AND MULTIPLE WORKING VACATION 92

4.1 Model Description.....	92
4.2 The Steady-state Analysis	95
4.3 Stochastic Decomposition	104
4.4 Numerical Study	107

**CHAPTER FIVE – QUEUEING MODELS WITH PHASE-TYPE SERVICE
AND *MAP* ARRIVAL FOR CROWDSOURCING 115**

5.1 <i>MAP/PH/1</i> Queueing Model with Crowdsourcing	115
5.1.1 Model Description	116
5.1.2 The Steady-state Analysis.....	118
5.1.3 The Stability Condition	122
5.1.4 The Steady-state Probability Vector.....	127
5.1.5 The System Performance Measures	129
5.1.6 Stationary Waiting Time Distribution of A Type 1 Customer.....	130
5.1.7 Numerical Examples	132
5.2 <i>MAP/PH/c</i> Queueing Model with Crowdsourcing	142
5.2.1 Model Description	142
5.2.2 Validation of The Simulated Model	143
5.2.3 Numerical Examples	146
5.3 <i>MAP/PH/1</i> Queueing Model with Crowdsourcing and Working Vacation.....	153
5.3.1 Model Description	154
5.3.2 The Steady-state Analysis.....	156
5.3.3 The Stability Condition	161
5.3.4 The Steady-state Probability Vector.....	164
5.3.5 <i>MAP/PH/1</i> Model with Crowdsourcing Under Pure Vacation	167
5.3.6 Phase-type Representation for Vacation/Working Vacation Duration	172
5.3.7 The System Performance Measures	174
5.3.8 Numerical Examples	176

CHAPTER SIX – CONCLUSION 188

REFERENCES..... 192

LIST OF FIGURES

	Page
Figure 1.1 A graphical representation of the literature.....	14
Figure 2.1 Transition-rate diagram for exponential distribution	20
Figure 2.2 Transition-rate diagram for Erlang- n distribution	21
Figure 2.3 Transition-rate diagram for hyperexponential- n distribution.....	23
Figure 2.4 Transition-rate diagram for Cox-2 distribution	24
Figure 2.5 Erlang, exponential and hyper-exponential distributions before normalization.....	27
Figure 2.6 Erlang, exponential and hyper-exponential distributions after normalization.....	28
Figure 2.7 Weibull distribution for various parameter values	30
Figure 2.8 Lognormal distribution for various parameter values	33
Figure 2.9 Description of Poisson process.....	36
Figure 2.10 Description of Markovian arrival process.....	38
Figure 2.11 The states of a system with two identical units and a repairman.....	41
Figure 2.12 Transition diagram of birth-death process	42
Figure 2.13 Transition diagram of quasi-birth-death process	43
Figure 3.1 A schematic representation of $M/C_2/c/c$ queue model	53
Figure 3.2 A schematic representation of $M/C_2/c$ queue model	58
Figure 3.3 Transition rate diagram of the queueing systems (a) $M/C_2/2$ and (b) $M/C_2/3$	59
Figure 3.4 Costs versus the number of beds for $M/C_2/c$ model	67
Figure 3.5 A schematic representation of $M/C_2/c/K$ queue model	68
Figure 3.6 Transition rate diagram of the queueing systems (a) $M/C_2/2/6$ and (b) $M/C_2/3/6$	70
Figure 3.7 The boxplot of the data about the length of stay for all patients	75
Figure 3.8 The boxplot of the data about the length of stay after removed outliers	77
Figure 3.9 The histogram of the data about the length of stay after removed outliers.....	77
Figure 3.10 Fitted and original cumulative distribution functions of the data.....	78

Figure 3.11 The comparison between the analytic solution values and the values obtained from simulation for buffer sizes 5 and 10	80
Figure 3.12 Cost functions for $\lambda = 1, 2, 3$ and for $K = 5, 10$	86
Figure 4.1 The effect of parameter μ_b on the performance measures	110
Figure 4.2 The effect of parameter μ_v on the performance measures	112
Figure 4.3 The effect of parameter λ on the performance measures.....	113
Figure 4.4 The effect of parameter η on the performance measures.....	114
Figure 5.1 Diagram of queue model with crowdsourcing	116
Figure 5.2 Probability that server is busy, above for $p = 0.5$ and below for $p = 1$	135
Figure 5.3 Probability of a Type 2 customer leaving with a Type 1 customer.....	136
Figure 5.4 Ratios for mean waiting time of Type 2 customers for $\rho = 0.10$	138
Figure 5.5 Ratios for mean waiting time of Type 2 customers for $\rho = 0.99$	139
Figure 5.6 Diagram of the multi-serve queue model with crowdsourcing	142
Figure 5.7 Ratios of various measures under different scenarios with <i>ERLS</i> and <i>HEXS</i> services.....	149
Figure 5.8 Comparison of the ratio, $\frac{\zeta(L=1)}{\zeta(L=0)}$, under different scenarios with <i>ERLS</i> and <i>HEXS</i> services.....	151
Figure 5.9 Diagram of the queue model with vacation and crowdsourcing	155
Figure 5.10 Ratios for mean waiting time of Type 2 customers for Erlang and Hyperexponential services for $\rho = 0.8$	179
Figure 5.11 Ratios for mean waiting time of Type 2 customers for Erlang and Hyperexponential services for $\rho = 0.99$	180
Figure 5.12 Comparison of the R_1 under different scenarios for $\rho = 0.8, 0.99$	183
Figure 5.13 Comparison of the R_2 under different scenarios for $\rho = 0.8, 0.99$	184

LIST OF TABLES

	Page
Table 2.1 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 2)$	31
Table 2.2 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 3)$	31
Table 2.3 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 4)$	31
Table 2.4 Log-likelihood values in fitting $Weibull(\beta < 1)$ to $Cox - n$	32
Table 2.5 Relative differences.....	32
Table 2.6 Log-likelihood values in fitting $log - normal(\sigma > 1)$ to $Cox - n$	34
Table 3.1 Performance measures of $M/C_2/c/c$ queue model.....	55
Table 3.2 The system costs for different number of beds	56
Table 3.3 Performance measures of the $M/C_2/c$ queue model	66
Table 3.4 Determination of optimal number of beds by the $M/C_2/c$ queue model	66
Table 3.5 Verification of the $M/C_2/c/K$ queueing system.....	79
Table 3.6 Analytic solution to the queue model for the data about the length of stay	81
Table 3.7 Comparison of the methods for computing loss probability and the mean bed usage	82
Table 3.8 Comparison of the performance measures calculated for $M/C_2/c/K$ and $M/C_2/c$ models	83
Table 3.9 Comparison of the costs obtained for $M/C_2/c/K$ and $M/C_2/c$ models	83
Table 3.10 Optimal c values for different arrival rates when the loss probability is 1%	85
Table 3.11 Optimal c values that minimize the expected cost for different arrival rates	85
Table 3.12 Comparison of the performance measures for the simulated data when the arrival rate is 1	89
Table 3.13 Comparison of the performance measures for the simulated data when the arrival rate is 2	90
Table 3.14 Comparison of the costs for the simulated data when $\lambda = 1$	91
Table 3.15 Comparison of the costs for the simulated data when $\lambda = 2$	91
Table 4.1 The effect of the parameters on the performance measures for $n=1$	109
Table 4.2 The effect of the parameters on the performance measures for $n=2,3$	111
Table 5.1 Optimum values of (L^*, p^*) for various scenarios.....	140

Table 5.2	Optimum values of L^* for various scenarios when $p = 0.5$	141
Table 5.3	Percentages of reduction in ETP at optimum values.....	141
Table 5.4	Values of λ_2 for various scenarios for $M/M/c$	144
Table 5.5	Error percentages of analytical and simulated models for $M/M/c$ crowdsourcing	144
Table 5.6	Values of λ_2 for various scenarios for $MAP/PH/1$	145
Table 5.7	Error percentages of analytical and simulated models for $MAP/PH/1$ crowdsourcing	146
Table 5.8	Optimum c^* values	153
Table 5.9	Optimum values of (L^*, p^*, N^*) for WV-model, (L^*, p^*) for PV-model	185
Table 5.10	Percentages of reduction in ETP at optimum values.....	186
Table 5.11	Percentages of reduction in ETP at optimum values for various N	187

CHAPTER ONE

INTRODUCTION

1.1 Scope of the Study

Exponential distributions are widely used in stochastic modelling because of their memoryless property and computational tractability. On the other hand, the coefficient of variation for exponential distribution is equal to one. Particularly, this property restricts the usage of the exponential distribution in some real life applications. In many cases, service time may follow heavy-tailed distribution. The phase type distributions are a generalization of the exponential distribution. This distribution class remove the limitations of exponential distribution as well as they have many properties of the exponential distribution. The studies which have been done with phase-type distributions can be classified into two main groups: (i.) the fitting phase-type distribution to any data or a probabilistic distribution, (ii.) the modeling a system consist of stages by phase-type distribution.

The $M/G/c/K$ queuing model is quite common in many real-life systems such as computer, telecommunication, healthcare systems and manufacturing systems. Regarding these systems, the Poisson process can be a reasonable assumption for the arrival stream, whereas service time follows a general distribution. The exact analytical solution to the $M/G/c/K$ model is only possible for special cases, such as for exponential service, a single server, or no waiting room. When a general distribution is replaced with a phase-type distribution, the phase-type distribution assumption for the service time makes it possible to achieve an exact solution. In addition, modeling the service process using phase-type distributions presents rational solutions to real problems, since it is possible that the service time exhibits high variability. For instance, exponential assumption for service times is not hold for both call center data and also data represent the length of patient's stay at special department of a hospital. The service times in these examples have heavy-tailed distributions. Heavy-tailed assumption for both inter-arrival and service times makes

the queueing model analytically intractable. In addition to these, some service processes may also include more than one stage before completion. All of them motivate the use of phase-type distributions to model service time in queueing systems.

Firstly, service process in a geriatric department is modeled by queueing system. The duration of elderly patients' stay at geriatric department has *great variability* and the Coxian distribution, which is a phase-type distribution, provides a good representation for the service time distribution. In this context, we have studied three multi-server queueing models. The patient (customer) arrivals occur according to Poisson process. c beds (servers) are available in the department. The beds are identical and service time in each bed follows 2-phase Cox distribution. The main difference among these models is the behaviour of patients when an arriving patient finds all beds are busy in the department. Note that these differences arise from the capacity of the system. The first model studied is $M/C_2/c/c$ model, which is also known as loss model. In this model an arrival leaves without having service if all beds are busy. The performance measures have been obtained by using Erlang's loss formula and an optimization study including the determination of the optimal number of beds in the department has been performed. The second one is $M/C_2/c$ model where there is no system capacity. The patients spend in waiting line for service until any bed becomes idle. Finally, in $M/C_2/c/K$ model where the system capacity is K , the system is analyzed under the assumption that limited patients can wait in the line. Besides, we have obtained the solutions of the last two queue models with infinite and finite capacities by using matrix-analytic methods and have represented similar optimization studies. The difference between these numerical studies proceeds from the usage of the real data belonging to a geriatric department for the finite capacity model.

Secondly, we have again considered Coxian distribution for service times, but now it has k -phase, and have studied the $M/C_k/1$ queue model with working vacation. Servers are always available in classical queueing models. However, servers may become unavailable for a period of time for a variety of reasons. The period of server absence is called a server vacation. There are many different vacation policies in the

literature. We have considered a working vacation policy which is extensively applied to real life problems. This is because in many practical systems encountered, the server does not remain idle during the vacation period due to the operational costs. Instead of being idle, the system operates at a slower rate to save energy and to reduce waste time. In this context, the vacation policy in the studied queueing model has been determined as follows. Once the system becomes empty, the server goes on the vacation in order to make another job. If there is at least one customer waiting in the queue when the server returns from the vacation, it starts service till the system becomes empty. On the other hand, if there is no customer in the queue on returning from the vacation, it takes another vacation and so on. We have obtained the steady state probabilities of the queue model by using matrix-analytic methods and have represented the corresponding performance measures in the stochastic decomposition structure. For example, the stationary queue length has been decomposed into the sum of two independent random variables, i.e. the queue length without any vacation and the additional queue length due to vacations.

As in above mentioned models, Poisson process is widely used to model an arrival process of queueing systems. In this context, only independent arrivals are allowed and inter-arrival distribution should be exponential distribution. These assumptions may not hold in real applications. It is also possible that the arrival process may not be a renewal process. For example, in a production line, arrivals from different sources form input to an assembly system. Markovian arrival process (*MAP*) is a generalization of some special processes such as Poisson process, renewal process, and *PH*-renewal process. In this process, the inter-arrival times may be positively/negatively correlated. Even for this reason only, Markovian arrival processes may be preferred to model some real-life systems.

Thus, we have studied a queueing model of *MAP/PH/1*-type in the context of crowdsourcing which is useful in service sector. We have used Markovian arrival process that allows correlated arrivals in the system and general phase-type distribution for the arrival and service processes, respectively. There are two types of arrivals to the system such that one group of customers may provide service to other

group after getting service from the system. For example, one type of customers visit the store to buy items while the other type of customers orders over some medium such as internet and phone, and expects them to be delivered. The store management can use the in-store customers as couriers to "serve" the other type of customers. Not all in-store customers may be volunteer to act as servers. Hence, a probability is introduced for in-store customers to opt for providing service for the other type. After we have obtained a solution for the model using matrix-analytic method, the effect of crowdsourcing has been represented under various combinations of the arrival and service processes by discussing some qualitative illustrative examples.

Finally, we have introduced vacation and working vacation policies for the above mentioned $MAP/PH/1$ queueing model with crowdsourcing. When the system becomes empty at the time of a completion of a service, the server will go on a vacation. In the first model, a vacation is interrupted only when a Type 1 customer arrives during that time. However, the server offers services to those (Type 1) customers arriving during a vacation at a lower rate. Note that only Type 1 customers can be served during a vacation. On the other hand, the server doesn't work during the vacation in the second model. In these two models, we have represented the impacts of vacation and working vacation in crowdsourcing.

In this thesis, service times in the studied queueing systems are modelled by using phase-type distributions. All queueing systems studied are in the structure of *quasi-birth-and-death process (QBD process)*. That is, the infinitesimal generator matrix that governs the system transitions has the block tridiagonal structure for each studied queueing model. By means of the advantage of the *QBD process*, the stationary distributions of the studied systems have the matrix geometric form under stability conditions and so are obtained by using *matrix geometric method (MGM)*. Only difference amongst the queueing models is the selected real-life applications consist of healthcare systems, working vacation, *MAP* arrival and crowdsourcing. Thus, we separate the next literature section into different subsections according to interested application areas.

1.2 Literature

A phase-type distribution is the distribution of time until absorption in a finite Markov chain defined on a state space where one state is absorbing and remaining states are transient. It was introduced by Neuts in 1975. Actually, Erlang first considered a generalization of exponential distribution which is called *the method of stages* in 1917. We can say that from this point of view, Neuts improved the Erlang's method. For a comprehensive treatment of phase-type distributions, see references in Neuts (1981), Latouche & Ramaswami (1999), He (2014), and Stewart (2009).

Phase-type distributions include many examples like all combinations of exponential phases in series or parallel and so are used in a wide range of stochastic modeling applications such as telecommunications, teletraffic modeling, healthcare systems, queuing theory, drug kinetics, and reliability theory. As a part of this thesis, the literature review related with queueing models has been summarized.

1.2.1 Queueing Models with Finite Capacity

The $M/G/c/K$ queueing model is quite common in many real-life systems such as computer, telecommunication, and manufacturing systems. Regarding these systems, the Poisson process can be a reasonable assumption for the arrival stream, whereas service time follows a general distribution. The exact analytical solution to the $M/G/c/K$ model is only possible for special cases, such as for exponential service, a single server, or no waiting room Asmussen (2008). There are some approximations for the multi-server, finite-buffer and general service time assumptions Nozaki & Ross (1978), Wolff (1989), Kimura (1996). An extensive review of the previous approximations is presented by Kimura (2000), and it is concluded that these approximations are difficult to evaluate computationally. Optimization of the multi-server $M/G/c/K$ queueing system with finite waiting rooms is examined by Smith (2003) and Smith (2007). Also, for multi-server and finite capacity queue models, we refer to the study in Hochrainer et al. (2014). They propose a algorithmic

method to calculate the the steady-state probabilities of $M/E_k/c/K$ queue model, not exact solution.

Since it is possible that the service time exhibits high variability, modeling the service process using phase-type distributions presents rational solutions to real problems. Phase-type distribution assumption for service time makes the analysis of the queuing system possible and an exact solution can be achieved.

In the queuing systems with a phase-type distribution, the generator matrix that governs the state transitions has structure in the *QBD* process. The advantage of the *QBD* process is that the stationary distribution of the process has the matrix geometric form under stability conditions, and the matrix geometric method developed in Neuts (1981) can be used to analyze the process. There are some studies about multi-server infinite buffer queuing systems assuming phase-type service distribution and using the *QBD* process approach for the solution Ramaswami & Lucantoni (1985b,a, 1993). These studies assume that the waiting room is *infinite*.

However, the existing results for an infinite *QBD* process are not valid for a *QBD* process with *finite* state space. In Hajek (1982), it is shown that the invariant vector can be written as a sum of two matrix geometric terms and a linear term. A system in which the state space is finite and where both the arrival and service processes have a phase-type distribution with only one server is examined by Gun & Makowski (1987). They formulate the queuing model $Ph/Ph/1/K$ by the *QBD* process, present a matrix geometric representation of finite state space, and propose a method to solve the system of balance equations. For the $M/Ph/1/K$ and $Ph/M/1/K$ queueing systems, Li (1989) proposes a closed-form solution to the queue-length distribution. For many examples in production lines, Fadiloğlu & Yeralan (2002) represents the solutions of single server queue models with finite capacity in the form of *QBD* process. An explicit matrix analytic solution to a finite-state *QBD* process is presented in De-Nitto Persone & Grassi (1996), whereas Akar & Sohraby (1997), Akar et al. (2000) propose a numerical method for finding the invariant probability distribution of a *QBD* process. The common idea underlying these proposed methods

is to reduce the global system to a smaller one and to express the steady-state probabilities as a function of the solution to this reduced system. A sophisticated comparison between these methods is presented in Elhafsi & Molle (2007). In addition, the solution to a *QBD* process with finite state space is studied in Elhafsi & Molle (2007) and it is shown that the stationary probability vector has the matrix geometric representation; moreover, the corresponding rate matrix is level-dependent.

To the best of our knowledge, the only explicit solution to finite *QBD* processes is limited to the processes obtained from the $M/Ph/1/K$ and $Ph/M/1/K$ queues Neuts (1981). There are many studies with multi-server queues; however, these models have infinite buffers. For example the $Ph/Ph/c$ queuing system is examined in Takahashi (1981), and an aggregation-disaggregation algorithm is presented. Furthermore, there are some studies on queuing systems with the assumption of the Cox distribution and with finite buffers. For instance, a work station of a flexible manufacturing system as a multi-server queue with finite waiting rooms is modeled in Yao & Buzacott (1985). $M/C_r/1$ queue model is considered in context of optimization in Bhulai (2006). The author gives a closed-form expression for the average cost. This cost arises from holding a customer in the system.

1.2.2 Phase-type Distributions in Healthcare Systems

Phase-type distributions are more popular distributions in healthcare field as indicated in Fackrell (2009). The studies which have been done in this area can be classified into two main groups. The studies discussing the representation of the data with phase-type distributions and fitting PH-distribution to these data are in the first group. Furthermore, the second group consist of some studies that handle the system as a queuing system and propose solutions to the problem of bed occupancy. Determination of the optimal number of beds is the subject matter in most of these studies.

For the first group, we can refer to the Faddy's studies in Faddy (1990) and Faddy

(1994) for medical experiments. The residency time the labelled cells spend in the body before being excreted is modelled with a generalized Erlang distribution by using two-compartment model in Faddy (1990). The length of treatment for patients at risk of suicide is represented with Cox distribution in Faddy (1994). The method of least squares and maximum likelihood estimation were used to fit the data in Faddy (1990) and Faddy (1994), respectively. Moreover, *the length of stay of patients in a department or hospital* has been usually studied to model data. Cox distributions are the most common ones to model data which represent the length of stay of patients in Faddy & McClean (1999, 2005), Marshall & McClean (2003, 2004), Marshall & M. (2009), Marshall & Zenga (2012), Payne et al. (2012). Because the duration of patients' stay at a specific department of a hospital has great variability, the Cox distribution provides a good representation for the service time distribution. In the formal definition, Cox distribution, a kind of phase-type distribution, arises from the convolution of exponential distributions with a bounded number of phases. Starting from the first phase, the process switches from phase i ($1 \leq i < m$) to the next phase with probability q_i and jumps to absorbing state $m + 1$ with probability $(1 - q_i)$. Detailed information about Cox distribution can be found in Augustin & Büsher (1982), Sinclair (2005) and Koole (2004).

Furthermore, the second group consist of some studies that handle the system as a *queuing system* and propose solutions to the problem of bed occupancy. Determination of the optimal number of beds and the bed management are the subject matter in most of these studies. As it is known, basic queueing models are described by three main components: arrival process, service distribution and the number of servers. When this principle is applied to a unit in hospital, these correspond to patient's arrivals, their length of stay in the unit and the number of staffed beds. Griffiths et al. (2006) developed a mathematical model for the *intensive care unit (ICU)* of a hospital by using $M/PH/c$ queueing system and used a hyper-exponential distribution for service times due to the duration of patients' stay at intensive care units has great variability. Namely, the service time distribution is highly skewed. The authors derived the steady-state equations of the system, and then calculated the

steady-state probabilities in an iterative approach (the pseudo code can be seen in Griffiths et al. (2006)). Another example for a service time distribution with high variability can be observed in any *geriatric care center* which provides service to patient aged more than 65 years. Patients are admitted to an initial acute care which may be followed by a number of different stages of treatment. Most of the patients are discharged (or died) after a short period of time. However, some patients may stay in the department over a long time period. Then, we can say that the distribution of length of stay in the department is very heterogeneous. Similar to the study in Griffiths et al. (2006), the modelling of a geriatric department by a queueing system can be seen in Bernardo et al. (2003), Gorunescu et al. (2002a,b). Also these studies are about the determination of the optimal number of beds. However, these studies give optimization results by minimizing the total cost or minimizing the loss probability. Exact analysis of the system has not been given any of these studies.

The number of beds is optimized by analyzing the geriatric department as $M/Ph/c/c$ queueing system in Gorunescu et al. (2002a). Although there is no queueing in this model, it is referred to as $M/Ph/c$ model in Gorunescu et al. (2002a). The authors indicate no specific distribution, the service time is considered to have a phase-type distribution. That is, as it is indicated by the authors the computation of loss probability does not depend on the distribution of the service time but only on the mean service time. Then, Gorunescu et al. (2002a) models the duration of stay in geriatric department by describing the hospital data with three phases; an acute, rehabilitative and long-stay service and by using *hyper-exponential distribution* for service time. For optimization is based on the probability of all beds being occupied (loss probability), they described two costs; the fixed *penalty cost* incurred for each patient turned away and the *holding cost* for each empty bed. In this study, it is assumed the number of patients served can be as the number of beds at most, which is an unrealistic assumption. In reality, elderly patients can be treated by obtaining beds from other departments instead of refusing them because of the unavailability of beds, or the limited waiting rooms can be considered. This provides a more suitable representation of real situations. A system in which patients can wait

was considered in Gorunescu et al. (2002b). Service time distribution is assumed to have k -phase Cox distribution and some scenarios are tested by using blocking probability. By these scenarios, the effect of the parameters including arrival rate, service rate, and number of servers, on the loss probability is analyzed, and no analytic solution is again presented. Moreover, Bernardo et al. (2003) develops a Bayesian methodology to make inference about the length of stay distribution for geriatric patient data. The system is considered as $M/G/c$ loss system and the service time distribution is approximated by mixed generalized Erlang distribution with unknown number of phases. The optimal number of beds minimizing the expected cost function formulated by using the loss probability has been determined. For bed-optimization, they also described *patient cost* based on the length of stay in addition to the costs defined by Gorunescu et al. (2002a).

1.2.3 Queueing Models with Working Vacation

Servers are always available in classical queue models. However, servers may become unavailable for a period of time for a variety of reasons. The period of server absence is called a server vacation. An extensive review on vacation policies in queue models can be found in Tian & Zhang (2006). The recent developments on vacation queue models are presented by Ke et al. (2010). Vacation queue models can be applied in performance evaluation and optimization of production, telecommunication, and service systems. The vacation policy assumed may differ according to the queueing system concerned. There are many different vacation policies in the literature. Of them, a working vacation is the one which is extensively applied to real life problems. This is because in many practical systems encountered, the server does not remain idle during the vacation period due to the operational costs. Instead of being idle, the system operates at a slower rate to save energy and to reduce waste time.

An $M/M/1$ queue model with multiple working vacations was *first* studied by Servi & Finn (2002). They presented formulae for mean, variance, and distribution

of the number and time in the system by using the probability generating function technique. In Tian et al. (2009), the performance measures and stochastic decomposition property belonging to working vacation queue models are examined by the matrix geometric method. Liu et al. (2007) considered the $M/M/1$ queue model with multiple working vacations. They described the quasi-birth-and-death process of the system and derived the analytic expression for the rate matrix. Later on, the authors obtained the performance measures of the system in the form of stochastic decomposition. That is, it is shown that the average number of customers in the system with a vacation equals the sum of the average number of customers in the system without any vacation and the average additional number of customers in the system. Tian et al. (2008) studied a queue model with only a single vacation period, unlike the previous one Liu et al. (2007). Using MGM, they obtained performance measures in the stochastic decomposition structure. The $M/M/1$ queue with a single working vacation serving at a slower rate during the start-up period is examined by Xiu et al. (2010). The $M/M/1$ queue model with multiple working vacations and N -policy is studied by Zhang & Xu (2008). They consider N -policy to avoid switching frequently from one speed to another speed. That is, if there are at least threshold N customers in the system at the end of the vacation, the system returns to the regular period while it is enough that there is one waiting customer in the multiple-vacation model without N -policy. Baba (2010) established the $M/PH/1$ queue with working vacations and vacation interruption. The queue model with phase-type service is examined and the vacation policy assumed has an interruption property, unlike the others. That is, the server returns to the normal busy period at the service completion even without completing the vacation. In addition, the performance measures in the stochastic decomposition structure are given conditionally. The condition is that the service time in the normal period has exponential distribution instead of phase-type distribution. Sreenivasan et al. (2013) studied a queueing model of the type $MAP/PH/1$ with working vacations, vacation interruptions and N policy. For the literature on queueing models with working vacations we also refer to the studies in Baba (2005), Kim et al. (2003), Li & Tian (2007), Wu & Takagi (2006), Zhang & Hou (2011).

1.2.4 Queueing Models with Crowdsourcing

Crowdsourcing, a concept tested in the past, has been used in different domains gaining significant exposure in many fields such as healthcare, computer science, environmental sciences, business and marketing. According to Howe (Howe), "Simply defined, *crowdsourcing* represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers."

We refer to a recent survey paper in Hosseini et al. (2015) and for a recent review of crowdsourcing literature in the context of manufacturing industry can be seen in Evans et al. (2015). Also, the meaning and interpretation of crowdsourcing is varied and despite its popularity among companies in many sectors it still remains little understood. So, for a number of examples of crowdsourcing in many sectors, we suggest the paper *Quora* (2015).

The literature on crowdsourcing dealing with various definitions and classification is huge. However, the literature on the analysis of the crowdsourcing with the help of mathematical models is limited. For example, quantitative models for crowdsourcing platforms will help businesses to understand the underlying processes and predict behavior when the system parameters change. The process of software development at a crowdsourced platform was analyzed in Hasteer et al. (2015). They modelled the development process by using Markov Decision Process and presented an illustration of their proposed model. The temporal behavior of crowd performance to achieve the best performance in terms of accuracy and completion time of large batches of tasks was studied in Bhattacharya et al. (2015) with the help of time-series models. The authors, using simulation testbed, constructed an automated system for optimally scheduling a large batch of tasks on multiple crowdsourcing platforms. In Faridani et al. (2011), based on the dataset from Mturk, the authors modelled the arrival

process of workers using a non-homogeneous Poisson process and derived the optimal pricing strategies for the employers. Based on survival analysis and using the dataset from MTurk, Wang, et.al., Wang et al. (2011) analyzed the completion time of crowdsourcing campaigns and showed the impact of time-independent factors such as the payment or the type of the task on the completion time. The material flow of crowdsourcing processes in manufacturing systems was studied by using stochastic petri nets in Wu et al. (2014).

Only recently stochastic models, more specifically queueing models, useful in crowdsourcing in the context of service sectors have been studied. We want to emphasise that the paper Bernstein et al. (2012) deals with optimizing a real-time crowdsourcing. A retainer model in which the system will try to recruit servers from the pool of workers who have been pre-determined and paid ahead of time to act as on-call servers as opposed to using the customers who visit the system to serve fellow customers is studied in Bernstein et al. (2012). The model was validated in experiment with users on MTurk. Schwartz et al. (2015) considered the crowdsourcing platform as a $M^{(X)}/M/c - \infty$ queueing model and used a dataset directly provided from the commercial crowdsourcing platform Microworkers. They further extended the study of Bernstein et al. (2012) by considering the time until the processing of any task in the campaign begins.

The crowdsourcing models considered in this thesis are motivated by the leading retailer, Walmart which considered a plan that would get its in-store customers to deliver packages to its online customers in *Forbes* (2013). To our knowledge, queueing model incorporating the crowdsourcing concept in the sense of using the customers as possible servers after these customers are served by one of the system servers was first studied in Chakravarthy & Dudin (2017).

Graphical representation of the literature in this section can be shown in Figure 1.1. That is, the literature is categorised in six parts; **WV**- queue models with working vacation, **PH-service**- queue models with PH-service, **CR**- crowdsourcing and queue models for crowdsourcing, **MAP-PH**- Markovian arrival process and

phase-type distribution, **MAM-QBD**- matrix geometric solution and quasi birth and death process, **Others**- other references including mathematical tools and general queue models.

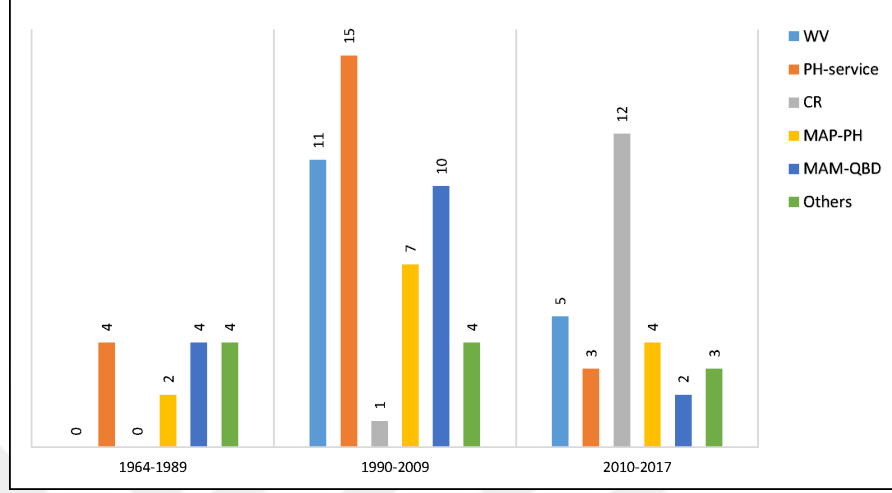


Figure 1.1 A graphical representation of the literature

1.3 Main Contributions

In this thesis, the queueing systems with phase-type service have been studied. The queueing systems have been analyzed by using *matrix geometric method* due to the all systems have the structure of *quasi-birth-and-death process*. The contributions of our study can be summarized as follows.

Model 1: The exact analytical solution to the $M/G/c/K$ model is only possible for special cases, such as for exponential service, a single server, or no waiting room. On the other hand, there are some approximations for the multi-server, finite-buffer and general service time assumptions. We have studied the $M/C_2/c/K$ queue model. We purpose an efficient analytical solution to calculate the steady-state probabilities by using matrix-geometric method. Then, we have performed a case study includes the determination of the optimal number of beds for geriatric patients. We have represented that Coxian distribution with two phases provides a good representation of the duration of elderly patients' stay at a geriatric department of a hospital.

Model 2: For the $M/PH/1$ queue with working vacations, the performance measures in the stochastic decomposition structure were given conditionally in Baba (2010). This condition is that the service times in the normal period have exponential distribution instead of phase-type distribution. We have studied $M/C_k/1$ queue model with working vacation. The performance measures of the queue model have been given in the stochastic decomposition structure.

Model 3: The queueing model incorporating the crowdsourcing concept in the sense of using the customers as possible servers after these customers are served by one of the system servers was first studied in Chakravarthy & Dudin (2017). The authors studied the $M/M/c$ model in which used exponential distribution to model arrival and service processes. We have generalized the model studied in Chakravarthy & Dudin (2017) by considering a more versatile point process, namely, Markovian arrival process (MAP) for Type 1 arrivals and use phase type distribution for services in the context of a single server. Thus, the model is of $MAP/PH/1$ -type with crowdsourcing. This model have been studied analytically using matrix-analytic method.

Model 4: We have developed our **Model 3** by considering in the context of a multi-server. Thus, the model is of $MAP/PH/c$ -type with crowdsourcing and it has been studied by using simulation because of huge state space.

Model 5: We have again improved our **Model 3** with vacation/working vacation policies. In this model, we considered a queueing system useful in crowdsourcing and with server taking (multiple) vacations. Further, the server can offer services to a special group of customers during vacationing but at a lower rate. Using a versatile class of point process for arrivals and phase type services, we showed the significant benefits in introducing this type of variants to the classical queueing models through illustrative numerical examples.

CHAPTER TWO

BACKGROUND MATERIAL

This chapter is organized to explain some fundamental concepts that are used in this thesis. We have used *phase-type distributions (PH-distributions)* to model the service processes in all studied queueing systems. This distribution class is given in Section 2.1. On the other hand, we have studied two different queueing models in which the arrival process is modeled by *Poisson process* and *Markovian arrival process (MAP)*. After brief definition of Poisson process, Markovian arrival process is particularly introduced in Section 2.2. We have utilized *matrix-geometric method (MGM)* to calculate the stationary probabilities of the all defined queueing systems in this thesis. Section 2.3 presents modeling the systems by *quasi-birth-death process* and obtaining their solutions by using *MGM*. We are interested in Markov processes along the thesis. So, it is convenient to start by reminding its definition.

The behaviour of any system can be modelled by describing all possible states and indicating the movements of among these states. While a system is modelled, it is assumed that the system occupies only one state at any moment and its evolution is performed by transitions from a state to another state. Also, another assumption is that these transitions instantaneously occur. A stochastic process is defined with $\{N(t), t \in T\}$ as given in Stewart (2009). That is,

- (i) the parameter t usually denotes time.
- (ii) T is called as *index set* or *parameter space*. The index set can be discrete or continuous. Depending on the index set, we call the process that *discrete parameter stochastic process* or *continuous parameter stochastic process*.
- (iii) $N(t)$ is a random variable that depends on time t . The values of the random variables are called *states* and the set of all possible states is called *state space*. The state space can be discrete or continuous, too. When the state space is discrete, the process is usually referred as *chain*.

A stochastic process is referred by **Markov process** if the future evolution of the modelled system depends on only its current state and is independent of its past history. In other words, if the conditional probability distribution function of a stochastic process satisfies the *Markov (or memoryless) property*, it is called as Markov process.

2.1 Phase-type Distributions

Exponential distribution is widely used in stochastic modelling because of its memoryless property and computational tractability. On the other hand, the coefficient of variation is equal to one for exponential distribution. This property restricts the usage of the exponential distribution in some real-life applications. In a generalization of exponential distribution, phase-type distributions were introduced by Neuts in 1975. As it is mentioned earlier, actually, Erlang first considered a generalization of exponential distribution which is called *the method of stages* in 1917. From this point of view, it can be said that Neuts improved the Erlang's method. This distribution class removes the limitations of exponential distribution as well as they have many properties of exponential distribution. The studies which have been done with phase-type distributions can be classified into two main groups: (i.) the fitting phase-type distribution to any data or a probabilistic distribution, (ii.) the modelling a system consisting of stages by phase-type distribution.

A phase-type distribution is the distribution of time until absorption in a finite Markov chain defined on state space where one state is absorbing and the other remaining states are transient. It is completely characterized by an infinitesimal generator matrix and an initial probability vector as following.

Let $\{I(t), t \geq 0\}$ be continuous time Markov chain, (CTMC), on the state space $E = \{1, 2, \dots, n, n+1\}$ with n transient states and an absorbing state. The infinitesimal

generator matrix of the *CTMC* has the following form,

$$Q = \begin{bmatrix} T & \mathbf{T}^0 \\ 0 & 0 \end{bmatrix}, \quad (2.1)$$

where T is $n \times n$ matrix with $T_{ii} < 0$, $T_{ij} \geq 0, \forall i \neq j$ and governs the transitions among the transient states, $\mathbf{T}^0$ is a column vector of order n and governs the transitions from transient states to absorbing state. It can be seen the relation $\mathbf{T}^0 = -T\mathbf{e}$ because of the definition of the generator matrix Q in (2.1). That is, sum of each rows in an infinitesimal generator matrix must equal to zero.

Let $(\beta_1, \dots, \beta_n, \beta_{n+1}) = (\boldsymbol{\beta}, \beta_{n+1})$ be a probability vector. That represents the initial probabilities. Namely, the probability that the *CTMC* is in state i at time 0 is given by $P\{I(0) = i\} = \beta_i$, $1 \leq i \leq n+1$. Note that, it is assumed that $\beta_{n+1} = 0$.

Let X denote the time until absorption into the absorbing state $n+1$. In this case, X is a continuous random variable distributed phase-type on $[0, \infty)$ and it is represented by $X \sim (\boldsymbol{\beta}, T)_n$. The key information belong to X , that is, the cumulative distribution function $F(t)$, the probability density function $f(t)$, the k^{th} noncentral moment $E(X^k)$, the variance σ^2 and the Laplace transform $f(s)$ are given as following. We omit the proofs of these formulas and refer to the notes in Bladt & Nielsen (2008) and book in He (2014).

$$F(t) = P(X \leq t) = 1 - \boldsymbol{\beta}e^{Tt}\mathbf{e}, \quad t \geq 0, \quad (2.2)$$

$$f(t) = \boldsymbol{\beta}e^{Tt}\mathbf{T}^0, \quad t \geq 0, \quad (2.3)$$

$$E(X^k) = (-1)^k k! \boldsymbol{\beta}(T)^{-k}\mathbf{e}, \quad (2.4)$$

$$\sigma^2 = 2\boldsymbol{\beta}(T)^{-2}\mathbf{e} - (\boldsymbol{\beta}T^{-1}\mathbf{e})^2, \quad (2.5)$$

$$f(s) = \boldsymbol{\beta}(sI - T)^{-1}\mathbf{T}^0, \quad (2.6)$$

where $e^{Tt} = \sum_{k=0}^{\infty} (Tt)^k / k!$ is a matrix exponential function which can also be written by using the inverse laplace transform as follows.

$$e^{Tt} = L^{-1}[(sI - T)^{-1}]. \quad (2.7)$$

Properties of phase-type distributions

Why the phase-type distributions are widely used in stochastic modelling? We consider that the following three properties can answer this question. For the other properties we suggest the books in Neuts (1981) and He (2014).

a. It has partial memoryless property.

Before the *CTMC*'s absorption, it is governed by transition rates in matrix T . The absorption rates from each transient state to the absorption state are given by the vector T^0 . Accordingly, by keeping track of the state $I(t)$ of the *CTMC*, the distribution of the residual time of absorption can be obtained. For example, if $I(t) = 1$, then the residual time has a phase-type distribution with representation $((1, 0, \dots, 0), T)$. This property is called **partial memoryless property based on the state of the *CTMC***.

b. It is closed under operations "min", "max", and "+".

Let assume that X_1 and X_2 have phase-type distributions with representations (α, T) and (β, S) , respectively, and they are independent.

i. Then $X = X_1 + X_2$ has a phase-type distribution with representation

$$\left(\left(\alpha, (1 - \alpha e)\beta \right), \begin{pmatrix} T & T^0\beta \\ 0 & S \end{pmatrix} \right).$$

ii. Then $X = \min[X_1, X_2]$ has a phase-type distribution with representation

$$(\alpha \otimes \beta, T \otimes I + I \otimes S).$$

iii. Then $X = \max[X_1, X_2]$ has a phase-type distribution with representation

$$\left(\begin{pmatrix} \alpha \otimes \beta, & (1 - \alpha e)\beta, & (1 - \beta e)\alpha \end{pmatrix}, \begin{pmatrix} T \otimes I + I \otimes S & T^0 \otimes I & I \otimes S^0 \\ 0 & S & 0 \\ 0 & 0 & T \end{pmatrix} \right).$$

c. It is dense in continuous probability distributions on $[0, \infty)$.

Phase-type distributions can approximate any probability distribution on the nonnegative half line.

2.1.1 Special Phase-type Distributions

In this section, a few special phase-type distributions used in this thesis will be explained by giving the definitions, transition diagrams and key formulas. One can obtain the key formulas for any phase-type distribution by using the common relations in (2.2)-(2.6). For example, we will only obtain the formulas of the Coxian distribution with two-phase at the end of this section. We refer to Stewart (2009), Bhat (2015) for information in detail (i.e., the proofs of the given formulas and the other special phase-type distributions).

Exponential distribution: Exponential distribution is a simplest phase-type distribution. Figure 2.1 represents that a customer enters the service process which has one phase from the left, spends an amount of time that is exponentially distributed with parameter μ in that phase, then exits to the right.

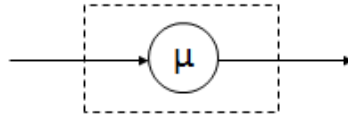


Figure 2.1 Transition-rate diagram for exponential distribution

Initial probability and transition rates for exponential distribution:

$$\beta = (1), T = (-\mu), \mathbf{T}^0 = (\mu). \quad (2.8)$$

Key formulas for the exponential distribution with parameter μ are given as follows.

$$F(t) = 1 - e^{-\mu t}, \quad t \geq 0,$$

$$f(t) = \mu e^{-\mu t}, \quad t \geq 0,$$

$$E(X) = \frac{1}{\mu},$$

$$\sigma_X^2 = \frac{1}{\mu^2},$$

$$f(s) = \frac{\mu}{s + \mu},$$

$$CV^2 = 1.$$

Erlang- n distribution: An Erlang- n distribution is the distribution of the sum of n exponential random variables. In Figure 2.2 the service process has now n sequential phases which are exponentially distributed with the same parameter. A customer enters the first phase from the left with probability 1, spends an amount of time that is exponentially distributed with parameter μ . Once completion of the customer's service in this phase, the customer enters the second phase and spends again an amount of time that is exponentially distributed with parameter μ , and so on. At the end of the n^{th} phase, the customer departs from the service process. There is only one customer in the service process at any time.

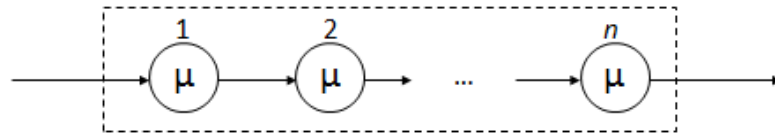


Figure 2.2 Transition-rate diagram for Erlang- n distribution

Initial probability distribution and transition rates for the Erlang- n distribution:

$$\beta = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}, T = \begin{bmatrix} -\mu & \mu & & \\ & \ddots & \ddots & \\ & & -\mu & \mu \\ & & & -\mu \end{bmatrix}, \mathbf{T}^0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu \end{bmatrix}. \quad (2.9)$$

Key formulas of the Erlang- n distribution are given below.

$$F(t) = 1 - e^{-\mu t} \sum_{i=0}^{n-1} \frac{(\mu t)^i}{i!}, \quad t \geq 0,$$

$$f(t) = \frac{\mu(\mu t)^{n-1} e^{-\mu t}}{(n-1)!}, \quad t \geq 0,$$

$$E(X) = \frac{n}{\mu},$$

$$\sigma_X^2 = \frac{n}{\mu^2},$$

$$f(s) = \left(\frac{\mu}{s + \mu} \right)^n,$$

$$CV^2 = \frac{1}{n}.$$

Hyper-exponential distribution: In the hyperexponential distribution, the service process has n parallel phases which are exponentially distributed with parameter μ_i , $i = 1, \dots, n$. The branch probabilities satisfy $\sum_{i=1}^n \alpha_i = 1$. A customer enters the phase i from the left with probability α_i , spends an amount of time that is exponentially distributed with parameter μ_i . Once completion of the customer's service in the phase i , the customer departs from the service process. Similar to Erlang distribution, only a single customer can be served at any time. The Figure 2.3 illustrates the distribution.

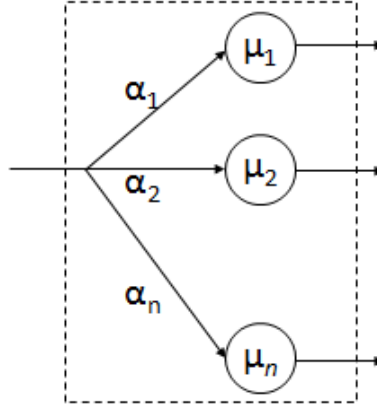


Figure 2.3 Transition-rate diagram for hyperexponential- n distribution

Initial probability distribution and transition rates for hyperexponential- n distribution:

$$\beta = \begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix}, T = \begin{bmatrix} -\mu_1 & & \\ & -\mu_2 & \\ & & \ddots \\ & & & -\mu_n \end{bmatrix}, \mathbf{T}^0 = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}. \quad (2.10)$$

Key formulas of the hyper-exponential- n distribution are given as following.

$$F(t) = \sum_{i=1}^n \alpha_i (1 - e^{-\mu_i t}), \quad t \geq 0,$$

$$f(t) = \sum_{i=1}^n \alpha_i \mu_i e^{-\mu_i t}, \quad t \geq 0,$$

$$E(X) = \sum_{i=1}^n \frac{\alpha_i}{\mu_i}, E(X^2) = 2 \sum_{i=1}^n \frac{\alpha_i}{\mu_i^2},$$

$$f(s) = \sum_{i=1}^n \alpha_i \frac{\mu_i}{s + \mu_i},$$

$$CV^2 = \frac{E(t^2)}{[E(t)]^2} - 1 \geq 1$$

(See Stewart (2009) for proof $CV^2 \geq 1$).

Coxian distribution: Coxian (or Cox) distribution has ordered two phases as given in Figure 2.4. Namely, customer can depart from the service process at the end of any phase. A customer enters the first phase from the left and service begins. In this case, the customer has two options, that is, either he continues service process to the second phase with probability α or departs from the service process with probability $1 - \alpha$.

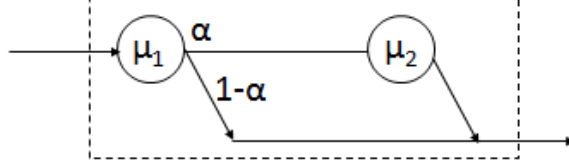


Figure 2.4 Transition-rate diagram for Cox-2 distribution

Initial probability distribution and transition rates for Cox distribution with n phases:

$$\beta = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix},$$

$$T = \begin{bmatrix} -\mu_1 & \alpha_1 \mu_1 & & \\ & \ddots & \ddots & \\ & & -\mu_{n-1} & \alpha_{n-1} \mu_{n-1} \\ & & & -\mu_n \end{bmatrix}, T^0 = \begin{bmatrix} (1 - \alpha_1) \mu_1 \\ \vdots \\ (1 - \alpha_{n-1} \mu_{n-1}) \\ \mu_n \end{bmatrix}. \quad (2.11)$$

Example 2.1 Let us obtain the equations given by (2.2)-(2.6) for Cox-2 distribution which has following parameters

$$T = \begin{bmatrix} -\mu_1 & \alpha \mu_1 \\ 0 & -\mu_2 \end{bmatrix}, T^0 = \begin{bmatrix} (1 - \alpha) \mu_1 \\ \mu_2 \end{bmatrix}, \beta = (1 \ 0).$$

Firstly the matrix exponential function is obtained by using (2.7) as follows:

$$e^{Tt} = L^{-1} \left[\left(\begin{bmatrix} s + \mu_1 & -\alpha \mu_1 \\ 0 & s + \mu_2 \end{bmatrix} \right)^{-1} \right] = L^{-1} \begin{bmatrix} \frac{1}{(s + \mu_1)} & \frac{\alpha \mu_1}{(s + \mu_1)(s + \mu_2)} \\ 0 & \frac{1}{(s + \mu_2)} \end{bmatrix},$$

The last equation can be rewritten by *partial fraction expansion*.

$$e^{Tt} = L^{-1} \begin{bmatrix} \frac{1}{(s+\mu_1)} & \frac{\alpha\mu_1}{(\mu_2-\mu_1)(s+\mu_1)} + \frac{\alpha\mu_1}{(\mu_1-\mu_2)(s+\mu_2)} \\ 0 & \frac{1}{(s+\mu_2)} \end{bmatrix}.$$

We know that if the Laplace transform of function $f(t)$ is $f(s) = \frac{1}{s-a}$, the inverse Laplace transform $f^{-1}(s) = f(t) = e^{at}$. Also, by using the property additive of the inverse Laplace transform $f^{-1}[af(s) + bg(s)] = af(t) + bg(t)$, we obtain the matrix exponential function as follows:

$$e^{Tt} = \begin{bmatrix} e^{-\mu_1 t} & \frac{\alpha\mu_1 e^{-\mu_1 t}}{(\mu_2-\mu_1)} + \frac{\alpha\mu_1 e^{-\mu_2 t}}{(\mu_1-\mu_2)} \\ 0 & e^{-\mu_2 t} \end{bmatrix}. \quad (2.12)$$

Substituting the matrix exponential function given by (2.12) into (2.3), we obtain the probability density function as

$$f(t) = (1 - \alpha)\mu_1 e^{-\mu_1 t} + \alpha \left(\frac{\mu_2}{\mu_2 - \mu_1} \mu_1 e^{-\mu_1 t} + \frac{\mu_1}{\mu_1 - \mu_2} \mu_2 e^{-\mu_2 t} \right). \quad (2.13)$$

The Laplace transform is obtained as

$$f(s) = \beta(sI - T)^{-1} \mathbf{T}^0 = (1 \ 0) \begin{bmatrix} s + \mu_1 & -\alpha\mu_1 \\ 0 & s + \mu_2 \end{bmatrix}^{-1} \begin{bmatrix} (1 - \alpha)\mu_1 \\ \mu_2 \end{bmatrix}$$

$$f(s) = \frac{(1 - \alpha)\mu_1}{s + \mu_1} + \frac{\alpha\mu_1\mu_2}{(s + \mu_1)(s + \mu_2)}.$$

$$f(s) = \frac{(1 - \alpha)\mu_1}{(s + \mu_1)} + \frac{\alpha\mu_1\mu_2}{(\mu_2 - \mu_1)(s + \mu_1)} + \frac{\alpha\mu_1\mu_2}{(\mu_1 - \mu_2)(s + \mu_2)}. \quad (2.14)$$

Note that the inverse Laplace transform $f'(s)$ gives $f(t)$ given by (2.13).

The mean and variance can be achieved by using (2.4) and (2.5).

$$E(X) = (1 \ 0) \begin{bmatrix} \mu_1 & -\alpha\mu_1 \\ 0 & \mu_2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\mu_1} + \frac{\alpha}{\mu_2}. \quad (2.15)$$

$$E(X^2) = 2(1 \ 0) \begin{bmatrix} \mu_1 & -\alpha\mu_1 \\ 0 & \mu_2 \end{bmatrix}^{-2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{2}{\mu_1^2} + \frac{2\alpha}{\mu_1\mu_2} + \frac{2\alpha}{\mu_2^2}.$$

$$\sigma_X^2 = E(X^2) - E^2(X) = \frac{1}{\mu_1^2} + \frac{2\alpha - \alpha^2}{\mu_2^2}. \quad (2.16)$$

Example 2.2: Suppose that the time between two phone calls has the following phase-type distribution with representation (β, S) . These related distributions are represented in Figure 2.5.

Erlang distribution:

$$\beta = \begin{bmatrix} 1 & 0 \end{bmatrix}, S = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

Exponential distribution:

$$\beta = \begin{bmatrix} 1 \end{bmatrix}, S = \begin{bmatrix} -1 \end{bmatrix}.$$

Hyper-exponential distribution:

$$\beta = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}, S = \begin{bmatrix} -10 & 0 \\ 0 & -1 \end{bmatrix}.$$

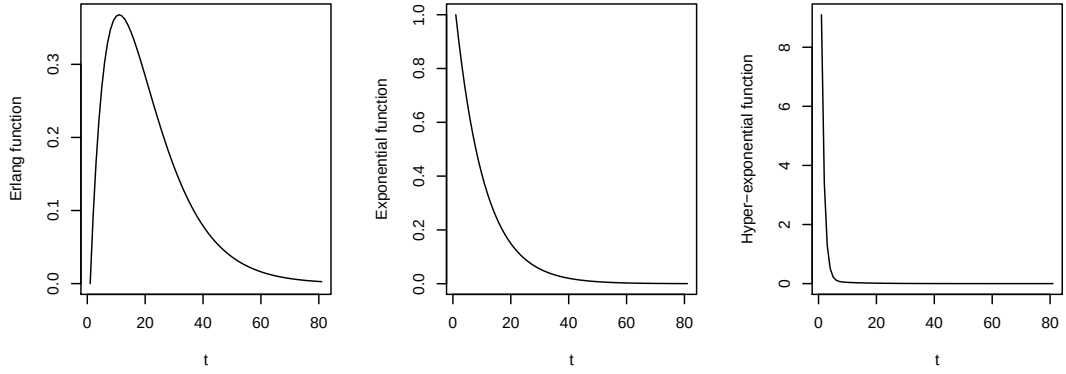


Figure 2.5 Erlang, exponential and hyper-exponential distributions before normalization

Erlang distribution has $E(X) = 2, \sigma = 1.414214$ and $CV^2 = 0.5$. Exponential distribution has $E(X) = 1, \sigma = 1$ and $CV^2 = 1$. Hyper-exponential distribution has $E(X) = 0.190, \sigma = 0.426497$ and $CV^2 = 5.038781$. These results can be obtained by using the equations (2.4) and (2.5).

Erlang, exponential and hyper-exponential distributions have the service rates 0.5, 1 and 5.263158, respectively. Let we want to normalize these three phase-type distributions to have the same specific service rate $\mu = 1.1$. Required results can be achieved by normalizing the matrices S in Bhat (2015).

The mean of Erlang distribution is $E(X) = \beta(-S)^{-1}\mathbf{e} = 2$. On the other hand, for $\mu = 1.1$ $E(X)$ becomes $\frac{10}{11}$. It can be seen that the normalizing constant is $\frac{(\frac{10}{11})}{2} = 2.2$. Multiplying the matrix S by the normalizing constant, we obtain the following parameters for Erlang distributions.

$$\beta = \begin{bmatrix} 1 & 0 \end{bmatrix}, S = \begin{bmatrix} -2.2 & 2.2 \\ 0 & -2.2 \end{bmatrix}.$$

Similarly, we can achieve the following results for exponential and hyper-exponential distributions.

Exponential distribution:

$$\beta = \begin{bmatrix} 1 \end{bmatrix}, S = \begin{bmatrix} -1.1 \end{bmatrix}.$$

Hyper-exponential distribution:

$$\beta = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}, S = \begin{bmatrix} -2.09 & 0 \\ 0 & -0.209 \end{bmatrix}.$$

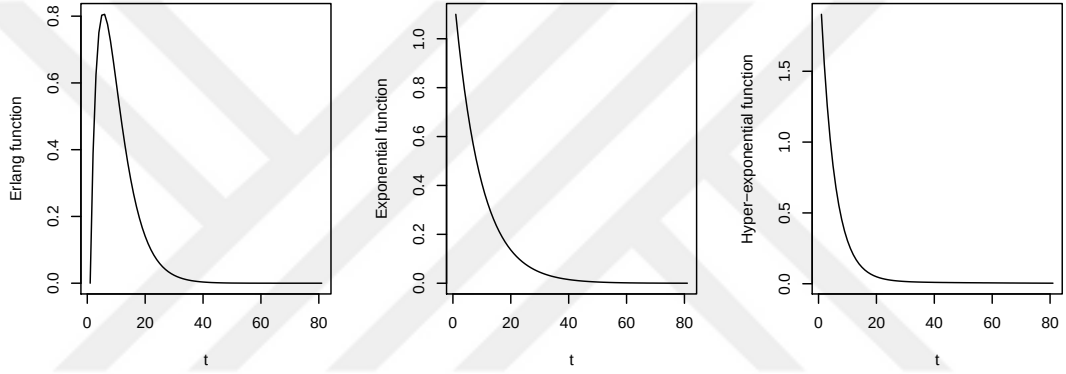


Figure 2.6 Erlang, exponential and hyper-exponential distributions after normalization

After normalizing, these three phase-type distributions have the same means, $E(X) = 0.9090909$, but, these distributions are qualitatively different because they have different variance structure. That is, Erlang, exponential and hyper-exponential distributions have the standard deviations 0.642824, 0.9090909, and 2.040657, respectively.

2.1.2 Fitting Phase-type Distributions

While an exponential distribution is characterized by only one parameter, a phase-type distribution is more versatile class of distributions that have popular properties of an exponential distribution. A phase-type distribution is closed under the formation

of finite mixtures, finite convolutions, finite minima and maxima. In addition to these closure properties, it is *dense* to any probability distribution on nonnegative half-line (see Theorem in He (2014)). The last property is the most important one among its other properties. It means that a phase-type distribution can be approximated to any common probability distribution on $[0, \infty)$. In other words, the basic idea of fitting is to replace nonnegative general distributions in a model by phase-type distributions.

For the fitting process, although there are many approaches in the literature, there are frequently a few methods such as *maximum likelihood estimation*, *moment matching method* and *the least squared method*. We strongly recommend the paper Asmussen et al. (1996). The authors developed an expectation-maximization algorithm (EM-algorithm) for calculating the maximum likelihood parameter estimates of phase-type distributions. One can fit phase-type distributions to any data or some probability distributions such as log-normal and Weibull distributions. The code that was written in *C programming language* can be downloaded from <http://home.math.au.dk/asmus/pspapers.htm>. We used this code for the following examples:

Case 1: We require to fit Coxian distribution to Weibull distribution that has the following density function.

$$f(x; \beta, \eta) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^{\beta}}, \quad x \geq 0, \beta > 0, \eta > 0,$$

where β and η are shape parameter and scale parameter, respectively.

The Weibull distribution represents a heavy-tail distribution in case $\beta < 1$ in Figure 2.7. We also know that Coxian distribution is a heavy-tailed distribution. So, we consider Weibull distribution with $\beta < 1$ for the following example.

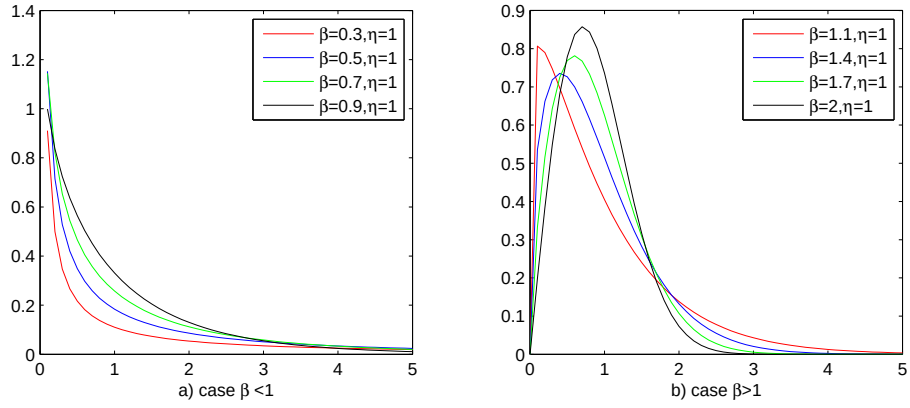


Figure 2.7 Weibull distribution for various parameter values

Example 2.3: The purpose of this example is to determine the number of phase while characterizing the distribution of different data sets. Weibull distributed data sets are used with different parameters. That is, for Weibull distribution we fix $\eta = 1$ and vary $\beta = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. On the other hand, different number of phases are considered such as $n = 2, 3, 4, 5$ for Coxian distribution.

Table 2.1, Table 2.2 and Table 2.3 present the estimated parameter values of Cox distributions which have the number of phase 2, 3, and 4, respectively. These estimation values are obtained by using Weibull distribution with parameters $\beta = 0.9$ and $\eta = 1$. The tables also represent the squared coefficient of variation CV^2 , the log likelihood values and the number of replications. Note that the log-likelihood values following bold rows in tables are the same. So, we have considered that the estimation values in the bold rows are sufficient because of small replication number. For example, we can say from Table 2.1 that Cox distribution with *two phases* ($\mu_1 = 3.2352, \mu_2 = 0.8315, q = 0.619931$) can be used instead of Weibull distribution with $\beta = 0.9, \eta = 1$. Similarly, from Table 2.2, three phase Cox distribution with parameters $\mu_1 = 10.1248, \mu_2 = 1.6425, \mu_3 = 0.7756, q_1 = 0.866496, q_2 = 0.383379$ can be used instead of Weibull distribution with parameters $\beta = 0.9, \eta = 1$.

Table 2.1 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 2)$

Replication	μ_1	μ_2	q	cv^2	Log – like
250	3.3221	0.8339	0.628398	1.1960	-1.042730
300	3.2661	0.8324	0.622975	1.1975	-1.042729
350	3.2352	0.8315	0.619931	1.1984	-1.042728
400	3.2203	0.8311	0.618421	1.1988	-1.042728
450	3.2139	0.8310	0.617754	1.1990	-1.042728

Table 2.2 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 3)$

Replication	μ_1	μ_2	μ_3	q_1	q_2	cv^2	Log – like
900	9.8118	1.6199	0.7740	0.862563	0.376999	1.2270	-1.042247
1000	9.6712	1.6096	0.7733	0.860720	0.374130	1.2272	-1.042248
1100	10.1248	1.6425	0.7756	0.866496	0.383379	1.2268	-1.042246
1200	10.2884	1.6542	0.7764	0.868454	0.386531	1.2268	-1.042246
1300	10.4167	1.6631	0.7769	0.869956	0.388972	1.2266	-1.042246

Table 2.3 Parameter values in fitting $Weibull(\beta = 0.9)$ to $Cox(n = 4)$

Replication	μ_1	μ_2	μ_3	μ_4	q_1	q_2	q_3	cv^2	Log – like
2100	9.0223	4.5006	1.6853	0.7780	0.851102	0.780073	0.422714	1.2263	-1.042251
2300	10.4249	4.5369	1.6187	0.7752	0.870416	0.774383	0.402978	1.2268	-1.042245
2500	11.7712	4.4970	1.5731	0.7731	0.884676	0.767000	0.389168	1.2270	-1.042241
2700	12.8852	4.4324	1.5393	0.7714	0.894243	0.760040	0.378808	1.2272	-1.042239
2900	13.6784	4.3776	1.5160	0.7702	0.900105	0.754820	0.371738	1.2272	-1.042239

We have given the estimated parameter values of Coxian distribution for Weibul distribution with parameter $\beta = 0.9$ in Table 2.1 through Table 2.3. Similar studies have been performed for Weibul distributions have various parameter values such as $\beta = 0.4, 0.5, 0.6, 0.7, 0.8$. At this point, we deal with the determination of the number of phase of Cox distribution. The results have been given in Table 2.4.

Table 2.4 Log-likelihood values in fitting $Weibull(\beta < 1)$ to $Cox - n$

$Weibull(\beta, \eta)$	cv^2	$Log - like$	n	$Rep.$	$Weibull(\beta, \eta)$	cv^2	$Log - like$	n	$Rep.$
(0.9,1)	1.1984	-1.042728	2	350	(0.6,1)	1.8174	-1.071095	2	50
	1.2268	-1.042246	3	1100		2.0967	-1.046620	3	300
	1.2272	-1.042239	4	2700		2.1730	-1.043089	4	600
(0.8,1)	1.4330	-1.084635	2	150	(0.5,1)	2.3581	-0.911492	3	250
	1.5361	-1.080935	3	450		2.4973	-0.902232	4	400
	1.5525	-1.080619	4	750		2.5376	-0.900950	5	900
(0.7,1)	1.6374	-1.098403	2	40	(0.4,1)	2.5921	-0.602095	3	200
	1.8136	-1.087791	3	350		2.8237	-0.580831	4	450
	1.8512	-1.086670	4	600		2.9033	-0.576267	5	1300

Now, we want to select the number of phase by using *Relative difference* equation in (2.17). That is, we have obtained the relative difference between log-likelihood values. The results can be seen in Table 2.5.

$$RD(x, y) = \left| \frac{x - y}{y} \right| \quad (2.17)$$

Table 2.5 Relative differences

$Weibull(\beta, \eta)$	$Phases$	RD	$Weibull(\beta, \eta)$	$Phases$	RD
(0.9,1)	$n = 2, n = 3$	0.000462	(0.6,1)	$n = 2, n = 3$	0.023385
	$n = 3, n = 4$	0.000007		$n = 3, n = 4$	0.003385
(0.8,1)	$n = 2, n = 3$	0.003423	(0.5,1)	$n = 3, n = 4$	0.010263
	$n = 3, n = 4$	0.000292		$n = 4, n = 5$	0.001423
(0.7,1)	$n = 2, n = 3$	0.009756	(0.4,1)	$n = 3, n = 4$	0.036609
	$n = 3, n = 4$	0.001032		$n = 4, n = 5$	0.007919

As a result, we can clearly say that one can use Coxian distribution instead of Weibull distribution according to the following results:

- Coxian distribution with *four-phase* for Weibul distribution with β in $[0.4, 0.5], \eta = 1$,

- Coxian distribution with *three-phase* for Weibul distribution with β in $(0.5, 0.7), \eta = 1$,
- Coxian distribution with *two-phase* for Weibul distribution with β in $[0.7, 0.9], \eta = 1$.

Case 2: We require to fit Coxian distribution to log-normal distribution has the following density function.

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad x > 0, \sigma > 0, -\infty < \mu < \infty.$$

where σ and μ are shape parameter and scale parameter, respectively.

The log-normal distribution represents a heavy-tail distribution in case $\sigma > 1$ in Figure 2.8. So, we consider log-normal distribution with $\sigma > 1$ for the following example.

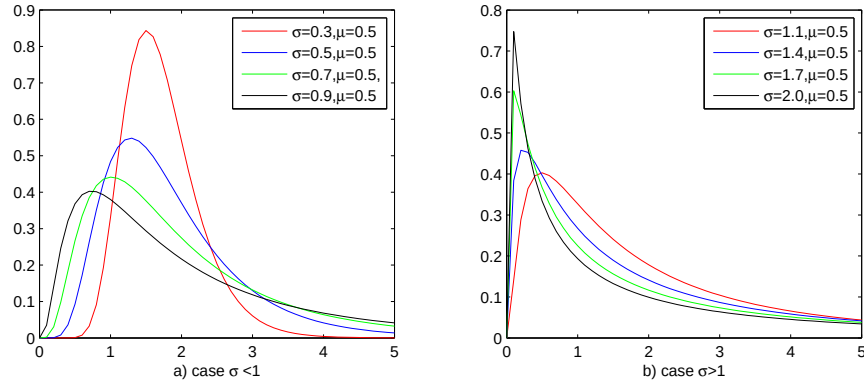


Figure 2.8 Lognormal distribution for various parameter values

Example 2.4: Similar to the previous example, the purpose of this example is to determine the required number of phase under the different scenarios of log-normal distribution. So, we fix $\mu = 0.5$ and vary $\sigma = 1.5, 2, 2.5$ for log-normal distribution. We also vary the number of phase, $n = 2, 3, 4$ for Coxian distribution. Log-likelihood values for estimation values and the relative differences between log-likelihood values by using (2.17) can be seen in Table 2.6.

Table 2.6 Log-likelihood values in fitting $\log - normal(\sigma > 1)$ to $Cox - n$

$Lognormal(\sigma, \mu)$	cv^2	$Log - like$	n	$Phases$	RD
(2.5,0.5)	2.6484	-2.083296	2	$n = 2, n = 3$	0.013245
	3.2030	-2.055885	3	$n = 3, n = 4$	0.000549
	3.2786	-2.054756	4		
(2.0,0.5)	2.6657	-2.230830	2	$n = 2, n = 3$	0.005488
	3.0727	-2.218620	3	$n = 3, n = 4$	0.000025
	3.0874	-2.218564	4		
(1.5,0.5)	2.5392	-2.228177	2	$n = 2, n = 3$	0.000793
	2.7250	-2.226412	3		
	-	-	4		

As a result, we can clearly say that one can use Coxian distribution instead of log-normal distribution according to the following results:

- Coxian distribution with *three-phase* for lognormal distribution with σ in (2,2.5] and $\mu = 0.5$.
- Coxian distribution with *two-phase* for lognormal distribution with σ in [1.5,2] and $\mu = 0.5$.

2.2 From Common Arrival Processes to Markovian Arrival Process

Poisson process is widely used to model an arrival process of queueing systems. That is, the inter-arrival distribution follows exponential distribution in this process. So, it only allows independent arrivals and arrival distribution should be Poisson distribution (or inter-arrival distribution should be exponential distribution). These assumptions may not hold in real applications. It is possible that the arrival process may not be a renewal process. For example, in a typical assembly line problem, the output from station 1 will form the input to station 2. If we consider non-Poisson arrivals, say, to the first node, the output process will not necessarily be a renewal

process. Also, in production line problems where arrivals from different sources form input to an assembly system, the arrival process may not necessarily be a renewal process. So, how do we model these point processes? The answer is to use Markovian Arrival process or simply referred to as *MAP*. In Neuts (1979), Neuts first introduced this process in the context of bulk arrivals as a Versatile Markovian Point Process (*VMPP*) with very heavy notation. Later on Lucantoni et al. (1990) simplified the original notation and introduced *Neuts' process* as Batch Markovian Arrival Process (*BMAP*). The notation are now standard in describing a *BMAP* as well as *MAP*, which is a special case of *BMAP* by setting the group size of the arrivals is 1.

2.2.1 Common Arrival Processes

Markovian arrival process is a generalization of some special processes such as Poisson process, renewal process, and *PH*-renewal process. To introduce the *MAP* process, we firstly recall the definitions of some stochastic processes below.

Counting process. A stochastic process $\{N(t), t \geq 0\}$ is a *counting process* if $N(t)$ represents the total number of occurred events by time t and satisfies the following axioms (see Ross (2014)):

- i. $N(t) \geq 0$.
- ii. $N(t)$ is integer valued.
- iii. If $s < t$, then $N(s) \leq N(t)$.
- iv. For $s < t$, $N(t) - N(s)$ equals the number of events that occur in the interval $(s, t]$.

Poisson process. A counting process $\{N(t), t \geq 0\}$ is a *Poisson process with rate λ* if it holds the following axioms (see Ross (2014)):

- i. $N(0) = 0$.

- ii. $\{N(t), t \geq 0\}$ has *independent increments*.
- iii. $P\{N(t+h) - N(t) = 1\} = \lambda h + o(h)$.
- iv. $P\{N(t+h) - N(t) \geq 2\} = o(h)$.

A Poisson process has *independent increments*. It means that the numbers of events in disjoint time intervals are independent. It has also *stationary increments*. It means that the number of events in the interval $(t, t+h)$ has Poisson distribution for all t .

Poisson process can be illustrated by Figure 2.9. X_i are independent exponential random variables with parameter λ . Also, let Y be the time elapsed until the first event after time t . Then Y has an exponential distribution with parameter λ . It is called the *memoryless property of Poisson process*.

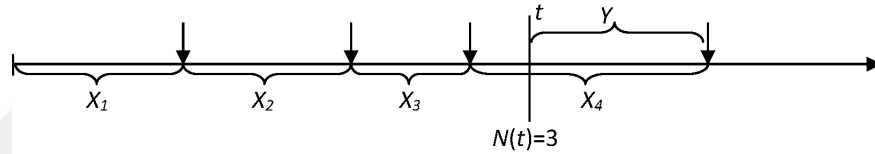


Figure 2.9 Description of Poisson process

The Poisson process with rate λ is a *pure birth process* whose the generator matrix is given by

$$Q = \begin{bmatrix} -\lambda & \lambda & & \\ & -\lambda & \lambda & \\ & & \ddots & \ddots \\ & & & \ddots & \ddots \end{bmatrix}.$$

Renewal process. A counting process $\{N(t), t \geq 0\}$ is called to be a *renewal process* if the times between successive events are independent and identically distributed with an arbitrary distribution. Namely, the distribution of the random variables X_i in Figure 2.9 is replaced by a common arbitrary distribution. Additionally, in this process, when an event occurs, it is said that a renewal happened. For example, let's use a lightbulb at a time, replace it with a new one when it fails. We assume that the lifetimes of the lightbulbs are independent and identically distributed.

If $N(t)$ is the number of lightbulbs that have failed by time t , we say that $\{N(t), t \geq 0\}$ is a renewal process. Note that, if the lifetimes of the lightbulbs are exponentially distributed, then $\{N(t), t \geq 0\}$ becomes a Poisson process.

PH-renewal process. The stochastic process is called to be a *PH-renewal process* if the times between successive events are independent and identically distributed with a common phase-type distribution. Namely, the distribution of the random variables X_i in Figure 2.9 is replaced by a common phase-type distribution.

We consider a *CTMC* with the infinitesimal generator matrix Q in (2.1). When an absorption occurs in the *CTMC*, we say that a renewal has occurred and the chain immediately starts from one of the n states according to the initial probability vector β . It can be seen that there are two-type of transitions in the process. i) The transitions without renewal are described by the matrix T , that is, the Markov chain transits from phase i to phase j with transition rate t_{ij} for $i \neq j$. ii) The transitions with renewal are described by the matrix $T^0\beta$, namely, the Markov chain transits from phase i to the absorption phase and is immediately reinitialized in phase j with transition rate $(T^0)_i(\beta)_j$.

$$T = \begin{bmatrix} -t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & -t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & -t_{nn} \end{bmatrix}, \quad T^0\beta = \begin{bmatrix} T_1^0\beta_1 & T_1^0\beta_2 & \cdots & T_1^0\beta_n \\ T_2^0\beta_1 & T_2^0\beta_2 & \cdots & T_2^0\beta_n \\ \vdots & \vdots & \ddots & \vdots \\ T_n^0\beta_1 & T_n^0\beta_2 & \cdots & T_n^0\beta_n \end{bmatrix}. \quad (2.18)$$

Note that the process restart in phase j with probability β_j . It is independent from past history, namely, the absorption phase is not important. See the columns of the matrix $T^0\beta$ in (2.18).

The stochastic process $\{N(t), I(t), t \geq 0\}$ is also a two dimensional *CTMC* with the infinitesimal generator matrix in (2.19). That is, $\{N(t), t \geq 0\}$ is a counting process where $N(t)$ denotes the number of renewals in $(0, t]$ while $\{I(t), t \geq 0\}$ is underlying

CTMC (also called phase process) with the generator matrix $T + T^0\beta$.

$$Q = \begin{bmatrix} T & T^0\beta & & & \\ & T & T^0\beta & & \\ & & T & T^0\beta & \\ & & & \ddots & \ddots \end{bmatrix}. \quad (2.19)$$

2.2.2 Markovian Arrival Process

We consider an irreducible *CTMC* has m transient states. The Markov chain stays in state i for the amount time that is exponentially distributed with parameter λ_i . At the end of this sojourn, there are two transient possibilities; transient with an arrival and transient with no arrival. The transient with an arrival means that the *CTMC* visits any of the m transient states with probability p_{ij} , while the transient corresponds to no arrival means that the *CTMC* visits any of the $m - 1$ transient states except i with probability q_{ij} . The mentioned transitions are illustrated in Figure 2.10.

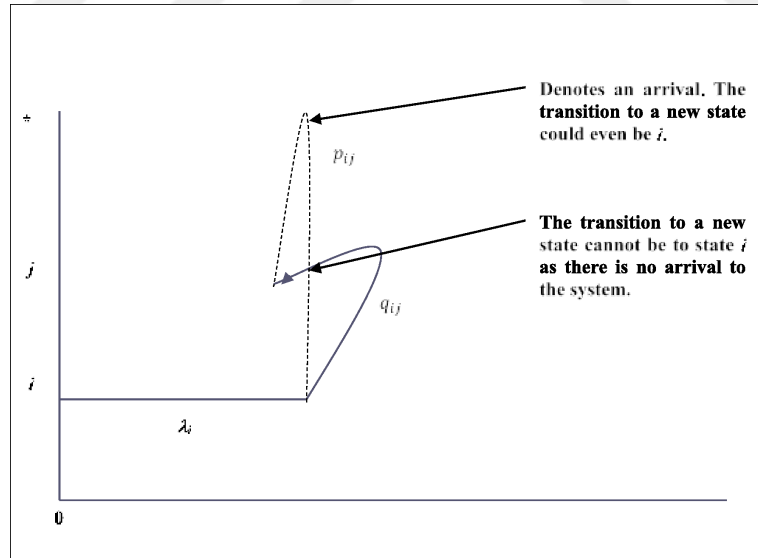


Figure 2.10 Description of Markovian arrival process

We also describe the transitions in following matrix forms. The matrix D_0 governs the transitions without an arrival while the transitions with an arrival are governed by the matrix D_1 . Thus, we say that a *MAP* is determined by using the two parameter

matrices (D_0, D_1) of order m . The matrix D_1 has the nonnegative elements whereas the diagonal elements of the matrix D_0 are negative and the other elements are nonnegative.

$$D_0 = \begin{bmatrix} d_{11}^{(0)} & d_{12}^{(0)} & \cdots & d_{1m}^{(0)} \\ d_{21}^{(0)} & d_{22}^{(0)} & \cdots & d_{2m}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}^{(0)} & d_{m2}^{(0)} & \cdots & d_{mm}^{(0)} \end{bmatrix} = \begin{bmatrix} -\lambda_1 & \lambda_1 q_{12} & \cdots & \lambda_1 q_{1m} \\ \lambda_2 q_{21} & -\lambda_2 & \cdots & \lambda_2 q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m q_{m1} & \lambda_m q_{m2} & \cdots & -\lambda_m \end{bmatrix}. \quad (2.20)$$

$$D_1 = \begin{bmatrix} d_{11}^{(1)} & d_{12}^{(1)} & \cdots & d_{1m}^{(1)} \\ d_{21}^{(1)} & d_{22}^{(1)} & \cdots & d_{2m}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}^{(1)} & d_{m2}^{(1)} & \cdots & d_{mm}^{(1)} \end{bmatrix} = \begin{bmatrix} \lambda_1 p_{11} & \lambda_1 p_{12} & \cdots & \lambda_1 p_{1m} \\ \lambda_2 p_{21} & \lambda_2 p_{22} & \cdots & \lambda_2 p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m p_{m1} & \lambda_m p_{m2} & \cdots & \lambda_m p_{mm} \end{bmatrix}. \quad (2.21)$$

Note that

$$\sum_{j=1}^m p_{ij} + \sum_{j=1, j \neq i}^m q_{ij} = 1, \quad 1 \leq i \leq m. \quad (2.22)$$

The matrices D_0 and D_1 are called *rate matrices*. $d_{ii}^{(0)}$, $1 \leq i \leq m$ denotes the total sojourn time in state i . $d_{ij}^{(0)}$, $j \neq i$, $1 \leq i, j \leq m$ denotes the transient rate from state i to state j when there is no arrival. $d_{ij}^{(1)}$, $1 \leq i, j \leq m$ represents the transient rate from state i to state j when an arrival occurs.

Note that the process is not renewal. After absorption from phase i , the process restarts from phase j with probability p_{ij} , $i = 1, \dots, m$. Namely, the restart of the process depends on from where the absorption occurs. See the columns of the matrix D_1 in (2.21).

MAP is a two-dimensional Markov process denoted by $\{(N(t), I(t)), t \geq 0\}$. Furthermore, $\{N(t), t \geq 0\}$ and $\{I(t), t \geq 0\}$ are called *counting process* and *phase process* (we will also say the underlying *CTMC* to the phase process), respectively. In this case, we can restate the definition of *MAP*. Suppose the process is in phase i ; $I(t) = i$. If an arrival occurs corresponding to $d_{ij}^{(0)}$, $j \neq i$, the variable $I(t)$ changes from phase i to phase j and the variable $N(t)$ does not change. On the other hand, if an arrival occurs corresponding to $d_{ij}^{(1)}$, the variable $I(t)$ changes from phase i to

phase j and the variable $N(t)$ also changes that it is increased by one.

The *MAP* $\{(N(t), I(t)), t \geq 0\}$ has the state space $\Omega = \{(i, j) : i \geq 0, 1 \leq j \leq m\}$ with the generator matrix

$$Q = \begin{bmatrix} D_0 & D_1 & & & \\ & D_0 & D_1 & & \\ & & \ddots & \ddots & \\ & & & D_0 & D_1 \end{bmatrix}. \quad (2.23)$$

The underlying *CTMC* is governed by the generator matrix $D = D_0 + D_1$. Let π denotes the steady-state probability vector of the generator matrix D . The vector satisfies

$$\pi D = 0, \quad \pi e = 1. \quad (2.24)$$

Let α denotes the initial probability vector of the underlying *CTMC* with generator matrix D . The stationary version of *MAP* can be obtain by $\alpha = \pi$.

The *fundamental rate* (also called the rate of arrivals per unit of time) is given as

$$\lambda = \pi D_1 e. \quad (2.25)$$

Example 2.5: We want to give this example from He (2010) for illustration of *MAP*. Consider a reliability system with two identical units and a repairman.

- If both units are functioning, then one is in work and the other one is on cold standby.
- If the unit in work fails, it is sent to the repairman for repair and the standby unit is put in work.
- If repair is completed before failure, the repaired unit is on cold standby.
- If failure occurs before repair completion, the failed unit has to wait for repair.

- A repaired unit is put in work immediately if the other unit has failed.
- The times to failure and repair times are exponentially distributed with parameters λ and μ , respectively.

The state of each component can be 0: in repair, 1: waiting for repair, 2: on cold standby, and 3: in work. The system has three phases: $\{(3, 2), (3, 0), (1, 0)\}$ since the two units are identical.

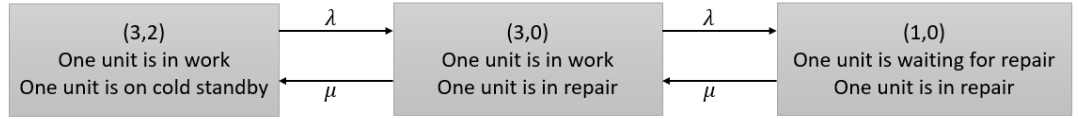


Figure 2.11 The states of a system with two identical units and a repairman

Let $I(t)$ be the status of the system at time t . Then, the process $\{I(t), t \geq 0\}$ has the following infinitesimal generator matrix.

$$D = \begin{bmatrix} -\lambda & \lambda & 0 \\ \mu & -\lambda - \mu & \lambda \\ 0 & \mu & -\mu \end{bmatrix}.$$

We are interested in the number of repairs in $[0, t]$. We define the arrivals as the events associated with repairs. Let $N(t)$ be the number of repairs completed in $[0, t]$. Then, $\{N(t), I(t), t \geq 0\}$ is an *MAP* with a matrix representation

$$D_0 = \begin{bmatrix} -\lambda & \lambda & 0 \\ 0 & -\lambda - \mu & \lambda \\ 0 & 0 & -\mu \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu & 0 \end{bmatrix}.$$

If $\lambda = 0.01$ and $\mu = 0.5$, the stationary distribution of underlying Markov chain $\{I(t), t \geq 0\}$ and the arrival rate are calculated by $\pi = (0.9800, 0.0196, 0.0004)$ and 0.009996 from (2.24) and (2.25), respectively. We can say that repair completion occurs 0.009996 times per unit time.

2.3 Quasi-birth-death Process and Matrix Geometric Solution

The birth-death processes are the stochastic processes that have Markov property. Each state can be associated with an element and transitions in this process should occur only to neighboring states. That is, if the process is in state n , it can have a transition only to state $n + 1$ (called birth) or state $n - 1$ (called death). Figure 2.12 represents the state transitions.



Figure 2.12 Transition diagram of birth-death process

While determining the stationary state probabilities of a Markov process, a set of linear equations are solved derived from the process. The solution can be obtained by using structural properties of the equations. For example, in $M/M/1$ queueing models that are birth-death processes, these equations have a *repetitive structure*. This repetition allows to determine a recursive solution for the stationary state probabilities. That is, if the stationary probability for any state i is known, then the stationary probability for state $i + 1$ can be determined. At this point, it can be said that the stationary state probabilities in the process have a *geometric form*.

The quasi-birth-death (*QBD*) process is a generalization of the birth-death process in Fadiloğlu & Yeralan (2002). In the case of the *QBD*, there is a group of states instead of one state. Namely, each state in Figure 2.12 has been replaced by the set of states as in Figure 2.13. This set of states is called *level*. Transitions are now restricted to neighboring levels. We say that if the process is in level n , it can have a transition only to level $n + 1$ or level $n - 1$. On the other hand, note that transitions in the process may occur within the same level as well as between the neighboring levels.

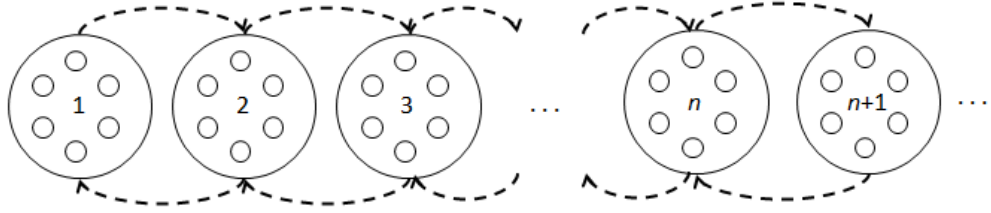


Figure 2.13 Transition diagram of quasi-birth-death process

In this case, a certain repetitive structure and a recursive procedure can be used to determine the stationary state probabilities of the $(i + 1)th$ level in terms of the probabilities for the ith level. The form of the solution is a generalization of the scalar case and is called *Neuts' matrix geometric form*.

Queueing systems can be determined with *QBD* process and the solution of the system is achieved by matrix-geometric method (*MGM*) by means of mentioned repetitive structure in the *QBD* process. The *QBD* process is configured based on the structure of the state space that can be occurred *finite* or *infinite*, also on whether the system has *one-server* or *multiple-server*.

One server queueing system with infinite state can be modeled by *QBD* process. This process is a bivariate Markov process $\{N(t), I(t), t \geq 0\}$ defined on the state space $\Omega = \{(i, j) : i \geq 0, 1 \leq j \leq n\}$ where i is the level of the process and j is the phase of the process. The *QBD* process has the infinitesimal generator matrix with *block tridiagonal structure* as follows.

$$Q = \begin{bmatrix} B_0 & C & & & \\ A_0 & B & C & & \\ & A & B & C & \\ & & A & B & C \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \quad (2.26)$$

The off-diagonal components of the matrix Q in (2.26) are nonnegative and its diagonal components are negative. Also, the components satisfy $B_0e + Ce = A_0e + Be + Ce = (A + B + C)e = 0$. The upper diagonal and lower diagonal components represent the

arrivals to the system and the departures from the system, respectively. The matrix Q is assumed to be *irreducible*. Necessary condition for this is that the matrices B_0 and B (all diagonal matrices) are *nonsingular*. The inverses B_0^{-1} and B^{-1} are then *nonpositive*. Also, the matrix $A+B+C$ is a finite generator matrix. The Theorem 2.3.1 in Neuts (1981) is given as follows.

Theorem 2.3.1. *The process Q is positive recurrent if and only if the minimal nonnegative solution R to the matrix-quadratic equation*

$$R^2 A + R B + C = 0 \quad (2.27)$$

has all its eigenvalues inside the unit disk and the finite system of equations

$$\begin{aligned} x_0(B_0 + R A_0) &= 0 \\ x_0(I - R)^{-1} \mathbf{e} &= 1 \end{aligned}$$

has a (unique) positive solution x_0 .

If the matrix $A + B + C$ is irreducible, then $\text{sp}(R) < 1$ if and only if

$$\pi C \mathbf{e} < \pi A \mathbf{e} \quad (2.28)$$

where π is the stationary probability vector of the matrix $A + B + C$.

The stationary probability vector of the generator matrix Q in (2.26) is $x = (x_0, x_1, \dots)$ and is obtained by

$$x_i = x_0 R^i, \quad i \geq 0.$$

2.3.1 $M/PH/1$ Queueing Model

In the $M/PH/1$ queue model, the arrival process is a Poisson process with parameter λ , and the service times have a common phase-type distribution with representation (β, T) of order n . Customers join a single queue and served on a *FCFS*

basis. Also, the waiting space is infinite size.

$\{(N(t), I_s(t)), t \geq 0\}$ is a stochastic process where $N(t)$ is the total number of customers in the system at time t ; $I_s(t)$ is the phase of the service process at time t , if the server is busy at time t (0, otherwise). The time until the next arrival is exponentially distributed with parameter λ at any time t , since the arrival process is a Poisson process. In addition to that the residual service time (time until the next service completion) is known probabilistically due to the service phase $I_s(t)$ is recorded. Namely, if we know $N(t)$ and $I_s(t)$ at time t , we can probabilistically determine the future arrival and service processes.

Then $\{(N(t), I_s(t)), t \geq 0\}$ is a CTMC with state space $\Omega = \{0\} \cup \{(i, j) : i \geq 1, 1 \leq j \leq n\}$. The state 0 represents the empty system, the state (i, j) corresponds to i customers in the system and the service process in the phase j . The infinitesimal generator matrix is given by

$$Q = \begin{bmatrix} -\lambda & \lambda\beta & & & \\ \mathbf{T}^0 & -\lambda I + T & \lambda I & & \\ & \mathbf{T}^0\beta & -\lambda I + T & \lambda I & \\ & & \mathbf{T}^0\beta & -\lambda I + T & \lambda I \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \quad (2.29)$$

The generator matrix Q in (2.29) has block-tridiagonal structure. So, we say that $\{(N(t), I_s(t)), t \geq 0\}$ is a QBD process. Some examples for the interpretations of the transition rates are given below.

- For $N(t) = 0$, the system is idle, that is, there is no service and the next event is an arrival. So, there are two possibilities for the next transition:
 - i. the arrival of a customer occurs by rate λ and for this arrival a service phase j is selected according to β . Then the rate for this transition is given by $\lambda\beta$ and $N(t)$ increases to 1.
 - ii. the arrival of no customer occurs. The $N(t)$ remains to be zero. The rate

for this transition is given by $-\lambda$.

- For $N(t) = 1$, there is only one customer in the system and the service process is in the phase j . Namely, the customer (or the arrival) is being served and he is in phase j of service; and also the waiting line is empty. So, there are three possibilities for the next transition:

- i. the arrival of a new customer occurs by rate λ and this arrival joins the waiting line. Then the rate for this transition is given by λI and $N(t)$ increases to 2.

- ii. the departure of the customer in service phase j occurs and the system becomes idle. The rate for this transition is given by T^0 and $N(t)$ decreases to 0.

- iii. no event occurs. The $N(t)$ remains the same. The rate for this transition is given by $-\lambda I + T$.

- For $N(t) = 2$, there are two customers in the system (one in the service and the other is in the queue) and the service process is in the phase j . So, there are three possibilities for the next transition:

- i. the arrival of a new customer occurs by rate λ and this arrival joins the waiting line. Then the rate for this transition is given by λI and $N(t)$ increases to 3.

- ii. the departure of the customer in service phase j occurs and a service phase is selected for the waiting customer in the queue. The rate for this transition is given by $T^0 \beta$ and $N(t)$ decreases to 1.

- iii. no event occurs. The $N(t)$ remains the same. The rate for this transition is given by $-\lambda I + T$.

In conclusion, the four components of the generator matrix Q can be interpreted as below.

- $\lambda\beta$: The system is empty. An arrival enters the system and immediately starts to service in phase j .
- λI : There are at least one customer in the system. An arriving customer joins the queue.
- T^0 : The queue is empty. The service is completed and the customer departs from the system.
- $T^0\beta$: The queue is not empty. After the service is completed, the customer departs from the system and the customer in the queue starts to service in phase j .

The stationary probability vector of the matrix Q in (2.29) is $x = (x_0, x_1, x_2, \dots)$ satisfies $xQ = 0$, $xe = 1$. Using the Theorem 2.3.2 given by Neuts (1981), the probabilities can be calculated.

Theorem 2.3.2. *The mean service time is $(-\beta T^{-1}e)$ and the arrival rate is λ . We assume that $\rho = \lambda(-\beta T^{-1}e) < 1$ and have*

$$\begin{aligned} x_0 &= 1 - \rho, \\ x_i &= (1 - \rho)\beta R^i, \quad i \geq 1. \end{aligned}$$

The matrix R is given by $R = \lambda(\lambda I - \lambda e\beta - T)^{-1}$.

For the proof of the theorem, we refer to books Neuts (1981) and He (2014). After the stationary probabilities are obtained, the related performance measures for the $M/PH/1$ model can be obtained.

2.3.2 $MAP/PH/1$ Queueing Model

In the $MAP/PH/1$ queue model, the arrival process is a Markovian arrival process with representation (D_0, D_1) of order m , and the service times have a common phase-

type distribution with representation (β, T) of order n . Customers join a single queue and are served on a *FCFS* basis. Also, the waiting space is infinite.

$\{(N(t), I_s(t), I_a(t)), t \geq 0\}$ is a stochastic process where $N(t)$ is the total number of customers in the system at time t ; $I_s(t)$ is the phase of the service process at time t , if the server is busy at time t (0, otherwise); $I_a(t)$ is the phase of the arrival process at time t . If we know $N(t)$, $I_s(t)$ and $I_a(t)$ at time t , we can probabilistically determine the future arrival and service processes.

Then $\{(N(t), I_s(t), I_a(t)), t \geq 0\}$ is a *CTMC* with state space $\Omega = \{(0, k) : 1 \leq k \leq m\} \cup \{(i, j, k) : i \geq 1, 1 \leq j \leq n, 1 \leq k \leq m\}$. The state $(0, k)$ represents that the system is empty and the arrival process is in phase k ; the state (i, j, k) corresponds to there are i customers in the system, the service process in phase j and the arrival process in phase k . The infinitesimal generator matrix is given by

$$Q = \begin{bmatrix} D_0 & \beta \otimes D_1 & & & \\ \mathbf{T}^0 \otimes I & T \oplus D_0 & I \otimes D_1 & & \\ & \mathbf{T}^0 \beta \otimes I & T \oplus D_0 & I \otimes D_1 & \\ & & \mathbf{T}^0 \beta \otimes I & T \oplus D_0 & I \otimes D_1 \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \quad (2.30)$$

The generator matrix Q in (2.30) has block-tridiagonal structure. So, we say that $\{(N(t), I_s(t), I_a(t)), t \geq 0\}$ is a *QBD* process. The transition rates in the matrix Q can be interpreted as follows.

- The system starts in state $(0, l)$. There are two possibilities for the next transition:
 - i. The system makes a transition to state $(1, j, k)$. An arrival occurs by rate $d_{lk}^{(1)}$ and for this arrival a service phase j is selected according to β . Then the rates for these transitions are governed by the matrix $\beta \otimes D_1$.
 - ii. No arrival occurs. The arrival process can change its phase by no arrival and the transition rates are governed by the matrix D_0 .
- The system is in state $(1, j, l)$. There are three possibilities for the next transition:

- i. The system makes a transition to state $(2, j, k)$. An arrival occurs by rate $d_{lk}^{(1)}$ and the service phase does not change. Then the rates for these transitions are governed by the matrix $I \otimes D_1$.
 - ii. The system makes a transition to state $(0, l)$. A departure from service phase j occurs and the system becomes idle. Note that the arrival phase does not change. The transitions are given by the matrix $T^0 \otimes I$.
 - iii. No event occurs. However, a change is possible in the arrival or service phase. The rates for these transitions are given by the matrix $T \oplus D_0$.
- The system is in state $(i, r, l), i \geq 2$. There are three possibilities for the next transition:
 - i. The system makes a transition to state $(i + 1, r, k)$. An arrival occurs by rate $d_{lk}^{(1)}$. The transitions are governed by the matrix $I \otimes D_1$.
 - ii. The system makes a transition to state $(i - 1, j, l)$. A departure from service phase r occurs and a service phase j is selected according to β for waiting customer. The transitions are given by the matrix $T^0 \beta \otimes I$.
 - iii. No event occurs. The rate for this transition is given by the matrix $T \oplus D_0$.

The stationary probability vector of the matrix Q in (2.30) is $x = (x_0, x_1, x_2, \dots)$ and satisfies $xQ = 0, xe = 1$. These equations can be expanded as the following:

$$\begin{aligned}
 x_0 D_0 + x_1 (T^0 \otimes I) &= 0, \\
 x_0 (\beta \otimes D_1) + x_1 (T \oplus D_0) + x_2 (T^0 \beta \otimes I) &= 0, \\
 x_{i-1} (I \otimes D_1) + x_i (T \oplus D_0) + x_{i+1} (T^0 \beta \otimes I) &= 0, \quad i \geq 2,
 \end{aligned}$$

with the normalizing equation $\sum_{i=0}^{\infty} x_i e = 1$.

In Bhat (2015), the solution to the above equations are given by

$$\begin{aligned}x_1[(\mathbf{T}^0 \boldsymbol{\beta} \otimes (-D_0^{-1} D_1)) + (T \oplus D_0) + R(\mathbf{T}^0 \boldsymbol{\beta} \otimes I)] &= 0, \\x_1[(\mathbf{T}^0 \otimes (-D_0^{-1} e)) + (I - R)^{-1} e] &= 1. \\x_i &= x_1 R^{i-1}, i \geq 2\end{aligned}$$

where the rate matrix R satisfies the matrix-quadratic equation $R^2(\mathbf{T}^0 \boldsymbol{\beta} \otimes I) + R(T \oplus D_0) + (I \otimes D_1) = 0$.

The matrix R is explicitly obtained for the $M/PH/1$ queue model (See in Section 2.3.1), but, usually the matrix can not be achieved explicitly. Now, an algorithm is needed for the calculation of the matrix R in the $MAP/PH/1$ queue model. The key steps for computing the matrix R that is the solution to the equation form $R^2 A + RB + C = 0$ in (2.27) are given below. For full details we refer to Latouche & Ramaswami (1999).

Logarithmic Reduction Algorithm for R :

Step 0: $H \leftarrow (-B)^{-1}C$, $L \leftarrow (-B)^{-1}A$, $G = L$, and $T = H$.

Step 1:

$$\begin{aligned}U &= HL + LH \\M &= H^2 \\H &\leftarrow (I - U)^{-1}M \\M &\leftarrow L^2 \\L &\leftarrow (I - U)^{-1}M \\G &\leftarrow G + TL \\T &\leftarrow TH\end{aligned}$$

Continue Step 1 until $\|e - Ge\|_\infty < \epsilon$, where ϵ is a very small prespecified value.

Step 2: $R = -C(B + CG)^{-1}$.

CHAPTER THREE

COX SERVICE QUEUEING MODELS AND OPTIMIZATION FOR A GERIATRIC CARE CENTER

In this chapter, three different multi-server queue models are studied. The main difference among these models is the behaviour of customers when an arrival finds all servers busy in the system. Note that these differences arise from the capacity of the system. The first model is $M/C_2/c/c$ model in which a customer leaves without having service if all servers are busy. The second one is $M/C_2/c$ model where there is no system capacity. The customers spend in waiting line for service until any server is idle. Finally, in $M/C_2/c/K$ model where the system capacity is K , any limited customers can wait in the line .

To determine the optimal number of bed in the geriatric department of a hospital, the above mentioned queueing models are analyzed. We describe patient arrivals to the geriatric department by Poisson process, beds corresponds to servers and length of stay in the department are modelled by using Cox distribution.

This chapter is structured as follows. In Section 3.1, we describe the $M/C_2/c/c$ model and perform the numerical studies including the performance of the system and optimization. In Section 3.2, we define the $M/C_2/c$ model and propose the methodology to calculate the steady-state probabilities and the average queue length by using QBD process. Then furthermore in this section we illustrate numerical studies as in Section 3.1. Finally, we define the $M/C_2/c/K$ model and propose methodology to analyze the model by QBD process in Section 3.3. This last section is ended by numerical studies including optimization and comparing our results with the ones that exists in the literature.

3.1 $M/C_2/c/c$ Queueing Model

In this section, the queueing systems in Gorunescu et al. (2002a) and Bernardo et al. (2003) have been used to model the length of patients' stay in geriatric department and determine the optimal number of beds in this department. The lengths of stay was represented by Cox distribution instead of the hyper-exponential distribution which is used in Gorunescu et al. (2002a) and mixed generalized Erlang distribution which is used in Bernardo et al. (2003). The Cox distribution is also useful tool to model the length of patients' stay in a geriatric department, because elderly patients may stay in the hospital for very long times as well as very short times. So, heterogeneity in lengths of stay is captured by Cox distribution.

3.1.1 Model Description

We consider a $M/C_2/c/c$ model in which *no queueing* is allowed to describe the hospital system. The patient arrivals occur according to a Poisson process with rate λ . c beds (servers) are available in the department and the probability of an arrival to be served by any idle bed is equal. Beds are identical and service time in each bed follows 2-phase Cox distribution. That is, service process consists of two phases connected in series, each with exponential distribution with parameters μ_1 and μ_2 . There are three possible cases in the system represented by Figure 3.1; i) the patient finding upon arrival that all c beds are occupied is lost, ii) the patient starts the service process from the first phase and proceeds to the second phase with probability q ; this means he stays for a long time in service, iii) the patient leaves from the service process with probability q' ; this means he occupies the bed for a short time.

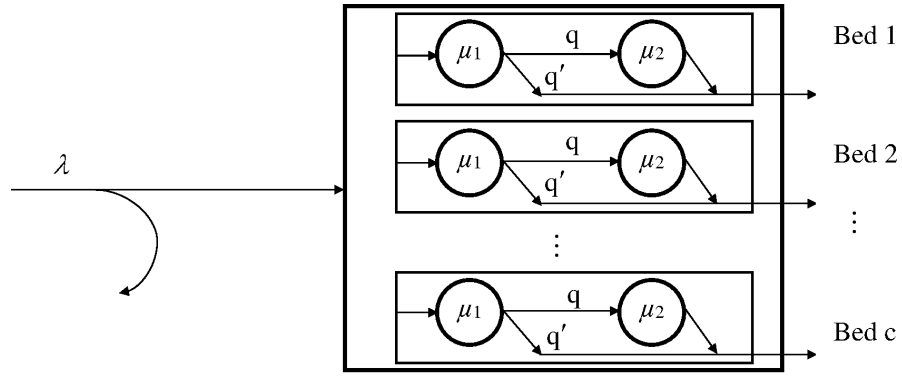


Figure 3.1 A schematic representation of $M/C_2/c/c$ queue model

3.1.2 Performance Measures and Optimization

We use the following performance measures as given in Gorunescu et al. (2002a) and Bernardo et al. (2003) to characterize the system.

- We used Cox distribution with 2-phase for the patients' lengths of stay . So, the average length of stay is given by $E(S) = \frac{\mu_2 + \mu_1 q}{\mu_1 \mu_2}$. We know that the average number of arrivals in an interval of length t is λt . Therefore, we can say that the average number of arrival during an average length of stay (or the *offered load*) is given by

$$a = \lambda E(S). \quad (3.1)$$

- From queueing theory, the probability that all c beds are occupied is given by the following Erlang's loss formula. We can also refer this measure as the probability that an arriving patient is lost.

$$B(c, a) = \frac{a^c / c!}{\sum_{k=0}^c \frac{a^k}{k!}}. \quad (3.2)$$

- Then, the average number of occupied beds (known as the *carried load*) is given as in (3.3). In other words it is called the average number of served patients.

$$a' = a[1 - B(c, a)]. \quad (3.3)$$

- The last quantity is defined for the bed occupancy as following. This measure is used to explore the degree of utilisation of beds.

$$\rho = \frac{a'}{c}. \quad (3.4)$$

To decide the optimal number of beds, Gorunescu et al. (2002a) describes two costs; *penalty cost* incurred for each patient that is turned away and *holding cost* for each empty bed per day. On the other hand Bernardo et al. (2003) uses *patient cost* for each patient (occupied bed) based on the length of stay in addition to the costs defined by Gorunescu et al. (2002a). We used these three costs for optimization as given in Bernardo et al. (2003). But note that our penalty cost includes two different costs unlike in Bernardo et al. (2003) because there are two phases in service process.

- π is the fixed *penalty cost* incurred for each patient that is not being served. Then, the average cost per day caused by a lost patient is

$$\pi \lambda B(c, a). \quad (3.5)$$

- h is the daily *holding cost* for each unoccupied bed. Note that the average number of unoccupied beds is equal to total number of beds minus the average number of occupied beds. Thus, the average cost per day incurred due to unoccupied beds is

$$h \left[c - a[1 - B(c, a)] \right]. \quad (3.6)$$

- r is the *patient cost* for each patient (occupied bed) based on the length of stay. That is, let's use r_S is short stay cost and r_L is long stay cost. Let p_S and p_L be the fraction of patients stay in the department for short time and long time, respectively. r satisfies $r = r_S p_S + r_L p_L$ and then the average cost per day incurred due to occupied beds is given by

$$r a [1 - B(c, a)]. \quad (3.7)$$

Finally, the total cost is given by

$$g(c) = \pi\lambda B(c, a) + h \left[c - a[1 - B(c, a)] \right] + ra[1 - B(c, a)]. \quad (3.8)$$

3.1.3 Numerical Study

The arrival rate of patients is assumed $\lambda = 5$ per day. Each phase of the service process is considered exponentially distributed with parameters $\mu_1 = 1.7$ and $\mu_2 = 1.2$. Namely, patients occupy in the system for a short time period or a long time period according to exponential distribution with parameter $\mu_1 = 1.7$ and $\mu_2 = 1.2$, respectively. Also, a patient is served in phase 1 may continue to the in phase 2 with probability $q = 0.2$ or the patient leaves the system with probability $q' = 1 - q = 0.8$. Thus, the average length of stay is obtained as $E(S) = 0.755$ days.

Example 3.1: This example illustrates how the performance of the system changes as the number of beds increases. After the offered load is obtained as $a = 3.775$ by using equation (3.1), the performance measures of the system can be obtained as given in Table 3.1.

Table 3.1 Performance measures of $M/C_2/c/c$ queue model

c	$B(c, a)$	$a[1 - B(c, a)]$	$a[1 - B(c, a)]q'$	$a[1 - B(c, a)]q$	Bed occupancy
2	0.5987	1.5147	1.2117	0.3029	0.7573
3	0.4296	2.1528	1.7223	0.4306	0.7176
4	0.2885	2.6857	2.1485	0.5371	0.6714
5	0.1788	3.0995	2.4796	0.6199	0.6199
6	0.1011	3.3928	2.7143	0.6786	0.5655
7	0.0517	3.5793	2.8635	0.7159	0.5113
8	0.0238	3.6846	2.9477	0.7369	0.4606

We can easily see from Table 3.1 that, while the number of beds is increased, the probability that a patient is turned away (in the second column) and the proportion of bed occupancy (in the last column) is decreased, and the average number of patients

in department is increased (as we would expect). For example, the table says that when there are 5 beds, 17.88% of patients are turned away from geriatric department and on average there are 3.0995 patients occupying beds, that is, 2.4796 of the patients shortly occupy beds while 0.6199 patients occupy the beds for a long time. We can also say from the table that the hospital wish to ensure that at most 17.88% of patients are turned away, it must provide at least 5 beds in the geriatric department.

Example 3.2: This example is performed to determine the optimal number of beds. We assume the penalty and holding costs are $\pi = 100$ and $h = 20$, respectively. Also, we calculate the patient cost as $r = 10(0.8) + 20(0.2) = 12$ by assuming the short stay cost $r_S = 10$ and long stay cost $r_L = 20$. The related costs for the system are represented in Table 3.2.

Table 3.2 The system costs for different number of beds

c	$\pi \lambda B(c, a)$	$h[c - a[1 - B(c, a)]]$	$ra[1 - B(c, a)]$	<i>Total Costs</i>
4	144.2347	26.2864	32.2281	202.7493
5	89.4121	38.0093	37.1944	164.6159
6	50.5600	52.1434	40.7140	143.4174
7	25.8531	68.4131	42.9521	137.2184
8	11.9074	86.3076	44.2155	142.4304
9	4.9444	105.2563	44.8462	155.0470
10	1.8793	124.7905	45.1257	171.7756

We can easily see from Table 3.2 that while the number of beds is increased, the average penalty cost that caused by lost patient is decreased, on the other hand, the average holding cost due to unoccupied beds and the average patient cost due to occupied beds are increased. We also observe that the increment in the patient cost slowly occurs comparing with the holding cost. From the column of the total cost, it can be seen that the minimum cost 137.2184 is obtained when the department has seven beds. Then, we say that the hospital must provide seven beds in department.

3.2 $M/C_2/c$ Queueing Model

We model the hospital system in Section 3.1 by using $M/C_2/c$ queueing system. Now, this queue model allow that the arriving patients wait until any stuffed bed is idle. Here, the waiting of patients means that these patients are served in unstaffed beds.

We propose a methodology to solve the $M/C_2/c$ queueing system analytically. After we provide a *QBD* process representation of the multi-server queueing system where the service time is modeled by the two-phase Cox distribution, we see that the system equations based on the $M/C_2/c$ model have a repetitive structure. This repetition allows to determine a recursive solution for the stationary state probabilities. In other words, once the infinitesimal generator matrix of the $M/C_2/c$ model is obtained, we recognize that the matrix has a block tri-diagonal form and mentioned repetitive structure. The block tri-diagonal structure allows us to use the matrix geometric method (*MGM*).

3.2.1 Model Description

We consider the multi-server queue system $M/C_2/c$ in which *waiting* is allowed to describe the hospital system. The definition of the queue model is the same as given in Section 3.1 except waiting space assumption. This time all beds are busy, an arriving patient will not lost. It is assumed these patients are served in unstaffed beds (in beds of another departments). If any staffed bed is idle, the patient in unstaffed bed is moved to geriatric department. Namely, we can say that a patient may behave in three manners in Figure 3.2; i) the patient finding upon arrival that all c beds are occupied waits until any bed is empty, ii) the patient starts the service process from the first phase and proceeds the second phase with probability q , that is he stays for a long time in service, iii) the patient leaves from the service process with probability q' , that is he occupies the bed for a short time.

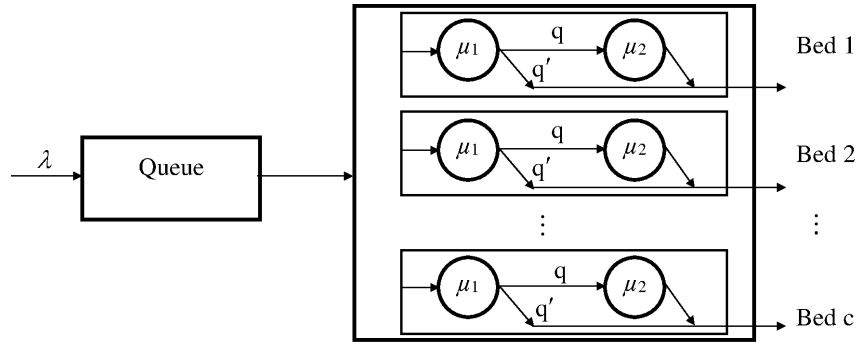


Figure 3.2 A schematic representation of $M/C_2/c$ queue model

3.2.2 QBD Process

The defined $M/C_2/c$ queuing system can be characterized by the bivariate Markov process $\{(k, i), k \geq 0, i \in I\}$, where k denotes the "level" of the process that corresponds to the number of customers in the system and i refers to the "states" at the corresponding level. State i is represented by two indices as the service unit is two-phase, and I is the set of possible values pairwise. Each of the indices contains the information regarding the number of customers in each phase of the service.

However, for the multi-server case, the customers in different servers may be in different phases. Hence each element in the state space is represented by two indices. The first index refers to the number of customers in the system and the second one contains the information regarding the number of customers in each phase. The general form of the transitions in the generator matrix for a multi-server queue can be examined in two main cases. For $0 \leq k \leq c - 1$, the number of states at each level would be different from each other and can be found by $[k, (k - j, j)]$, where $j = 0, 1, \dots, k$. Furthermore, for $c \leq k$, the number of states at each level is fixed and can be represented by $[k, (c - j, j)]$, where $j = 0, 1, \dots, c$. Then, the Markov process is defined on the *infinite state space* as follows:

$$\Omega = \begin{cases} (k, (k - j, j)) & \text{if } 0 \leq k \leq c - 1, \text{ and } j = 0, 1, \dots, k \\ (k, (c - j, j)) & \text{if } c \leq k, \text{ and } j = 0, 1, \dots, c \end{cases}$$

For instance, let us take the case of three servers. For level one, because there is only one customer in the system, the number of customers in phase one or the number of customers in phase two is one. That is, $k = 1$ and $i \in \{(1, 0), (0, 1)\}$; furthermore, the state of the system are $(1, 1, 0)$ and $(1, 0, 1)$. The transition rate diagrams of two-server and three-server queueing systems with buffer size 6 are given in Figure 3.3. It can be seen that the number of states at each level increases as the number of servers increases.

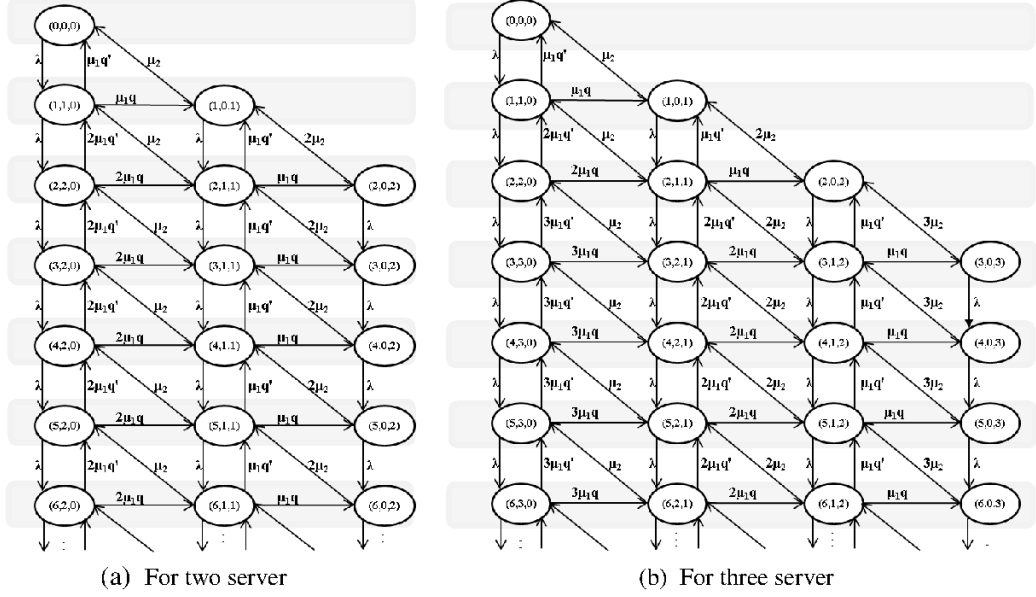


Figure 3.3 Transition rate diagram of the queueing systems (a) $M/C_2/2$ and (b) $M/C_2/3$

The Markov process that has the given state space is a QBD process. The infinitesimal generator matrix Q of this process has the block tri-diagonal form as presented in (3.9). The matrix has the main diagonal consisting of the matrices B_k and B ; the super-diagonal consisting of matrices C_k and C ; and the sub-diagonal consisting of matrices A_{k+1} and A for $0 \leq k \leq c - 1$.

$$Q = \begin{pmatrix} B_0 & C_0 & & & \\ A_1 & B_1 & C_1 & & \\ & \ddots & \ddots & \ddots & \\ & & A_{c-1} & B_{c-1} & C_{c-1} \\ & & & A_c & B & C \\ & & & & A & B & C \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \quad (3.9)$$

The matrices belonging to the main diagonal of the Q matrix are given below. It should be noted that all of these matrices are non-singular.

$$B_0 = [-\lambda]. \quad (3.10)$$

Let $m_j = -(\lambda + (k-j)\mu_1 + j\mu_2)$ for $0 \leq j \leq k$, and let $n_i = (k-i)\mu_1 q$ for $0 \leq i \leq k-1$. Then, matrix B_k can be written as follows:

$$B_k = \begin{pmatrix} m_0 & n_0 & & & \\ & m_1 & n_1 & & \\ & & \ddots & \ddots & \\ & & & m_{k-1} & n_{k-1} \\ & & & & m_k \end{pmatrix}_{(k+1) \times (k+1)}, \quad 1 \leq k \leq c-1. \quad (3.11)$$

Let $r_j = -(\lambda + (c-j)\mu_1 + j\mu_2)$ for $0 \leq j \leq c$, and let $s_i = (c-i)\mu_1 q$ for $0 \leq i \leq c-1$. Then, matrix B is given by

$$B = \begin{pmatrix} r_0 & s_0 & & & \\ & r_1 & s_1 & & \\ & & \ddots & \ddots & \\ & & & r_{c-1} & s_{c-1} \\ & & & & r_c \end{pmatrix}_{(c+1) \times (c+1)}. \quad (3.12)$$

The matrices that constitute the subdiagonal of the Q matrix are given in the succeeding

text. First define $a_j = (k - j)\mu_1 q'$ for $0 \leq j \leq k - 1$, and $b_i = i\mu_2$ for $1 \leq i \leq k$. Then,

$$A_k = \begin{pmatrix} a_0 & & & & \\ b_1 & a_1 & & & \\ & \ddots & \ddots & & \\ & & b_{k-1} & a_{k-1} & \\ & & & b_k & \end{pmatrix}_{(k+1) \times k}, 1 \leq k \leq c. \quad (3.13)$$

Define $d_j = (c - j)\mu_1 q'$ for $0 \leq j \leq c - 1$, and $e_i = i\mu_2$ for $1 \leq i \leq c$. Then,

$$A = \begin{pmatrix} d_0 & & & & \\ e_1 & d_1 & & & \\ & \ddots & \ddots & & \\ & & e_{c-1} & d_{c-1} & \\ & & & e_c & 0 \end{pmatrix}_{(c+1) \times (c+1)}. \quad (3.14)$$

The matrices belonging to the super-digonal of the Q matrix are given as follows:

$$C_k = \begin{pmatrix} \lambda & 0 & & & \\ & \lambda & 0 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 0 \end{pmatrix}_{(k+1) \times (k+2)}, 0 \leq k \leq c - 1. \quad (3.15)$$

and finally

$$C = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix}_{(c+1) \times (c+1)}. \quad (3.16)$$

3.2.3 Steady-state Probabilities

It is shown in Neuts (1981) that if the matrices along the main diagonal of the generator matrix Q are non-singular, then the underlying Markov chain is irreducible

and a steady-state distribution exists. Since $B_0, B_1, \dots, B_{c-1}, B$ are non-singular, the generator matrix Q is irreducible. The stationary probability vector of the QBD process that defines the related queueing system is

$$\pi = (\pi_0 \ \pi_1 \ \dots \ \pi_{c-1} \ \pi_c \ \pi_{c+1} \ \dots) \quad (3.17)$$

where π_k is the subvector of π with the dimension of $(k+1)$ for $0 \leq k \leq c-1$ and $(c+1)$ for $c \leq k$. The norm of the vector π_k represents the stationary probability of being k customers in the system.

Theorem 3.2.1. *For a QBD process with infinite state space, having an infinitesimal generator Q given in (3.9), the steady-state probabilities can be given as follows by using the matrix-geometric form:*

$$\pi_k = \pi_c R_k^* \text{ for } 0 \leq k \leq c-1,$$

where $R_k^* = R_{c-1}R_{c-2}\dots R_k$. The matrices R_k are computed by using $R_k = -A_{k+1}(R_{k-1}C_{k-1} + B_k)^{-1}$ and $R_0 = -A_1B_0^{-1}$.

$$\pi_k = \pi_c R^{k-c} \text{ for } c+1 \leq k,$$

where the matrix R is minimal non-negative solution of matrix equation $R^2A + RB + RC = 0$.

Proof. From $\pi Q = 0$, the following set of equations is obtained in matrix form:

$$\pi_0 B_0 + \pi_1 A_1 = 0 \quad (3.18)$$

$$\pi_{k-1} C_{k-1} + \pi_k B_k + \pi_{k+1} A_{k+1} = 0, \quad 1 \leq k \leq c-1 \quad (3.19)$$

$$\pi_{c-1} C_{c-1} + \pi_c B + \pi_{c+1} A = 0 \quad (3.20)$$

$$\pi_{k-1} C + \pi_k B + \pi_{k+1} A = 0, \quad c+1 \leq k \quad (3.21)$$

Note that there is a geometric relation between the steady state probabilities in (3.21).

Thus, the solution of c server queue models in QBD process is given $\pi_k = \pi_c R^{k-c}$, $k \geq c + 1$. We omit the proof of this part and refer to Neuts (1981) for this solution. The proof for $k \leq c$ continue as follows.

From Equation (3.18), we obtain

$$\begin{aligned}\pi_0 &= -\pi_1 A_1 B_0^{-1}, \\ \pi_0 &= \pi_1 R_0, \text{ where } R_0 = -A_1 B_0^{-1}.\end{aligned}\tag{3.22}$$

For $k = 1$ and $k = 2$ in Equation (3.19), we obtain the following:

$$\begin{aligned}\pi_1 &= -\pi_2 A_2 (R_0 C_0 + B_1)^{-1}, \\ \pi_1 &= \pi_2 R_1, \text{ where } R_1 = -A_2 (R_0 C_0 + B_1)^{-1}. \\ \pi_2 &= -\pi_3 A_3 (R_1 C_1 + B_2)^{-1}, \\ \pi_2 &= \pi_3 R_2, \text{ where } R_2 = -A_3 (R_1 C_1 + B_2)^{-1}.\end{aligned}$$

This can be generalized as

$$\pi_k = \pi_{k+1} R_k, \text{ where } R_k = -A_{k+1} (R_{k-1} C_{k-1} + B_k)^{-1} \text{ for } 1 \leq k \leq c-1. \tag{3.23}$$

However, the probability vector π_c still needs to be determined. Therefore, Equation (3.20) can be rewritten by using $\pi_{c-1} = \pi_c R_{c-1}^*$ and $\pi_{c+1} = \pi_c R$ from Theorem 3.2.1 as follows:

$$\pi_c [R_{c-1}^* C_{c-1} + B + RA] = 0 \tag{3.24}$$

The normalizing condition $\pi e = 1$ is written as

$$\pi_0 e_1 + \pi_1 e_2 + \dots + \pi_{c-1} e_c + \sum_{k=c}^{\infty} \pi_k e_{c+1} = 1 \tag{3.25}$$

where e_k are column vector of ones with dimension k .

The last equation can be rewritten in accordance with Theorem 3.2.1 as follows:

$$\pi_c R_0^* e_1 + \pi_c R_1^* e_2 + \dots + \pi_c R_{c-1}^* e_c + \pi_c \sum_{k=c}^{\infty} R^{k-c} e_{c+1} = 1$$

Then, we continue with some mathematical manipulations

$$\pi_c \left[\sum_{k=0}^{c-1} R_k^* e_{k+1} + (I - R)^{-1} e_{c+1} \right] = 1. \quad (3.26)$$

π_c can be computed from the solution to the system of the equations given in (3.24) and (3.26). After the computation of π_c , the other components of the steady-state probability vector can be calculated by using Theorem 3.2.1.

3.2.4 Performance Measures and Optimization

After calculating π_c and R , it is possible to obtain an analytic expression for the mean number in the queue. Let the random variable N_q represent the number of customers waiting in the queue in steady state. The average queue length for the $M/C_2/c$ queueing system can be computed from the expectation of N_q as follows:

$$\begin{aligned} EN_q &= \sum_{k=c+1}^{\infty} (k-c) \|\pi_k\|_1 = \sum_{k=c+1}^{\infty} (k-c) \|\pi_c R^{k-c}\|_1 \\ &= \left\| \pi_c \sum_{k=c+1}^{\infty} (k-c) R^{k-c} \right\|_1 = \left\| \pi_c R \sum_{k=c+1}^{\infty} (k-c) R^{k-c-1} \right\|_1 \\ EN_q &= \left\| \pi_c R (I - R)^{-2} \right\|_1 \end{aligned} \quad (3.27)$$

where $\|\cdot\|_1$ is one norm. Mean queue length EN_q is one of the performance measures of the queueing system. It is expected to increase with respect to system load ρ .

Once the average number of customers in the queue is calculated, we can obtain the other measures. That is, the average waiting time in the queue EW_q , the average waiting time in the system EW and the average number of customers in the system

EN are obtained by using the equations $EN_q = EW_q\lambda$, $EW = E(S) + EW_q$ and $EN = EW\lambda$, respectively.

Let Y be a random variable that represents the number of unoccupied beds. Then the average number of unoccupied beds is given by

$$EY = \|\pi_{c-1}\|_1 + 2\|\pi_{c-1}\|_1 + 3\|\pi_{c-3}\|_1 + \dots + (c-1)\|\pi_1\|_1 + c\|\pi_0\|_1, \quad (3.28)$$

The mean bed usage (MBU) is equal the total number of beds minus the average number of unoccupied beds and it is calculated by

$$MBU = (c - EY). \quad (3.29)$$

The bed occupancy rate (BOR) is given by

$$BOR = \frac{MBU}{c}. \quad (3.30)$$

Here, the optimal number of beds is determined by balancing the cost of unoccupied beds and the waiting cost of patients.

- we assume a *waiting cost*, C_w for each patient per day. Hence an average waiting cost per day is

$$C_w EN_q, \quad (3.31)$$

where EN_q is the mean number of patients waiting, which is computed with (3.27).

- we assume a *holding cost*, h for each unoccupied bed per day. Because EY is the mean number of unoccupied beds, the average daily cost due to unoccupied beds is

$$hEY. \quad (3.32)$$

By adding the cost functions provided in (3.31) and (3.32), the expected total cost

function yielding the average cost per day is obtained as presented in (3.33) when there are c beds. Note that as c increases, the average waiting cost decreases, while the average cost of unoccupied beds increases.

$$g(c) = C_w EN_q + hEY. \quad (3.33)$$

3.2.5 Numerical Study

The calculated values for these measures are summarized in Table 3.3.

Table 3.3 Performance measures of the $M/C_2/c$ queue model

c	$\lambda = 4$					$\lambda = 5$				
	ρ	MBU	BOR	EY	EN_q	ρ	MBU	BOR	EY	EN_q
4	0.75490	3.01961	0.75490	0.98039	1.62678	0.94363	3.77451	0.94363	0.22549	15.02513
5	0.60392	3.01961	0.60392	1.98039	0.37379	0.75490	3.77451	0.75490	1.22549	1.47758
6	0.50327	3.01961	0.50327	2.98039	0.10448	0.62908	3.77451	0.62908	2.22549	0.40130
7	0.43137	3.01961	0.43137	3.98039	0.02979	0.53922	3.77451	0.53922	3.22549	0.12556
8	0.37745	3.01961	0.37745	4.98039	0.00822	0.47181	3.77451	0.47181	4.22549	0.03990
9	0.33551	3.01961	0.33551	5.98039	0.00215	0.41939	3.77451	0.41939	5.22549	0.01234
10	0.30196	3.01961	0.30196	6.98039	0.00053	0.37745	3.77451	0.37745	6.22549	0.00364

Table 3.4 Determination of optimal number of beds by the $M/C_2/c$ queue model

c	$\lambda = 4$			$\lambda = 5$		
	<i>Waiting Cost</i>	<i>Holding Cost</i>	<i>Total Cost</i>	<i>Waiting Cost</i>	<i>Holding Cost</i>	<i>Total Cost</i>
4	81.33924	0.98039	82.31964	751.25626	0.22549	751.48175
5	18.68958	1.98039	20.66998	73.87909	1.22549	75.10458
6	5.22414	2.98039	8.20453	20.06507	2.22549	22.29056
7	1.48956	3.98039	5.46995	6.27808	3.22549	9.50357
8	0.41099	4.98039	5.39139	1.99498	4.22549	6.22048
9	0.10751	5.98039	6.08790	0.61678	5.22549	5.84227
10	0.02645	6.98039	7.00684	0.18218	6.22549	6.40767
11	0.00610	7.98039	7.98650	0.05100	7.22549	7.27649

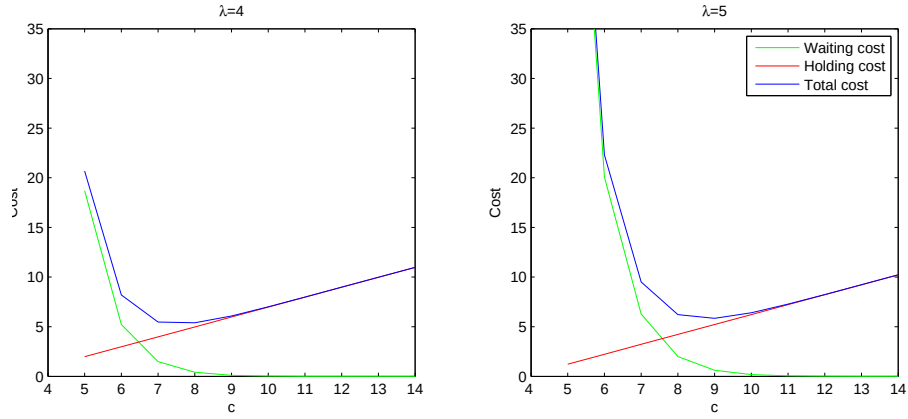


Figure 3.4 Costs versus the number of beds for $M/C_2/c$ model

3.3 $M/C_2/c/K$ Queueing Model

Recall that if all stuffed beds are busy, an arriving patient is lost in Section 3.1 and an arrival can wait until any stuffed bed is idle in Section 3.2. $M/C_2/c/K$ queueing system presented in this section allows that a **limited waiting space**. Patients can wait until any stuffed bed is idle and an arriving patient when this limit exceeded is lost. This case is more reasonable when it is considered that these patients are served in the unstaffed beds at another department.

We propose a methodology to solve the $M/C_2/c/K$ queueing system analytically. Similar to the solution given in Section 3.2 this queueing system is represented by QBD process but this time it has finite state space. So, the infinitesimal generator matrix which has block tri-diagonal structure is also finite and the matrix geometric method according to finite QBD process is used for the solution of the steady-state probabilities.

As a second objective, our methodology is applied to a *real case study* concerning the geriatric patients. Here, in the problem of deciding the optimal number of beds, two possible costs have considered; these are holding cost and waiting cost, as given in Section 3.2.

3.3.1 Model Description

A multi-server and finite buffer $M/C_2/c/K$ queuing system is considered. Recall the definition of the queue model from Sections 3.1 and 3.2. Arrivals occur according to the Poisson process with rate λ . There are available c servers in the system, and the probability of an arrival to be served by any of the idle servers is equal. Servers are identical, and the service time in each server follows the two-phase Cox distribution. That is, the service process for any arrival consists of two phases connected in series, each with the exponential service time distribution with parameters μ_1 and μ_2 . An arrival starts the service process from the first phase and proceeds to the second phase with probability q , or leaves the service process with probability $q' (= 1 - q)$. The mean service time for a single server is $E(S) = \frac{\mu_2 + \mu_1 q}{\mu_1 \mu_2}$.

The system capacity, which is the sum of the number of waiting customers in the queue and the number of customers being served (if any) is represented by K , where $K > c$. The stationary probability that there are K customers in the system is the probability that arriving customers are lost. Figure 3.5 provides a representation of the $M/C_2/c/K$ queuing system.

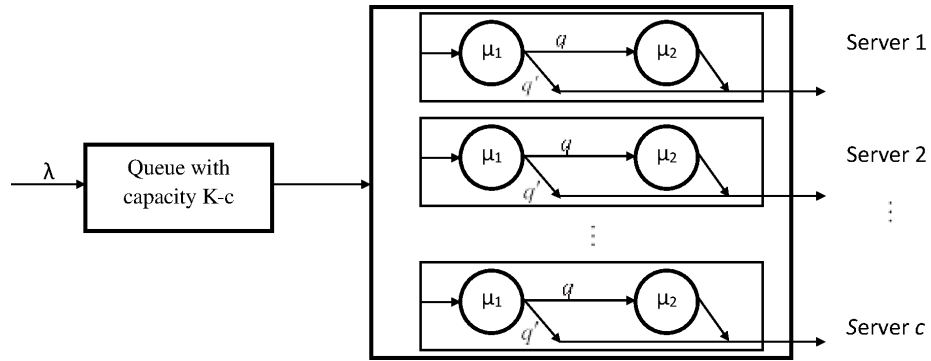


Figure 3.5 A schematic representation of $M/C_2/c/K$ queue model

3.3.2 QBD Process

The defined $M/C_2/c/K$ queuing system can be characterized by the bivariate Markov process $\{(k, i), 0 \leq k \leq K, i \in I\}$, where k denotes the "level" of the process

that corresponds to the number of customers in the system and i refers to the "states" at the corresponding level. State i is represented by two indices as the service unit is two-phase, and I is the set of possible values pairwise. Each of the indices contains the information regarding the number of customers in each phase of the service.

However, for the multi-server case, the customers in different servers may be in different phases. Hence each element in the state space is represented by two indices. The first index refers to the number of customers in the system and the second one contains the information regarding the number of customers in each phase. The general form of the transitions in the generator matrix for a multi-server queue can be examined in two main cases. For $0 \leq k \leq c - 1$, the number of states at each level would be different from each other and can be found by $[k, (k - j, j)]$, where $j = 0, 1, \dots, k$. Furthermore, for $c \leq k \leq K$, the number of states at each level is fixed and can be represented by $[k, (c - j, j)]$, where $j = 0, 1, \dots, c$. Then, the Markov process is defined on the finite state space as follows:

$$\Omega = \begin{cases} (k, (k - j, j)) & \text{if } 0 \leq k \leq c - 1, \text{ and } j = 0, 1, \dots, k \\ (k, (c - j, j)) & \text{if } c \leq k \leq K, \text{ and } j = 0, 1, \dots, c \end{cases}$$

The transition rate diagrams of two-server and three-server queueing systems with buffer size 6 are given in Figure 3.6.

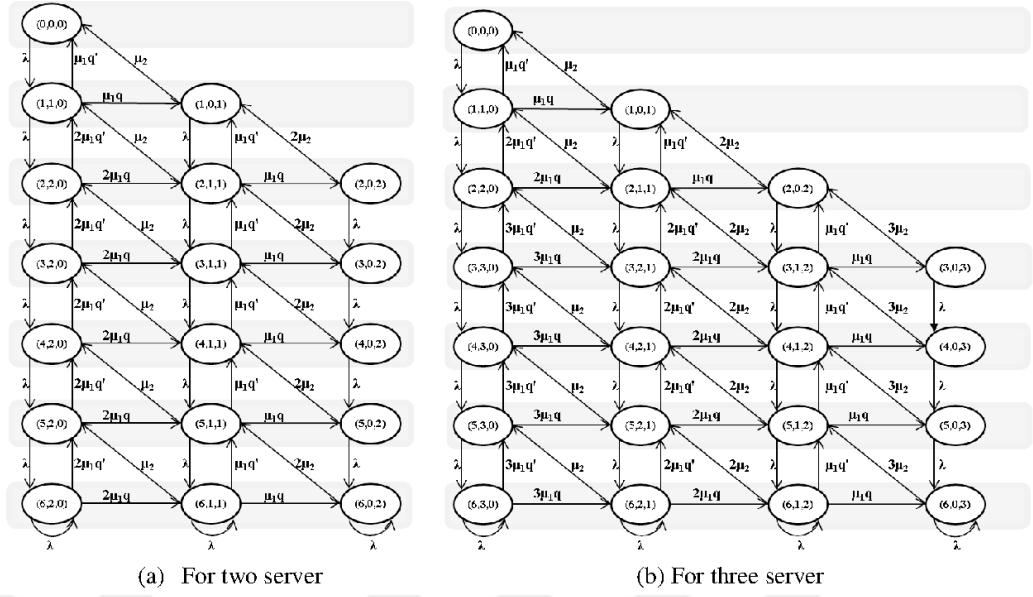


Figure 3.6 Transition rate diagram of the queueing systems (a) $M/C_2/2/6$ and (b) $M/C_2/3/6$

The Markov process that has the given state space is a QBD process. The infinitesimal generator matrix Q of this process has the block tri-diagonal form as presented in (3.34). The matrix has the main diagonal consisting of the matrices B_k, B_{00} , and B ; the super-diagonal consisting of matrices C_k and C ; and the sub-diagonal consisting of matrices A_{k+1} and A for $0 \leq k \leq c-1$.

$$Q = \begin{pmatrix} B_0 & C_0 & & & & \\ A_1 & B_1 & C_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & A_{c-1} & B_{c-1} & C_{c-1} & \\ & & & A_c & B & C \\ & & & & A & B & C \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & A & B & C \\ & & & & & & & A & B_{00} \end{pmatrix}. \quad (3.34)$$

The submatrices that the generator matrix Q includes are given by (3.10)-(3.16) in Section 3.2. The matrix B_{00} is the summation of matrices B and C .

3.3.3 Steady-state Probabilities

It is shown in Neuts (1981) that if the matrices along the main diagonal of generator matrix Q are non-singular, then the underlying Markov chain is irreducible and a steady-state distribution exists. Since $B_0, B_1, \dots, B_{c-1}, B$ and B_{00} are non-singular, generator matrix Q is irreducible. The stationary probability vector of the QBD process that defines the related queueing system is

$$\pi = (\pi_0 \ \pi_1 \ \dots \ \pi_{c-1} \ \pi_c \ \pi_{c+1} \ \dots \ \pi_{K-1} \ \pi_K) \quad (3.35)$$

where π_k is the subvector of π with the dimension of $(k+1)$ for $0 \leq k \leq c-1$ and $(c+1)$ for $c \leq k \leq K$. The norm of the vector π_k represents the stationary probability of being k customers in the system.

Theorem 3.3.1. *For a QBD process with finite state space, having an infinitesimal generator Q given in (3.34), the steady-state probabilities can be given as follows by using the matrix-geometric form:*

$$\pi_k = \pi_c R_k^* \text{ for } 0 \leq k \leq c-1,$$

where $R_k^* = R_{c-1}R_{c-2}\dots R_k$. The matrices R_k are computed by using $R_k = -A_{k+1}(R_{k-1}C_{k-1} + B_k)^{-1}$ and $R_0 = -A_1B_0^{-1}$.

$$\pi_k = \pi_c R_k^{**} \text{ for } c+1 \leq k \leq K,$$

where $R_k^{**} = R_{c+1}R_{c+2}\dots R_k$. The matrices R_k are computed by using $R_k = -C(B + R_{k+1}A)^{-1}$ and $R_K = -CB_{00}^{-1}$.

Proof. Note that the first result in Theorem 3.3.1 is as in Theorem 3.2.1. So, see to Section 3.2.3 for the proof of the probability vectors π_k , $0 \leq k \leq c-1$.

From $\pi Q = 0$, the following set of equations is obtained in matrix form:

$$\pi_0 B_0 + \pi_1 A_1 = 0 \quad (3.36)$$

$$\pi_{k-1} C_{k-1} + \pi_k B_k + \pi_{k+1} A_{k+1} = 0, \quad 1 \leq k \leq c-1 \quad (3.37)$$

$$\pi_{c-1} C_{c-1} + \pi_c B + \pi_{c+1} A = 0 \quad (3.38)$$

$$\pi_{k-1} C + \pi_k B + \pi_{k+1} A = 0, \quad c+1 \leq k \leq K-1 \quad (3.39)$$

$$\pi_{K-1} C + \pi_K B_{00} = 0 \quad (3.40)$$

From Equation (3.40), we obtain

$$\begin{aligned} \pi_K &= -\pi_{K-1} C B_{00}^{-1}, \\ \pi_K &= \pi_{K-1} R_K, \text{ where } R_K = -C B_{00}^{-1} \end{aligned} \quad (3.41)$$

For $k = K-1$ and from Equation (3.39), we obtain

$$\begin{aligned} \pi_{K-1} &= -\pi_{K-2} C (B + R_K A)^{-1}, \\ \pi_{K-1} &= -\pi_{K-2} R_{K-1} \text{ where } R_{K-1} = -C (B + R_K A)^{-1}. \end{aligned}$$

This can be generalized as

$$\pi_k = -\pi_{k-1} R_k \text{ where } R_k = -C (B + R_k A)^{-1} \text{ for } c+1 \leq k \leq K. \quad (3.42)$$

The components of the probability vector that have the matrix-geometric form in (3.41)-(3.42) can further be expressed by using π_c . This completes the proof of the Theorem 3.3.1.

However, the probability vector π_c still needs to be determined. Therefore, Equation (3.38) can be rewritten by using $\pi_{c-1} = \pi_c R_{c-1}^*$ and $\pi_{c+1} = \pi_c R_{c+1}^{**}$ from Theorem 3.3.1 as follows:

$$\pi_c [R_{c-1}^* C_{c-1} + B + R_{c+1}^{**} A] = 0 \quad (3.43)$$

The normalizing condition $\pi e = 1$ is written as

$$\pi_0 e_1 + \pi_1 e_2 + \dots + \pi_{c-1} e_c + \pi_c e_{c+1} + \pi_{c+1} e_{c+1} + \dots + \pi_K e_{c+1} = 1 \quad (3.44)$$

where e_k are column vector of ones with dimension k .

The last equation can be rewritten in accordance with Theorem 3.2.1 as follows:

$$\pi_c R_0^* e_1 + \pi_c R_1^* e_2 + \dots + \pi_c R_{c-1}^* e_c + \pi_c e_{c+1} + \pi_c R_{c+1}^{**} e_{c+1} + \dots + \pi_c R_K^{**} e_{c+1} = 1$$

Then, we continue with some mathematical manipulations

$$\begin{aligned} \pi_c [R_0^* e_1 + R_1^* e_2 + \dots + R_{c-1}^* e_c + e_{c+1} + R_{c+1}^{**} e_{c+1} + \dots + R_K^{**} e_{c+1}] &= 1 \\ \pi_c \left[\sum_{k=0}^{c-1} R_k^* e_{k+1} + e_{c+1} + \sum_{k=c+1}^K R_k^{**} e_{c+1} \right] &= 1. \end{aligned}$$

We can see $R_c^* = I$ from $\pi_c = \pi_c R_c^*$, where I is the identity matrix of dimension $c+1$.

Thus, the last equation becomes as follows:

$$\pi_c \left[\sum_{k=0}^c R_k^* e_{k+1} + \sum_{k=c+1}^K R_k^{**} e_{c+1} \right] = 1. \quad (3.45)$$

π_c can be computed from the solution to the system of the equations given in (3.43) and (3.45). After the computation of π_c , the other components of the invariant probability vector can be calculated by using Theorem 3.3.1.

3.3.4 Performance Measures and Optimization

After calculating π_c and R_k^{**} , it is possible to obtain an analytic expression for the mean number in the queue. Let the random variable N_q represent the number of customers waiting in the queue in steady state. The average queue length for the

$M/C_2/c/K$ queueing system can be computed from the expectation of N_q as follows:

$$EN_q = \sum_{k=c+1}^K (k-c) \|\pi_k\|_1 = \left\| \pi_c \sum_{k=c+1}^K (k-c) R_k^{**} \right\|_1 \quad (3.46)$$

where $\|\cdot\|_1$ is one norm. Mean queue length EN_q is one of the performance measures of the queueing system. It is expected to increase with respect to system load ρ .

The average number of unoccupied beds (EY), the mean bed usage (MBU) and bed occupancy rate (BOR) are given by (3.28), (3.29) and (3.30), respectively. In addition, we use π_K to calculate the loss probability.

Here, the optimal number of beds is determined by balancing the cost of unoccupied beds and the waiting cost of patients which is similar to Section 3.2. To calculate the average waiting cost per day and the average daily cost due to unoccupied beds, the quantities (3.31) and (3.32) are used, respectively. Finally, the average cost per day is obtained as presented in (3.33).

3.3.5 Case Study- The Planning of A Geriatric Care Center

Determination of the optimal number of beds at geriatric centers is a challenging problem. The refusal of an elderly patient who has serious health problems because of the lack of bed capacity is desirable for neither the patient nor the hospital. On the other hand, providing many more beds than demanded is not desirable owing to the limited amount of equipment and staff. As a result, specifying the optimal number of beds is an important, yet difficult problem, as the patients' length of stay at a geriatric center often has great variability. While some patient stay for a short period, some patients have to stay much longer because their treatment is complicated. Thus, the distribution of length of stay is heavier tailed than exponential distribution. On the other hand, as in the other previous studies Gorunescu et al. (2002a,b), Bernardo et al. (2003) that model geriatric patients' length of stay, it is assumed that patient arrivals follow the Poisson process. As a result, the movement of patients at a geriatric center can be modeled as

a multi-channel queueing system in which beds correspond to servers, and the arrival process is Poisson. In this case, because length of stay corresponds to service time, the service time distribution occurs as a heavy-tailed distribution.

3.3.5.1 Data

The data collected from St. George's Hospital in London for the period of 1965-1984 have been used in this study. The data gathered and organized in Taylor et al. (2000) may be downloaded from the Royal Statistical Society at <http://www.blackwellpublishers.co.uk/rss/>. The boxplot for the data about patients' length of stay is given in Figure 3.7. It can clearly be seen in Figure 3.7 that the data are highly skewed to the right. The mean and standard deviation of the data have been found to be 438.57 and 649.12, respectively; moreover, the goodness-of-fit of the data for exponential distribution has been tested by means of the Kolmogorov-Smirnov (KS) test. Now that the calculated test statistic (0.3042) is higher than the value $D_{1092;0.05} = 0.0411$, found for $\alpha = 0.05$ from the KS table, it is concluded that the data do not follow exponential distribution. Considering the result of the KS test and the coefficient of variation, it might be stated that the distribution of the data is heavier tailed than exponential distribution.

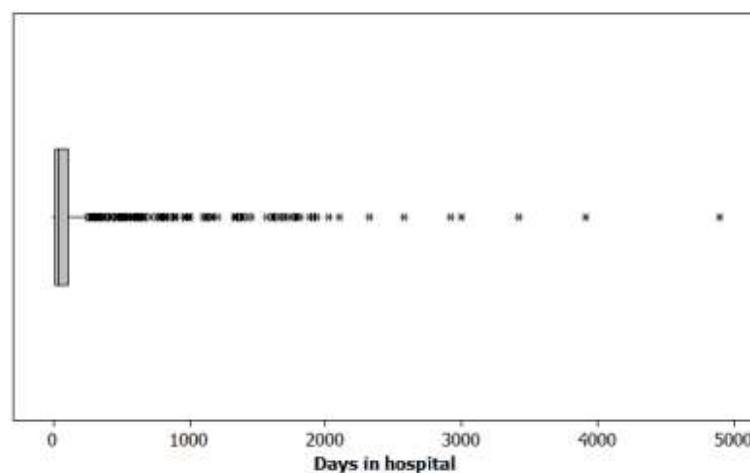


Figure 3.7 The boxplot of the data about the length of stay for all patients

After we have seen that the service time has a heavy-tailed distribution, that is, it can be characterized by a phase-type distribution, we now need to decide on the number of phases and the values of the parameters. Therefore, the maximum likelihood estimation, the most popular method to fit a certain phase-type distribution to data, is used. An expectation-maximization (EM) algorithm is developed in Asmussen et al. (1996) to compute the maximum likelihood parameter estimates for general phase-type distributions when fitted to empirical data. In our case, the data about the length of stay are modeled as the Cox distribution, and the parameters are estimated by using the EM algorithm. The data about the length of stay have been fitted to both two-phase and three-phase Cox distributions by using the EM algorithm, and the distribution parameters have been determined. The goodness-of-fit test has been carried out individually for two-phase and three-phase Cox distributions. When the hypothesis that the data follow the two-phase Cox distribution with the specified parameters is tested by means of the KS test, the test statistic is calculated as 0.035, whereas when the hypothesis that they follow the three-phase Cox distribution is tested, the test statistic is computed as 0.019. From the KS table, the critical value is found to be $D_{1092,0.05} = 0.041$ for the significance level $\alpha = 0.05$. As the calculated test statistic remains at the limit in this case, we hesitate to say that the data fit the two-phase Cox distribution. At the significance level of 0.05, it might be stated that the data about the length of stay can be better represented with the three-phase Cox distribution.

Because data are highly skewed to the right, it can be concluded statistically that data that remain outside the outer fence are considered outliers and can be ignored. The boxplot and the histogram of the data after the outliers have been removed are presented in Figure 3.8 and Figure 3.9, respectively.

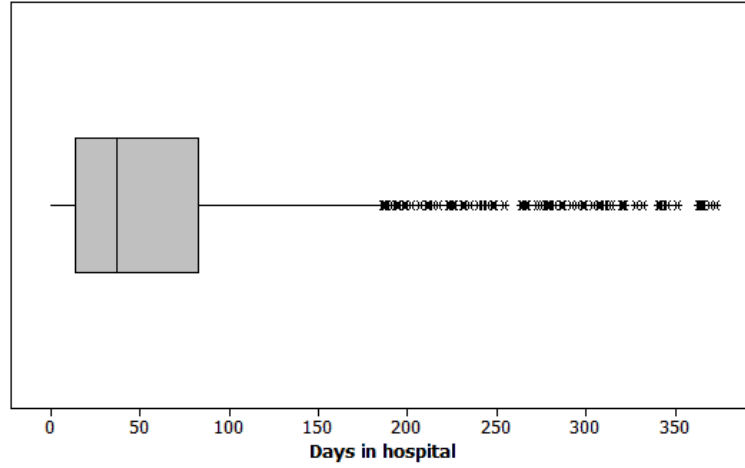


Figure 3.8 The boxplot of the data about the length of stay after removed outliers

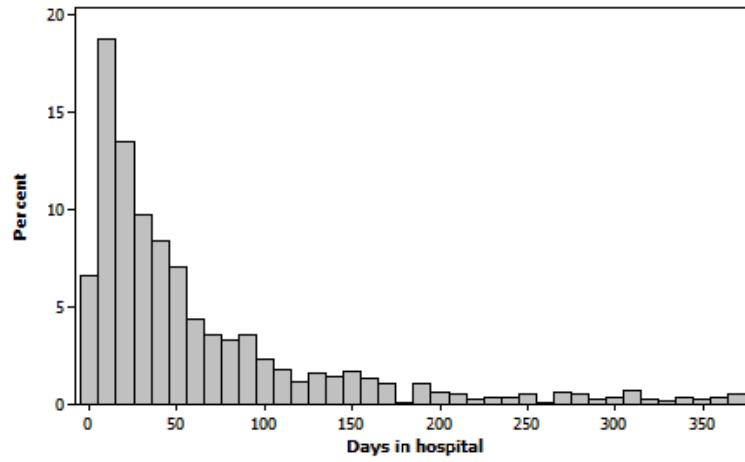


Figure 3.9 The histogram of the data about the length of stay after removed outliers

After the outliers have been removed, the mean, median, and standard deviation of the data are calculated as 64.45, 37, and 76.30, respectively. The hypothesis that the data follow exponential distribution is tested by the KS test, and the test statistic is calculated as 0.083, and because the critical value for the significance level $\alpha = 0.05$ is $D_{1000,0.05} = 0.043$, it can be concluded that the data do not follow exponential distribution. This time the parameters of the two-phase and three-phase Cox distributions have been estimated by running the EM algorithm for the truncated data set in which the outliers have been removed. By means of the KS test performed, it has been tested that the data come from a population, which has a two-phase Cox distribution, and the test statistic has been found to be 0.021. Because the calculated

test statistic is smaller than the critical value, we conclude that the data fit the two-phase Cox distribution reasonably well. There is no significant difference between the data we used in the study and the data that fit the two-phase Cox distribution with the specified parameters. The test statistic has been obtained as 0.025 for the test of hypothesis, which tests that data follow the three-phase Cox distribution. This test statistic is also smaller than the critical value. That is, although the three-phase Cox distribution fits the data, we require that the phase-type distribution, which fits the data, has the minimum number of phases. This is because the number of parameters increases with an increase in the number of phases. Furthermore, it can be seen from the cumulative distribution function given in Figure 3.10 that the two-phase Cox distribution with specified parameters represents the data well. Consequently, it is concluded that the data might be characterized by the two-phase Cox distribution by ignoring the outliers in the data set. The maximum likelihood estimates of the parameters are obtained as $\mu_1 = 0.03264$, $\mu_2 = 0.01033$, and $q = 0.35$.

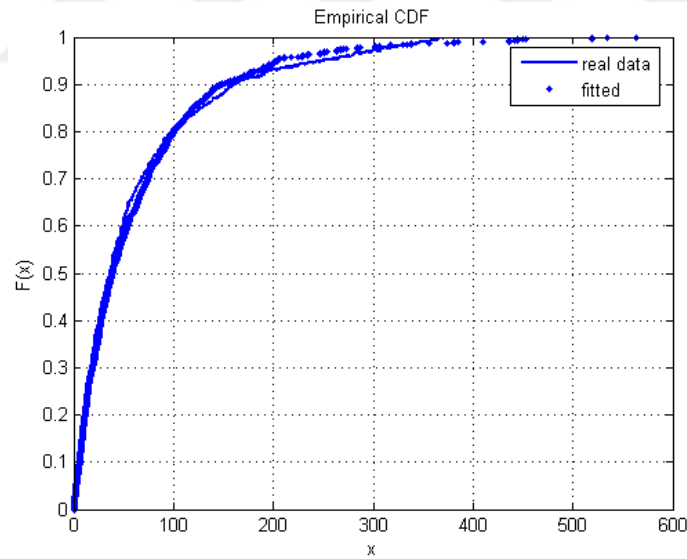


Figure 3.10 Fitted and original cumulative distribution functions of the data

3.3.5.2 Verification

The proposed methodology, explained in Section, is implemented by Matlab 7.12; the level-dependent rate matrices and steady-state probabilities are computed. Before presenting the numerical results, the verification of the model implemented by Matlab is discussed.

To verify the model implemented by Matlab, the $M/C_2/c/K$ queueing model is simulated using Arena 14.0. The model is simulated with the estimated parameters, and the mean service time $E(S)$ is computed as 64.44. The simulation model is run for 50 000 customers. The point and interval estimates for the mean queue length obtained as a result of the 20 repetitions are given in Table 3.5 for different values of ρ and c .

Table 3.5 Verification of the $M/C_2/c/K$ queueing system

λ	ρ	Waiting room 5			Waiting room 10		
		Analytic result	Simulation result		Analytic result	Simulation result	
		EN_q	EN_q	95% CI	EN_q	EN_q	95% CI
1	1	1.0047	1.0000	(0.976;1.030)	2.7882	2.782	(2.691;2.872)
	0.90	0.3310	0.3310	(0.311;0.349)	0.8195	0.823	(0.768;0.877)
	0.80	0.0439	0.0448	(0.040;0.049)	0.0873	0.087	(0.075;0.099)
2	1	0.7541	0.7411	(0.712;0.771)	2.1996	2.170	(2.082;2.263)
	0.90	0.1753	0.1741	(0.159;0.190)	0.4551	0.449	(0.406;0.491)
	0.80	0.0056	0.0057	(0.004;0.007)	0.0113	0.012	(0.008;0.015)
3	1	0.6702	0.6552	(0.626;0.684)	2.0143	1.965	(1.874;2.043)
	0.90	0.0919	0.0931	(0.081;0.105)	0.2412	0.241	(0.203;0.278)
	0.80	0.0006	0.0009	(0.0003;0.0015)	0.0013	0.002	(0.0004;0.004)

The comparison between the analytic solution values and the values obtained as a result of the simulation are given in Table 3.5 for buffer sizes 5 and 10. This comparison can be seen in Figure 3.11 more clearly. The left side Figure 3.11 is for buffer size 5, whereas the figure in right side is for buffer size 10. Both from the values given in Table 3.5 and Figure 3.11, it can be concluded that the simulation results confirm our analytic results.

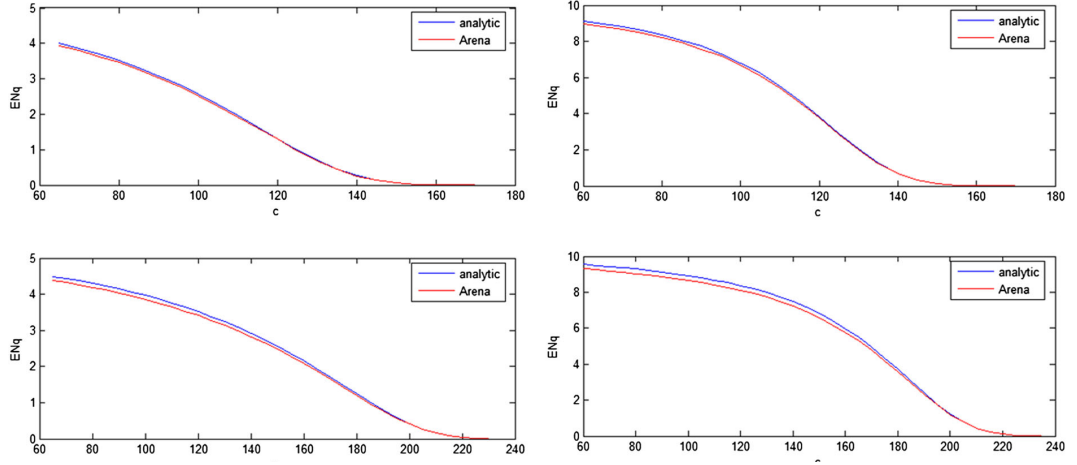


Figure 3.11 The comparison between the analytic solution values and the values obtained from simulation for buffer sizes 5 and 10

3.3.5.3 Analysis of Model for Geriatric Data

By using fitted parameters for the data about length of stay, μ_1, μ_2 , and q , the submatrices forming infinitesimal generator matrix Q are constructed by using the Matlab program and the stationary probabilities are computed.

Whereas the average length of stay $E(S)$ is computed as 64.44 for the given parameters, there is no information about the arrival rate of the patient in the data obtained from St. George's Hospital. Therefore, we assume different values for the arrival rates. The loss probability π_K , the mean bed usage MBU , the bed occupancy rate (BOR), the mean number of unoccupied beds EY and the mean queue length EN_q are computed and some of the calculated values for these measures are summarized in Table 3.6.

Table 3.6 Analytic solution to the queue model for the data about the length of stay

λ	Waiting room 5							Waiting room 10						
	c	ρ	π_K	MBU	BOR	EY	EN_q	c	ρ	π_K	MBU	BOR	EY	EN_q
1	39	1.65	0.397	39	0.997	0.11	3.59	48	1.34	0.258	48	0.997	0.13	7.21
	56	1.15	0.155	55	0.974	1.48	1.93	60	1.08	0.094	58	0.974	1.56	4.07
	62	1.04	0.088	59	0.949	3.12	1.22	64	1.01	0.055	61	0.953	3.00	2.79
	69	0.94	0.033	62	0.903	6.63	0.54	70	0.92	0.018	63	0.905	6.63	1.19
	75	0.86	0.010	64	0.851	11.14	0.19	76	0.85	0.004	64	0.845	11.73	0.34
	80	0.81	0.003	64	0.804	15.66	0.06	81	0.80	0.001	65	0.795	16.53	0.09
2	87	1.48	0.328	87	0.997	0.26	3.22	102	1.27	0.212	102	0.997	0.30	6.59
	117	1.10	0.115	114	0.976	2.78	1.51	125	1.03	0.060	121	0.970	3.68	2.92
	129	1.00	0.051	123	0.949	6.52	0.75	131	0.98	0.034	125	0.951	6.29	1.87
	141	0.92	0.014	128	0.903	13.72	0.23	142	0.91	0.007	128	0.902	13.91	0.53
	151	0.86	0.003	129	0.852	22.30	0.05	151	0.86	0.001	129	0.853	22.12	0.11
	161	0.80	0.000	129	0.801	31.99	0.01	161	0.80	0.000	129	0.801	31.97	0.01
3	137	1.41	0.294	137	0.997	0.39	3.02	157	1.23	0.191	157	0.997	0.44	6.26
	181	1.07	0.087	177	0.975	4.35	1.19	190	1.02	0.047	185	0.970	5.54	2.39
	197	0.98	0.034	187	0.949	10.03	0.52	199	0.97	0.023	189	0.950	9.85	1.33
	213	0.91	0.007	192	0.902	20.78	0.12	214	0.90	0.004	193	0.901	21.16	0.28
	227	0.85	0.001	193	0.852	33.60	0.02	227	0.85	0.000	194	0.852	33.51	0.03
	241	0.80	0.000	194	0.803	47.45	0.00	241	0.80	0.000	194	0.803	47.45	0.00

From Table 3.6, it can be stated that the c values with high BOR are the values with high loss probabilities, which is an expected result. This is because a higher BOR means a higher system load. For instance, BOR is 99.7%, ρ is calculated as 1.65 (because the system is capacitated, there is no such requirement as $\rho < 1$) for the system with waiting room 5 and $\lambda = 1$. When ρ drops below 1.0, the loss probability also goes down sharply and BOR corresponds to the system load. In this case, the mean number of unoccupied beds grows dramatically, whereas the loss probability approaches zero. In addition, it is seen in Table 3.6 that as the number of beds increases, the BOR goes down, the difference between the results for buffer sizes 5 and 10 dies down.

We compare our results with another method that has been employed for $M/G/c/K$ queueing system in Gorunescu et al. (2002b). We have applied our methodology to the data that Gorunescu et al. have used in their study, such as the mean length of stay and the mean arrival rate, and calculated both loss probability and the MBU . We compare

the results obtained by our methodology with the results presented in Gorunescu et al. (2002b). This comparison is given in Table 3.7, that is, (II) represents our results, whereas (I) represents the results in Gorunescu et al. (2002b). It is concluded that the difference between the values obtained by the two methods is not significant.

Table 3.7 Comparison of the methods for computing loss probability and the mean bed usage

<i>Bed</i>	Loss probability				Mean bed usage			
	+5		+10		+5		+10	
	I	II	I	II	I	II	I	II
130	0.1308	0.1321	0.1231	0.1256	128	127	129	128
135	0.1031	0.1045	0.0942	0.0974	132	131	133	132
140	0.0777	0.0793	0.0682	0.0718	136	135	137	136
145	0.0554	0.0569	0.0459	0.0496	139	138	140	139
150	0.0369	0.0382	0.0283	0.0316	142	141	143	142
155	0.0225	0.0236	0.0157	0.0182	144	143	145	144
160	0.0124	0.0132	0.0077	0.0094	145	145	146	145
165	0.0061	0.0065	0.0033	0.0042	146	146	146	146

Table 3.8 Comparison of the performance measures calculated for $M/C_2/c/K$ and $M/C_2/c$ models

c	ρ	$M/C_2/c$				$M/C_2/c/c + 5$				$M/C_2/c/c + 10$			
		MBU	BOR	EY	EN_q	MBU	BOR	EY	EN_q	MBU	BOR	EY	EN_q
4	0.94363	3.77451	0.94363	0.22549	15.02513	3.40299	0.85075	0.59701	1.58333	3.57571	0.89393	0.42429	3.41952
5	0.75490	3.77451	0.75490	1.22549	1.47758	3.65534	0.73107	1.34466	0.72447	3.74642	0.74928	1.25358	1.18513
6	0.62908	3.77451	0.62908	2.22549	0.40130	3.74108	0.62351	2.25892	0.28994	3.77110	0.62852	2.22890	0.38223
7	0.53922	3.77451	0.53922	3.22549	0.12556	3.76584	0.53798	3.23416	0.10634	3.77409	0.53916	3.22591	0.12398
8	0.47181	3.77451	0.47181	4.22549	0.03990	3.77239	0.47155	4.22761	0.03640	3.77446	0.47181	4.22554	0.03975
9	0.41939	3.77451	0.41939	5.22549	0.01234	3.77401	0.41933	5.22599	0.01169	3.77450	0.41939	5.22550	0.01232
10	0.37745	3.77451	0.37745	6.22549	0.00364	3.77440	0.37744	6.22560	0.00352	3.77451	0.37745	6.22549	0.00364
11	0.34314	3.77451	0.34314	7.22549	0.00102	3.77449	0.34314	7.22551	0.00100	3.77451	0.34314	7.22549	0.00102

Table 3.9 Comparison of the costs obtained for $M/C_2/c/K$ and $M/C_2/c$ models

c	$M/C_2/c$			$M/C_2/c/c + 5$			$M/C_2/c/c + 10$		
	<i>Waiting Cost</i>	<i>Holding Cost</i>	<i>Total Cost</i>	<i>Waiting Cost</i>	<i>Holding Cost</i>	<i>Total Cost</i>	<i>Waiting Cost</i>	<i>Holding Cost</i>	<i>Total Cost</i>
4	751.25626	0.22549	751.48175	79.16654	0.59701	79.76355	170.97610	0.42429	171.40039
5	73.87909	1.22549	75.10458	36.22338	1.34466	37.56804	59.25659	1.25358	60.51017
6	20.06507	2.22549	22.29056	14.49710	2.25892	16.75602	19.11147	2.22890	21.34037
7	6.27808	3.22549	9.50357	5.31697	3.23416	8.55113	6.19899	3.22591	9.42490
8	1.99498	4.22549	6.22048	1.82010	4.22761	6.04771	1.98745	4.22554	6.21299
9	0.61678	5.22549	5.84227	0.58446	5.22599	5.81044	0.61599	5.22550	5.84149
10	0.18218	6.22549	6.40767	0.17623	6.22560	6.40183	0.18209	6.22549	6.40758
11	0.05100	7.22549	7.27649	0.04992	7.22551	7.27544	0.05099	7.22549	7.27648

3.3.5.4 The Optimization

In this section, the waiting space is assumed as finite. That is, if all of the beds are occupied, the limited number of patients is not allowed to return. The main objective is to optimize the number of beds available at geriatric center so that a patient can obtain the best service immediately.

Optimization of the cost function over the number of servers is a classical problem. Two costs take part in the cost function Smith (2007). These are unit cost of providing service and expected waiting cost of a customer. In our case, the number of beds is determined by balancing the cost of unoccupied beds and the waiting cost of patients.

A patient's waiting, especially at a geriatric department, is a case, which causes a high cost. Thus, the unit waiting cost is assumed to be $C_w = 50$, which is 25% of the cost of the lost customer that is generally accepted in the literature Bernardo et al. (2003), Gorunescu et al. (1999). The unit cost of an unoccupied bed is assumed to be $h = 1$, based on that in Bernardo et al. (2003).

To determine the optimal policy, we can make the evaluation in two different ways. First, the decision is made by considering the loss probabilities. The optimal number of beds and the *BORs* when the loss probability is 1% for five different arrival rates and for two buffer sizes are summarized in Table 3.10.

As seen from Table 3.10, when the loss probability is fixed around 1% and the buffer size is 5, the mean queue length is computed as 0.18 patients. In this case, the system load varies between 86% and 93% depending on the arrival rate. When the buffer size is 10, the system load ranges from 90% to 94%, which is quite high, and the mean queue length is higher than the former case. Because the system allows more patients to wait, a decrease is seen in the optimal number of beds.

Table 3.10 Optimal c values for different arrival rates when the loss probability is 1%

λ	Waiting room 5			Waiting room 10		
	<i>optimal value of c</i>	<i>System load $\rho(\%)$</i>	<i>EN_q</i>	<i>optimal value of c</i>	<i>System load $\rho(\%)$</i>	<i>EN_q</i>
1.0	75	86	0.186	72	90	0.819
1.5	109	89	0.186	106	91	0.766
2.0	143	90	0.175	140	92	0.697
2.5	176	92	0.183	173	93	0.710
3.0	209	93	0.185	206	94	0.708

When considered in terms of cost optimization, the optimal number is the value that minimizes the function given in (3.33) and these values are given in Table 3.11 for different arrival rates.

Table 3.11 Optimal c values that minimize the expected cost for different arrival rates

λ	Waiting room 5					Waiting room 10				
	<i>optimal c</i>	ρ	<i>MBU</i>	<i>EN_q</i>	<i>Total cost</i>	<i>optimal c</i>	ρ	<i>MBU</i>	<i>EN_q</i>	<i>Total cost</i>
1.0	72	0.86	63	0.186	51.45	76	0.85	64	0.340	63.87
1.5	103	0.94	94	0.435	57.12	110	0.88	96	0.386	74.06
2.0	134	0.96	125	0.495	60.94	143	0.90	128	0.455	81.78
2.5	165	0.98	156	0.531	63.75	176	0.92	160	0.493	87.96
3.0	195	0.99	186	0.593	65.91	208	0.93	192	0.573	64.37

Regarding the care of elderly patients, the cost of a patient's waiting is quite high. In this case, the expected total cost function yields results to minimize the number of patients waiting, which improves the service quality. When the buffer size is increased from 5 to 10, the mean queue lengths do not change much; however, unlike Table 3.10, the optimal cost is found in the c values with a low system load. Moreover, when the buffer size increases, the expected cost increases as well because of the increment in the number of beds. Graphs of the total expected cost functions for arrival rates 1, 2, and 3 and for buffer size 5 and 10 are presented in Figure 3.12, where it is seen that as the number of beds increases, the cost of unoccupied beds also increases, whereas the waiting cost decreases. As a result, it is found that the total cost function is convex.

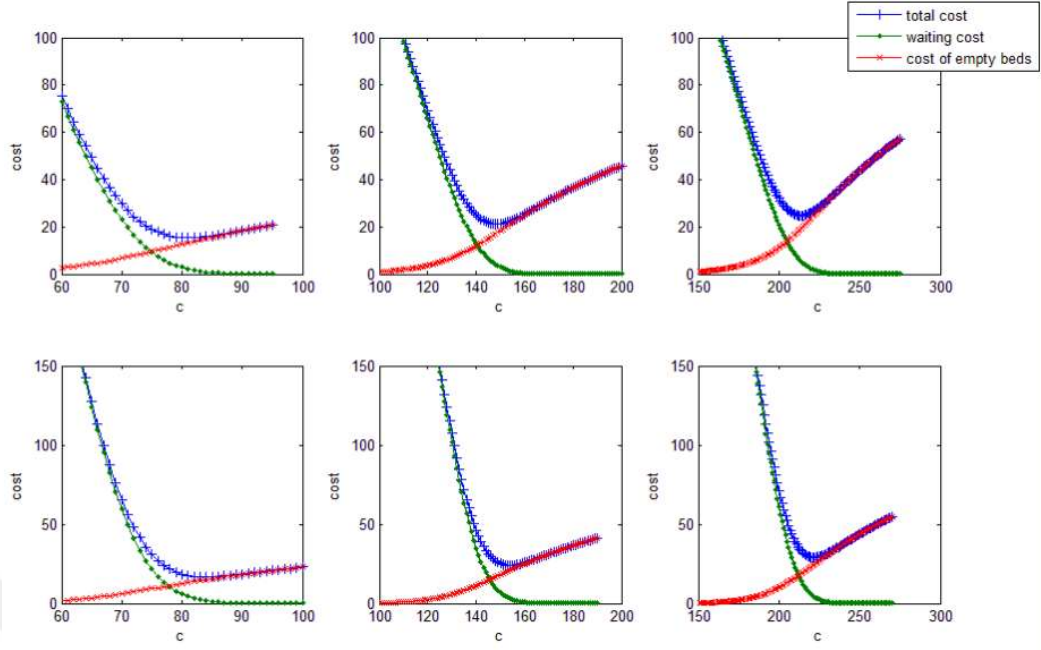


Figure 3.12 Cost functions for $\lambda = 1, 2, 3$ and for $K = 5, 10$

3.3.5.5 Compare the Studied Model with the models in Literature

In this section, a comparison between the methods is presented using simulated data sets. The following three different simulated data sets (DS) are used:

- DS 1: $\tau = 46.861$, $\sigma = 46.245$ and $CV^2 = 0.974$,
- DS 2: $\tau = 49.380$, $\sigma = 69.908$ and $CV^2 = 2.004$,
- DS 3: $\tau = 64.208$, $\sigma = 79.674$ and $CV^2 = 1.539$.

In the literature, Gorunescu et al. (2002b) is the study, which makes the closest assumptions to our study. Even though no analytic solution is provided in this study, the system is characterized by the capacitated and multi-server $M/Ph/c/N$ queueing model. Loss probabilities have been computed, and the MBU and bed occupancy values have been calculated using the loss probabilities. All calculations are provided for the states in which the buffer sizes are 5 and 10. However, there is no waiting room in Gorunescu et al. (2002a). In both studies, the loss probabilities are calculated

depending on Erlang's loss formula. For each of the three simulated data sets, these calculations are made considering the three different values of arrival rate λ . A summary of the results is presented in Table 3.12. The calculations with the method provided in Gorunescu et al. (2002a) when $\lambda = 1$ are in the columns indicated with (I), and the calculations with method provided in Gorunescu et al. (2002b) when $\lambda = 1$ in the columns indicated with (II). Column (III) contains a summary of the results obtained with the analytic method provided in our study for the same data sets and $\lambda = 1$. The steady-state probabilities and the loss probabilities for all states have been obtained for the buffer sizes are 5 and 10. The mean number of unoccupied beds (EY) has been computed, from which the MBU has been found by $MBU = c - EY$. The system load denotes the traffic density of the queueing system and is found by $\rho = \lambda\tau/c$. The BOR has been obtained with the ratio MBU/c , as computed in Gorunescu et al. (2002a) and Gorunescu et al. (2002b).

It is appropriate to compare the results in column (III) with the results in column (II) because they assume the same model. When MBU in column (III) and the values obtained with the method described in (II) are compared, it is seen that the difference in between is too small to be considered significant. A similar comment may also be made for the loss probabilities. The same calculations have been made for the state in which the arrival rate is $\lambda = 2$. The summary results are presented in Table 3.13. The results in these tables are also parallel with the previous results.

The cost function is considered in two ways in the two studies that discuss the optimal number of beds Gorunescu et al. (2002a), Bernardo et al. (2003). The total cost is expected as the sum of the cost of loss and the holding cost of the unoccupied beds in Gorunescu et al. (2002a)- the first of the first two studies both with no waiting room. In Bernardo et al. (2003), the other study, however, the cost of occupancy of beds is also considered in addition to the two costs in Gorunescu et al. (2002a) when computing the total cost. On the other hand, in Gorunescu et al. (2002b), which assumed the presence of finite waiting room, the total cost has been found as in Bernardo et al. (2003). For the simulated data set, the cost calculations performed by using the total cost functions provided in these studies are presented for $\lambda = 1$

in Table 3.14. The last column (III) represents the results of the our model. In these calculations, it is assumed that the unit penalty cost for lost customers is 200, that the unit cost of an unoccupied bed is $h = 1$, and that each unit cost due to occupied bed is 4. These values are based on those in Bernardo et al. (2003). The numbers of beds with the minimum cost are indicated in bold. Those shown with (IV) are the results obtained by using the cost function in Bernardo et al. (2003).

When we examine the tables collectively, we see that our model overlaps (II) in terms of the queueing system; however, the cost function we adopt overlaps (I). Hence, there appears some parallelism between (I) and our study in terms of the optimal number of beds. Nevertheless, now that the cost of waiting is considered and not the cost of loss, a component of the total cost function in our study, and that the queueing model differs from that in (I), some differences are also seen numerically. Some parallelism for the optimal number of beds is seen when the buffer sizes are 5 and 10, and the same parallelism is also seen in the results obtained in (II). Similar results have been obtained for $\lambda = 2$ as well as and are summarized in Table 3.15.

Table 3.12 Comparison of the performance measures for the simulated data when the arrival rate is 1

		(I)			(II)						(III)					
<i>c</i>	<i>Data set</i>	No waiting room			Waiting room 5			Waiting room 10			Waiting room 5			Waiting room 10		
		P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR
40	1	0.209	37	0.926	0.170	39	0.973	0.156	40	0.989	0.158	39	0.970	0.144	39	0.987
	2	0.399	37	0.936	0.205	39	0.981	0.195	40	0.994	0.214	39	0.979	0.205	40	0.991
	3	0.242	39	0.965	0.379	40	0.997	0.377	40	1.000	0.382	40	0.997	0.380	40	1.000
50	1	0.072	43	0.870	0.040	45	0.900	0.025	46	0.914	0.034	45	0.891	0.020	45	0.904
	2	0.260	45	0.891	0.063	46	0.926	0.045	47	0.943	0.070	46	0.927	0.056	47	0.941
	3	0.098	48	0.951	0.231	49	0.987	0.224	50	0.997	0.235	49	0.987	0.228	50	0.996
60	1	0.010	46	0.773	0.003	47	0.779	0.001	47	0.780	0.002	46	0.767	0.001	46	0.768
	2	0.136	48	0.808	0.007	49	0.817	0.002	49	0.821	0.009	49	0.824	0.004	50	0.827
	3	0.019	55	0.924	0.104	58	0.959	0.089	58	0.975	0.108	58	0.959	0.094	58	0.974
70	1	0.000	47	0.669	0.000	47	0.669	0.000	47	0.669	0.000	46	0.659	0.000	46	0.659
	2	0.047	49	0.705	0.000	49	0.705	0.000	49	0.705	0.000	50	0.712	0.000	50	0.712
	3	0.001	61	0.874	0.026	63	0.894	0.015	63	0.903	0.028	63	0.896	0.018	63	0.905
80	1	0.000	47	0.586	0.000	47	0.586	0.000	47	0.586	0.000	46	0.576	0.000	46	0.576
	2	0.008	49	0.617	0.000	49	0.617	0.000	49	0.617	0.000	50	0.623	0.000	50	0.623
	3	0.000	64	0.797	0.002	64	0.801	0.001	64	0.802	0.003	64	0.804	0.001	64	0.806

Table 3.13 Comparison of the performance measures for the simulated data when the arrival rate is 2

		(I)			(II)						(III)					
<i>c</i>	<i>Data set</i>	No waiting room			Waiting room 5			Waiting room 10			Waiting room 5			Waiting room 10		
		P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR	P_{loss}	MBU	BOR
80	1	0.185	76	0.954	0.162	79	0.982	0.153	79	0.992	0.150	78	0.980	0.140	79	0.991
	2	0.221	77	0.962	0.200	79	0.988	0.193	80	0.996	0.208	79	0.987	0.202	80	0.995
	3	0.389	78	0.981	0.378	80	0.998	0.377	80	1.000	0.381	80	0.998	0.380	80	1.000
90	1	0.104	84	0.933	0.080	86	0.958	0.067	87	0.971	0.070	86	0.952	0.057	87	0.966
	2	0.138	85	0.946	0.114	87	0.972	0.103	89	0.984	0.122	88	0.972	0.113	88	0.983
	3	0.315	88	0.977	0.302	90	0.996	0.300	90	0.999	0.305	90	0.996	0.303	90	0.999
100	1	0.043	90	0.897	0.026	91	0.913	0.017	92	0.921	0.021	90	0.903	0.013	91	0.910
	2	0.069	92	0.920	0.048	94	0.940	0.037	95	0.951	0.055	94	0.942	0.045	95	0.952
	3	0.243	97	0.972	0.227	99	0.992	0.223	100	0.998	0.231	99	0.992	0.227	100	0.998
110	1	0.010	93	0.843	0.005	93	0.848	0.002	94	0.850	0.003	92	0.835	0.001	92	0.837
	2	0.023	96	0.877	0.013	98	0.887	0.007	98	0.892	0.015	98	0.892	0.010	99	0.897
	3	0.175	106	0.963	0.156	108	0.985	0.149	109	0.993	0.160	108	0.985	0.153	109	0.993
120	1	0.001	94	0.780	0.000	94	0.781	0.000	94	0.781	0.000	92	0.768	0.000	92	0.768
	2	0.004	98	0.819	0.002	99	0.822	0.001	99	0.822	0.002	99	0.829	0.001	100	0.830
	3	0.112	114	0.950	0.093	116	0.971	0.083	118	0.981	0.097	117	0.971	0.087	118	0.981
130	1	0.000	94	0.721	0.000	94	0.721	0.000	94	0.721	0.000	92	0.709	0.000	92	0.709
	2	0.000	99	0.759	0.000	99	0.760	0.000	99	0.760	0.000	100	0.767	0.000	100	0.767
	3	0.079	118	0.910	0.056	121	0.932	0.041	123	0.947	0.047	123	0.946	0.037	124	0.956

Table 3.14 Comparison of the costs for the simulated data when $\lambda = 1$

	(I)			(IV)			(II)						(III)					
	No waiting room			No waiting room			Waiting room 5			Waiting room 10			Waiting room 5			Waiting room 10		
c	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3
40	44.85	50.98	81.15	90.76	96.12	123.94	77.16	83.36	116.17	72.57	79.77	115.48	105.49	124.08	176.74	276.44	314.98	417.27
45	30.93	36.54	67.42	84.66	88.98	116.11	71.05	75.27	106.60	64.81	69.59	105.13	72.11	92.56	156.69	174.71	233.24	385.85
50	20.84	25.06	54.37	83.85	85.99	109.28	73.04	73.64	98.09	67.87	67.58	95.32	48.87	64.72	133.54	89.53	145.79	341.77
55	15.73	17.77	42.33	89.89	88.82	103.96	83.70	80.26	91.52	81.27	76.36	87.05	41.14	47.24	108.29	51.18	80.18	282.47
60	15.55	15.31	31.81	102.74	98.41	100.90	100.36	94.21	88.22	99.69	92.73	82.32	45.31	43.14	83.31	46.92	52.00	210.42
65	18.68	16.94	23.51	120.16	113.71	101.13	119.56	112.32	89.82	119.45	111.98	83.93	54.03	48.68	62.46	54.21	50.26	137.89
70	23.22	20.88	18.21	139.53	132.23	105.83	139.44	131.92	97.51	139.42	131.87	93.35	63.01	57.77	49.94	63.02	57.96	83.85
75	28.15	25.66	16.40	159.43	151.91	115.70	159.42	151.87	110.96	159.42	151.86	108.97	71.06	66.91	47.65	71.06	66.93	58.71
80	33.14	30.62	17.79	179.42	171.87	130.34	179.42	171.86	128.34	179.42	171.86	127.69	78.14	75.16	53.40	78.14	75.16	56.02
85	38.14	35.62	21.33	199.42	191.86	148.18	199.42	191.86	147.57	199.42	191.86	147.42	84.39	82.46	62.97	84.39	82.46	63.43

Table 3.15 Comparison of the costs for the simulated data when $\lambda = 2$

	(I)			(IV)			(II)						(III)					
	No waiting room			No waiting room			Waiting room 5			Waiting room 10			Waiting room 5			Waiting room 10		
c	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3	DS 1	DS 2	DS 3
75	95.24	108.79	171.73	176.06	188.77	249.32	161.19	175.34	241.92	156.90	172.25	241.43	120.70	139.02	185.66	317.52	351.78	430.09
80	77.84	91.48	157.00	165.15	177.62	240.01	149.08	162.80	231.66	143.06	158.28	230.90	102.16	122.85	176.79	270.64	316.87	418.13
85	61.76	75.08	142.43	156.07	167.74	230.93	139.44	151.85	221.56	131.78	145.68	220.41	84.18	106.06	167.02	216.40	275.28	403.88
90	47.53	59.94	128.06	149.55	159.62	222.13	133.40	143.28	211.69	124.82	135.59	210.01	68.63	89.52	156.35	160.25	228.05	386.83
95	35.82	46.54	113.93	146.50	153.94	213.71	132.17	138.12	202.16	124.06	129.60	199.77	57.77	74.55	144.78	111.51	178.34	366.42
100	27.33	35.50	100.13	147.88	151.54	205.77	136.69	137.48	193.13	130.49	129.43	189.82	53.42	62.87	132.40	79.14	131.82	342.04
105	22.55	27.45	86.75	154.39	153.33	198.46	146.98	142.23	184.82	143.26	135.99	180.38	55.92	56.23	119.37	65.76	95.44	313.23
110	21.43	22.87	73.94	165.94	159.95	191.99	161.93	152.44	177.53	160.21	148.59	171.84	63.72	55.73	106.01	66.72	74.23	279.77
115	23.24	21.73	61.89	181.54	171.38	186.65	179.80	167.18	171.69	179.18	165.31	164.76	74.35	61.01	92.81	75.09	68.10	242.06
120	26.89	23.42	50.85	199.69	186.76	182.82	199.09	184.86	167.87	198.91	184.15	159.98	85.75	70.34	80.50	85.90	72.54	201.52
125	31.44	26.96	41.16	219.22	205.32	180.99	218.89	204.03	166.76	218.85	203.81	158.53	96.85	81.59	70.04	96.87	82.14	160.94
130	36.31	31.44	43.32	238.89	223.99	181.76	238.85	223.79	178.97	238.84	223.74	167.03	107.26	93.18	62.61	107.26	93.29	124.43

CHAPTER FOUR

QUEUEING MODELS WITH COX SERVICE AND MULTIPLE WORKING VACATION

After Cox distribution with two phases is considered for the modelling of the service times in previous chapter, now k -phase Cox distribution is studied in another real life application.

Servers are always available in classical queueing models. However, servers may become unavailable for a period of time for a variety of reasons. The period of server absence is called a server vacation. There are many different vacation policies in the literature. In this chapter, we consider a working vacation policy which is extensively applied to real life problems. In many practical systems encountered, the server does not remain idle during the vacation period due to the operational costs. Instead of being idle, the system operates at a slower rate to save energy and to reduce waste time.

In this chapter, the $M/C_k/1$ queueing model with multiple working vacation is examined. The queueing model whose service time is assumed to be k -phase Cox distribution is described in Section 4.1. After the stationary probability distribution is obtained by using the *QBD* process and the matrix geometric method in Section 4.2, Section 4.3 gives the average number of customers in the system in the *stochastic decomposition structure*. To see the effect of parameters on performance measures, a numerical study is conducted and the results are presented in Section 4.4.

4.1 Model Description

We consider an $M/C_k/1$ queueing system with multiple working vacation policy. Customers arrive at the system according to the Poisson process with arrival rate λ and are moved from the waiting line to the service station with service discipline *FCFS*. There is a single server at the service station. The customers are served according to k -phase Cox distribution with different parameters on the basis of the status of the server.

The server can be either in the working vacation period or in the regular busy period. Then, the parameters belonging to service distribution can be given as follows.

The server is in the regular busy period. Vector $\mathbf{t}$ is interpreted as the initial probability vector among the k transient phases, while matrix T and vector $\mathbf{T}^0$ are interpreted as the transition rate matrix and the absorption rate vector, respectively. That is, starting from the first phase, the process switches to phase i ($1 < i \leq k$) with probability α_i^b and jumps to absorbing state $k + 1$ with probability $\bar{\alpha}_i^b$. Moreover, the sojourn time in phase i is exponentially distributed with parameter μ_i^b . The service rate in the busy period is $\mu_b = -1/\mathbf{t}T^{-1}\mathbf{e}_k$, where $\mathbf{e}_k$ is a column vector with all elements being 1.

$$Q = \begin{pmatrix} T & \mathbf{T}^0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}, \quad (4.1)$$

where

$$T = \begin{pmatrix} -\mu_1^b & \alpha_1^b\mu_1^b & & \\ & -\mu_2^b & \ddots & \\ & & \ddots & \alpha_{k-1}^b\mu_{k-1}^b \\ & & & -\mu_k^b \end{pmatrix}, \quad \mathbf{T}^0 = \begin{pmatrix} \bar{\alpha}_1^b\mu_1^b \\ \vdots \\ \bar{\alpha}_{k-1}^b\mu_{k-1}^b \\ \mu_k^b \end{pmatrix}.$$

The server is in the working vacation period. Vector $\mathbf{s}$ is interpreted as the initial probability vector among the k transient phases, while matrix S and vector $\mathbf{S}^0$ are interpreted as the transition rate matrix and the absorption rate vector, respectively. That is, starting from the first phase, the process switches to phase i ($1 < i \leq k$) with probability α_i^v and jumps to absorbing state $k + 1$ with probability $\bar{\alpha}_i^v$. Moreover, the sojourn time in phase i is exponentially distributed with parameter μ_i^v . The service rate in the busy period is $\mu_v = -1/\mathbf{s}S^{-1}\mathbf{e}_k$, where $\mathbf{e}_k$ is a column vector with all elements being 1.

$$Q = \begin{pmatrix} S & \mathbf{S}^0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{s} = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}, \quad (4.2)$$

where

$$S = \begin{pmatrix} -\mu_1^v & \alpha_1^v \mu_1^v & & \\ & -\mu_2^v & \ddots & \\ & & \ddots & \alpha_{k-1}^v \mu_{k-1}^v \\ & & & -\mu_k^v \end{pmatrix}, \quad S^0 = \begin{pmatrix} \bar{\alpha}_1^v \mu_1^v \\ \vdots \\ \bar{\alpha}_{k-1}^v \mu_{k-1}^v \\ \mu_k^v \end{pmatrix}.$$

The vacation policy in the model is determined as follows. Once the system becomes empty, the server goes on the vacation in order to make another job. If there is at least one customer waiting in the queue when the server returns from the vacation, it starts service till the system becomes empty. On the other hand, if there is no customer in the queue on returning from the vacation, it takes another vacation and so on. Finally, the vacation time of the service station follows the exponential distribution with parameter η .

For an illustration, let's take a service given from a ticket counter of a specific firm at an airport.

- Passengers arrive to the counter according to a Poisson process with rate λ and wait in a single line.
- Service process includes two phases such as the process related with the tickets and the luggage. The service time for both the first phase (related with the tickets) and the second phase (related with the luggage) are exponential with rate μ_1^b and μ_2^b , respectively. Passengers have luggage with probability α^b , so they leave the service after they get their tickets with probability $1 - \alpha^b$, or proceed the second phase and check their baggage with probability α^b .
- When there is no customer in waiting line, the server goes on a working vacation. The durations of the working vacations are i.i.d. exponential random variables with rate η .
- In working vacation period, the employee performing another activity whose

duration can be represented by Cox distribution. If there is no customer in the queue on returning from the vacation, it takes another vacation and so on.

4.2 The Steady-state Analysis

The system can be modeled by a three-dimensional continuous-time Markov chain. Let the state of the system be denoted by triplet (N_t, I_t, J_t) , where N_t represents the total number of customers in the system at time t . I_t denotes the service phase of the customer being served at time t , and finally J_t represents the service status with

$$J_t = \begin{cases} 0, & \text{if the server is in the working vacation period at time } t \\ 1, & \text{if the server is in the regular busy period at time } t. \end{cases}$$

By using the lexicographical ordering for the states, the state space of the Markov chain $\{(N_t, I_t, J_t), t \geq 0\}$ is

$$\Omega = \{0\} \cup \{(n, i, j); n \geq 1, i = 1, \dots, k, j = 0, 1\}. \quad (4.3)$$

Let us define the steady-state probability for the defined Markov chain as

$$\pi_0 = P\{N_t = 0, I_t = 0, J_t = 0\},$$

$$\pi_{nij} = P\{N_t = n, I_t = i, J_t = j\}, \text{ for } n \geq 1, 1 \leq i \leq k, j = 0, 1.$$

Then, let the steady-state probability vector be denoted by

$$\boldsymbol{\pi} = \left[\pi_0 \quad \boldsymbol{\pi}_1 \quad \boldsymbol{\pi}_2 \quad \cdots \right], \quad (4.4)$$

where π_0 is scalar and $\boldsymbol{\pi}_n = [\pi_{n10} \ \pi_{n20} \ \cdots \ \pi_{nk0} \ \pi_{n11} \ \pi_{n21} \ \cdots \ \pi_{nk1}]$ is $1 \times 2k$ -dimensional row vector for $n \geq 1$. For example, π_{n10} is the probability that there are n customers in the system, that the customer being served is in phase 1, and that the server works at the vacation rate, whereas π_{n11} is the probability that there are n customers in the system, that the customer being served is in phase 1, and that the

server works at the normal rate.

The infinitesimal generator matrix of the system has the block-tridiagonal matrix structure in (4.5).

$$Q = \begin{bmatrix} B^0 & C^0 & & & \\ A^0 & B & C & & \\ & A & B & C & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \quad (4.5)$$

where

$$A^0 = \begin{bmatrix} S^0 \\ T^0 \end{bmatrix}_{2k \times 1}, \quad B^0 = [-\lambda]_{1 \times 1}, \quad C^0 = [\lambda s \quad 0]_{1 \times 2k},$$

$$A = \begin{bmatrix} S^0 s & 0 \\ 0 & T^0 t \end{bmatrix}, \quad B = \begin{bmatrix} S - \lambda I - \eta I & \eta e_k t \\ 0 & T - \lambda I \end{bmatrix}, \quad C = \begin{bmatrix} \lambda I & 0 \\ 0 & \lambda I \end{bmatrix}.$$

The submatrices of matrix Q also have the block matrix structure. Note that each of matrices A, B and C are square matrices of dimension $2k$. In this case, the steady-state probability vector $\pi = [\pi_0 \ \pi_1 \ \cdots]$ can be partitioned as $\pi_n = [\pi_{n0} \ \pi_{n1}]$, where π_n is a $1 \times 2k$ -dimensional row vector for $n \geq 1$. Each component of the vector is now a k -dimensional row vector. For instance, π_{n0} is a probability vector in which there are n customers in the system and the server works at the vacation rate and can be denoted as $\pi_{n0} = [\pi_{n10} \ \pi_{n20} \ \cdots \ \pi_{nk0}]$, for $n \geq 1$.

The minimal non-negative solution of the matrix quadratic equation in (4.6) must be obtained to analyze the QBD process, namely, to calculate vector π in (4.4).

$$R^2 A + R B + C = 0 \quad (4.6)$$

The solution is called *rate matrix* R and it has an upper-triangular block matrix structure since the matrices A, B , and C have an upper-triangular block matrix structure.

Lemma 4.2.1. *The matrix quadratic equation (4.6) has the minimal non-negative solution*

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}_{2k \times 2k}, \quad (4.7)$$

where $R_{22} = \lambda[\lambda I - \lambda e_k \mathbf{t} - T]^{-1}$; $R_{12} = R_{11}(I - R_{11})^{-1} \eta e_k \mathbf{t} \lambda^{-1}$; and R_{11} is obtained by the following algorithm:

$$R_{11}(0) = 0$$

$$R_{11}(n+1) = R_{11}^2(n) \mathbf{S}^0 \mathbf{s} (-S + \lambda I + \eta I)^{-1} + \lambda I (-S + \lambda I + \eta I)^{-1}, \quad n \geq 0.$$

Lemma 4.2.2. *k -dimensional square matrices R_{11} , R_{12} , and R_{22} satisfy the following equations*

$$R_{11} \mathbf{S}^0 + R_{12} \mathbf{T}^0 = R_{22} \mathbf{T}^0 = \lambda e_k \quad (4.8)$$

$$(I - R_{11})^{-1} R_{11} \eta = \lambda I + R_{11} S. \quad (4.9)$$

Proof of Lemma 4.2.2 We know that the balance equations $\pi Q = 0$ procure $\pi_n = \pi_1 R^{n-1}$ and $[\pi_0 \ \pi_1]$ satisfies (4.10) by Neuts (1981).

$$[\pi_0 \ \pi_1] B[R] = 0 \quad (4.10)$$

where

$$B[R] = \begin{bmatrix} B^0 & C^0 \\ A^0 & B + RA \end{bmatrix}$$

$$B[R] = \begin{bmatrix} -\lambda & \lambda \mathbf{s} & 0 \\ \mathbf{S}^0 & (S - \lambda I - \eta I) + R_{11} \mathbf{S}^0 \mathbf{s} & \eta e_k \mathbf{t} + R_{12} \mathbf{T}^0 \mathbf{t} \\ \mathbf{T}^0 & 0 & (T - \lambda I) + R_{22} \mathbf{T}^0 \mathbf{t} \end{bmatrix}. \quad (4.11)$$

Since $B[R]$ is a generator matrix, the following equations can be written from (4.11):

$$\mathbf{T}^0 + (T - \lambda I + R_{22} \mathbf{T}^0 \mathbf{t}) e_k = 0$$

$$\mathbf{S}^0 + (S - \lambda I - \eta I + R_{11} \mathbf{S}^0 \mathbf{s}) e_k + (\eta e_k \mathbf{t} + R_{12} \mathbf{T}^0 \mathbf{t}) e_k = 0$$

Equation (4.8) in Lemma 4.2.2 is obtained by using $Te_k = -T^0$ in (4.1) and $Se_k = -S^0$ in (4.2).

$$T^0 + Te_k - \lambda Ie_k + R_{22}T^0te_k = 0$$

$$T^0 + (-T^0) - \lambda e_k + R_{22}T^0 = 0$$

$$R_{22}T^0 = \lambda e_k.$$

$$S^0 + Se_k - \lambda Ie_k - \eta Ie_k + R_{11}S^0se_k + \eta e_kte_k + R_{12}T^0te_k = 0$$

$$S^0 + (-S^0) - \lambda e_k - \eta e_k + R_{11}S^0 + \eta e_k + R_{12}T^0 = 0$$

$$R_{11}S^0 + R_{12}T^0 = \lambda e_k.$$

To obtain equation (4.9) in Lemma 4.2.2, we substitute the matrix R into (4.6) and obtain the following three equations:

$$R_{11}^2S^0s + R_{11}(S - \lambda I - \eta I) + \lambda I = 0 \quad (4.12)$$

$$R_{22}^2T^0t + R_{22}(T - \lambda I) + \lambda I = 0 \quad (4.13)$$

$$(R_{11}R_{12} + R_{12}R_{22})T^0t + R_{11}\eta e_kt + R_{12}(T - \lambda I) = 0 \quad (4.14)$$

Equation (4.12) is rearranged and multiplied from left by $(I - R_{11})^{-1}$ as follows:

$$R_{11}^2S^0s + R_{11}S - R_{11}\eta I = R_{11}\lambda I - \lambda I$$

$$-R_{11}^2S^0s - R_{11}S + R_{11}\eta I = (I - R_{11})\lambda I$$

$$-(I - R_{11})^{-1}R_{11}^2S^0s - (I - R_{11})^{-1}R_{11}S + (I - R_{11})^{-1}R_{11}\eta I = \lambda I$$

$$(I - R_{11})^{-1}R_{11}[-R_{11}S^0s - S] + (I - R_{11})^{-1}R_{11}\eta I = \lambda I$$

The last equation is multiplied from right by column vector $\mathbf{e}_k$:

$$(I - R_{11})^{-1}R_{11}[-R_{11}\mathbf{S}^0 - S\mathbf{e}_k] + (I - R_{11})^{-1}R_{11}\eta\mathbf{e}_k = \lambda\mathbf{e}_k$$

$$(I - R_{11})^{-1}R_{11}[(I - R_{11})\mathbf{S}^0] + (I - R_{11})^{-1}R_{11}\eta\mathbf{e}_k = \lambda\mathbf{e}_k$$

$$(I - R_{11})^{-1}R_{11}\eta = \lambda I + R_{11}S.$$

Proof of Lemma 4.2.1 To obtain matrix R_{11} , the algorithm given in Lemma 4.2.1 is used.

To calculate matrix R_{22} , the equation (4.13) is used as follows:

$$R_{22}^2\mathbf{T}^0\mathbf{t} + R_{22}(T - \lambda I) + \lambda I = 0$$

$$R_{22}\mathbf{T}^0\mathbf{t} + (T - \lambda I) + R_{22}^{-1}\lambda I = 0$$

Equation $R_{22}\mathbf{T}^0 = \lambda\mathbf{e}_k$ in Lemma 4.2.2 is substituted and multiplied from left by R_{22} :

$$\lambda\mathbf{e}_k\mathbf{t} + (T - \lambda I) + R_{22}^{-1}\lambda I = 0$$

$$R_{22}^{-1}\lambda I = (\lambda I - \lambda\mathbf{e}_k\mathbf{t} - T)$$

$$\lambda I = R_{22}(\lambda I - \lambda\mathbf{e}_k\mathbf{t} - T)$$

$$R_{22} = \lambda I(\lambda I - \lambda\mathbf{e}_k\mathbf{t} - T)^{-1}.$$

To calculate matrix R_{12} , the equation (4.14) is rewritten by using the equation (4.8) in Lemma 4.2.2:

$$R_{11}R_{12}\mathbf{T}^0\mathbf{t} + R_{12}R_{22}\mathbf{T}^0\mathbf{t} + R_{11}\eta\mathbf{e}_k\mathbf{t} + R_{12}T - R_{12}\lambda I = 0$$

$$R_{11}(\lambda\mathbf{e}_k\mathbf{t} - R_{11}\mathbf{S}^0\mathbf{t}) + R_{11}\eta\mathbf{e}_k\mathbf{t} = R_{12}\lambda I - R_{12}\lambda\mathbf{e}_k\mathbf{t} - R_{12}T$$

$$R_{11}(\lambda\mathbf{e}_k\mathbf{t} - R_{11}\mathbf{S}^0\mathbf{t} + \eta\mathbf{e}_k\mathbf{t}) = R_{12}(\lambda I - \lambda\mathbf{e}_k\mathbf{t} - T)$$

$$R_{12} = R_{11}(\lambda \mathbf{e}_k \mathbf{t} - R_{11} \mathbf{S}^0 \mathbf{t} + \eta \mathbf{e}_k \mathbf{t})(\lambda I - \lambda \mathbf{e}_k \mathbf{t} - T)^{-1}$$

From the equation (4.9) in Lemma 4.2.2 and the matrix R_{22} , the equation belonging to R_{12} is given as

$$\begin{aligned} R_{12} &= R_{11}[(I - R_{11})^{-1} R_{11} \eta \mathbf{e}_k + \eta \mathbf{e}_k] \mathbf{t} \lambda^{-1} R_{22} \\ &= R_{11}[(I - R_{11})^{-1} - I] \eta \mathbf{e}_k + \eta \mathbf{e}_k \mathbf{t} \lambda^{-1} R_{22} \\ &= R_{11}(I - R_{11})^{-1} \eta \mathbf{e}_k \mathbf{t} \lambda^{-1} R_{22}. \end{aligned}$$

Theorem 4.2.3. π_{n0} and π_{n1} are stationary probability vectors representing that there are n customers in the system when the server is in the working vacation period and when the server is in the regular busy period, respectively.

$$\pi_{n0} = \pi_{00} \mathbf{s} R_{11}^n, \quad n \geq 1$$

$$\pi_{n1} = \pi_{00} \mathbf{s} \sum_{i=0}^{n-1} R_{11}^i R_{12} R_{22}^{n-1-i}, \quad n \geq 1 \quad (4.15)$$

$\pi_{00} = (1 - \rho) [\mathbf{s}(I - R_{11})^{-2} R_{11} \mathbf{e}_k \frac{\eta}{\mu_b} + \mathbf{s}(I - R_{11})^{-1} \mathbf{e}_k (1 - \rho)]^{-1}$ where $\rho = -\lambda \mathbf{t} T^{-1} \mathbf{e}_k$ is traffic intensity; $\mu_b = 1 / -\mathbf{t} T^{-1} \mathbf{e}_k$ is the service rate in the regular busy period.

Proof. First the following system of equations was obtained from equation (4.10):

$$-\lambda \pi_{00} + \pi_{10} \mathbf{S}^0 + \pi_{11} \mathbf{T}^0 = 0$$

$$\pi_{00} \lambda \mathbf{s} + \pi_{10} (S - \lambda I - \eta I + R_{11} \mathbf{S}^0 \mathbf{s}) = 0$$

$$\pi_{10} (\eta \mathbf{e}_k \mathbf{t} + R_{12} \mathbf{T}^0 \mathbf{t}) + \pi_{11} (T - \lambda I + R_{22} \mathbf{T}^0 \mathbf{t}) = 0$$

Firstly, the equation (4.12) is multiplied from left by R_{11}^{-1} and the equation $R_{11} \mathbf{S}^0 \mathbf{s} + S = (\lambda I + \eta I) - R_{11}^{-1} \lambda I$ is obtained. Substituting this equation into the second equation of the above system of equations, π_{10} is obtained as follows.

$$\pi_{00} \lambda \mathbf{s} - \pi_{10} R_{11}^{-1} \lambda I = 0$$

$$\pi_{10} = \pi_{00} s R_{11}.$$

Secondly, by using the equations $R_{22} \mathbf{T}^0 \mathbf{t} = \lambda \mathbf{e}_k \mathbf{t}$ and $R_{12} \mathbf{T}^0 = \lambda \mathbf{e}_k + R_{11} S \mathbf{e}_k$ in Lemma 4.2.2, the third equation of the above system of equation becomes the following equation.

$$\pi_{10}(\eta I + \lambda I + R_{11} S) \mathbf{e}_k \mathbf{t} + \pi_{11}(T - \lambda I + \lambda \mathbf{e}_k \mathbf{t}) = 0$$

The equation $\lambda I + R_{11} S = R_{11}(I - R_{11})^{-1} \eta = [(I - R_{11})^{-1} - I] \eta$ in Lemma 4.2.2 is substituted into the last equation.

$$\pi_{11}(\lambda I - \lambda \mathbf{e}_k \mathbf{t} - T) = \pi_{10}(I - R_{11})^{-1} \eta \mathbf{e}_k \mathbf{t}$$

$$\pi_{11} = \pi_{10}(I - R_{11})^{-1} \eta \mathbf{e}_k \mathbf{t} (\lambda I - \lambda \mathbf{e}_k \mathbf{t} - T)^{-1}$$

After π_{10} is substituted in the last equation, the equations R_{22} and R_{12} in Lemma 4.2.1 are used and π_{11} is obtained as follows.

$$\pi_{11} = \pi_{00} s R_{12}.$$

Thirdly, matrix R^{n-1} is calculated iteratively as

$$R^{n-1} = \begin{bmatrix} R_{11}^{n-1} & \sum_{i=0}^{n-2} R_{11}^i R_{12} R_{22}^{n-2-i} \\ 0 & R_{22}^{n-1} \end{bmatrix}, \quad n \geq 2.$$

By substituting vectors π_{10} and π_{11} and the matrix R^{n-1} into equation $\pi_n = \pi_1 R^{n-1}$, the steady-state probability vectors in Theorem 4.2.3 are obtained in terms of π_{00} .

To obtain probability π_{00} , the normalization condition $1 = \pi_0 + \pi_1(I - R)^{-1} \mathbf{e}_k$, provided in Neuts (1981), is rearranged as in (4.16):

$$\begin{aligned}
1 &= \pi_0 + \pi_1(I - R)^{-1}e_k \\
1 &= \pi_{00} + \begin{bmatrix} \pi_{10} & \pi_{11} \end{bmatrix} \left[\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix} \right]^{-1} e_k \\
1 &= \pi_{00} + \begin{bmatrix} \pi_{10} & \pi_{11} \end{bmatrix} \begin{bmatrix} (I - R_{11})^{-1} & (I - R_{11})^{-1}R_{12}(I - R_{22})^{-1} \\ 0 & (I - R_{22})^{-1} \end{bmatrix} e_k \\
1 &= \pi_{00}s[I + R_{11}(I - R_{11})^{-1} + R_{11}(I - R_{11})^{-1}R_{12}(I - R_{22})^{-1} + R_{12}(I - R_{22})^{-1}]e_k.
\end{aligned}$$

The matrices R_{11} and $(I - R_{11})^{-1}$ are interchangeable. Also we know that $I + R_{11}(I - R_{11})^{-1} = (I - R_{11})^{-1}$. In this case, the mathematical operation is continued as follows.

$$\begin{aligned}
1 &= \pi_{00}s(I - R_{11})^{-1}[I + R_{11}R_{12}(I - R_{22})^{-1} + (I - R_{11})R_{12}(I - R_{22})^{-1}]e_k \\
1 &= \pi_{00}s(I - R_{11})^{-1}[(I - R_{22}) + R_{11}R_{12} + (I - R_{11})R_{12}](I - R_{22})^{-1}e_k \\
1 &= \pi_{00}s(I - R_{11})^{-1}[I - R_{22} + R_{12}](I - R_{22})^{-1}e_k. \tag{4.16}
\end{aligned}$$

By substituting matrix R_{12} into the equation (4.16), the following is found:

$$1 = \pi_{00}s(I - R_{11})^{-1}[I + R_{11}(I - R_{11})^{-1}\eta e_k t \lambda^{-1} R_{22}(I - R_{22})^{-1}]e_k.$$

Note that we have the following equation:

$$\begin{aligned}
\lambda^{-1}R_{22}(I - R_{22})^{-1} &= (\lambda I - \lambda e_k t - T)^{-1} \left(I - \lambda(\lambda I - \lambda e_k t - T)^{-1} \right)^{-1} \\
\lambda^{-1}R_{22}(I - R_{22})^{-1} &= \left[\left(I - \lambda(\lambda I - \lambda e_k t - T)^{-1} \right) (\lambda I - \lambda e_k t - T) \right]^{-1} \\
\lambda^{-1}R_{22}(I - R_{22})^{-1} &= -(\lambda e_k t + T)^{-1}. \tag{4.17}
\end{aligned}$$

By substituting the equation (4.17) into the last equation, we obtain the following equation.

$$1 = \pi_{00}s(I - R_{11})^{-1} \left[I - R_{11}(I - R_{11})^{-1}\eta e_k t (\lambda e_k t + T)^{-1} \right] e_k$$

The inverse of $(\lambda \mathbf{e}_k \mathbf{t} + T)$ is now stated with square matrix $\rho = -\lambda \mathbf{e}_k \mathbf{t} T^{-1}$ as follows:

$$\begin{aligned}
 (\lambda \mathbf{e}_k \mathbf{t} + T)^{-1} &= \left((\lambda \mathbf{e}_k \mathbf{t} T^{-1} + I) T \right)^{-1} \\
 (\lambda \mathbf{e}_k \mathbf{t} + T)^{-1} &= T^{-1} \left(I - (-\lambda \mathbf{e}_k \mathbf{t} T^{-1}) \right)^{-1} \\
 (\lambda \mathbf{e}_k \mathbf{t} + T)^{-1} &= T^{-1} \sum_{n=0}^{\infty} (-\lambda \mathbf{e}_k \mathbf{t} T^{-1})^n \\
 (\lambda \mathbf{e}_k \mathbf{t} + T)^{-1} &= T^{-1} (I - \rho)^{-1}.
 \end{aligned} \tag{4.18}$$

Substituting the inverse in (4.18) into the last equation completes the proof of Theorem 4.2.3.

$$\begin{aligned}
 1 &= \pi_{00} \mathbf{s} (I - R_{11})^{-1} \left[I - R_{11} (I - R_{11})^{-1} \eta \mathbf{e}_k \mathbf{t} T^{-1} (I - \rho)^{-1} \right] \mathbf{e}_k \\
 &= \pi_{00} \left[\mathbf{s} (I - R_{11})^{-1} \mathbf{e}_k + \mathbf{s} R_{11} (I - R_{11})^{-2} \mathbf{e}_k \frac{\eta}{\lambda} (-\lambda \mathbf{t} T^{-1}) \sum_{n=0}^{\infty} (-\lambda \mathbf{e}_k \mathbf{t} T^{-1})^n \mathbf{e}_k \right] \\
 &= \pi_{00} \left[\mathbf{s} (I - R_{11})^{-1} \mathbf{e}_k + \mathbf{s} R_{11} (I - R_{11})^{-2} \mathbf{e}_k \frac{\eta}{\lambda} \rho (1 - \rho)^{-1} \right].
 \end{aligned}$$

Theorem 4.2.4. *The probabilities that the system is in the working vacation period and that the system is in the regular busy period are given as follows, respectively:*

$$P\{j = 0\} = K_1 (1 - \rho) \mathbf{s} (I - R_{11})^{-1} \mathbf{e}_k. \tag{4.19}$$

$$P\{j = 1\} = K_1 \frac{\eta}{\mu_b} \mathbf{s} (I - R_{11})^{-2} R_{11} \mathbf{e}_k \tag{4.20}$$

where

$$K_1 = \left[\mathbf{s} (I - R_{11})^{-1} \mathbf{e}_k (1 - \rho) + \mathbf{s} (I - R_{11})^{-2} R_{11} \mathbf{e}_k \frac{\eta}{\mu_b} \right]^{-1}.$$

Proof. The probability that the system is in the working vacation period is obtained by Theorem 4.2.3.

$$P\{j = 0\} = \sum_{n=0}^{\infty} \pi_{n0} = \pi_{00} \mathbf{s} \left(\sum_{n=0}^{\infty} R_{11}^n \right) \mathbf{e}_k = K_1 (1 - \rho) \mathbf{s} (I - R_{11})^{-1} \mathbf{e}_k.$$

The probability that the system is in the regular busy period is obtained by Theorem 4.2.3.

$$\begin{aligned}
P\{j = 1\} &= \sum_{n=1}^{\infty} \boldsymbol{\pi}_{n1} = \pi_{00} \mathbf{s} \left[\sum_{n=1}^{\infty} \sum_{i=0}^{n-1} R_{11}^i R_{12} R_{22}^{n-1-i} \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[\sum_{n=0}^{\infty} R_{11}^n R_{12} \sum_{i=n+1}^{\infty} R_{22}^{i-n-1} \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[(I - R_{11})^{-1} R_{12} (I - R_{22})^{-1} \right] \mathbf{e}_k.
\end{aligned}$$

After the matrix R_{12} is substituted, we obtain the following equation by (4.17).

$$P\{j = 1\} = \pi_{00} \mathbf{s} \left[(I - R_{11})^{-2} R_{11} \eta \mathbf{e}_k \mathbf{t} (-1) (\lambda \mathbf{e}_k \mathbf{t} + T)^{-1} \right] \mathbf{e}_k$$

Now, we substitute the equation (4.18) in the last equation.

$$\begin{aligned}
P\{j = 1\} &= \pi_{00} \mathbf{s} \left[(I - R_{11})^{-2} R_{11} \eta \mathbf{e}_k \mathbf{t} (-1) \frac{\lambda}{\lambda} T^{-1} \sum_{n=0}^{\infty} (-\lambda \mathbf{e}_k \mathbf{t} T^{-1})^n \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[(I - R_{11})^{-2} R_{11} \frac{\eta}{\lambda} \mathbf{e}_k (-\lambda \mathbf{t} T^{-1}) \sum_{n=0}^{\infty} (-\lambda \mathbf{e}_k \mathbf{t} T^{-1})^n \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} (I - R_{11})^{-2} R_{11} \mathbf{e}_k \frac{\eta}{\lambda} \rho (1 - \rho)^{-1} \\
&= (1 - \rho) K_1 \mathbf{s} (I - R_{11})^{-2} R_{11} \mathbf{e}_k \frac{\eta}{\mu_b} (1 - \rho)^{-1} = K_1 \mathbf{s} (I - R_{11})^{-2} R_{11} \mathbf{e}_k \frac{\eta}{\mu_b}.
\end{aligned}$$

4.3 Stochastic Decomposition

In vacation queue models, the stationary queue length can be decomposed into the sum of two independent random variables, i.e. the queue length of the queuing system without any vacation and the additional queue length due to vacations. Stochastic decomposition for many vacation models can be seen in Tian & Zhang (2006). Let L , L_d , and L_v denote the total number of customers in the system for the classical queue

model, the additional number of customers in the system due to vacations, and the number of customers in the system for the vacation queue model, respectively. Let $Q(z)$, $Q_d(z)$, and $Q_v(z)$ be defined as the corresponding probability generating functions (p.g.f.). The stochastic decomposition properties are written as follows

$$L_v = L + L_d, \quad (4.21)$$

$$Q_v(z) = Q(z)Q_d(z). \quad (4.22)$$

Probability generating function $Q_v(z)$ is obtained by using Theorem 4.2.3 as follows:

$$\begin{aligned}
Q_v(z) &= \sum_{n=0}^{\infty} z^n \pi_{n0} \mathbf{e}_k + \sum_{n=1}^{\infty} z^n \pi_{n1} \mathbf{e}_k \\
&= \pi_{00} + \sum_{n=1}^{\infty} z^n \pi_{00} \mathbf{s} R_{11}^n \mathbf{e}_k + \sum_{n=1}^{\infty} z^n \pi_{00} \mathbf{s} \sum_{i=0}^{n-1} R_{11}^i R_{12} R_{22}^{n-1-i} \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[I + \sum_{n=1}^{\infty} z^n R_{11}^n + \sum_{n=1}^{\infty} z^n \sum_{i=0}^{n-1} R_{11}^i R_{12} R_{22}^{n-1-i} \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[I + \sum_{n=1}^{\infty} z^n R_{11}^n + \sum_{n=0}^{\infty} R_{11}^n R_{12} \sum_{i=n+1}^{\infty} z^i R_{22}^{i-n-1} \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} \left[I + \sum_{n=1}^{\infty} z^n R_{11}^n + \sum_{n=0}^{\infty} z^n R_{11}^n z R_{12} (I - z R_{22})^{-1} \right] \mathbf{e}_k \\
&= \pi_{00} \mathbf{s} (I - z R_{11})^{-1} [I + z R_{12} (I - z R_{22})^{-1}] \mathbf{e}_k. \\
&= \pi_{00} \mathbf{s} (I - z R_{11})^{-1} [I - z R_{22} + z R_{12}] (I - z R_{22})^{-1} \mathbf{e}_k.
\end{aligned}$$

By substituting matrix R_{12} into the last equation, the following is obtained:

$$\begin{aligned}
Q_v(z) &= \pi_{00} \mathbf{s} (I - z R_{11})^{-1} \left[\left(I - R_{11} (I - R_{11})^{-1} \frac{\eta}{\lambda} \right) \mathbf{e}_k \left(\mathbf{t} (I - z R_{22})^{-1} \mathbf{e}_k \right)^{-1} \right. \\
&\quad \left. + R_{11} (I - R_{11})^{-1} \frac{\eta}{\lambda} \mathbf{e}_k \right] \mathbf{t} (I - z R_{22})^{-1} \mathbf{e}_k.
\end{aligned}$$

We substitute π_{00} in Theorem 4.2.3 and obtain $Q_v(z)$ as follows:

$$Q_v(z) = (1 - \rho)t(I - zR_{22})^{-1}e_k K_1 s(I - zR_{11})^{-1} \left[\left(I - R_{11}(I - R_{11})^{-1} \frac{\eta}{\lambda} \right) \times \left((t(I - zR_{22})^{-1}e_k)^{-1} \right) + R_{11}(I - R_{11})^{-1} \frac{\eta}{\lambda} \right] e_k$$

where

$$K_1 = \left[s(I - R_{11})^{-1}e_k(1 - \rho) + s(I - R_{11})^{-2}R_{11}e_k \frac{\eta}{\mu_b} \right]^{-1}. \quad (4.23)$$

$Q_v(z)$ in (4.23) satisfies stochastic decomposition in (4.22) since $Q(z) = (1 - \rho)t(I - zR_{22})^{-1}e_k$ is p.g.f. for the $M/C_k/1$ queue model without any vacation in Tian et al. (2009). Then, $Q_d(z)$ can be written as follows:

$$Q_d(z) = K_1 s(I - zR_{11})^{-1} \left[\left(I - R_{11}(I - R_{11})^{-1} \frac{\eta}{\lambda} \right) \left(t(I - zR_{22})^{-1}e_k \right)^{-1} + R_{11}(I - R_{11})^{-1} \frac{\eta}{\lambda} \right] e_k.$$

By substituting the equation (4.9) in Lemma 4.2.2 into the last equation (also, into K_1), $Q_d(z)$ is given as in (4.24):

$$Q_d(z) = K_2 s(I - zR_{11})^{-1} \left[\left((I + R_{11}S) \frac{1}{\lambda} \right) - R_{11}S \frac{1}{\lambda} \left(t(I - zR_{22})^{-1}e_k \right)^{-1} \right] e_k \quad (4.24)$$

where

$$K_2 = \left[s(I - R_{11})^{-1} \left(I + R_{11}S \frac{1}{\mu_b} \right) e_k \right]^{-1}.$$

From the first-order derivatives of $Q_d(z)$, $Q(z)$ and $Q_v(z)$, the performance measures are obtained as follows:

- The average number of customers in the system for the $M/C_k/1$ queue model without any vacation is calculated from $EN = Q'(z = 1)$ as:

$$EN = (1 - \rho)t(I - R_{22})^{-2}R_{22}e_k \quad (4.25)$$

- The average additional number of customers in the system due to vacations is

computed from $EN_d = Q'_d(z = 1)$ as:

$$EN_d = K_2 \mathbf{s}(I - R_{11})^{-1} R_{11} \left[(I - R_{11})^{-1} \left(I + R_{11} S \frac{1}{\mu_b} \right) + S \frac{1}{\lambda} (1 - \rho)^2 \mathbf{t}(I - R_{22})^{-2} R_{22} \mathbf{e}_k \right] \mathbf{e}_k$$

where

$$K_2 = \left[\mathbf{s}(I - R_{11})^{-1} \left(I + R_{11} S \frac{1}{\mu_b} \right) \mathbf{e}_k \right]^{-1}. \quad (4.26)$$

- The average number of customers in the system for the $M/C_k/1$ queue model with a vacation is obtained as:

$$EN_v = Q'_v(z = 1) = EN + EN_d. \quad (4.27)$$

4.4 Numerical Study

The effect of parameters μ_b , μ_v , λ , and η on the performance measures, such as (i) the expected number of customers in the system without any vacation (EN), (ii) the expected additional number of customers in the system due to vacations (EN_d), and (iii) the expected number of customers in the system with a vacation (EN_v), is investigated for the queue model concerned. In addition, the effect of the variation in the parameters concerned on both probabilities that the server was in the working vacation period (P_v) and that the server was in the regular busy period (P_b) were presented. Service times with different CV^2 may be characterized by different kinds of phase-type distributions. In this context, Cox distribution is of great importance as it is the generalized form of many phase-type distributions. That is, k -phase Cox distribution is equivalent to

- the exponential distribution when k is one,
- Erlang distribution when the sojourn times in each state are the same, $\mu_1 = \dots = \mu_k = \mu$, and the probability that the service is ended in any state equals zero, $\alpha_1 = \dots = \alpha_{k-1} = 1$,

- the hypo-exponential (generalized Erlang) distribution when the probability that the service is ended in any state equals zero, $\alpha_1 = \dots = \alpha_{k-1} = 1$.

A numerical study is performed by MATLAB. To examine the effect of service rates μ_v and μ_b , we first assume that the arrival rate $\lambda = 1.5$ and that the vacation rate $\eta = 0.5$. That is, parameter $\mu_v = 1.4286$ is also fixed to examine the change in parameter μ_b . Additionally, $\mu_b = 2.4301$ is fixed to see the change in parameter μ_v . Secondly, we fix both parameters $\mu_b = 2.4301$ and $\mu_v = 1.4286$ to analyze the effect of parameters λ and η . Similarly, a sensitivity analysis is given for parameters λ and η by using values $\eta = 0.5$ and $\lambda = 1.5$, respectively.

We know that the $M/C_k/1$ queue model with $k = 1$ becomes the $M/M/1$ queue model. Liu et al. (2007) investigated the $M/M/1$ queue model with multiple working vacations, as stated previously. Our results in Table 4.1 support their study when there is a single phase.

The increments in parameters μ_b and λ affect traffic intensity ρ since $\rho = \lambda/\mu_b$. Thus, P_b is expected to decrease with an increment in μ_b and to increase for the decreasing values of μ_b . An increment in parameter μ_b causes the number in the system to decrease. When λ increases, the number in the system is expected to increase. This result can be seen from columns EN and EN_v in Table 4.1. Moreover, if μ_b and λ increase, the additional number due to vacation EN_d also increases. Parameters μ_v and η in the vacation period do not have any effect on ρ , unlike parameters μ_b and λ . P_b decreases as μ_v increases; however, it increases as η increases. Increments in parameters μ_v and η do not affect EN , while these increments cause EN_d and EN_v to decrease.

The square of coefficient of variation is $CV_b^2 = 1$ in Table 4.1 since the service times are exponentially distributed. The number of phases increases as CV_b^2 increases in Table 4.2, for Cox distribution is a heavy-tailed distribution and $CV_b^2 > 1$. The comments given for Table 4.1 can also be stated for the results in Table 4.2. The effect of the number of phases on the performance measures can be inferred by

comparing Table 4.1 and Table 4.2.

Table 4.1 The effect of the parameters on the performance measures for $n=1$

<i>Parameters</i>	ρ	P_b	P_v	EN	EN_d	EN_v
μ_b						
2.0489	0.73	0.5525	0.4475	2.7327	0.6852	3.4180
2.1145	0.71	0.5244	0.4756	2.4410	0.7193	3.1603
2.1735	0.69	0.5015	0.4985	2.2272	0.7470	2.9742
2.2266	0.67	0.4826	0.5174	2.0644	0.7700	2.8344
λ						
0.5	0.21	0.0755	0.9245	0.2591	0.1512	0.4103
1.0	0.41	0.2106	0.7894	0.6993	0.4218	1.1211
1.5	0.62	0.4215	0.5785	1.6127	0.8441	2.4568
2.0	0.82	0.7067	0.2933	4.6501	1.4151	6.0652
μ_v						
1.3846	0.62	0.4275	0.5725	1.6127	0.8938	2.5066
1.4286	0.62	0.4215	0.5785	1.6127	0.8441	2.4568
1.4667	0.62	0.4163	0.5837	1.6127	0.8020	2.4147
1.5000	0.62	0.4118	0.5882	1.6127	0.7659	2.3786
η						
0.25	0.62	0.3626	0.6374	1.6127	1.4522	3.0649
0.50	0.62	0.4215	0.5785	1.6127	0.8441	2.4568
0.75	0.62	0.4542	0.5458	1.6127	0.6065	2.2192
1.00	0.62	0.4761	0.5239	1.6127	0.4768	2.0896

When μ_b is fixed (2.0489) in Table 4.1 and Table 4.2, we expect the average number in the system to increase as the number of phases increases. For example, EN (EN_v) is 2.7327, 3.2606, and 4.1414 (3.4180, 3.7975, and 4.4180) for $k = 1, 2, 3$, respectively. On the other hand, we expect the average additional number due to vacations to increase as the number of phases decreases. For instance, EN_d is 0.6852, 0.5368, and 0.2766 for $k = 1, 2, 3$, respectively. These observations are also illustrated in Figure 4.1 for the increasing values of μ_b .

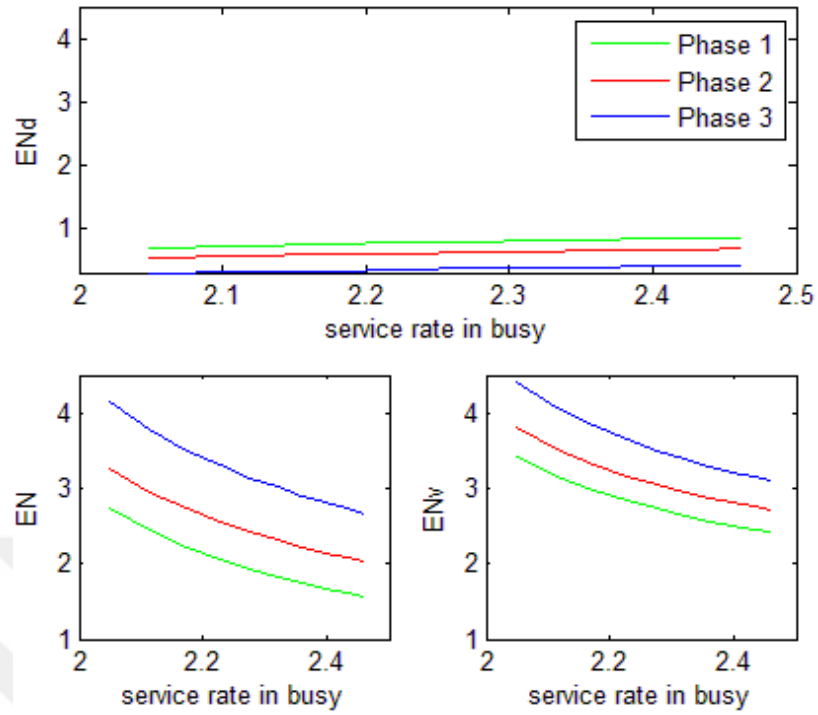


Figure 4.1 The effect of parameter μ_b on the performance measures

When μ_v is fixed (1.3846) in Table 4.1 and Table 4.2, the average number in the system increases as the number of phases increases. For example, EN (EN_v) is 1.6127, 2.0819, and 2.7366 (2.5066, 2.8051, and 3.2062) for $k = 1, 2, 3$, respectively. On the other hand, the average additional number due to vacations increases as the number of phases decreases. For instance, EN_d is 0.8938, 0.7233, and 0.4696 for $k = 1, 2, 3$, respectively. The observations for the increasing values of μ_v are represented in Figure 4.2.

Table 4.2 The effect of the parameters on the performance measures for $n=2,3$

<i>Parameters</i>	ρ	Phase=2					Phase=3				
		P_b	P_v	EN	EN_d	EN_v	P_b	P_v	EN	EN_d	EN_v
μ_b											
2.0489	0.73	0.5433	0.4567	3.2606	0.5368	3.7975	0.5314	0.4686	4.1414	0.2766	4.4180
2.1145	0.71	0.5151	0.4849	2.9528	0.5654	3.5182	0.5032	0.4968	3.7752	0.3058	4.0810
2.1735	0.69	0.4922	0.5078	2.7280	0.5886	3.3166	0.4803	0.5197	3.5074	0.3295	3.8369
2.2266	0.67	0.4732	0.5268	2.5568	0.6078	3.1646	0.4614	0.5386	3.3029	0.3491	3.6520
λ											
0.5	0.21	0.0726	0.9274	0.2842	0.1219	0.4061	0.0688	0.9312	0.3192	0.0807	0.3999
1.0	0.41	0.2046	0.7954	0.8349	0.3354	1.1703	0.1968	0.8032	1.0241	0.2108	1.2350
1.5	0.62	0.4124	0.5876	2.0819	0.6695	2.7513	0.4009	0.5991	2.7366	0.4118	3.1484
2.0	0.82	0.6986	0.3014	6.4537	1.1310	7.5847	0.6884	0.3116	8.9712	0.6958	9.6670
μ_v											
1.3846	0.62	0.4191	0.5809	2.0819	0.7233	2.8051	0.4078	0.5922	2.7366	0.4696	3.2062
1.4286	0.62	0.4124	0.5876	2.0819	0.6695	2.7513	0.4009	0.5991	2.7366	0.4118	3.1484
1.4667	0.62	0.4067	0.5933	2.0819	0.6243	2.7062	0.3951	0.6049	2.7366	0.3637	3.1003
1.5000	0.62	0.4017	0.5983	2.0819	0.5859	2.6677	0.3901	0.6099	2.7366	0.3232	3.0598
η											
0.25	0.62	0.3611	0.6389	2.0819	1.3142	3.3961	0.3596	0.6404	2.7366	1.1150	3.8516
0.50	0.62	0.4124	0.5876	2.0819	0.6695	2.7513	0.4009	0.5991	2.7366	0.4118	3.1484
0.75	0.62	0.4412	0.5588	2.0819	0.4264	2.5083	0.4237	0.5763	2.7366	0.1504	2.8870
1.00	0.62	0.4610	0.5390	2.0819	0.2996	2.3815	0.4396	0.5604	2.7366	0.0182	2.7549

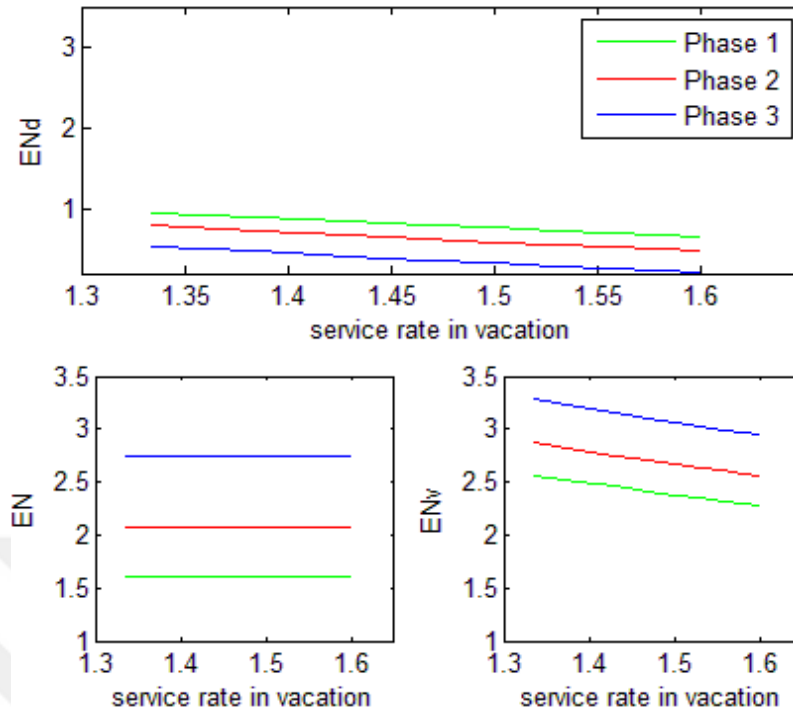


Figure 4.2 The effect of parameter μ_v on the performance measures

When arrival rate λ is fixed as $\lambda = 1.5$ in Table 4.1 and Table 4.2, it is seen that the average numbers in the system are considerably different as the number of phases increases. For example, EN (EN_v) is 1.6127, 2.0819, and 2.7366 (2.4568, 2.7513, and 3.1484) for $k = 1, 2, 3$, respectively. The values of EN_d are 0.8441, 0.6695, and 0.4118 for $k = 1, 2, 3$, respectively. Figure 4.3 illustrates the comments mentioned for the increasing values of λ .

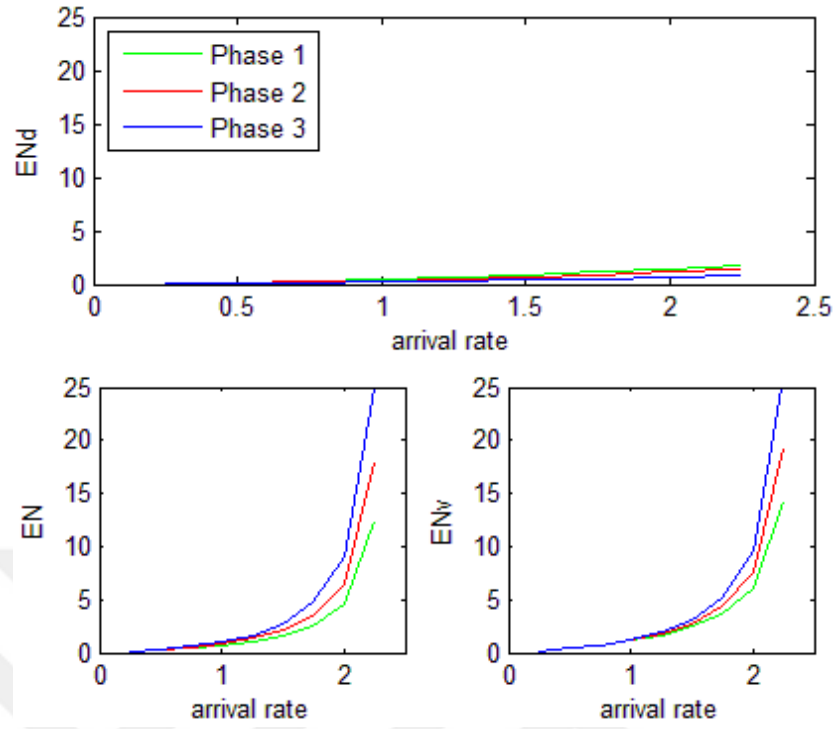


Figure 4.3 The effect of parameter λ on the performance measures

Finally, for the fixed values of η as $\eta = 0.5$ in Table 4.1 and Table 4.2, it can be seen that a high average number in the system is obtained as the number of phases increases. For example, EN (EN_v) is 1.6127, 2.0819, and 2.7366 (2.4568, 2.7513, and 3.1484) for $k = 1, 2, 3$, respectively. On the other hand, a high average additional number due to vacations is expected as the number of phases decreases. For instance, EN_d is 0.8441, 0.6695, and 0.4118 for $k = 1, 2, 3$, respectively. The observations for the increasing values of η are represented in Figure 4.4.

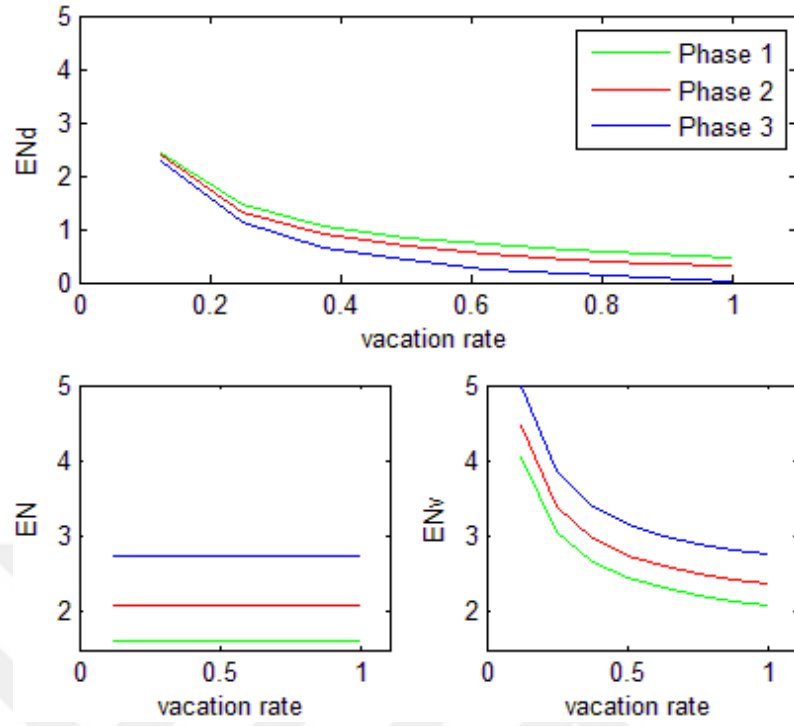


Figure 4.4 The effect of parameter η on the performance measures

As stated before, depending on CV_b^2 of service time distribution, the phase-type distribution chosen to characterize the service time differs. That is, by using the solution given for the $M/C_k/1$ queue model with multiple working vacations in Sect. “Stochastic Decomposition”, we can analyze the queue models with different service time distributions such as $M/E_k/1$ and $M/H_k/1$.

CHAPTER FIVE

QUEUEING MODELS WITH PHASE-TYPE SERVICE AND *MAP* ARRIVAL FOR CROWDSOURCING

For the studied queueing models in the previous chapters, Cox distribution (a special phase-type distribution) and Poisson process are used for service times and arrival process, respectively. This chapter represents the usage of general phase-type distribution for service times and also the utilization of Markovian arrival process as the generalization of Poisson process.

Three different queueing models with crowdsourcing in the context of service sectors will be presented in this chapter. In these models, there are two types of arrivals; namely Type 1 and Type 2 such that one group of customers (Type 1) after getting service from the system may volunteer with a certain probability for providing service for the other group of customers (Type 2). For all models, we keep to consider a more versatile point process, namely, Markovian arrival process (*MAP*) for Type 1 arrivals, Poisson arrivals for Type 2 and use phase type distribution for services. Firstly, we focus our attention to single server case of *MAP/PH/1*-type queue model with crowdsourcing. This model is studied analytically using matrix geometric method in Section 5.1. Additionally, in this chapter *MAP/PH/c*-type queueing model with crowdsourcing is carried out through simulation in Section 5.2. The multi-server model is studied by simulation approach because of the size of the state space. Finally, this chapter is finished by studying the model *MAP/PH/1* in the context of vacation in Section 5.3.

5.1 *MAP/PH/1* Queueing Model with Crowdsourcing

The section is organized as follows. In Section 5.1.1, the model *MAP/PH/1* is described. The steady-state analysis of the model is performed and some key system performance measures are presented in Section 5.1.2- Section 5.1.5. The stationary waiting time distribution of an admitted Type 1 customer in the queue is presented in

Section 5.1.6. After illustrative numerical examples, including an optimization problem, are discussed in Section 5.1.7, the analysis of the first model is ended.

5.1.1 Model Description

We consider a single server queueing system in which two types, say, Type 1 and Type 2, of customers arrive. The interested model is illustrated in Figure 5.1. We assume that Type 1 customers arrive according to a *MAP* with representation (D_0, D_1) of order m . An arriving Type 1 customer finding the server idle will get into service immediately. Otherwise, the customer will enter into a finite buffer of size $L, 1 \leq L < \infty$, to be served on a First-Come-First-Served (FCFS) basis when the server becomes free. Thus, it is possible for a Type 1 customer to be lost at the time of arrival due to the buffer being full. Let D , defined by $D = D_0 + D_1$, govern the underlying Markov chain of the *MAP* such that D_0 accounts for the transitions corresponding to no arrival; D_1 governs those corresponding to an arrival of a Type 1 customer. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and the arrival process does not terminate. Hence, we see that D_0 is a stable matrix. Let λ_1 denote the arrival rate of Type 1 customers. The arrivals of Type 2 customers are assumed to follow a Poisson process with rate λ_2 . There is no restriction on how many Type 2 customers can be in the system. That is, there is an infinite buffer space for Type 2 customers.

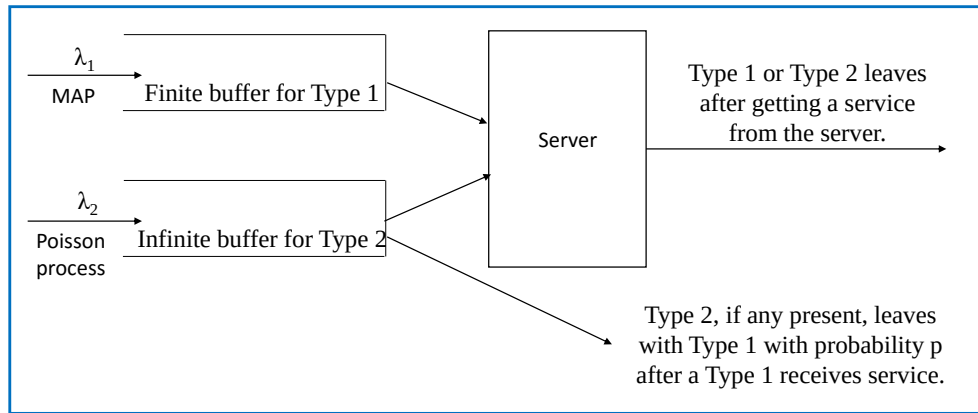


Figure 5.1 Diagram of queue model with crowdsourcing

While Type 1 customers are to be served by the server, Type 2 customers may be served by a Type 1 customer having already been served and also available to act as a server or by the (system) server. For example, one type of customers visit the store to procure items while the other type of customers orders over some medium such as Internet and phone and expects them to be delivered. The store management use the customers visiting them as couriers to "serve" the other type of customers. Not all in-store customers may be willing and in some cases not possible to act as servers for the store. Hence a probability is introduced for in-store customers to opt for servicing the other type. A Type 2 customer getting serviced by a Type 1 customer depends on the following conditions. First, that Type 1 customer should have just finished getting a service and opts to service a Type 2 customer. Secondly, at the time of opting to serve there is at least one Type 2 customer waiting to get a service. We assume that a served Type 1 customer will be available to act as a server for a Type 2 customer under the conditions mentioned above with probability p , $0 \leq p \leq 1$. With probability $q = 1 - p$, the served Type 1 customer will leave the system. Upon completion of a service the free server will offer service to a Type 1 customer on a *FCFS* basis; however, if there are no Type 1 customers waiting, the server will serve a Type 2 customer if there is one present in the queue. If a Type 1 customer decides to serve a Type 2 customer, for our analysis purposes that Type 2 customer will be removed from the system immediately. This is due to the fact that the system no longer needs to track that Type 2 customer.

We assume that the system server offers services to either type on a *FCFS* within the type; however, Type 1 customers have *non-preemptive priority* over Type 2 customers. The service times are assumed to be of phase type with representation (β, S) of order n with mean $\frac{1}{\mu} = \beta(-S)^{-1}\mathbf{e}$.

In the sequel we need the following notations. By $\mathbf{e}$ we will denote a column vector (of appropriate dimension) of 1's; $\mathbf{e}_i$ we will denote a unit column vector (of appropriate dimension) with 1 in the i^{th} position and 0 elsewhere; and I an identity matrix (of appropriate dimension). We will display the dimension should there be a need to emphasize it. For example, if there is a need to display the dimension of an identity matrix of order m , we will do so by writing I_m rather than I ; a unit vector of

dimension m will be denoted as $\mathbf{e}(m)$ rather than $\mathbf{e}$. The notation $\mathbf{S}^0$ is such that $\mathbf{S}\mathbf{e} + \mathbf{S}^0 = \mathbf{0}$. The symbols, $\otimes$ and $\oplus$, respectively, will stand for the Kronecker product and sum of matrices. Thus, if A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose $(i, j)^{th}$ block matrix is given by $a_{ij}B$; the Kronecker sum of two square matrices, say, G of order g and H of h , is given by $G \otimes I + I \otimes H$, a square matrix of order gh . For details and other properties on Kronecker products and Kronecker sums we refer the reader to Graham (1981), Marcus & Minc (1964), Steeb & Hardy (2011).

The *MAP*, first introduced by Neuts Neuts (1979) as a versatile Markovian point process, is a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-modulated Poisson process. For further details on *MAP* and their usefulness in stochastic modelling, we refer to Lucantoni et al. (1990), Lucantoni (1991), Neuts (1989, 1992) and for a review and recent work on *MAP* we refer the reader to Artalejo et al. (2010), Chakravarthy (2001, 2010).

Let δ be the stationary probability vector of the Markov process with irreducible generator D . That is, δ is the unique (positive) probability vector satisfying.

$$\delta D = \mathbf{0}, \delta \mathbf{e} = 1. \quad (5.1)$$

The constant $\lambda_1 = \delta D_1 \mathbf{e}$, referred to as the **fundamental rate**, gives the expected number of arrivals per unit of time in the stationary version of the *MAP*. Often, in model comparisons, it is convenient to select the time scale of the *MAP* so that λ_1 has a certain value. That is accomplished, in the continuous *MAP* case, by multiplying the coefficient matrices D_0 and D_1 , by the appropriate common constant.

5.1.2 The Steady-state Analysis

The steady-state analysis of the queueing model described in Section 5.1.1 will be represented in this section. First we need to define a few notations. Let $N_1(t)$, $N_2(t)$,

$N_3(t)$, $N_4(t)$, and $N_5(t)$ denote, respectively, the number of Type 2 customers in the queue, the number of Type 1 customers in the queue, the phase of the service (if any), the phase of the arrival process, and type of customer in service, at time t . Note that when the server is idle we need to track only the variable $N_4(t)$. The process $\{(N_1(t), N_2(t), N_3(t), N_4(t), N_5(t)) : t \geq 0\}$ is a continuous-time Markov chain with state space given by

$$\Omega = \{k : 1 \leq k \leq m\} \cup \{(i, i_1, j, k, r) : i \geq 0, 0 \leq i_1 \leq L, 1 \leq j \leq n, 1 \leq k \leq m, r = 1, 2\},$$

which is regrouped in terms of set of states as $\Omega = \{*\} \cup \{\underline{i}, i \geq 0\}$.

Note that the level $*$ of dimension m corresponds to the case when the server is idle and the arrival process is in one of m phases. The set of states, $\underline{i}$, partitioned as $\underline{i} = \{\hat{\underline{i}}\} \cup \{\hat{\hat{\underline{i}}}\}$, for $i \geq 0$. That is,

the level $\hat{\underline{i}} = \{(i, i_1, j, k, 1) : i \geq 0, 0 \leq i_1 \leq L, 1 \leq j \leq n, 1 \leq k \leq m\}$ of dimension $(L+1)mn$ corresponds to the case when the server is busy serving a Type 1 customer with i Type 2 customers waiting in the queue, $i_1, 0 \leq i_1 \leq L$, Type 1 customers waiting in the queue, the service is in one of n phases and the arrival process is in one of m phases;

the level $\hat{\hat{\underline{i}}} = \{(i, i_1, j, k, 2) : i \geq 0, 0 \leq i_1 \leq L, 1 \leq j \leq n, 1 \leq k \leq m\}$ of dimension $(L+1)mn$ corresponds to the case when the server is busy serving a Type 2 customer with i Type 2 customers waiting in the queue, $i_1, 0 \leq i_1 \leq L$, Type 1 customers waiting in the queue, the service is in one of n phases and the arrival process is in one of m phases.

The infinitesimal generator of the Markov chain governing the system is given by

$$Q = \begin{pmatrix} B_1 & B_0 & & & \\ B_2 & C_1 & A_0 & & \\ B_3 & A_2 & A_1 & A_0 & \\ & A_3 & A_2 & A_1 & A_0 \\ & & A_3 & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (5.2)$$

where the matrices appearing in Q are as follows (first in partitioned form and then the form of the blocks in those partitioned matrices).

$$B_0 = \mathbf{e}'_1(2L+2) \otimes \boldsymbol{\beta} \otimes D_1 + \lambda_2 \mathbf{e}'_{L+2}(2L+2) \otimes \boldsymbol{\beta} \otimes I, \quad B_1 = D_0 - \lambda_2 I,$$

$$B_2 = [\mathbf{e}_1(2L+2) + \mathbf{e}_{L+2}(2L+2)] \otimes \mathbf{S}^0 \otimes I, \quad B_3 = p(\mathbf{e}_1(2L+2) \otimes \mathbf{S}^0 \otimes I)$$

$$C_1 = \begin{pmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{pmatrix},$$

with

$$C_{11} = \begin{pmatrix} S \oplus D_0 - \lambda_2 I & I \otimes D_1 & & \\ \mathbf{S}^0 \boldsymbol{\beta} \otimes I & S \oplus D_0 - \lambda_2 I & I \otimes D_1 & \\ \ddots & \ddots & \ddots & \\ & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & S \oplus D_0 - \lambda_2 I & I \otimes D_1 \\ & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & S \oplus D - \lambda_2 I \end{pmatrix},$$

$$C_{21} = \begin{pmatrix} 0 & & & \\ \mathbf{S}^0 \boldsymbol{\beta} \otimes I & & & \\ & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & & \\ & & \ddots & \ddots \\ & & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \end{pmatrix},$$

$$C_{22} = \begin{pmatrix} S \oplus D_0 - \lambda_2 I & I \otimes D_1 & & & \\ & S \oplus D_0 - \lambda_2 I & I \otimes D_1 & & \\ & & \ddots & \ddots & \\ & & & S \oplus D_0 - \lambda_2 I & I \otimes D_1 \\ & & & & S \oplus D - \lambda_2 I \end{pmatrix},$$

$$A_0 = \lambda_2 I, \quad A_1 = \begin{pmatrix} C_{11} - pC_{21} & 0 \\ C_{21} & C_{22} \end{pmatrix}, \quad A_3 = p[\mathbf{e}_1(2L+2) \otimes \mathbf{e}'_{L+2}(2L+2) \otimes \mathbf{S}^0 \boldsymbol{\beta} \otimes I],$$

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & q\mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & \cdots & 0 \\ p\mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & p\mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & \cdots & 0 & 0 & \cdots & 0 & & \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & p\mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

The generator Q given in (5.2) is of the $GI/M/1$ type in Neuts (1981) and we can apply the well-know matrix-analytic method to perform the steady-state analysis. In this set up the rate matrix, R , of dimension $(2L + 2)mn$, is the minimal nonnegative solution to a cubic matrix equation in Neuts (1981). However, by combining the set of states as follows:

$$\tilde{\mathbf{i}} = \{\underline{2\mathbf{i}}, \underline{2\mathbf{i}+1}\}, \quad i \geq 0,$$

the model under study can be studied as QBD -process with the generator, $\tilde{Q}$ on state space $\tilde{\Omega} = \{\underline{*}\} \cup \{\tilde{\mathbf{i}}, \tilde{i} \geq 0\}$ given by

$$\tilde{Q} = \begin{pmatrix} B_1 & \mathbf{e}'_1(2) \otimes B_0 & & & \\ \tilde{B}_2 & E_1 & F_0 & & \\ & F_2 & F_1 & F_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (5.3)$$

where the block entries appearing in $\tilde{Q}$ are obtained from those of Q as follows.

$$\tilde{B}_2 = \begin{pmatrix} B_2 \\ B_3 \end{pmatrix}, E_1 = \begin{pmatrix} C_1 & A_0 \\ A_2 & A_1 \end{pmatrix},$$

$$F_0 = \begin{pmatrix} 0 & 0 \\ A_0 & 0 \end{pmatrix}, F_1 = \begin{pmatrix} A_1 & A_0 \\ A_2 & A_1 \end{pmatrix}, F_2 = \begin{pmatrix} A_3 & A_2 \\ 0 & A_3 \end{pmatrix}.$$

For the rest of the study we will analyze the model in steady-state using QBD approach.

5.1.3 The Stability Condition

Let $\xi = (\xi(1), \xi(2))$ be the steady state probability vector of the generator $F = F_0 + F_1 + F_2$. That is, ξ satisfies

$$\xi F = \mathbf{0}, \xi e = 1. \quad (5.4)$$

On noting that

$$F = \begin{pmatrix} A_1 + A_3 & A_0 + A_2 \\ A_0 + A_2 & A_1 + A_3 \end{pmatrix}, \quad (5.5)$$

is a circulant matrix, the vector ξ is of the form given by

$$\xi = 0.5(e' \otimes \pi), \quad (5.6)$$

where π satisfies (with $A = A_0 + A_1 + A_2 + A_3$)

$$\pi A = \mathbf{0}, \pi e = 1. \quad (5.7)$$

In the sequel, let $\pi = (\hat{\pi}, \tilde{\pi})$, with $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_1, \dots, \hat{\pi}_L)$ and $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \dots, \tilde{\pi}_L)$.

We now establish a few results that are needed for later use as well as serve as internal accuracy checks in our numerical computation.

Lemma 5.1.1. *We have*

$$\sum_{j=0}^L [\hat{\pi}_j + \tilde{\pi}_j](\mathbf{e} \otimes I) = \boldsymbol{\delta} \text{ and } \sum_{j=0}^L [\hat{\pi}_j + \tilde{\pi}_j](I \otimes \mathbf{e}) = \mu \boldsymbol{\beta}(-S)^{-1}. \quad (5.8)$$

where $\boldsymbol{\delta}$ is as given in (5.1) and μ is the service rate.

Proof: The steady-state equations in (5.7) can be rewritten as

$$\hat{\pi}_0(S \oplus D_0) + (\hat{\pi}_1 + \tilde{\pi}_1)(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) = \mathbf{0},$$

$$\hat{\pi}_{j-1}(I \otimes D_1) + \hat{\pi}_j(S \oplus D_0) + (\hat{\pi}_{j+1} + \tilde{\pi}_{j+1})(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) = \mathbf{0}, \quad 1 \leq j \leq L-1,$$

$$\hat{\pi}_{L-1}(I \otimes D_1) + \hat{\pi}_L(S \oplus D) = \mathbf{0},$$

$$\hat{\pi}_0(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) + \tilde{\pi}_0[(S + \mathbf{S}^0 \boldsymbol{\beta}) \oplus D_0] = \mathbf{0},$$

$$\tilde{\pi}_{j-1}(I \otimes D_1) + \tilde{\pi}_j(S \oplus D_0) = \mathbf{0}, \quad 1 \leq j \leq L-1,$$

$$\tilde{\pi}_{L-1}(I \otimes D_1) + \tilde{\pi}_L(S \oplus D) = \mathbf{0},$$

(5.9)

with the normalizing condition

$$\sum_{j=0}^L [\hat{\pi}_j + \tilde{\pi}_j] \mathbf{e} = 1. \quad (5.10)$$

Now adding the equations given in (5.9) we get

$$\begin{aligned} & \sum_{j=0}^{L-1} (\hat{\pi}_j + \tilde{\pi}_j) [(S \otimes I) \\ & + (I \otimes D_0)] + \sum_{j=0}^{L-1} (\hat{\pi}_j + \tilde{\pi}_j) (I \otimes D_1) + (\hat{\pi}_L + \tilde{\pi}_L) [(S \otimes I) + (I \otimes D)] \\ & + \sum_{j=1}^L (\hat{\pi}_j + \tilde{\pi}_j) (\mathbf{S}^0 \boldsymbol{\beta} \otimes I) + \hat{\pi}_0(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) + \tilde{\pi}_0[(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) + (I \otimes D_0)] = 0 \end{aligned}$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(S \otimes I) + \sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(\mathbf{S}^0 \boldsymbol{\beta} \otimes I) + \sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(I \otimes D) = 0$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(S + \mathbf{S}^0 \boldsymbol{\beta}) \otimes I + (I \otimes D)] = 0$$

$$\sum_{j=0}^L [\hat{\pi}_j + \tilde{\pi}_j] [(S + \mathbf{S}^0 \boldsymbol{\beta}) \oplus D] = \mathbf{0}.$$

Post-multiplying the last equation by $(\mathbf{e} \otimes I)$:

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(S \otimes I)(\mathbf{e} \otimes I) + (\mathbf{S}^0 \boldsymbol{\beta} \otimes I)(\mathbf{e} \otimes I) + (I \otimes D)(\mathbf{e} \otimes I)] = 0$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(S\mathbf{e} \otimes I) + (\mathbf{S}^0 \boldsymbol{\beta} \mathbf{e} \otimes I) + (\mathbf{e} \otimes D)] = 0$$

By substituting the equation $S\mathbf{e} = -\mathbf{S}^0$,

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(\mathbf{e} \otimes D) = 0 \quad \text{or} \quad \sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(\mathbf{e} \otimes I)D = 0.$$

From the uniqueness of the steady-state vectors of the generator D , the result gives the first equation in (5.8).

Post-multiplying the last equation by $(I \otimes \mathbf{e})$:

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(S \otimes I)(I \otimes \mathbf{e}) + (\mathbf{S}^0 \boldsymbol{\beta} \otimes I)(I \otimes \mathbf{e}) + (I \otimes D)(I \otimes \mathbf{e})] = 0$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(S \otimes \mathbf{e}) + (\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e})] = 0$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j) [(I \otimes \mathbf{e})S + (I \otimes \mathbf{e})\mathbf{S}^0 \boldsymbol{\beta}] = 0$$

Multiplying the above equation by S^{-1} ,

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(I \otimes \mathbf{e}) + \sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(I \otimes \mathbf{e}) \mathbf{S}^0 \beta S^{-1} = 0$$

$$\sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(I \otimes \mathbf{e}) = \sum_{j=0}^L (\hat{\pi}_j + \tilde{\pi}_j)(I \otimes \mathbf{e}) \mathbf{S}^0 \beta (-S)^{-1}.$$

From the uniqueness of the steady-state vectors of the generator $S + \mathbf{S}^0 \beta$ along with the normalizing condition in (5.10), the result gives the second equation in (5.8).

The following theorem establishes the stability condition of the queueing system under study.

Theorem 5.1.2. *The queueing system under study is stable if and only if the following condition is satisfied.*

$$\lambda_2 < p\mu + q \sum_{j=0}^L \tilde{\pi}_j (\mathbf{S}^0 \otimes \mathbf{e}). \quad (5.11)$$

Proof: The queueing system under study with the *QBD* type generator given in (5.3) is stable (see, e.g., Neuts (1981)) if and only if $\xi F_0 \mathbf{e} < \xi F_2 \mathbf{e}$. First note that from the elements of the matrices A_0 , A_1 , A_2 , and F_2 , we see that

$$\xi F_0 \mathbf{e} = 0.5\lambda_2 \quad \text{and} \quad \xi F_2 \mathbf{e} = 0.5[\pi A_2 \mathbf{e} + 2\pi A_3 \mathbf{e}]. \quad (5.12)$$

It is easy to verify that

$$\pi A_2 \mathbf{e} = [q\hat{\pi}_0 + \tilde{\pi}_0 + p \sum_{j=1}^L \hat{\pi}_j](\mathbf{S}^0 \otimes \mathbf{e}), \quad (5.13)$$

$$\pi A_3 \mathbf{e} = p\hat{\pi}_0(\mathbf{S}^0 \otimes \mathbf{e}).$$

Using equations (5.12) and (5.13), the stability condition becomes

$$\lambda_2 < [\hat{\pi}_0 + \tilde{\pi}_0](\mathbf{S}^0 \otimes \mathbf{e}) + p \sum_{j=0}^L \hat{\pi}_j (\mathbf{S}^0 \otimes \mathbf{e}). \quad (5.14)$$

Post-multiplying the last three equations in (5.9) by $\mathbf{e}$ we get

$$\hat{\pi}_0(\mathbf{S}^0 \otimes \mathbf{e}) + \tilde{\pi}_0[(S\mathbf{e} \otimes \mathbf{e}) + (\mathbf{S}^0 \otimes \mathbf{e}) + (\mathbf{e} \otimes D_0\mathbf{e})] = 0$$

$$\sum_{j=1}^{L-1} \tilde{\pi}_{j-1}(\mathbf{e} \otimes D_1\mathbf{e}) + \sum_{j=1}^{L-1} \tilde{\pi}_j[(S\mathbf{e} \otimes \mathbf{e}) + (\mathbf{e} \otimes D_0\mathbf{e})] = 0$$

$$\tilde{\pi}_{L-1}(\mathbf{e} \otimes D_1\mathbf{e}) + \tilde{\pi}_L[(S\mathbf{e} \otimes \mathbf{e}) + (\mathbf{e} \otimes D\mathbf{e})] = 0$$

Now adding these last equations we obtain the equation in (5.15).

$$\hat{\pi}_0(\mathbf{S}^0 \otimes \mathbf{e}) = \sum_{j=1}^L \tilde{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}). \quad (5.15)$$

Substituting (5.15) into (5.14) and using the fact (from Lemma 5.1.1) that $\sum_{j=0}^L [\hat{\pi}_j + \tilde{\pi}_j](\mathbf{S}^0 \otimes \mathbf{e}) = \mu$, the operations are continued as follows.

$$\begin{aligned} \lambda_2 &< \sum_{j=0}^L \tilde{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}) + p \sum_{j=0}^L \hat{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}) \\ \lambda_2 &< p \sum_{j=0}^L \tilde{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}) + q \sum_{j=0}^L \tilde{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}) + p \sum_{j=0}^L \hat{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}) \\ \lambda_2 &< p \sum_{j=0}^L (\tilde{\pi}_j + \hat{\pi}_j)(\mathbf{S}^0 \otimes \mathbf{e}) + q \sum_{j=0}^L \tilde{\pi}_j(\mathbf{S}^0 \otimes \mathbf{e}). \end{aligned}$$

Note: In the case when $p = 1$, the stability condition reduces to $\lambda_2 < \mu$. This is intuitively obvious since for all practical purposes a Type 2 customer present in the system is bound to leave the system upon completion of any type of a service. Thus, the acceptance of Type 1 customers into the system will not increase the traffic load of a system which has only Type 2 arrivals (at a rate of λ_2 and a service rate of μ).

5.1.4 The Steady-state Probability Vector

Let $\mathbf{x} = (\mathbf{x}^*, \mathbf{x}(0), \mathbf{x}(1), \dots)$ denote the steady-state probability vector of $\tilde{Q}$. That is, $\mathbf{x}$ satisfies

$$\mathbf{x} \tilde{Q} = \mathbf{0}, \mathbf{x} \mathbf{e} = 1. \quad (5.16)$$

Note that the vector $\mathbf{x}^*$ of dimension m is such that the k^{th} component of $\mathbf{x}^*$ gives the steady-state probability that the system is idle with the arrival process being in phase k , $1 \leq k \leq m$. The vector $\mathbf{x}(i)$, $i \geq 0$, of dimension $4(L+1)mn$, is further partitioned as $\mathbf{x}(i) = (\hat{\mathbf{x}}(i), \tilde{\mathbf{x}}(i))$, $i \geq 0$. That is,

The vector $\hat{\mathbf{x}}(i)$, of dimension $2(L+1)mn$, is further partitioned into vectors of dimension mn as, $\hat{\mathbf{x}}(i) = (\hat{\mathbf{x}}_{1,0}(i), \dots, \hat{\mathbf{x}}_{1,L}(i), \hat{\mathbf{x}}_{2,0}(i), \dots, \hat{\mathbf{x}}_{2,L}(i))$. Note that $\hat{\mathbf{x}}_{r,s}(i)$, $i \geq 0, 0 \leq s \leq L, r = 1, 2$ gives the steady-state probability vector that there are s Type 1 customers and $2i$ Type 2 customers waiting in the queue with the server busy serving a Type r customer and the arrival and services are in various phases (accounting for the dimension to be mn).

Similarly we partition $\tilde{\mathbf{x}}(i)$, of dimension $2(L+1)mn$, into vectors of dimension mn as, $\tilde{\mathbf{x}}(i) = (\tilde{\mathbf{x}}_{1,0}(i), \dots, \tilde{\mathbf{x}}_{1,L}(i), \tilde{\mathbf{x}}_{2,0}(i), \dots, \tilde{\mathbf{x}}_{2,L}(i))$. The vector, $\tilde{\mathbf{x}}_{r,s}(i)$, $i \geq 0, 0 \leq s \leq L, r = 1, 2$ has similar interpretation except that the number of Type 2 customers in the queue is now given by $2i + 1$.

Under the stability condition given in (5.11) the steady-state probability vector $\mathbf{x}$ is obtained (see, e.g., Neuts (1981)) as follows

$$\mathbf{x}(i) = \mathbf{x}(0) R^i, i \geq 0, \quad (5.17)$$

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

$$R^2 F_2 + R F_1 + F_0 = 0, \quad (5.18)$$

and the vectors, $\mathbf{x}^*$ and $\mathbf{x}(0)$ are obtained by solving

$$\mathbf{x}^* B_1 + \mathbf{x}(0) \tilde{B}_2 = \mathbf{0}, \quad (5.19)$$

$$\mathbf{x}^* (\mathbf{e}'_1 \otimes B_0) + \mathbf{x}(0) [E_1 + R F_2] = \mathbf{0},$$

subject to the normalizing condition

$$\mathbf{x}^* \mathbf{e} + \mathbf{x}(0) (I - R)^{-1} \mathbf{e} = 1. \quad (5.20)$$

The computation of the R matrix can be carried out using a number of well-known methods such as logarithmic reduction and (block) Gauss-Seidel iterative by exploiting the special structure of the coefficient matrices F_0 , F_1 , and F_2 , which are of dimension $4(L + 1)mn$. This is very important especially when one is dealing with large values of L , m and n . For example, one can exploit the structure of R matrix:

$$R = \begin{pmatrix} 0 & 0 \\ R_2 & R_1 \end{pmatrix}. \quad (5.21)$$

The details of other exploitation will be omitted; however, the key steps in the logarithmic reduction algorithm are given in 2.3 and for full details on this we refer the reader to Latouche & Ramaswami (1999).

For use in the sequel we define the $4(L + 1)mn$ -dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ in partitioned form as

$$\mathbf{a} = (\hat{\mathbf{a}}_{1,0}, \dots, \hat{\mathbf{a}}_{1,L}, \hat{\mathbf{a}}_{2,0}, \dots, \hat{\mathbf{a}}_{2,L}, \tilde{\mathbf{a}}_{1,0}, \dots, \tilde{\mathbf{a}}_{1,L}, \tilde{\mathbf{a}}_{2,0}, \dots, \tilde{\mathbf{a}}_{2,L}),$$

$$\mathbf{b} = (\hat{\mathbf{b}}_{1,0}, \dots, \hat{\mathbf{b}}_{1,L}, \hat{\mathbf{b}}_{2,0}, \dots, \hat{\mathbf{b}}_{2,L}, \tilde{\mathbf{b}}_{1,0}, \dots, \tilde{\mathbf{b}}_{1,L}, \tilde{\mathbf{b}}_{2,0}, \dots, \tilde{\mathbf{b}}_{2,L}),$$

where

$$\mathbf{a} = \sum_{i=0}^{\infty} x(i) = \mathbf{x}(0) (I - R)^{-1}, \quad (5.22)$$

$$\mathbf{b} = \sum_{i=1}^{\infty} x(i) = \mathbf{x}(0) R (I - R)^{-1}.$$

We now list some useful results which are intuitively obvious and serve as accuracy checks in our numerical computation.

Lemma 5.1.3. *We have*

$$\sum_{s=0}^L [p(\hat{\mathbf{b}}_{1,s} + \tilde{\mathbf{b}}_{1,s} + \tilde{\mathbf{x}}_{1,s}(0)) + (\hat{\mathbf{a}}_{2,s} + \tilde{\mathbf{a}}_{2,s})](\mathbf{S}^0 \otimes \mathbf{e}) = \lambda_2, \quad (5.23)$$

$$\sum_{s=0}^L [(\hat{\mathbf{a}}_{1,s} + \tilde{\mathbf{a}}_{1,s})](\mathbf{S}^0 \otimes \mathbf{e}) = \lambda_1(1 - P_{lost}),$$

where

$$P_{lost} = \frac{1}{\lambda_1} [(\hat{\mathbf{a}}_{1,L} + \tilde{\mathbf{a}}_{1,L}) + (\hat{\mathbf{a}}_{2,L} + \tilde{\mathbf{a}}_{2,L})](\mathbf{e} \otimes D_1 \mathbf{e}). \quad (5.24)$$

5.1.5 The System Performance Measures

In this section we will list a number of system performance measures of interest along with their expressions. In the sequel, we denote

1. Probability that the system is idle.

$$P_{idle}^{System} = \mathbf{x}^* \mathbf{e}.$$

2. The probability that the server is busy serving Type r customer.

$$P_{busy}^{(r)} = \begin{cases} \sum_{s=0}^L [\hat{\mathbf{a}}_{1,s} + \tilde{\mathbf{a}}_{1,s}] \mathbf{e}, & r = 1, \\ \sum_{s=0}^L [\hat{\mathbf{a}}_{2,s} + \tilde{\mathbf{a}}_{2,s}] \mathbf{e}, & r = 2. \end{cases}$$

3. Probability an arriving Type 1 customer is lost. The probability, P_{lost} , that an arriving customer is lost due to lack of buffer space is given in (5.24).

4. Probability of a Type 2 customer leaving by getting service from the server.

$$P_{T2GS} = \frac{1}{\lambda_2} \sum_{s=0}^L [\hat{\mathbf{a}}_{2,s} + \tilde{\mathbf{a}}_{2,s}](\mathbf{S}^0 \otimes \mathbf{e}).$$

5. Probability of a Type 2 customer leaving with a Type 1 customer.

$$P_{T2LT1} = \frac{p}{\lambda_2} \sum_{s=0}^L [\hat{\mathbf{b}}_{1,s} + \tilde{\mathbf{b}}_{1,s} + \tilde{\mathbf{x}}_{1,s}(0)] (\mathbf{S}^0 \otimes \mathbf{e}).$$

6. Mean number of Type 1 customers in the queue.

$$\mu_{MNQ1} = \sum_{s=1}^L s [(\hat{\mathbf{a}}_{1,s} + \tilde{\mathbf{a}}_{1,s}) + (\hat{\mathbf{a}}_{2,s} + \tilde{\mathbf{a}}_{2,s})] \mathbf{e}.$$

7. Mean number of Type 2 customers in the queue.

$$\mu_{MNQ2} = \sum_{i=0}^{\infty} [2i \hat{x}(i) \mathbf{e} + (2i + 1) \tilde{x}(i) \mathbf{e}]$$

From the equation in (5.17) and the structure of the rate matrix R we get

$$x(i) = [\hat{x}(i), \tilde{x}(i)] = x(0) R^i = [\hat{x}(0), \tilde{x}(0)] \begin{bmatrix} 0 & 0 \\ R_1^{i-1} R_2 & R_1^i \end{bmatrix} = [\tilde{x}(0) R_1^{i-1} R_2, \tilde{x}(0) R_1^i].$$

By substituting $\hat{x}(i)$ and $\tilde{x}(i)$ we get

$$\begin{aligned} \mu_{MNQ2} &= \sum_{i=0}^{\infty} [2i \tilde{x}(0) R_1^{i-1} R_2 \mathbf{e} + (2i + 1) \tilde{x}(0) R_1^i \mathbf{e}] \\ &= \tilde{x}(0) \left[2 \left(\sum_{i=0}^{\infty} i R_1^{i-1} \right) R_2 \mathbf{e} + 2 \left(\sum_{i=0}^{\infty} i R_1^i \right) \mathbf{e} + \left(\sum_{i=0}^{\infty} R_1^i \right) \mathbf{e} + \right] \\ &= \tilde{x}(0) [2(I - R_1)^{-2} R_2 \mathbf{e} + 2(I - R_1)^{-2} R_1 \mathbf{e} + (I - R_1)^{-1} \mathbf{e}] \end{aligned}$$

$$\mu_{MNQ2} = 2\tilde{x}(0)(I - R_1)^{-2} R_2 \mathbf{e} + \tilde{x}(0)(I - R_1)^{-2} (I + R_1) \mathbf{e}.$$

5.1.6 Stationary Waiting Time Distribution of A Type 1 Customer

In this section we will show that the stationary waiting time distribution of an admitted customer at an arrival epoch follows a phase type distribution.

Theorem 5.1.4. *Suppose X denotes the waiting time in the queue of an admitted Type 1*

customer in steady-state. Then X follows a PH-distribution with representation (α, T) of order Ln , where

$$\alpha = (\alpha_1, \dots, \alpha_L), \quad T = \begin{pmatrix} S & & & \\ S^0\beta & S & & \\ & S^0\beta & S & \\ & & \ddots & \ddots \\ & & & S^0\beta & S \end{pmatrix}, \quad (5.25)$$

where the vector $\alpha_k, 1 \leq k \leq L$, is given by

$$\alpha_k = \frac{1}{\lambda_1(1 - P_{lost})} [\hat{a}_{1,k-1} + \hat{a}_{2,k-1} + \tilde{a}_{1,k-1} + \tilde{a}_{2,k-1}] (I \otimes D_1 \mathbf{e}), \quad 1 \leq k \leq L. \quad (5.26)$$

Proof: First note that there is no need to keep track of the phase of the arrival process soon after a tagged Type 1 customer is admitted into the system. Secondly, due to the nature of (identical) services for Type 1 and Type 2 customers, there is no need to keep track of the type of customer in service. Now observe that the j^{th} component of the n -dimensional vector $\alpha_k, 1 \leq k \leq L$, gives the probability that an admitted Type 1 customer finds $(k - 1)$ Type 1 customers waiting in the queue and the current service phase is in $j, 1 \leq j \leq n$. Since Type 1 customers have non-preemptive priority over Type 2 customers, there is no need to keep track of the number of Type 2 customers in the queue as well as any future arrivals (of any type). The stated result follows immediately from the law of total probability. The probability that the waiting time in the queue is zero is given by $1 - \alpha \mathbf{e}$.

Corollary: The mean, $\mu_{WQ}^{(1)}$, waiting time in the queue of an admitted Type 1 customer is given by

$$\mu_{WQ}^{(1)} = \alpha(-T)^{-1} \mathbf{e} = \sum_{k=1}^L \alpha_k (-S)^{-1} \mathbf{e} + \frac{1}{\mu} \sum_{k=2}^L (k-1) \alpha_k \mathbf{e}. \quad (5.27)$$

Proof: Follows immediately using the special form of the matrix T given in (5.25).

NOTE: One can also use the Little's law which in our model is given by $\mu_{MNQ1} = \lambda_1(1 - P_{lost})\mu_{WQ}^{(1)}$, and use (5.27) as another internal accuracy check in numerical computation.

5.1.7 Numerical Examples

In this section we discuss the qualitative aspects of the crowdsourcing queueing model under consideration through illustrative numerical examples. First we note that it is very important to verify the correctness and the accuracy of the code written to compute various system performance measures. This is accomplished through verifying a number of results. For example, one can use the results of Lemmas 5.1.1 and 5.1.3 as well as the Little's law for Type 1 customers. We also obtained the numerical solution for the Poisson arrivals in its simple form. Next, we implemented the general algorithm, but using the following *MAP* representation: Let D_0 be an irreducible, stable matrix with eigenvalue of maximum real part $-\eta < 0$. Let κ denote the corresponding left eigenvector, normalized by $\kappa e = 1$. Taking $D_1 = -pD_0 e \kappa$, the *MAP* representation reduces to the Poisson arrival process with intensity rate η Neuts (1995). The general algorithm does not utilize this fact in any manner, but the numerical results were identical.

For the arrival process, we consider the following five sets of values for D_0 and D_1 as follows.

1. Erlang distribution (*ERLA*):

$$D_0 = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}$$

2. Exponential distribution (*EXPA*):

$$D_0 = \begin{pmatrix} -1 \end{pmatrix}, D_1 = \begin{pmatrix} 1 \end{pmatrix}$$

3. Hyperexponential distribution (*HEXA*):

$$D_0 = \begin{pmatrix} -1.90 & 0 \\ 0 & -0.19 \end{pmatrix}, D_1 = \begin{pmatrix} 1.71 & 0.19 \\ 0.171 & 0.019 \end{pmatrix}$$

4. *MAP* with negative correlation (*MNCA*):

$$D_0 = \begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.9922 \\ 223.4925 & 0 & 2.2575 \end{pmatrix}$$

5. *MAP* with positive correlation (*MPCA*):

$$D_0 = \begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.9922 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{pmatrix}.$$

The above *MAP* processes will be normalized so as to have a specific arrival rate. However, these are qualitatively different in that they have different variance and correlation structure. The first three arrival processes, namely, *ERLA*, *EXPA*, and *HEXA*, have zero correlation for two successive inter-arrival times. The arrival processes labeled *MNCA* and *MPCA*, respectively, have negative and positive correlation for two successive inter-arrival times with values -0.4889 and 0.4889. The ratio of the standard deviation of the inter-arrival times of these five arrival processes with respect to *ERLA* are, respectively, 1, 1.41421, 3.17451, 1.99336, and 1.99336.

For the service times (β, S) we consider the following three phase-type distributions. These distributions will be normalized so as to arrive at a desired value for μ .

1. Erlang distribution (*ERLS*):

$$\beta = (1, 0), \quad S = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}.$$

2. Exponential distribution (*EXPS*):

$$\beta = (1), \quad S = \begin{pmatrix} -1 \end{pmatrix}.$$

3. Hyperexponential distribution (*HEXS*):

$$\beta = (0.9, 0.1), \quad S = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix}.$$

In the following we define ρ as

$$\rho = \frac{\lambda_2}{\left[p\mu + q \sum_{j=0}^L \tilde{\pi}_j (\mathbf{S}^0 \otimes \mathbf{e}) \right]}.$$

Whenever we need to fix a specific value for ρ , we can vary any of the system parameters, $\lambda_1, \lambda_2, \mu, L$, and p to arrive at that value. However, the vector $\tilde{\pi}_j, 0 \leq j \leq L$, is independent of λ_2 but has a complex dependence structure on λ_1, μ, L , and p . Thus, to set ρ to have a specific values it will be easy to adjust λ_2 as opposed to changing any of the other parameters. Thus, in the following we will adapt this approach.

Example 5.1: The purpose of this example is to see how the measures, $P_{busy}^{(r)}, r = 1, 2$, and P_{T2LT1} behave under various scenarios to see the effectiveness of crowdsourcing. We fix $\lambda_1 = 1, \mu = 1.1, \rho = 0.99$; vary p and L . As mentioned earlier λ_2 will vary based on the values of ρ, L and p . We display the busy probabilities in Figure 5.2 and P_{T2LT1} in Figure 5.3 under various scenarios.

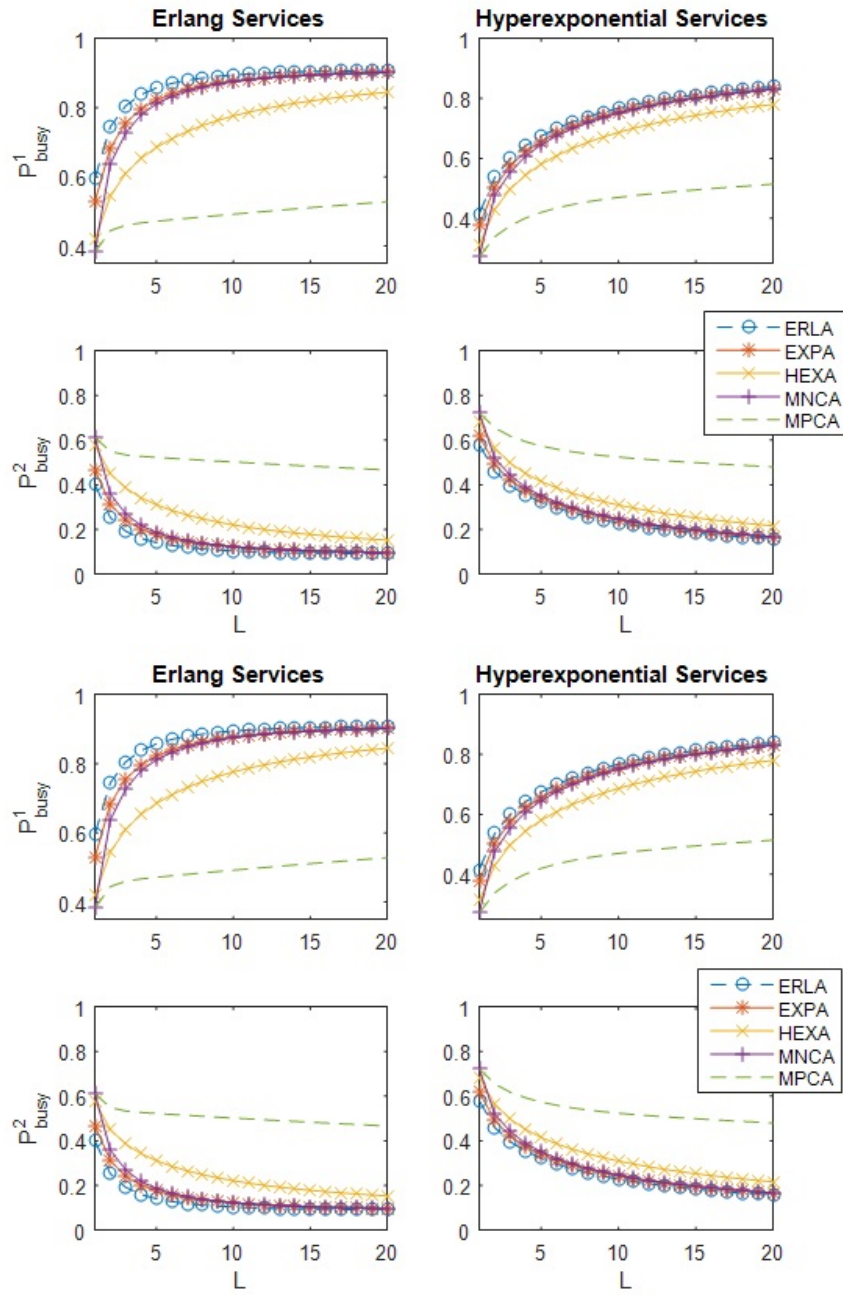


Figure 5.2 Probability that server is busy, above for $p = 0.5$ and below for $p = 1$

A quick look at these two graphs reveal the significant benefit in using Type 1 customers to help serving Type 2 customers. Also these graphs reveal the impact of variation present in both the arrival and service times as well as correlation (if any) present in the times between two successive arrivals. Some general observations are listed below.

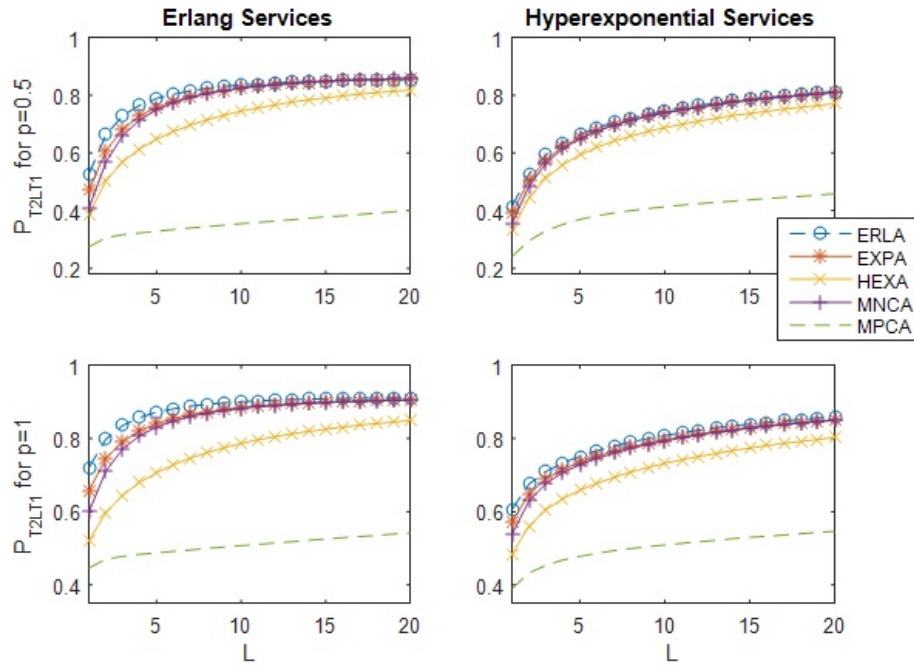


Figure 5.3 Probability of a Type 2 customer leaving with a Type 1 customer

- First note that the busy probabilities vary based on the type of arrival and service processes as is to be expected; however, in all cases, they appear to be insignificant to p (See Figure 5.2).
- It appears that P_{T2LT1} is more sensitive to the service time distribution for small values of L .
- While the busy probabilities are insignificant to p , the significant effect of p on P_{T2LT1} can be seen in Figure 5.3. This is an interesting and important observation for the following reason. Just because the busy probabilities are insignificant to p does not imply that p does not play a role in the model. In fact, by looking at Figure 5.3, it is obvious that the fraction of Type 2 customers served by Type 1 customers is very significant. This would not have been discovered by looking only at the busy probabilities.

Example 5.2: The effect of crowdsourcing is studied again in this example by comparing the model with $p = 0$ that corresponds to having two independent arrival processes governed by a single server on a non-preemptive priority basis to the

crowdsourcing model (i.e., $p > 0$). In the latter case there is a possibility for Type 1 customers to act as servers for Type 2. Towards this end we look at the ratio of the mean waiting time in the queue of a Type 2 customer. First note that the mean waiting time in the queue of a Type 2 customer, $\mu_{WTQ2}(p = 0)$ when $p = 0$ is obviously larger than that of, $\mu_{WTQ2}(p > 0)$, for the case $p > 0$. Thus, the ratio, $\frac{\mu_{WTQ2}(p>0)}{\mu_{WTQ2}(p=0)}$, will be in the interval $(0,1)$. This ratio, by fixing $\lambda_1 = 1, \mu = 1.1$; varying p on the interval $(0, 1]$; L on $[1, 20]$, for two values of $\rho = 0.1, 0.99$, under various scenarios, is plotted in Figure 5.4 and Figure 5.5 displayed below. Some key observations are summarized as follows. First observe that the smaller the ratio the better the system in terms of the mean waiting time in the queue of a Type 2 customers.

- Even when the system has a very low saturation level (i.e., when ρ is small - see Figure 5.4) we see a marked improvement in having some fraction of Type 1 customers acting as servers for Type 2 customers.
- For very small values of p (i.e., when $0 \leq p < 0.2$) we do not see significant drop in this ratio as L is increased; however, when $p \geq 0.2$, we notice a significant rate of decrease as L increases. The rate of decrease is much higher when the service times have a higher variability like hyperexponential one.

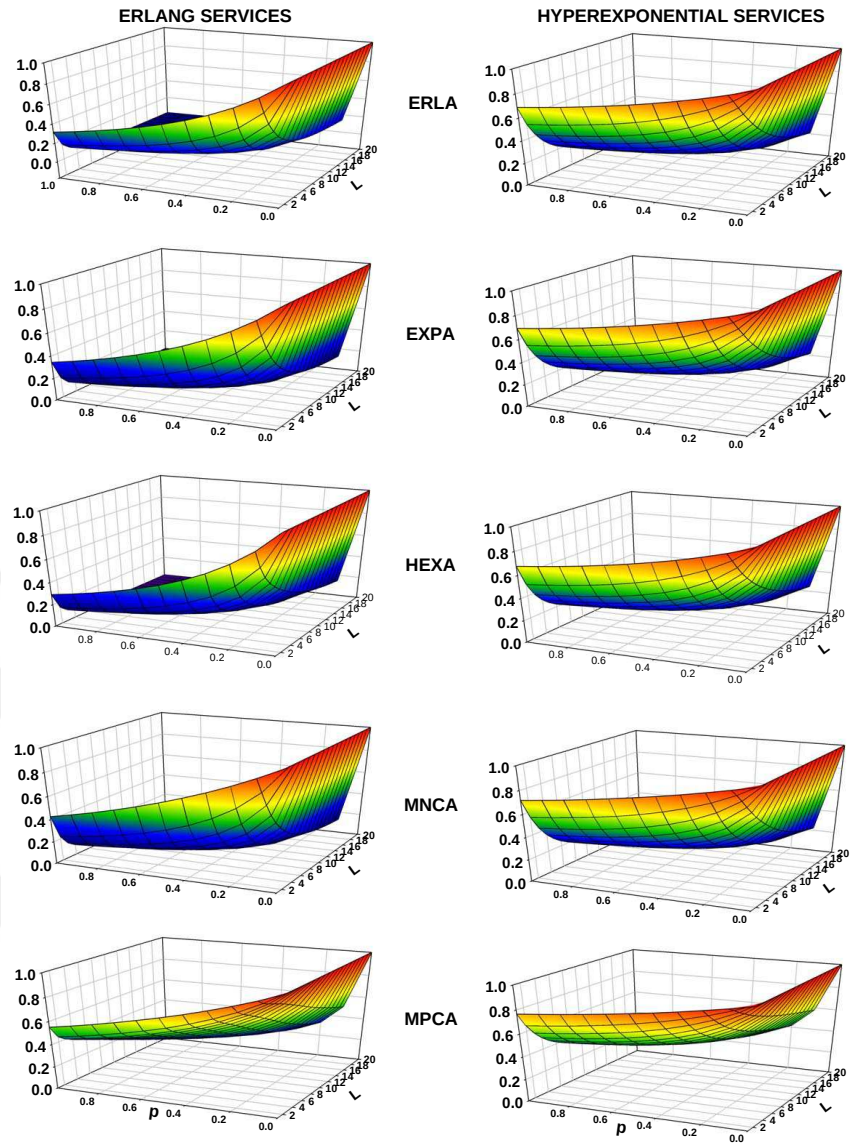


Figure 5.4 Ratios for mean waiting time of Type 2 customers for $\rho = 0.10$

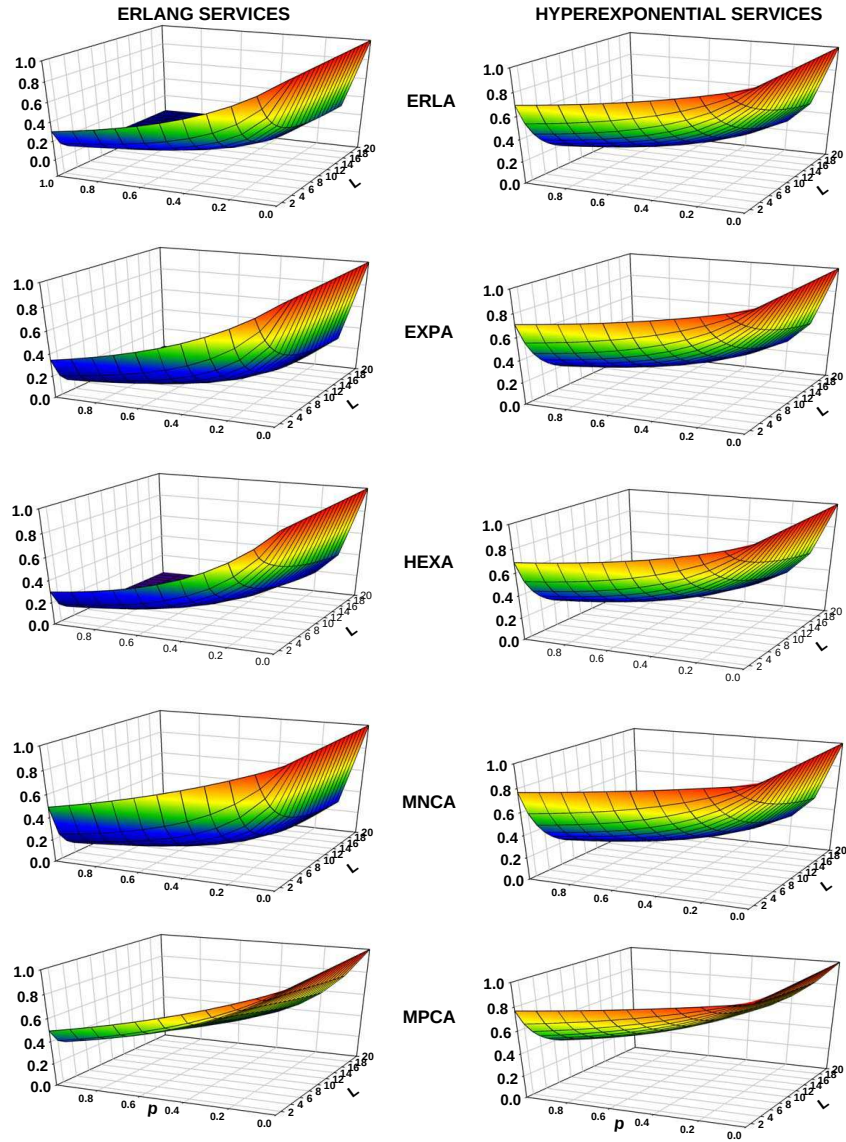


Figure 5.5 Ratios for mean waiting time of Type 2 customers for $\rho = 0.99$

Example 5.3: In this example we will discuss an optimization problem of interest. Suppose that

- r = revenue per processed customer per unit of time
- c_1 = cost per unit of time of holding a Type 1 customer in the queue
- c_2 = cost per unit of time of holding a Type 2 customer in the queue
- c_3 = cost per unit of time of using a Type 1 customer to serve a Type 2 customer

The goal of the optimization problem is to search for optimum values of p and L for which the following objective function, ETP , giving the expected total profit per unit of time, is maximized.

$$ETP = \lambda_1(1 - P_{loss}) + \lambda_2 - c_1\mu_{MNQ1} - c_2\mu_{MNQ2} - c_3\lambda_2P_{T2LT1}. \quad (5.28)$$

We fix $\lambda_1 = 1$, $\mu = 1.1$, and $\rho = 0.99$. The following table displays the optimum, (L^*, p^*) , under various scenarios.

Table 5.1 Optimum values of (L^*, p^*) for various scenarios

PH	(c_1, c_2, c_3)	$ERLA$	$EXPA$	$HEXA$	$MNCA$	$MPCA$
$ERLS$	(0.05, 0.001, 0.1)	(5,1)	(5,1)	(6,1)	(6,1)	(11,1)
	(0.05, 0.001, 0.2)	(4,1)	(5,1)	(5,1)	(5,1)	(9,1)
	(0.1, 0.001, 0.1)	(3,1)	(3,1)	(3,1)	(4,1)	(5,1)
	(0.1, 0.001, 0.2)	(3,1)	(3,1)	(3,1)	(4,1)	(5,1)
$HEXS$	(0.05, 0.001, 0.1)	(6,1)	(7,1)	(7,1)	(7,1)	(10,1)
	(0.05, 0.001, 0.2)	(6,1)	(6,1)	(6,1)	(6,1)	(9,1)
	(0.1, 0.001, 0.1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)
	(0.1, 0.001, 0.2)	(3,1)	(3,1)	(3,1)	(3,1)	(3,1)

A quick look at the table reveals the following observations.

1. It is not surprising to see that all optimum occurs when $p = 1$ indicating that all Type 1 customers should offer services to Type 2 customers provided there is at least one Type 2 customer waiting at the time of the completion of a service of a Type 1 customer.
2. $MPCA$ arrivals appear to need a larger L value compared to the other arrivals.

Since in Table 5.1 we saw that the optimum expected total profit requires all Type 1 customers to provide (when eligible criteria are met) services to Type 2, one question

that naturally arises is what happens if not all Type 1 customers are not willing to serve Type 2 customers? That is to say, what should be L^* for a given p , $0 < p < 1$? Towards this end, we looked at the case when $p = 0.5$. That is, we looked at the case where only 50% of Type 1 customers meeting the eligibility criteria are actually willing to serve. In this case, the following Table 5.2 summarizes the optimum L^* . As is to be expected, L^* values are below the corresponding ones for the case when $p = 1.0$. Also, in Table 5.3 we display the reduction in the percentages of ETP when going from the optimal point when $p = 1$ to $p = 0.5$. The table of this type will help the business managers to figure out the appropriate ranges for the profits and costs so as to attain a given profit margin.

Table 5.2 Optimum values of L^* for various scenarios when $p = 0.5$

PH	(c_1, c_2, c_3)	$ERLA$	$EXPA$	$HEXA$	$MNCA$	$MPCA$
$ERLS$	(0.05, 0.001, 0.1)	3	3	3	4	4
	(0.05, 0.001, 0.2)	3	3	3	3	3
	(0.1, 0.001, 0.1)	2	2	2	2	2
	(0.1, 0.001, 0.2)	2	2	2	2	2
$HEXS$	(0.05, 0.001, 0.1)	2	3	3	3	2
	(0.05, 0.001, 0.2)	2	2	2	3	2
	(0.1, 0.001, 0.1)	2	2	2	2	1
	(0.1, 0.001, 0.2)	1	2	1	2	1

Table 5.3 Percentages of reduction in ETP at optimum values

PH	(c_1, c_2, c_3)	$ERLA$	$EXPA$	$HEXA$	$MNCA$	$MPCA$
$ERLS$	(0.05, 0.001, 0.1)	24.64	24.40	22.99	24.60	17.56
	(0.05, 0.001, 0.2)	23.07	22.78	21.27	22.92	16.21
	(0.1, 0.001, 0.1)	24.08	23.46	21.11	23.85	17.25
	(0.1, 0.001, 0.2)	22.45	21.80	19.53	21.90	16.02
$HEXS$	(0.05, 0.001, 0.1)	23.51	23.61	23.05	23.88	19.72
	(0.05, 0.001, 0.2)	21.20	21.51	20.96	21.68	17.68
	(0.1, 0.001, 0.1)	21.11	20.93	19.93	21.11	15.87
	(0.1, 0.001, 0.2)	18.85	19.00	17.97	18.86	13.80

5.2 $MAP/PH/c$ Queueing Model with Crowdsourcing

In Section 5.2.1 the description of the model $MAP/PH/c$ is given. In Section 5.2.2, the validation of the simulated model is represented. Finally, this Section is finished by the discussion of the simulated results in Section 5.2.3.

5.2.1 Model Description

In this section we consider $MAP/PH/c$ queueing model with crowdsourcing. The model assumptions are as outlined in Section 5.1 except that here we assume that there are c servers in the system.

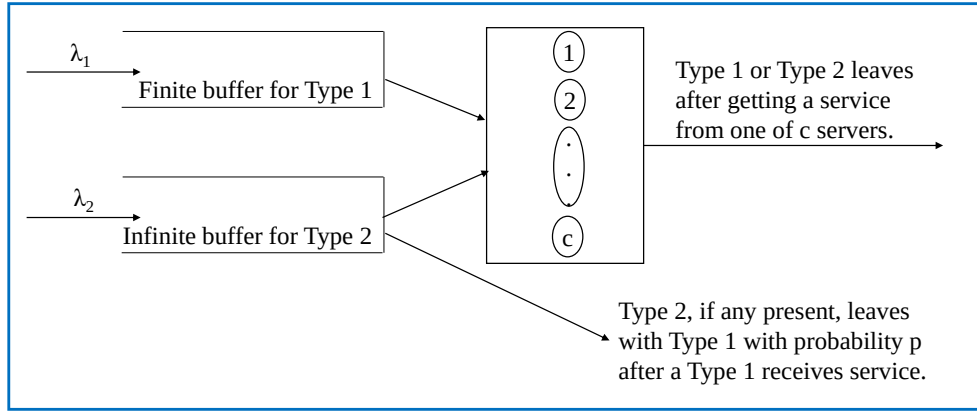


Figure 5.6 Diagram of the multi-serve queue model with crowdsourcing

This model can be studied as a Markov process by keeping track of (a) the number, $K_1(t)$, of Type 2 customers in the queue; (b) the number, $K_2(t)$, Type 1 customers in the queue; (c) the number, $K_3(t)$, of servers busy serving Type 1 customers; (d) the number, $K_4(t)$, of servers busy with Type 2 customers; (e) the phase, $J_r(t)$, of the r^{th} server, for $1 \leq r \leq c$, and (f) the phase, $J(t)$, of the arrival process at time t .

The process $\{(K_1(t), K_2(t), K_3(t), K_4(t), J_1(t), \dots, J_{K_3(t)+K_4(t)}(t), J(t)) : t \geq 0\}$ is

a CTMC with state space given by

$$\Omega = \{(i_1, i_2, i_3, i_4, j_1, \dots, j_c, k) : i_1 \geq 0, 0 \leq i_2 \leq L, 0 \leq i_3 \leq c, 0 \leq i_4 \leq c, \\ 0 \leq i_3 + i_4 \leq c, 0 \leq j_r \leq n, 1 \leq r \leq c, 1 \leq k \leq m\}.$$

Note that we take $J_r = 0$ when the r^{th} server is idle. The generator of this Markov process can be set up with the help of Kronecker products and sums of matrices. However, it is clear that the analysis of this model analytically requires a large state space to account for all the states described above. However, our goal in this section is to see how the impact of introducing *crowdsourcing* in the context of multi-server queueing system with *MAP* arrivals through simulation. The rest of the section is based on simulating the *crowdsourcing* queueing model described here with the help of ARENA Kelton et al. (2010). Unless otherwise mentioned, we ran our simulation models using 3 replications and for 1,000,000 units (which in our case is minutes) for each replicate.

5.2.2 Validation of The Simulated Model

In any simulation work, it is important to validate the simulated model by comparing the results with any known analytical results. Hence, we will do that in this section. The only cases for which analytical results are available for the model under study are for $M/M/c$ in Chakravarthy & Dudin (2017) and for $MAP/PH/1$ in Section 5.1.

i. M/M/c crowdsourcing

In Chakravarthy & Dudin (2017), the authors studied an $M/M/c$ type queueing models with *crowdsourcing*. We will compare our simulated results against their numerical ones, which were generated through analytical study by employing matrix-analytic methods. Specifically, we fix $\lambda_1 = 1, \mu = 1.1, L = 10$, and vary other parameters as follows: $p = 0, 0.5, 1.0, \rho = 0.8, 0.99, c = 1, 2, 5$, and $c = 10$. The values of λ_2 , which depend on the (fixed) value of ρ (see Chakravarthy & Dudin

(2017)), are displayed in Table 5.4 below. The error percentage, which is calculated as $\left\{ \frac{|Analytical - Simulated|}{Analytical} 100\% \right\}$, for various scenarios are displayed in Table 5.5.

Table 5.4 Values of λ_2 for various scenarios for $M/M/c$

c	$\rho = 0.80$			$\rho = 0.99$		
	$p = 0.0$	$p = 0.50$	$p = 1.0$	$p = 0.0$	$p = 0.50$	$p = 1.0$
1	0.1232	0.5016	0.8800	0.1524	0.6207	1.0890
2	0.9602	1.3601	1.7600	1.1882	1.6831	2.1780
5	3.6000	4.0000	4.4000	4.4550	4.9500	5.4450
10	8.0000	8.4000	8.8000	9.9000	10.3950	10.8900

By looking at Table 5.5 we notice that our simulated results are very close to the ones obtained using analytical results presented in Chakravarthy & Dudin (2017) for all except in a couple of scenarios. For these scenarios we ran the simulation again but with 10,000,000 minutes and 3 replicates and found the error percentages for these cases to drop significantly.

Table 5.5 Error percentages of analytical and simulated models for $M/M/c$ crowdsourcing

c	ρ p	$MN1Q$		$MN2Q$		$PLOS$		$PT2L1$	
		0.8	0.99	0.8	0.99	0.8	0.99	0.8	0.99
1	0	0.13%	0.20%	0.66%	7.24%	0.00%	0.00%	0.00%	0.00%
	0.5	0.42%	0.22%	1.04%	11.98%	0.39%	0.93%	0.11%	0.13%
	1	0.30%	0.16%	0.10%	22.47%	1.16%	0.00%	0.10%	0.07%
2	0	0.10%	0.05%	0.51%	7.05%	0.00%	0.00%	0.00%	0.00%
	0.5	0.17%	0.07%	0.88%	8.68%	0.00%	0.00%	0.00%	0.03%
	1	0.09%	0.00%	0.07%	3.05%	0.00%	0.00%	0.02%	0.02%
5	0	0.22%	0.18%	0.28%	0.80%	0.00%	0.00%	0.00%	0.00%
	0.5	0.07%	0.00%	0.43%	7.02%	0.00%	0.00%	0.12%	0.10%
	1	0.07%	0.05%	0.20%	2.77%	0.00%	0.00%	0.07%	0.11%
10	0	0.22%	0.10%	0.19%	1.16%	0.00%	0.00%	0.00%	0.00%
	0.5	0.00%	0.00%	0.12%	0.59%	0.00%	0.00%	0.00%	0.21%
	1	0.20%	0.00%	0.26%	1.00%	0.00%	0.00%	0.35%	0.00%

ii. MAP/PH/1 crowdsourcing

In Section 5.1, we studied *MAP/PH/1 crowdsourcing* queue using matrix-analytic method and presented some illustrative numerical examples to bring out the qualitative nature of the model. Here we will use this (analytical) approach to validate our simulated model by looking at different combinations of arrival and service processes. For the arrival process and the service times, we consider five sets of values for D_0 and D_1 and three *PH*-distributions as given in *MAP/PH/1* model.

To validate our simulated model for this case, we fix $\lambda_1 = 1$, $\mu = 1.1$, and $L = 10$ and vary other parameters as follows: $p = 0, 0.5, 1.0$, and $\rho = 0.8, 0.99$. The values of λ_2 are displayed in the Table 5.6 as ρ is fixed to be a certain value.

Table 5.6 Values of λ_2 for various scenarios for *MAP/PH/1*

ρ	p <i>MAP</i>	<i>ERLS</i>			<i>EXPS</i>			<i>HEXS</i>		
		0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
0.8	<i>ERLA</i>	0.093	0.487	0.880	0.108	0.494	0.880	0.205	0.543	0.880
	<i>EXPA</i>	0.108	0.494	0.880	0.123	0.502	0.880	0.217	0.548	0.880
	<i>HEXA</i>	0.197	0.539	0.880	0.210	0.545	0.880	0.276	0.578	0.880
	<i>MNCA</i>	0.111	0.495	0.880	0.126	0.503	0.880	0.220	0.550	0.880
	<i>MPCA</i>	0.447	0.664	0.880	0.448	0.664	0.880	0.467	0.673	0.880
0.99	<i>ERLA</i>	0.115	0.602	1.089	0.133	0.611	1.089	0.254	0.672	1.089
	<i>EXPA</i>	0.134	0.611	1.089	0.152	0.621	1.089	0.268	0.679	1.089
	<i>HEXA</i>	0.244	0.667	1.089	0.260	0.674	1.089	0.342	0.715	1.089
	<i>MNCA</i>	0.137	0.613	1.089	0.156	0.623	1.089	0.273	0.681	1.089
	<i>MPCA</i>	0.553	0.821	1.089	0.555	0.822	1.089	0.578	0.833	1.089

In this case, we ran our simulation model using 3 replications and for 5,000,000 minutes for each replicate. Due to a large number of combinations we will display in Table 5.7 the error percentages for only one measure, *MN2Q*, which appear to have a larger error percentage compared to the others. For example, we noticed when

looking at all scenarios considered in this paper, the maximum error percentages for $PLOS$, $PT2L1$, and $MN1Q$, respectively, to be 0.80%, 0.17%, and 0.50%. In the case of the measure $MN2Q$ while for most scenarios the error percentages are reasonable, there are some cases for which this percentage is large. When we simulated the model again but with 10,000,000 minutes and 3 replicates these errors reduced significantly. Further, the 95% confidence intervals included the true analytical means in almost all cases. Only for this measure the simulation need to be run for a large period of time for some scenarios. In any case, the analytical and simulation models appear to be close enough validate the simulated model for $MAP/PH/1$ case.

Table 5.7 Error percentages of analytical and simulated models for $MAP/PH/1$ crowdsourcing

ρ	p MAP	$ERLS$			$EXPS$			$HEXS$		
		0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
0.8	$ERLA$	0.34	0.09	0.0	2.35	0.16	0.29	0.04	0.26	0.20
	$EXPA$	1.34	0.53	0.07	0.02	0.38	0.12	0.36	1.33	0.14
	$HEXA$	0.71	0.15	0.20	0.13	0.15	0.32	0.52	1.00	0.63
	$MNCA$	0.92	0.53	0.25	0.96	0.13	0.27	0.34	0.14	0.53
	$MPCA$	0.31	0.53	0.08	0.94	0.15	0.14	1.07	0.90	0.35
0.99	$ERLA$	13.41	3.35	3.84	22.47	5.11	2.61	11.91	7.45	7.80
	$EXPA$	4.62	0.55	6.48	6.92	2.65	3.90	15.82	0.50	3.31
	$HEXA$	5.13	4.63	5.48	9.75	9.36	0.65	15.39	16.17	7.04
	$MNCA$	12.12	5.09	1.44	14.27	18.58	5.15	12.48	4.48	7.40
	$MPCA$	0.62	1.53	1.42	6.32	3.35	1.17	24.60	4.90	4.60

5.2.3 Numerical Examples

Having validated our simulated *crowdsourcing* model for known cases in the previous sections, we will now present a few illustrate examples to bring out qualitative nature of $MAP/PH/c$ *crowdsourcing* model under study. We will discuss three examples here. In the sequel, we let Y denote the waiting time in the system of a Type 2 customer. We consider the following the measures listed .

- The probability, $PLOS$, that an arriving Type 1 customer is lost due to the buffer being full.
- The probability, $PT2L1$, that a Type 2 customer leaves (served) with a Type 1 customer.
- The mean, $MNQ1$, number of Type 1 customers waiting time in the queue.
- The mean, $MN2Q$, number of Type 2 customers waiting time in the queue.
- The probability, $PIDL$, the system is idle.
- The mean, $MWTS_1$, waiting time in the system of Type 1 customer.
- The mean, $MWTS_2$, waiting time in the system of Type 2 customer.
- The fraction, $FATH$, of Type 2 customers whose waiting time in the system exceeds r , $r \geq 2$, times the average service time by one of the system servers. That is, $P(Y > \frac{r}{\mu})$, $r \geq 2$. Since there is no analytical expression available for the measure, $FATH$, dealing with a specific tail probability of the waiting time in the system of a Type 2 customer, we used the simulated result instead. It should be pointed out that one can compute algorithmically the tail probability for classical single-server model, $MAP/PH/1$ using the matrix-analytic methods Neuts (1981) but for a multi-server case it is highly complicated and hence we resort to simulation only for this particular measure.

The effect of *crowdsourcing* in the context of $MAP/PH/1$ was studied analytically and observed the marked improvements in such a system. The purpose of our next example is to investigate the level of such effect $MAP/PH/c$ case using simulated results. Similar to $MAP/PH/1$ we will look at the ratio of a few of the measures under study here. The ratio, $\frac{\eta(p>0)}{\eta(p=0)}$, will be of interest for a given measure η .

Example 5.4: The effect of *crowdsourcing* is studied in this example by comparing the models: (a) $p = 0$ that corresponds to having two independent arrival processes and (b) $p > 0$ that corresponds to the *crowdsourcing* model in the context of a

multi-server system. In the latter case there is a possibility for Type 1 customers to act as servers for Type 2. We fix $\lambda_1 = 1, \lambda_2 = 2, \mu = 1.1, c = 3, L = 10$ and vary p on the interval $(0, 1]$ under different combinations of arrival and service distributions. In Figure 5.7 below we display the ratios of all but *PBUS1* since the ratio for *PBUS1* for all scenarios are almost 1. Some key observations are summarized as follows. First observe that the smaller the ratio the better the system in terms of all measures except *PIDL* in which case it should be the larger the better. Having more idle time for the system will enable the management to use that time for other activities without having to lower the quality of service provided to the customers.

In Figure 5.7, we display the ratios of the key measures under different scenarios for *ERLS* and *HEXS* services. Note that for lack of space we display only selected combinations; however, our observations summarized below are valid for other combinations not displayed in this figure.

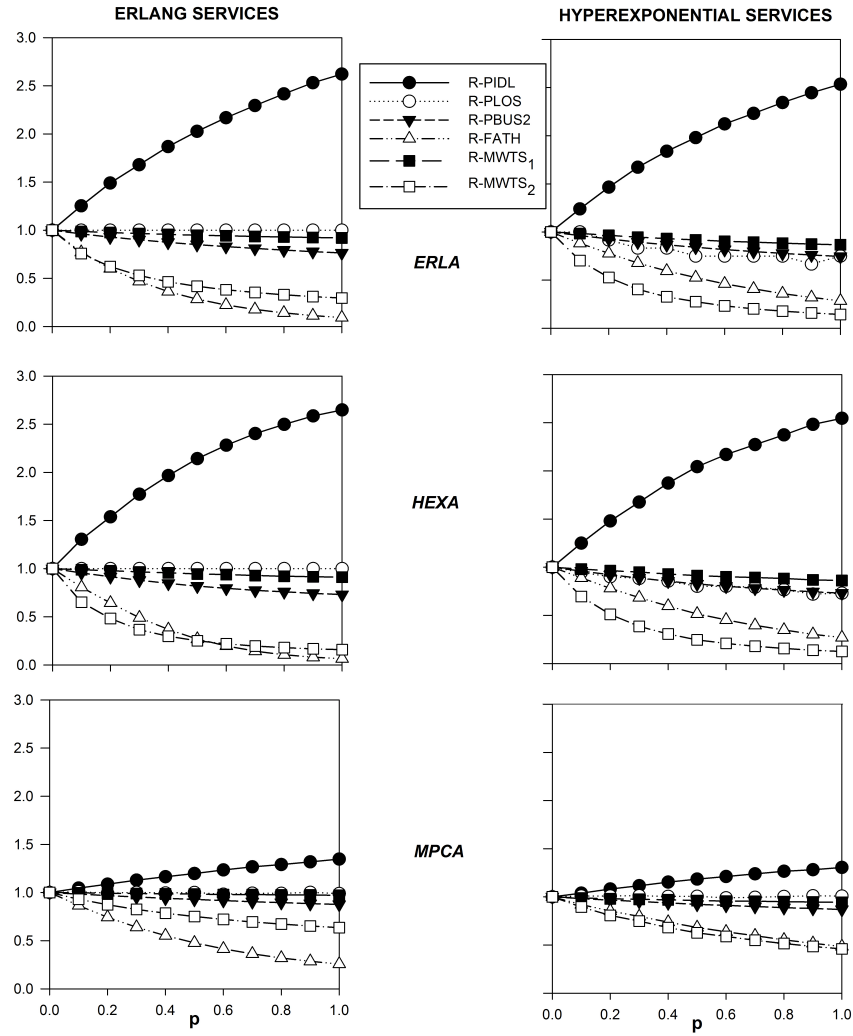


Figure 5.7 Ratios of various measures under different scenarios with *ERLS* and *HEXS* services

A quick look at these figures reveal the following interesting and important observations.

- The ratio for the measure, *PIDL*, is greater than 1 and increases as p is increased indicating that the server becomes idle more often when $p > 0$ when compared to that of $p = 0$. This is true for all combinations of arrivals and services.
- The ratios of all other measures are less than 1 and decreases as p is increased. The rate of decrease depends on the type of arrivals and services.

- The ratio for $MWTS_2$ decreases significantly as p increases. This is true for all cases. This is very important from both management as well as customers' points of view.
- The ratio for $MWTS_1$ decreases as p increases but not as significantly as compared to $MWTS_2$. This is somewhat surprising since one would expect the ratio to be close to 1 since Type 1 service is not affected by the value of p due to non-preemptive nature of services. However, as p increases, Type 1 customers have a higher probability of serving Type 2 customers resulting in a relatively fewer Type 2 customers to be served by one of the system servers and hence a reduction in the mean waiting time in the system.
- The ratio for $FATH$ is decreasing at a significantly at a higher rate as p increases (for all scenarios) which is again very important from both management as well as customers' points of view and can also be used by the management to guarantee some kind of a guarantee on the service times of Type 2 customers.
- The above observations shows the significant advantage in introducing this type of variants, namely, *crowdsourcing*, to the classical queueing models.

In Chakravarthy & Dudin (2017) it was shown that even in the case of small p there is a significant advantage in considering *crowdsourcing* by offering more traffic load through increasing the rate of Type 2 customers into the system. The rate of increase in the offered load to the system is much higher for small values of c . Here, we will investigate a similar advantage from a different point of view by considering the cases when $L = 0$ and $L = 1$. In the former case Type 1 customers are allowed only when at least one server is idle. Thus, the maximum number of Type 1 customers that can be present at any time in the system is c and $c + 1$, respectively.

Example 5.5: This example is very similar to Example 5.4 except that we look at the cases when (a) $L = 0$ and (b) $L = 1$. Note that in these two cases the model can be considered as a slight variation of $M/PH/c$ model since Type 2 arrivals arrive to a multi-server system with phase type arrivals and occasionally Type 1 customers are

allowed to enter into the system. Hence, it will be interesting to see how having only a small number of Type 1 customers, namely, c and $c + 1$ when $L = 0$ and $L = 1$, respectively, at any given time will have an impact on the selected system performance measures. Towards this end, we will fix $\lambda_1 = 1, \lambda_2 = 2, \mu = 1.1, c = 3$, and vary p on the interval $(0, 1]$ under different combinations of arrival and service distributions. Note that the queue is stable for all combinations under these values. In order to properly compare, we now look at the ratio $\frac{\zeta(L=1)}{\zeta(L=0)}$, where $\zeta(L = r) = \frac{\eta(p>0, L=r)}{\eta(p=0, L=r)}, r = 0, 1$. In Figure 5.8 below we display the ratios $\frac{\zeta(L=1)}{\zeta(L=0)}$ for selected measures and for representative scenarios. First observe that the smaller the ratio the better the system with $L = 1$ as compared to $L = 0$ in terms of all measures except $PIDL$ in which case it should be the larger the better. Having more idle time for the system will enable the management to use that time for other activities without having to lower the quality of service provided to the customers.

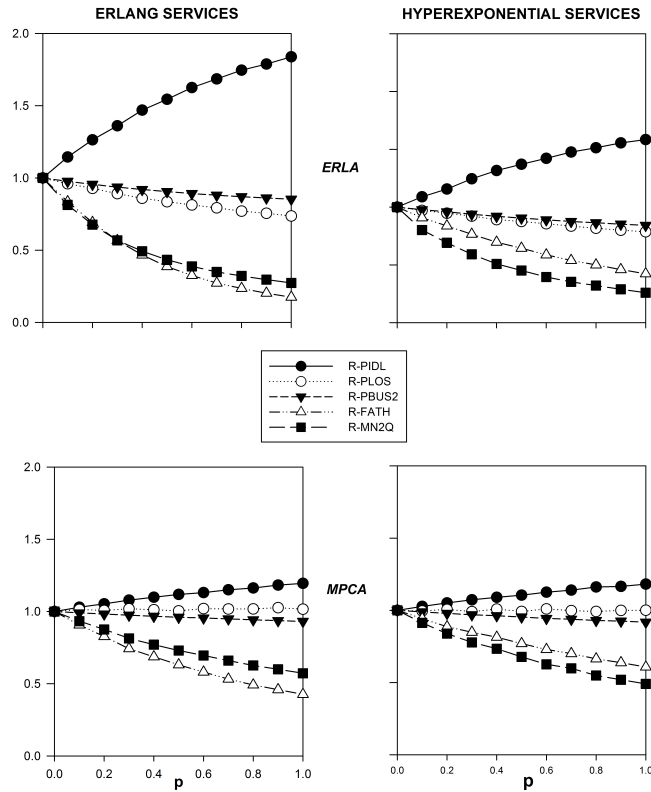


Figure 5.8 Comparison of the ratio, $\frac{\zeta(L=1)}{\zeta(L=0)}$, under different scenarios with *ERLS* and *HEXS* services

A quick look at this Figure reveals a very significant advantage of having an extra Type 1 customers in the system. This is the case for all scenarios (even the ones not displayed here for lack of space) and for all values of p . So, having Type 1 customers even if they are willing to offer services rarely plays a significant role in *crowdsourcing* applications.

In the next example, we try to find the optimum c^* such that the proportion of Type 2 customers whose waiting time in the system exceeds $r, r \geq 2$, times the average service time by one of the system servers does not exceed a pre-determined value, such as 5%. Recalling that Y denotes the waiting time in the system of a Type 2 customer, then for any $c \geq c^*$ when all other parameters are fixed, the following holds good.

$$P(Y > \frac{r}{\mu}) < 0.05, r \geq 2.$$

The purpose of this is to identify the regions where for a given p the minimum value of c that will guarantee that only certain (pre-determined) percentage of Type 2 customers has to longer than a (pre-determined) multiple of the average service time. Similarly, we can fix c and identify if there is any p that will yield a similar guarantee.

Example 5.6: Here we fix $\lambda_1 = 1, \mu = 1.1, L = 10$, and vary other parameters as follows: $\rho = 0.8, 0.99, p = 0, 0.5, 1$. In order to achieve a specific ρ , λ_2 will be adjusted (by taking $c = 1$ and the same λ_2 is used for other values of $c \geq 2$) as pointed out in *MAP/PH/1* model.

In Table 5.8 we display the optimum c^* for various combinations. We ran our simulation for c up to 50 and if an optimum is not found in that range, we will denote this by simply displaying with "> 50". A quick at this table suggests that for *HEXS*, which has a large variation compared to the other two service distributions, one need c to be larger than 50 for all values of p . Only in the case of positively correlated arrivals and with non-Erlang service times we see a relatively large c when Type 1 customers are not willing to serve Type 2 customers (i.e., when $p = 0$) which is not surprising as positively correlated arrivals have known to show such "odd" behavior

with regard to other system performance measures in the literature (see e.g., Chakravarthy (2010)).

Table 5.8 Optimum c^* values

ρ	p MAP	<i>ERLS</i>			<i>EXPS</i>			<i>HEXS</i>		
		0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
0.8	<i>ERLA</i>	3	2	2	4	4	3	>50	>50	>50
	<i>EXPA</i>	3	2	2	5	4	3	>50	>50	>50
	<i>HEXA</i>	3	3	2	5	4	3	>50	>50	>50
	<i>MNCA</i>	3	2	2	5	4	3	>50	>50	>50
	<i>MPCA</i>	3	2	2	20	4	3	>50	>50	>50
0.99	<i>ERLA</i>	3	3	2	4	4	3	>50	>50	>50
	<i>EXPA</i>	3	3	2	10	4	3	>50	>50	>50
	<i>HEXA</i>	3	3	2	5	4	3	>50	>50	>50
	<i>MNCA</i>	3	3	2	4	4	3	>50	>50	>50
	<i>MPCA</i>	3	3	2	20	4	3	>50	>50	>50

5.3 *MAP/PH/1* Queueing Model with Crowdsourcing and Working Vacation

In this section we introduce vacation and working vacation in the context of *MAP/PH/1* with crowdsourcing. The matrix-analytic method is employed to study the model in steady-state analysis.

The section is organized as follows. In Section 5.3.1 the model under study is described. The steady-state analysis of the model is performed in Section 5.3.2 and in Section 5.3.5 we discuss *crowdsourcing* with pure vacation (i.e., the server doesn't offer any services during vacation time). After the phase type representation of the duration of vacation and working vacation is represented in Section 5.3.6, some key system performance measures are given in the follows Section 5.3.7. Finally, Through illustrative numerical examples we demonstrate the significant benefits in introducing this type of variants to the classical queueing models. in Section 5.3.8.

5.3.1 Model Description

We consider a single server queueing system in which two types of customers, say, Type 1 arrive according to a *MAP* with representation (D_0, D_1) of order m . Let λ_1 denote the arrival rate of Type 1 customers. An arriving Type 1 customer finding the server idle (i.e., on vacation) will get into service immediately but at a lower rate. Otherwise, the customer will enter into a finite buffer of size $L, 1 \leq L < \infty$, to be served on FCFS basis when the server becomes available. Thus, it is possible for a Type 1 customer to be lost at the time of arrival due to the buffer being full. Let D , defined by $D = D_0 + D_1$, govern the underlying Markov chain of the *MAP* such that D_0 accounts for the transitions corresponding to no arrival; D_1 governs those corresponding to an arrival of a Type 1 customer. On the other hand, the arrivals of Type 2 customers are assumed to follow a Poisson process with rate λ_2 . There is no restriction on how many Type 2 customers can be in the system. That is, there is an infinite buffer space for Type 2 customers. An arriving Type 2 customer finding the server idle will have to wait until the server becomes available to offer service.

The service times follow a phase type distribution with representation (β, S) of order n . When the system becomes empty at the time of a completion of a service, the server will go on a vacation. The duration of a vacation is assumed to be exponentially distributed with parameter τ . A vacation is interrupted only when a Type 1 customer arrives during that time. However, the server offers services to those (Type 1) customers arriving during a vacation at a lower rate. Note that only Type 1 customers can be served during a vacation mode. We assume that the service times of those customers (served at a lower rate) are also of phase type but with representation $(\beta, \theta S)$ of order n , where $0 \leq \theta \leq 1$. The server continues to serve at this rate until either the vacation expires or the number of Type 1 customers in the system hits a pre-determined threshold, say, $N, 1 \leq N \leq L + 1$. At this instant, the server instantaneously switches over to the normal rate and continues to serve at this rate until the system becomes empty. Note that once the service resumes at a normal rate (either through vacation getting over or when the number of Type 1 customers in the

system hits N) the server will serve both types of customers. At the end of a vacation if there is no customer waiting for service, the server takes another vacation. Let μ denote the regular service rate. It is easy to verify that $\mu = [\beta(-S)^{-1}e]^{-1}$ and the vacation mode of service has rate $\theta\mu$.

While Type 1 customers are to be served by the server, Type 2 customers may be served by a Type 1 customer having already been served and also available to act as a server or by the (system) server. A Type 2 customer getting serviced by a Type 1 customer depends on the following conditions. First, that Type 1 customer should have just finished getting a service and opts to service a Type 2 customer. Secondly, at the time of opting to serve there is at least one Type 2 customer waiting to get a service. That is, this Type 2 customer should not have started getting a service from the server. We assume that a served Type 1 customer will be available to act as a server for a Type 2 customer under the conditions mentioned above with probability p , $0 \leq p \leq 1$. With probability $q = 1 - p$, the served Type 1 customer will leave the system. Upon completion of a service the free server will offer service to a Type 1 customer on a *FCFS* basis; however, if there are no Type 1 customers waiting, the server will serve a Type 2 customer if there is one present in the queue. We assume that Type 1 customers have a non-preemptive priority over Type 2 customers. If a Type 1 customer decides to serve a Type 2 customer, for our analysis purposes that Type 2 customer will be removed from the system immediately. This is due to the fact that the system no longer needs to track that Type 2 customer.

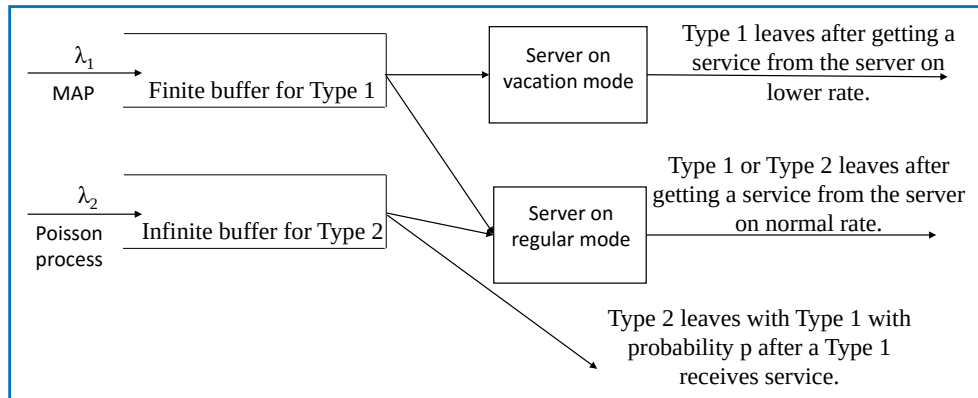


Figure 5.9 Diagram of the queue model with vacation and crowdsourcing

Let η be the steady state probability vector of the Markov process with irreducible generator D . That is, η is the unique probability vector satisfying

$$\eta D = 0, \quad \eta e = 1. \quad (5.29)$$

5.3.2 The Steady-state Analysis

Let $N_1(t), N_2(t), N_3(t), N_4(t)$, and $N_5(t)$ denote, respectively, the number of Type 2 customers in the queue, the number of Type 1 customers in the system, the phase of the service (if any), the phase of the arrival process and the state of the server, at time t . The state of the server is given by

$$N_5(t) = \begin{cases} 0_v, & \text{if the server is on vacation mode} \\ 0_w, & \text{if the server is on working vacation mode} \\ 1, & \text{if the server is on regular mode and busy with Type 1 customer} \\ 2, & \text{if the server is on regular mode and busy with Type 2 customer.} \end{cases}$$

The process $\{(N_1(t), N_2(t), N_3(t), N_4(t), N_5(t)) : t \geq 0\}$ is a continuous-time Markov chain with state space given by

$$\begin{aligned} \Omega = & \{(i, k_2, 0_v), 1 \leq k_2 \leq m\} \\ & \cup \{(i, j, k_1, k_2, 0_w), 1 \leq j \leq N-1, 1 \leq k_1 \leq n, 1 \leq k_2 \leq m\} \\ & \cup \{(i, j, k_1, k_2, 1), 1 \leq j \leq L+1, 1 \leq k_1 \leq n, 1 \leq k_2 \leq m\} \\ & \cup \{(i, j, k_1, k_2, 2), 0 \leq j \leq L, 1 \leq k_1 \leq n, 1 \leq k_2 \leq m\}, i \geq 0\}, \end{aligned}$$

which is rewritten in terms of set of states as

$$\Omega = \{\underline{i}, i \geq 0\},$$

with the set of states, $\underline{i}$, partitioned as

$$\underline{i} = \{\underline{i}_v\} \cup \{\underline{i}_w\} \cup \{\underline{i}_1\} \cup \{\underline{i}_2\}.$$

Note that the level $\underline{i}_v$ is of dimension m and corresponds to the case when the server is on vacation mode with i Type 2 customers waiting in the queue and the arrival process is in one of m phases; the level $\underline{i}_w$, of dimension $(N - 1)mn$, corresponds to the case when the server is on working vacation mode with i Type 2 customers waiting in the queue, $j, 1 \leq j \leq N - 1$, Type 1 customers in the system, the service is in one of n phases and the arrival process is in one of m phases; the level $\underline{i}_1$, of dimension $(L + 1)mn$, corresponds to case when the server is on regular mode serving a Type 1 customer with i Type 2 customers waiting in the queue, $j, 1 \leq j \leq L + 1$, Type 1 customers in the system, the service is in one of n phases and the arrival process is in one of m phases; the level $\underline{i}_2$ is of dimension $(L + 1)mn$ and corresponds to the case when the server is on regular mode and busy serving a Type 2 customer with i Type 2 customers waiting in the queue, $j, 0 \leq j \leq L$, Type 1 customers in the system, the service is in one of n phases and the arrival process is in of m phases.

The infinitesimal generator of the Markov chain governing the system is given by

$$Q = \begin{pmatrix} B_1 & A_0 & & & \\ B_2 & A_1 & A_0 & & \\ A_3 & A_2 & A_1 & A_0 & \\ & A_3 & A_2 & A_1 & A_0 \\ & & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (5.30)$$

where the matrices appearing in Q are as follows (first in partitioned form and then the form of the blocks in those partitioned matrices).

$$B_1 = \begin{pmatrix} B_{11}^{(1)} & B_{12}^{(1)} & 0 & 0 \\ B_{21}^{(1)} & B_{22}^{(1)} & B_{23}^{(1)} & 0 \\ B_{31}^{(1)} & 0 & B_{33}^{(1)} & 0 \\ B_{41}^{(1)} & 0 & B_{43}^{(1)} & B_{44}^{(1)} \end{pmatrix},$$

with

$$B_{11}^{(1)} = D_0 - \lambda_2 I, \quad B_{21}^{(1)} = \mathbf{e}_1(N-1) \otimes (\theta \mathbf{S}^0 \otimes I), \quad B_{31}^{(1)} = B_{41}^{(1)} = \mathbf{e}_1(L+1) \otimes (\mathbf{S}^0 \otimes I),$$

$$B_{12}^{(1)} = \mathbf{e}_1'(N-1) \otimes (\boldsymbol{\beta} \otimes D_1),$$

$$B_{22}^{(1)} = \begin{pmatrix} (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & & \\ \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 \\ & & & \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I \end{pmatrix},$$

$$B_{33}^{(1)} = \begin{pmatrix} (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & & & \\ \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D_0) - \lambda_2 I & I \otimes D_1 \\ & & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D) - \lambda_2 I \end{pmatrix},$$

$$B_{44}^{(1)} = \begin{pmatrix} (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & & & \\ & \ddots & \ddots & \ddots & \\ & & (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & \\ & & & (S \oplus D) - \lambda_2 I & \end{pmatrix},$$

$$B_{43}^{(1)} = \begin{pmatrix} 0 & & & & \\ \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & \\ & & & \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \end{pmatrix},$$

$$B_{23}^{(1)} = \begin{pmatrix} \tau I & & & \\ & \tau I & & \\ & & \ddots & \\ & & & \tau I & I \otimes D_1 & 0 & \cdots \end{pmatrix}.$$

$$B_2 = \begin{pmatrix} 0 & 0 & 0 & B_{14}^{(2)} \\ B_{21}^{(2)} & B_{22}^{(2)} & 0 & 0 \\ B_{31}^{(2)} & 0 & B_{33}^{(2)} & B_{34}^{(2)} \\ 0 & 0 & 0 & B_{44}^{(2)} \end{pmatrix},$$

with

$$B_{14}^{(2)} = \mathbf{e}_1'(L+1) \otimes (\tau \boldsymbol{\beta} \otimes I), \quad B_{21}^{(2)} = \mathbf{e}_1(N-1) \otimes (\theta p \mathbf{S}^0 \otimes I), \quad B_{31}^{(2)} = \mathbf{e}_1(L+1) \otimes (p \mathbf{S}^0 \otimes I),$$

$$B_{34}^{(2)} = \mathbf{e}_1(L+1) \otimes \mathbf{e}_1'(L+1) \otimes (q \mathbf{S}^0 \boldsymbol{\beta} \otimes I), \quad B_{44}^{(2)} = \mathbf{e}_1(L+1) \otimes \mathbf{e}_1'(L+1) \otimes (\mathbf{S}^0 \boldsymbol{\beta} \otimes I),$$

$$B_{22}^{(2)} = \begin{pmatrix} 0 & & & \\ \theta p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & & \\ & \ddots & \ddots & \\ & & \theta p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \\ & & & \theta p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \end{pmatrix},$$

$$B_{33}^{(2)} = \begin{pmatrix} 0 & & & \\ p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 & & \\ & \ddots & \ddots & \\ & & p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \\ & & & p \mathbf{S}^0 \boldsymbol{\beta} \otimes I & 0 \end{pmatrix}.$$

$$A_1 = \begin{pmatrix} A_{11} & B_{12}^{(1)} & 0 & 0 \\ A_{21} & A_{22} & B_{23}^{(1)} & 0 \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & B_{43}^{(1)} & B_{44}^{(1)} \end{pmatrix},$$

with

$$A_{11} = D_0 - \tau I - \lambda_2 I, \quad A_{21} = \mathbf{e}_1(N-1) \otimes (\theta q \mathbf{S}^0 \otimes I),$$

$$A_{22} = \begin{pmatrix} (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & & \\ \theta q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & \\ & \ddots & & \ddots & \\ & & \theta q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 \\ & & & \theta q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I \end{pmatrix},$$

$$A_{33} = \begin{pmatrix} (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & & & \\ q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D_0) - \lambda_2 I & I \otimes D_1 & & \\ & \ddots & & \ddots & \\ & & q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D_0) - \lambda_2 I & I \otimes D_1 \\ & & & q \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (S \oplus D) - \lambda_2 I \end{pmatrix}.$$

$$A_2 = \begin{pmatrix} 0 & 0 & 0 & B_{14}^{(2)} \\ B_{21}^{(2)} & B_{22}^{(2)} & 0 & 0 \\ 0 & 0 & B_{33}^{(2)} & B_{34}^{(2)} \\ 0 & 0 & 0 & B_{44}^{(2)} \end{pmatrix}, \quad A_0 = \lambda_2 I,$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{A} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \hat{A} = \mathbf{e}_1(L+1) \otimes \mathbf{e}_1'(L+1) \otimes (p \mathbf{S}^0 \boldsymbol{\beta} \otimes I),$$

The generator Q in (5.30) is of the $GI/M/1$ -type and we can apply the well-known matrix-analytic methods Neuts Neuts (1981) to perform the steady-state analysis. In this set up the rate matrix, R , of dimension $m + (N - 1)mn + 2(L + 1)mn$, is the minimal nonnegative solution to a matrix cubic equation. However, by combining the set of states, $\{\underline{2i}, \underline{2i + 1}\}$, for $i \geq 0$ as

$$\tilde{i} = \{\underline{2i}, \underline{2i + 1}\},$$

the model under study can be studied as a QBD -process with generator, $\tilde{Q}$, on the state space $\tilde{\Omega} = \{\tilde{i}, \tilde{i} \geq 0\}$ given by

$$\tilde{Q} = \begin{pmatrix} \tilde{B}_0 & F_0 & & & \\ F_2 & F_1 & F_0 & & \\ & F_2 & F_1 & F_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \quad (5.31)$$

where the block entries appearing in $\tilde{Q}$ are obtained from those of Q as follows.

$$\tilde{B}_0 = \begin{pmatrix} B_1 & A_0 \\ B_2 & A_1 \end{pmatrix}, F_0 = \begin{pmatrix} 0 & 0 \\ A_0 & 0 \end{pmatrix}, F_1 = \begin{pmatrix} A_1 & A_0 \\ A_2 & A_1 \end{pmatrix}, F_2 = \begin{pmatrix} A_3 & A_2 \\ 0 & A_3 \end{pmatrix}.$$

In the rest of the study we will analyze the model under study in steady-state using QBD approach.

5.3.3 The Stability Condition

Let $\xi = (\xi(1), \xi(2))$ be the steady state probability vector of the generator $F = F_0 + F_1 + F_2$. That is, ξ satisfies

$$\xi F = \mathbf{0}, \xi e = 1. \quad (5.32)$$

On noting that

$$F = \begin{pmatrix} A_1 + A_3 & A_0 + A_2 \\ A_0 + A_2 & A_1 + A_3 \end{pmatrix}, \quad (5.33)$$

is a circulant matrix, the vector ξ is of the form given by

$$\xi = 0.5(e' \otimes \pi), \quad (5.34)$$

where π satisfies (with $A = A_0 + A_1 + A_2 + A_3$)

$$\pi A = \mathbf{0}, \pi e = 1. \quad (5.35)$$

First note that the matrix A is reducible and hence π is of the form

$$\pi = (\mathbf{0}, \pi_1, \pi_2), \text{ with } \pi_1 = (\pi_{11}, \pi_{12}, \dots, \pi_{1,L+1}) \text{ and } \pi_2 = (\pi_{20}, \pi_{21}, \dots, \pi_{2L}).$$

We now establish a few results that are needed for later use as well as serve as internal accuracy checks in our numerical results.

Lemma 5.3.1. *We have*

$$\left(\sum_{j=1}^{L+1} \pi_{1j} + \sum_{j=0}^L \pi_{2j} \right) (I \otimes e) = \mu \beta (-S)^{-1}, \quad \left(\sum_{j=1}^{L+1} \pi_{1j} + \sum_{j=0}^L \pi_{2j} \right) (e \otimes I) = \eta, \quad (5.36)$$

where η is as given in (5.29) and μ is the service rate.

Proof: The steady state equations in (5.35) can be rewritten as

$$\pi_{11}(S \oplus D_0) + (\pi_{12} + \pi_{21})(S^0 \beta \otimes I) = 0$$

$$\pi_{1j}(S \oplus D_0) + \pi_{1,j-1}(I \otimes D_1) + (\pi_{1,j+1} + \pi_{2j})(S^0 \beta \otimes I) = 0, \quad 2 \leq j \leq L$$

$$\pi_{1,L+1}(S \oplus D) + \pi_{1L}(I \otimes D_1) = 0$$

$$\pi_{20}((S + \mathbf{S}^0\beta) \oplus D_0) + \pi_{11}(\mathbf{S}^0\beta \otimes I) = 0 \quad (5.37)$$

$$\pi_{2j}(S \oplus D_0) + \pi_{2,j-1}(I \otimes D_1) = 0 \quad 1 \leq j \leq L-1$$

$$\pi_{2L}(S \oplus D) + \pi_{2,L-1}(I \otimes D_1) = 0$$

with the normalizing condition

$$\left(\sum_{j=1}^{L+1} \pi_{1j} + \sum_{j=0}^L \pi_{2j} \right) e = 1. \quad (5.38)$$

Now adding the equations given in (5.37) we get

$$\left(\sum_{j=1}^{L+1} \pi_{1j} + \sum_{j=0}^L \pi_{2j} \right) [(S + \mathbf{S}^0\beta) \oplus D] = 0.$$

Post-multiplying the equation by $(e \otimes I)$ and using the uniqueness of the steady-state vector of the generator D along with the normalizing condition in (5.38) yields the first result in (5.36). Similarly, by post-multiplying the equation by $(I \otimes e)$ one can obtain the second result in (5.36) from the uniqueness of the steady-state vector of the generator $S + \mathbf{S}^0\beta$ and the normalizing condition in (5.38).

The following theorem establishes the stability condition of the queueing system under study.

Theorem 5.3.2. *The queueing system under study is stable if and only if the following condition is satisfied.*

$$\lambda_2 < p\mu + q \sum_{j=0}^L \pi_{2j}(\mathbf{S}^0 \otimes e). \quad (5.39)$$

Proof: The queueing system under study with the QBD type generator given in (5.31) is stable if and only if $\xi F_0 e < \xi F_2 e$. From the entries of the generator $\tilde{Q}$ and the equation in (5.34), we see that

$$\xi F_0 e = 0.5\lambda_2 \quad \text{and} \quad \xi F_2 e = 0.5[\pi A_2 e + 2\pi A_3 e]. \quad (5.40)$$

It is easy to verify that

$$\pi A_2 \mathbf{e} = q\pi_{11}(\mathbf{S}^0 \boldsymbol{\beta} \otimes I)\mathbf{e} + \sum_{j=2}^{L+1} p\pi_{1j}(\mathbf{S}^0 \boldsymbol{\beta} \otimes I)\mathbf{e} + \pi_{20}(\mathbf{S}^0 \boldsymbol{\beta} \otimes I)\mathbf{e}, \quad (5.41)$$

$$\pi A_3 \mathbf{e} = p\pi_{11}(\mathbf{S}^0 \boldsymbol{\beta} \otimes I)\mathbf{e}.$$

Using equations (5.40) and (5.40), the stability condition becomes

$$\lambda_2 < (\pi_{11} + \pi_{20})(\mathbf{S}^0 \otimes \mathbf{e}) + \sum_{j=1}^{L+1} p\pi_{1j}(\mathbf{S}^0 \otimes \mathbf{e}).$$

Now post-multiplying the last three equations in (5.37) by $\mathbf{e}$ and adding them we get

$$\pi_{11}(\mathbf{S}^0 \otimes \mathbf{e}) = \sum_{j=1}^L \pi_{2j}(\mathbf{S}^0 \otimes \mathbf{e}). \quad (5.42)$$

From (5.42) and the fact (from Lemma 5.3.1) that $(\sum_{j=0}^L \pi_{2j} + \sum_{j=1}^{L+1} \pi_{1j})(\mathbf{S}^0 \otimes \mathbf{e}) = \mu$, the stated result follows immediately.

5.3.4 The Steady-state Probability Vector

Let $\mathbf{x}=(x(0), x(1), \dots)$ denote the steady-state probability vector of $\tilde{Q}$. That is, $\mathbf{x}$ satisfies

$$\mathbf{x}\tilde{Q} = 0, \mathbf{x}\mathbf{e} = 1. \quad (5.43)$$

The vector $\mathbf{x}(i), i \geq 0$, of dimension $2m + 2(N-1)mn + 4(L+1)mn$, is further partitioned as $\mathbf{x}(i) = (\hat{\mathbf{x}}(i), \tilde{\mathbf{x}}(i)), i \geq 0$. The vector $\hat{\mathbf{x}}(i)$, of dimension $m + (N -$

1) $mn + 2(L + 1)mn$, is further partitioned into

$$\hat{\mathbf{x}}(i) = (\hat{\mathbf{x}}_v(i), \hat{\mathbf{x}}_{w1}(i), \dots, \hat{\mathbf{x}}_{w,N-1}(i), \hat{\mathbf{x}}_{11}(i), \dots, \hat{\mathbf{x}}_{1,L+1}(i), \hat{\mathbf{x}}_{20}(i), \dots, \hat{\mathbf{x}}_{2,L}(i)).$$

Note that $\hat{\mathbf{x}}_v(i)$, of dimension m , gives the steady-state probability that the server is on vacation with no Type 1 customers in the system and $2i$ Type 2 customers waiting in the queue with arrival process in one of the m phases; $\hat{\mathbf{x}}_{wj}(i)$, $1 \leq j \leq N - 1$, of dimension mn , gives the steady-state probability that there are j Type 1 customers in the system and $2i$ Type 2 customers in the queue with the server busy with a Type 1 customer during a vacation, the arrival process in one of the m phases, and the service phase is in one of n phases; $\hat{\mathbf{x}}_{1j}(i)$, $1 \leq j \leq L + 1$, of dimension mn , gives the steady-state probability that there are j Type 1 customers in the system, $2i$ Type 2 customers in the queue with the server on a regular mode busy with a Type 1 customer, and the arrival process in one of the m phases along with the service in one of n phases; $\hat{\mathbf{x}}_{2j}(i)$, $0 \leq j \leq L$, of dimension mn , gives the steady-state probability that there are j Type 1 customers in the system, $2i$ Type 2 customers in the queue with the server on a regular mode busy with a Type 2 customer, and the arrival process in one of the m phases along with the service in one of n phases. Similarly, the vector $\tilde{\mathbf{x}}(i)$, of dimension $m + (N - 1)mn + 2(L + 1)mn$, is partitioned into

$$\tilde{\mathbf{x}}(i) = (\tilde{\mathbf{x}}_v(i), \tilde{\mathbf{x}}_{w1}(i), \dots, \tilde{\mathbf{x}}_{w,N-1}(i), \tilde{\mathbf{x}}_{11}(i), \dots, \tilde{\mathbf{x}}_{1,L+1}(i), \tilde{\mathbf{x}}_{20}(i), \dots, \tilde{\mathbf{x}}_{2,L}(i)).$$

The components of $\tilde{\mathbf{x}}(i)$ have similar interpretations except that the number of Type 2 customers in the queue is now given $2i + 1$.

Under the stability condition given in (5.39) the steady-state probability vector $\mathbf{x}$ is obtained as follows

$$\mathbf{x}(i) = \mathbf{x}(0)R^i, i \geq 0, \quad (5.44)$$

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

$$R^2 F_2 + R F_1 + F_0 = 0, \quad (5.45)$$

and $\mathbf{x}(0)$ is obtained by solving

$$\mathbf{x}(0)[\tilde{B}_0 + RF_2] = 0 \quad (5.46)$$

subject to the normalizing condition

$$\mathbf{x}(0)(I - R)^{-1}\mathbf{e} = 1. \quad (5.47)$$

The computation of the R matrix can be carried out using a number of well-known methods such as logarithmic reduction and (block) Gauss-Seidel iterative by exploiting the special structure of the coefficient matrices F_0 , F_1 , and F_2 , which are of dimension $2m + 2(N - 1)mn + 4(L + 1)mn$. This is very important especially when one is dealing with large values of L , N , m and n . For example, one can exploit the structure of R matrix:

$$R = \begin{pmatrix} 0 & 0 \\ R_2 & R_1 \end{pmatrix}. \quad (5.48)$$

Note that the above structure is due to the fact that F_0 has first $m + (N - 1)mn + 2(L + 1)mn$ rows full of zeros. It is worth pointing out that $R_1 = R_2^2$ and that R_2 satisfies the matrix-cubic equation (in $GI/M/1$ set up as pointed out Section 3): $R_2^3A_3 + R_2^2A_2 + R_2A_1 + A_0 = 0$.

The details of other exploitation will be omitted; however, the key steps in the logarithmic reduction are given below and for full details on this we refer the reader to Latouche & Ramaswami (1999).

For use in the sequel we define

$$\mathbf{a} = \mathbf{x}(0)(I - R)^{-1}, \quad \mathbf{b} = \mathbf{x}(0)R(I - R)^{-1}. \quad (5.49)$$

We partition the $2m + 2(N - 1)mn + 4(L + 1)mn$ -dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ as

$$\mathbf{a} = (\hat{\mathbf{a}}_v, \hat{\mathbf{a}}_{w1}, \dots, \hat{\mathbf{a}}_{w,N-1}, \hat{\mathbf{a}}_{1,1}, \dots, \hat{\mathbf{a}}_{1,L+1}, \hat{\mathbf{a}}_{2,0}, \dots, \hat{\mathbf{a}}_{2,L}, \tilde{\mathbf{a}}_v, \tilde{\mathbf{a}}_{w1}, \dots, \tilde{\mathbf{a}}_{w,N-1}, \tilde{\mathbf{a}}_{1,1}, \dots, \tilde{\mathbf{a}}_{1,L+1}, \tilde{\mathbf{a}}_{2,0}, \dots, \tilde{\mathbf{a}}_{2,L}),$$

$$\mathbf{b} = (\hat{\mathbf{b}}_v, \hat{\mathbf{b}}_{w1}, \dots, \hat{\mathbf{b}}_{w,N-1}, \hat{\mathbf{b}}_{1,1}, \dots, \hat{\mathbf{b}}_{1,L+1}, \hat{\mathbf{b}}_{2,0}, \dots, \hat{\mathbf{b}}_{2,L}, \tilde{\mathbf{b}}_v, \tilde{\mathbf{b}}_{w1}, \dots, \tilde{\mathbf{b}}_{w,N-1}, \tilde{\mathbf{b}}_{1,1}, \dots, \tilde{\mathbf{b}}_{1,L+1}, \tilde{\mathbf{b}}_{2,0}, \dots, \tilde{\mathbf{b}}_{2,L}),$$

We now list some useful results in the following Lemma which are intuitively obvious and serve as accuracy checks in our numerical computation.

Lemma 5.3.3. *We have*

$$\begin{aligned} & \sum_{j=0}^L (\hat{\mathbf{a}}_{2j} + \tilde{\mathbf{a}}_{2j})(\mathbf{S}^0 \otimes \mathbf{e}) + \sum_{j=1}^{L+1} p(\tilde{\mathbf{x}}_{1j}(0) + \hat{\mathbf{b}}_{1j} + \tilde{\mathbf{b}}_{1j})(\mathbf{S}^0 \otimes \mathbf{e}) \\ & + \sum_{j=1}^{N-1} p(\tilde{\mathbf{x}}_{wj}(0) + \hat{\mathbf{b}}_{wj} + \tilde{\mathbf{b}}_{wj})(\theta \mathbf{S}^0 \otimes \mathbf{e}) = \lambda_2, \end{aligned} \quad (5.50)$$

$$\sum_{j=1}^{N-1} (\hat{\mathbf{a}}_{wj} + \tilde{\mathbf{a}}_{wj})(\theta \mathbf{S}^0 \otimes \mathbf{e}) + \sum_{j=1}^{L+1} (\hat{\mathbf{a}}_{1j} + \tilde{\mathbf{a}}_{1j})(\mathbf{S}^0 \otimes \mathbf{e}) = \lambda_1(1 - P_{lost}), \quad (5.51)$$

where

$$P_{lost} = \frac{1}{\lambda_1} [(\hat{\mathbf{a}}_{1,L+1} + \tilde{\mathbf{a}}_{1,L+1}) + (\hat{\mathbf{a}}_{2,L} + \tilde{\mathbf{a}}_{2,L})](\mathbf{e} \otimes D_1 \mathbf{e}). \quad (5.52)$$

5.3.5 MAP/PH/1 Model with Crowdsourcing Under Pure Vacation

First note that MAP/PH/1 queue with *crowdsourcing* with vacation and working vacation cannot be reduced to MAP/PH/1 queue with *crowdsourcing* by suitably choosing the values for θ and N , which appear in working vacation model. Secondly, we will compare the working vacation model with pure vacation model. Thus, we will very briefly outline the needed quantities for MAP/PH/1 *crowdsourcing* queue with vacation in this section.

The process $\{(N_1(t), N_2(t), N_3(t), N_4(t), N_5^*(t)) : t \geq 0\}$ is a continuous-time Markov chain, where the state of the server, $N_5^*(t)$, is now given by

$$N_5^*(t) = \begin{cases} v, & \text{if the server is on vacation with no Type 1 customer present} \\ v1, & \text{if the server is on vacation with Type 1 customer present} \\ 1, & \text{if the server is on regular mode and busy with Type 1 customer} \\ 2, & \text{if the server is on regular mode and busy with Type 2 customer.} \end{cases}$$

The state space of the Markov chain is given by

$$\begin{aligned} \Omega_v = & \{(i, k_2, v), 1 \leq k_2 \leq m\} \cup \{(i, j, k_2, v1), 1 \leq j \leq L, 1 \leq k_2 \leq m\} \\ & \cup \{(i, j, k_1, k_2, 1), 1 \leq j \leq L+1, 1 \leq k_1 \leq n, 1 \leq k_2 \leq m\} \\ & \cup \{(i, j, k_1, k_2, 2), 0 \leq j \leq L, 1 \leq k_1 \leq n, 1 \leq k_2 \leq m\}, i \geq 0\}, \end{aligned}$$

which is rewritten in terms of set of states as $\Omega_v = \{\underline{i}, i \geq 0\}$, with the set of states, $\underline{i}$, partitioned as

$$\underline{i} = \{\underline{i}_v\} \cup \{\underline{i}_{v1}\} \cup \{\underline{i}_1\} \cup \{\underline{i}_2\}.$$

Note that the level $\underline{i}_v$ is of dimension m and corresponds to the case when the server is on vacation mode with i Type 2 customers and no Type 1 customer waiting in the queue, and the arrival process is in one of m phases; the level $\underline{i}_{v1}$ is of dimension Lm and corresponds to the case when the server is on vacation mode with i Type 2 customers and j , $1 \leq j \leq L$, Type 1 customers waiting in the queue, and the arrival process is in one of m phases.

The infinitesimal generator of the Markov chain governing the system is given by

$$Q_v = \begin{pmatrix} B_{1v} & A_{0v} & & & \\ B_{2v} & A_{1v} & A_{0v} & & \\ A_{3v} & A_{2v} & A_{1v} & A_{0v} & \\ & A_{3v} & A_{2v} & A_{1v} & A_{0v} \\ & & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where the matrices appearing in Q_v are as follows (first in partitioned form and then the form of the blocks in those partitioned matrices). Since most of the blocks are common with Q (see equation (5.30)), we will display only those that are new here.

$$B_{1v} = \begin{pmatrix} B_{11}^{(1)} & B_{12}^{(1v)} & 0 & 0 \\ 0 & B_{22}^{(1v)} & B_{23}^{(1v)} & 0 \\ B_{31}^{(1)} & 0 & B_{33}^{(1)} & B_{34}^{(1)} \\ B_{41}^{(1)} & 0 & B_{43}^{(1)} & B_{44}^{(1)} \end{pmatrix},$$

with

$$B_{22}^{(1v)} = \begin{pmatrix} D_0 - \tau I - \lambda_2 I & D_1 & & & \\ 0 & D_0 - \tau I - \lambda_2 I & D_1 & & \\ & \ddots & & \ddots & \\ & & 0 & D_0 - \tau I - \lambda_2 I & D_1 \\ & & & 0 & D_0 - \tau I - \lambda_2 I \end{pmatrix},$$

$$B_{23}^{(1v)} = \begin{pmatrix} \tau(\beta \otimes I) & & & & \\ & \tau(\beta \otimes I) & & & \\ & & \ddots & & \\ & & & \tau(\beta \otimes I) & 0 \dots \end{pmatrix}, \quad B_{12}^{(1v)} = \mathbf{e}'_1(L) \otimes D_1.$$

$$B_{2v} = \begin{pmatrix} 0 & 0 & 0 & B_{14}^{(2)} \\ 0 & 0 & 0 & 0 \\ B_{31}^{(2)} & 0 & B_{33}^{(2)} & B_{34}^{(2)} \\ 0 & 0 & 0 & B_{44}^{(2)} \end{pmatrix}, \quad A_{0v} = \lambda_2 I,$$

$$A_{1v} = \begin{pmatrix} A_{11} & B_{12}^{(1v)} & 0 & 0 \\ 0 & B_{22}^{(1v)} & B_{23}^{(1v)} & 0 \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & B_{43}^{(1)} & B_{44}^{(1)} \end{pmatrix},$$

$$A_{2v} = \begin{pmatrix} 0 & 0 & 0 & B_{14}^{(2)} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & B_{33}^{(2)} & B_{34}^{(2)} \\ 0 & 0 & 0 & B_{44}^{(2)} \end{pmatrix}, \quad A_{3v} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{A} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Combining the set of states $\{2i, 2i+1\}$, for $i \geq 0$, like earlier, this model under study can be studied as a QBD -process with generator, $\tilde{Q}_v$, on the state space $\tilde{\Omega}_v = \{\tilde{i}, \tilde{i} \geq 0\}$ given by

$$\tilde{Q}_v = \begin{pmatrix} \tilde{B}_{0v} & F_{0v} & & & \\ F_{2v} & F_{1v} & F_{0v} & & \\ & F_{2v} & F_{1v} & F_{0v} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

where the block entries appearing in $\tilde{Q}_v$ are obtained from those of Q_v as follows.

$$\tilde{B}_{0v} = \begin{pmatrix} B_{1v} & A_{0v} \\ B_{2v} & A_{1v} \end{pmatrix}, \quad F_{0v} = \begin{pmatrix} 0 & 0 \\ A_{0v} & 0 \end{pmatrix},$$

$$F_{1v} = \begin{pmatrix} A_{1v} & A_{0v} \\ A_{2v} & A_{1v} \end{pmatrix}, \quad F_{2v} = \begin{pmatrix} A_{3v} & A_{2v} \\ 0 & A_{3v} \end{pmatrix}.$$

Let $\mathbf{y}=(\mathbf{y}(0), \mathbf{y}(1),...)$ denote the steady-state probability vector of $\tilde{Q}_v$. That is, $\mathbf{y}$

satisfies $\mathbf{y}\tilde{Q}_v = 0$ and $\mathbf{y}e = 1$. The vector $\mathbf{y}(i), i \geq 0$, of dimension $2(L+1)m + 4(L+1)mn$, is further partitioned as $\mathbf{y}(i) = (\hat{\mathbf{y}}(i), \tilde{\mathbf{y}}(i)), i \geq 0$. The vectors $\hat{\mathbf{y}}(i)$ and $\tilde{\mathbf{y}}(i)$, of dimension $(L+1)m + 2(L+1)mn$, are further partitioned into

$$\hat{\mathbf{y}}(i) = (\hat{\mathbf{x}}_v(i), \hat{\mathbf{y}}_{v1,1}(i), \dots, \hat{\mathbf{y}}_{v1,L}(i), \hat{\mathbf{x}}_{11}(i), \dots, \hat{\mathbf{x}}_{1,L+1}(i), \hat{\mathbf{x}}_{20}(i), \dots, \hat{\mathbf{x}}_{2,L}(i)).$$

$$\tilde{\mathbf{y}}(i) = (\tilde{\mathbf{x}}_v(i), \tilde{\mathbf{y}}_{v1,1}(i), \dots, \tilde{\mathbf{y}}_{v1,L}(i), \tilde{\mathbf{x}}_{11}(i), \dots, \tilde{\mathbf{x}}_{1,L+1}(i), \tilde{\mathbf{x}}_{20}(i), \dots, \tilde{\mathbf{x}}_{2,L}(i)).$$

Note that $\hat{\mathbf{y}}_{v1,j}(i), 1 \leq j \leq L$, of dimension m , gives the steady-state probability that there are j Type 1 customers in the queue and $2i$ Type 2 customers in the queue when the server is on vacation mode with arrival process in one of the m phases. The components $\tilde{\mathbf{y}}_{v1,j}(i), 1 \leq j \leq L$, have similar interpretations except that the number of Type 2 customers in the queue is now given $2i + 1$.

The rate matrix, R_v , of dimension $2(L+1)m + 4(L+1)mn$ is the minimal nonnegative solution to the matrix quadratic equation $R_v^2 F_{2v} + R_v F_{1v} + F_{0v} = 0$. The stability condition for this case is the same as that for working vacation (see equation (5.39)).

For use in the performance measures we define $\mathbf{a}_v = \mathbf{y}(0)(I - R_v)^{-1}$ and $\mathbf{b}_v = \mathbf{y}(0)R_v(I - R_v)^{-1}$. We partition the $2(L+1)m + 4(L+1)mn$ -dimensional vectors $\mathbf{a}_v$ and $\mathbf{b}_v$ as

$$\mathbf{a}_v = (\hat{\mathbf{a}}_v, \hat{\mathbf{a}}_{v1,1}, \dots, \hat{\mathbf{a}}_{v1,L}, \hat{\mathbf{a}}_{1,1}, \dots, \hat{\mathbf{a}}_{1,L+1}, \hat{\mathbf{a}}_{2,0}, \dots, \hat{\mathbf{a}}_{2,L}, \tilde{\mathbf{a}}_v, \tilde{\mathbf{a}}_{w1}, \dots, \tilde{\mathbf{a}}_{w,N-1}, \\ \tilde{\mathbf{a}}_{1,1}, \dots, \tilde{\mathbf{a}}_{1,L+1}, \tilde{\mathbf{a}}_{2,0}, \dots, \tilde{\mathbf{a}}_{2,L}),$$

$$\mathbf{b}_v = (\hat{\mathbf{b}}_v, \hat{\mathbf{b}}_{v1,1}, \dots, \hat{\mathbf{b}}_{v1,L}, \hat{\mathbf{b}}_{1,1}, \dots, \hat{\mathbf{b}}_{1,L+1}, \hat{\mathbf{b}}_{2,0}, \dots, \hat{\mathbf{b}}_{2,L}, \tilde{\mathbf{b}}_v, \tilde{\mathbf{b}}_{w1}, \dots, \tilde{\mathbf{b}}_{w,N-1}, \\ \tilde{\mathbf{b}}_{1,1}, \dots, \tilde{\mathbf{b}}_{1,L+1}, \tilde{\mathbf{b}}_{2,0}, \dots, \tilde{\mathbf{b}}_{2,L}).$$

5.3.6 Phase-type Representation for Vacation/Working Vacation Duration

In this section we will prove that the duration of the vacation/working vacation can be modeled as a phase type distribution. The generator matrix of a Markov chain with a single absorbing state is the form

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

is further partitioned by considering two subsets (1: pure vacation & 2: working vacation) as

$$Q^* = \begin{pmatrix} T(1,1) & T(1,2) & T^0(1) \\ T(2,1) & T(2,2) & T^0(2) \\ 0 & 0 & 0 \end{pmatrix},$$

with

$$T^0(1) = 0, \quad T^0(2) = (\tau \mathbf{e}, \dots, \tau \mathbf{e}, \mathbf{e} \otimes D_1 \mathbf{e} + \tau \mathbf{e})',$$

$$T(1,1) = D_0, \quad T(1,2) = \mathbf{e}_1'(N-1) \otimes (\boldsymbol{\beta} \otimes D_1), \quad T(2,1) = \mathbf{e}_1(N-1) \otimes (\theta \mathbf{S}^0 \otimes I),$$

$$T(2,2) = \begin{pmatrix} (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & & \\ \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 & & \\ & \ddots & & \ddots & \\ & & \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I & I \otimes D_1 \\ & & & \theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I & (\theta S \oplus D_0) - \tau I - \lambda_2 I \end{pmatrix},$$

where the square matrices $T(1,1)$ and $T(2,2)$ correspond to transitions within subsets 1 and 2 of transient states. The submatrices $T(1,2)$ and $T(2,1)$ correspond to transitions from set 1 to set 2 and vice versa.

The initial probability vector is written in the correspondingly partitioned form

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, 0, 0),$$

with $\boldsymbol{\alpha}_0 = \delta(\mathbf{x}_{11}(0) + \mathbf{x}_{02}(0) + p\mathbf{x}_{11}(1))(\mathbf{S}^0 \otimes I)$ where δ is the normalizing constant.

The total time spent in the subsets 1 and 2 prior to absorption (server works at normal rate) is of phase type with representation (α, T) of order $m + (N - 1)mn$.

$$\mu'_{v/wv} = \alpha(-T)^{-1}e.$$

The computation of this mean is achieved by exploiting the special structure of α and T . We will briefly outline the steps involved in this computation.

Defining $\alpha(-T)^{-1} = c$ and partitioning the vector c as

$$c = (c_0, \dots, c_{N-1}),$$

where c_0 is of dimension m and $c_i, 1 \leq i \leq N - 1$, is of dimension mn , the mean $\mu'_{v/wv}$ is given by

$$\mu'_{v/wv} = \sum_{i=0}^{N-1} c_i e.$$

The vector c is calculated using the following set of equations.

$$c_0 D_0 + c_1 (\theta S^0 \otimes I) = -\alpha_0,$$

$$c_0 (\beta \otimes D_1) + c_1 (\theta S \oplus D_0 - \tau I) + c_2 (\theta S^0 \beta \otimes I) = 0,$$

$$c_{i-1} (I \otimes D_1) + c_i (\theta S \oplus D_0 - \tau I) + c_{i+1} (\theta S^0 \beta \otimes I) = 0, \quad 2 \leq i \leq N - 2,$$

$$c_{N-2} (I \otimes D_1) + c_{N-1} (\theta S \oplus D_0 - \tau I) = 0.$$

subject to the condition

$$\tau \sum_{i=1}^{N-1} c_i e + c_{N-1} (e \otimes D_1 e) = 1.$$

The time spent *only* in the subset 1 prior to absorption is of phase type with representation (α_0, K) of order m .

$$\mu'_{pv} = \alpha_0(-K)^{-1}e,$$

where

$$K = T(1, 1) + T(1, 2)(-T(2, 2))^{-1}T(2, 1).$$

Defining $T(1, 2)(-T(2, 2))^{-1} = \mathbf{d}$ and partitioning the vector $\mathbf{d}$ as

$$\mathbf{d} = (\mathbf{d}_1, \dots, \mathbf{d}_{N-1}).$$

where $\mathbf{d}_i, 1 \leq i \leq N-1$, is of dimension mn , K is rewritten by exploiting the special structure of $T(2, 1)$ as $K = T(1, 1) + \mathbf{d}_1(\theta \mathbf{S}^0 \otimes I)$.

The vector $\mathbf{d}$ is calculated using the following system of equations.

$$\mathbf{d}_1(\tau I - (\theta S \oplus D_0)) - \mathbf{d}_2(\theta \mathbf{S}^0 \beta \otimes I) = (\beta \otimes D_1),$$

$$\mathbf{d}_{i-1}(I \otimes D_1) + \mathbf{d}_i((\theta S \oplus D_0) - \tau I) + \mathbf{d}_{i+1}(\theta \mathbf{S}^0 \beta \otimes I) = 0, \quad 2 \leq i \leq N-2,$$

$$\mathbf{d}_{N-1}(\tau I - (\theta S \oplus D_0)) - \mathbf{d}_{N-2}(I \otimes D_1) = 0.$$

5.3.7 The System Performance Measures

In this section we will list a number of system performance measures of interest along with their expressions. Since we are looking at two models, one with pure vacation and the other with working vacations, it should be pointed out that most measures are common and have similar expressions. However, the values of the measures will be different depending on the model. Where the expressions for the measures differ we will distinguish accordingly by using the symbols " v " or " wv ".

1. Probability that the server is on vacation.

$$P_{vac}^{(wv)} = (\hat{\mathbf{a}}_v + \tilde{\mathbf{a}}_v)\mathbf{e}.$$

$$P_{vac}^{(v)} = (\hat{\mathbf{a}}_v + \tilde{\mathbf{a}}_v)\mathbf{e} + \sum_{j=1}^L (\hat{\mathbf{a}}_{v1,j} + \tilde{\mathbf{a}}_{v1,j})\mathbf{e}.$$

2. Probability that the server is on working vacation.

$$P_{workvac} = \sum_{j=1}^{N-1} (\hat{a}_{wj} + \tilde{a}_{wj})e.$$

3. Probability that the server is on regular mode.

$$P_{busy}^{(reg)} = \sum_{j=1}^{L+1} (\hat{a}_{1j} + \tilde{a}_{1j})e + \sum_{j=0}^L (\hat{a}_{2j} + \tilde{a}_{2j})e.$$

4. Probability that the server is busy with a Type 1 customer.

$$P_{busyT_1}^{(wv)} = \sum_{j=1}^{N-1} (\hat{a}_{wj} + \tilde{a}_{wj})e + \sum_{j=1}^{L+1} (\hat{a}_{1j} + \tilde{a}_{1j})e.$$

$$P_{busyT_1}^{(v)} = \sum_{j=1}^{L+1} (\hat{a}_{1j} + \tilde{a}_{1j})e.$$

5. Probability that the server is busy with a Type 2 customer.

$$P_{busy}^{(2)} = \sum_{j=0}^L (\hat{a}_{2j} + \tilde{a}_{2j})e.$$

6. Probability an arriving Type 1 customer is lost.

$$P_{lost} = \frac{1}{\lambda_1} [(\hat{a}_{1,L+1} + \tilde{a}_{1,L+1}) + (\hat{a}_{2,L} + \tilde{a}_{2,L})](e \otimes D_1 e).$$

7. Rate of a Type 2 customer leaving by getting service from the server.

$$R_{T_2LS} = \sum_{j=0}^L (\hat{a}_{2j} + \tilde{a}_{2j})(S^0 \otimes e).$$

8. Rate of a Type 2 customer leaving with a Type 1 customer.

$$R_{T_2L1}^{(wv)} = \sum_{j=1}^{N-1} p(\tilde{x}_{wj}(0) + \hat{b}_{wj} + \tilde{b}_{wj})(\theta S^0 \otimes e) + \sum_{j=1}^{L+1} p(\tilde{x}_{1j}(0) + \hat{b}_{1j} + \tilde{b}_{1j})(S^0 \otimes e).$$

$$R_{T_2L1}^{(v)} = \sum_{j=1}^{L+1} p(\tilde{x}_{1j}(0) + \hat{b}_{1j} + \tilde{b}_{1j})(S^0 \otimes e).$$

9. Mean number of Type 1 customers in the queue.

$$\mu_{T1Q}^{(wv)} = \sum_{j=1}^{N-1} (j-1)(\hat{a}_{wj} + \tilde{a}_{wj})e + \sum_{j=1}^{L+1} (j-1)(\hat{a}_{1j} + \tilde{a}_{1j})e + \sum_{j=0}^L j(\hat{a}_{2j} + \tilde{a}_{2j})e.$$

$$\mu_{T1Q}^{(v)} = \sum_{j=1}^L j(\hat{a}_{v1,j} + \tilde{a}_{v1,j})e + \sum_{j=1}^{L+1} (j-1)(\hat{a}_{1j} + \tilde{a}_{1j})e + \sum_{j=0}^L j(\hat{a}_{2j} + \tilde{a}_{2j})e.$$

10. Mean number of Type 2 customers in the queue.

$$\mu_{T2Q} = 2\tilde{x}(0)(I - R_1)^{-2}R_2e + \tilde{x}(0)(I - R_1)^{-2}(I + R_1)e.$$

Using Little's law we can obtain the mean waiting times in the queue of an admitted Type 1 customer and a Type 2 customer as follows.

11. Mean waiting time of an admitted Type 1 customer in the queue.

$$\mu_{WTQ}^{(1)} = \frac{\mu_{T1Q}}{\lambda_1(1 - P_{lost})}.$$

12. Mean waiting time of a Type 2 customer in the queue.

$$\mu_{WTQ}^{(2)} = \frac{\mu_{T2Q}}{\lambda_2}.$$

5.3.8 Numerical Examples

In this section we discuss the qualitative aspects of the *crowdsourcing* queueing model with working vacation under consideration through illustrative numerical examples. To verify the correctness and the accuracy of the code written to compute various system performance measures the results of Lemmas 5.3.1 and 5.3.3 were used. For the arrival process and service times, we consider the five sets of values for D_0 and D_1 and three *PH*-distributions with parameter (β, S) , respectively in Section 5.1.7.

The effect of *crowdsourcing* in the context of *MAP/PH/1* was studied

analytically and significant improvements of such a system were presented in Section. The purpose of our next example is to investigate such a significance for our model under study here. We will look at the ratio of the mean waiting time of Type 2 customers in the queue measure under study here. The ratio, $\frac{\mu_{WTQ}^{(2)}(p>0)}{\mu_{WTQ}^{(2)}(p=0)}$, will be of interest for the following example. In the sequel we will use the notation $CSQ - WV$ and $CSQ - PV$, respectively, for crowdsource with working vacation and crowdsource for pure vacation models.

Example 5.7: The effect of *crowdsourcing* in the presence of pure vacation and working vacation is studied in this example. Towards this we look at the two scenarios: (a) $p = 0$ which corresponds to having two independent arrival processes and (b) $p > 0$ which corresponds to the *crowdsourcing* model. Note that in the case of $CSQ - WV$, the server offers services (to only Type 1 customers) at a reduced rate during vacation in both the scenarios; however, for $CSQ - PV$ case the server does not offer services. First note that the mean waiting time of Type 2 customers in the queue, $\mu_{WTQ}^{(2,p=0)}$, when $p = 0$ is obviously larger than that of $\mu_{WTQ}^{(2,p>0)}$, for the case $p > 0$. Thus, the ratio $\frac{\mu_{WTQ}^{(2,p>0)}}{\mu_{WTQ}^{(2,p=0)}}$, will be in the interval $(0,1]$. We fix $\lambda_1 = 1, \mu = 1.1, \tau = 0.05, 0.1$ and $0.5, L = 10$ and vary p on the interval $(0, 1]$. Additionally, for $CSQ - WV$ case we fix $\theta = 0.2$ and vary N from 1 to 11. The values of λ_2 are chosen based on ρ and we look at two values: $\rho = 0.8$ and $\rho = 0.99$. The ratio is plotted in Figure 5.10 and Figure 5.11 under different combinations of arrival and service distributions. These figures reveal the following observations.

- As the vacation rate, τ , increases, we see the ratios for $CSQ - WV$ and $CSQ - PV$ approach those of CSQ model. This is to be expected as the vacation will get over faster. This is even more pronounced when $\rho = 0.99$.
- We notice the ratios are lower when services are of Erlang type compared to hyperexponential ones and shows the significant role played by the variability in the service times.
- As expected, the $CSQ - PV$ model appears to have a higher ratio when the

traffic intensity is not overly saturated.

- In the case of *MPCA* arrivals we notice that one needs a larger value for the vacation rate so that the ratio values will be closer to those of the *CSQ* model. This is especially when the traffic intensity is moderate and when services are of Erlang type.
- The effect of N (for the *CSQ* – *WV* case) for the various scenarios considered is seen when the inter-arrival times have more variability or are (positively) correlated.



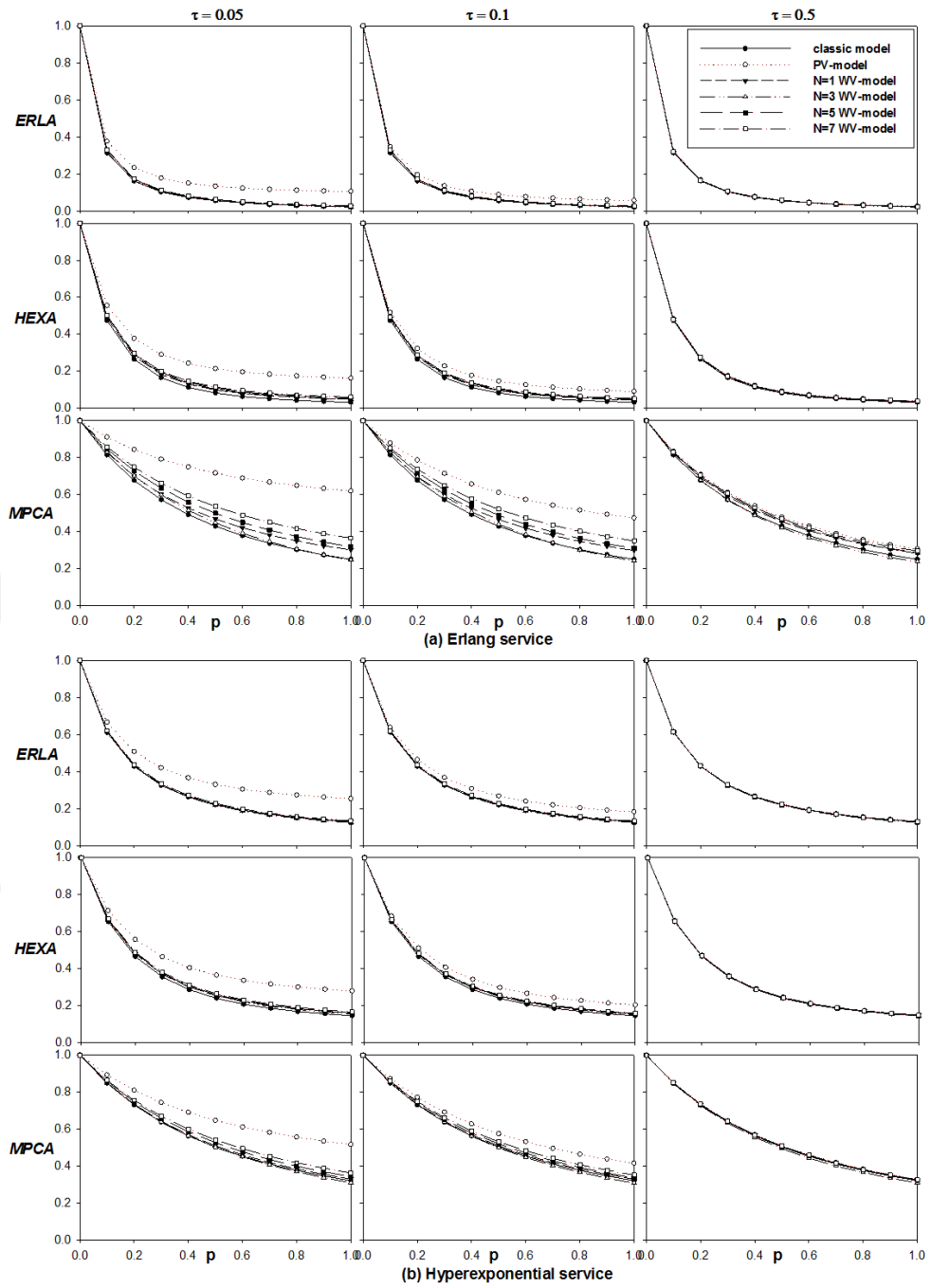


Figure 5.10 Ratios for mean waiting time of Type 2 customers for Erlang and Hyperexponential services for $\rho = 0.8$

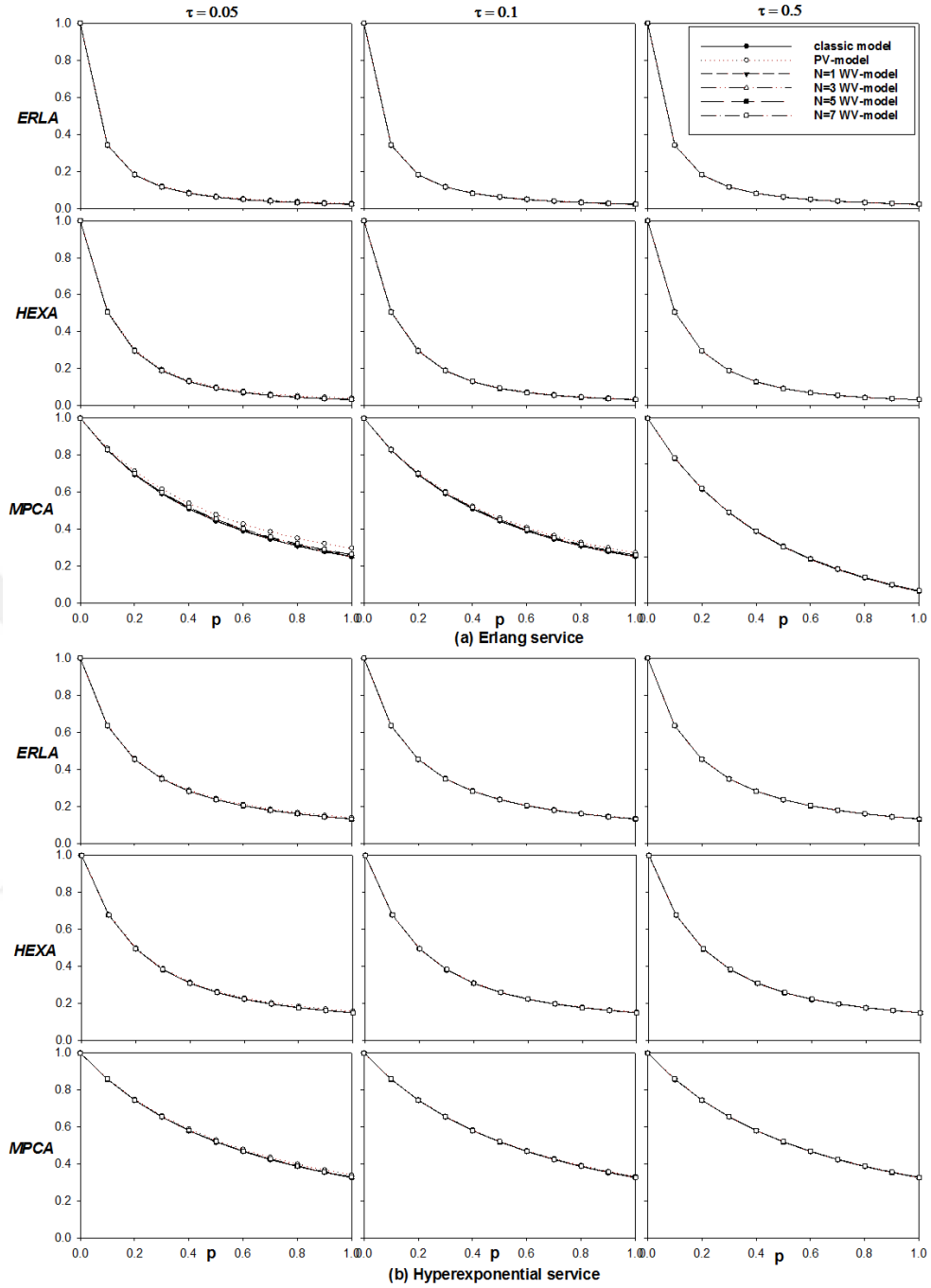


Figure 5.11 Ratios for mean waiting time of Type 2 customers for Erlang and Hyperexponential services for $\rho = 0.99$

In above sections, mentioned models it was shown that even in the case of small p there is a significant advantage in considering *crowdsourcing* by offering more traffic load through increasing the rate of Type 2 customers into the system. Here, we

will investigate a similar advantage for the current model under study by considering the cases when $L = 0$ and $L = 1$. In the former case, Type 1 customer is allowed only when the server is on vacation. Thus, the maximum number of Type 1 customers that can be present at any time in the system is 1 and 2, respectively.

Example 5.8: This example is very similar to Example 5.7 except that we look at the cases when (i) $L = 0, N = 1$, (ii) $L = 1, N = 1$ and (iii) $L = 1, N = 2$ for a few measures. Note that in case (i), Type 1 customer enters the system only when the server is on vacation and when that happens the customer is immediately served at normal rate. Also, there can be at most only one Type 1 customer in the system at any time. On the other hand, for case (ii) the system can have up to two Type 1 customers at any time. The server here only works at normal rate as in case (i). In contrast to two cases, (i) and (ii), the system can have the server work at a lower rate under case (iii). Towards this end, we will fix $\lambda_1 = 1, \mu = 1.1, \rho = 0.8, 0.99, \theta = 0.2$ and vary p on the interval $(0, 1]$ under different combinations of arrival and service distributions, and for $\tau = 0.1, 1.0$. In order to properly compare, we now look at the ratios $R_1 = \frac{\zeta(L=1, N=1, p)}{\zeta(L=0, N=1, p)}$ and $R_2 = \frac{\zeta(L=1, N=2, p)}{\zeta(L=0, N=1, p)}$, for some measure ζ . For example, if we are comparing the ratio R_1 for the measure μ_{T2Q} when $p = 0$, then we will look at the value of this measure for the combinations, $\{L = 0, N = 1, p = 0\}$ and $\{L = 1, N = 1, p = 0\}$, and compute the appropriate ratio.

In Figure 5.12 and Figure 5.13 below we display the ratios R_1 and R_2 , respectively, for selected measures and for representative scenarios. A quick look at these figures reveal the following observations.

- The impact of having a buffer for Type 1 customer (as opposed to not having) is clearly seen with regard to the two measures: μ_{T2Q} and $P_{T2LS} = \frac{R_{T2LS}}{\lambda_2}$, and as p increases to 1, the significance of the impact becomes less. This is the case for both moderate ($\rho = 0.8$) and heavy ($\rho = 0.99$) traffic intensities.
- Similarly the impact of server serving at a lower rate (this happens when $N = 2$) is clearly seen for the two measures.

- When comparing the impact of Erlang and hyperexponential services, we see some interesting trend. When the service times are of Erlang type, the ratio appears to decrease as p increases, whereas for hyperexponential type services, the ratio appears to increase as p increases. This is the case for all types of arrival processes.
- The ratios for the measures, P_{lost} and $P_{busy}^{(2)}$, are insignificant to p but are significant to the types of arrival and service times. This implies that having a buffer space, that is when $L = 1$, for Type 1 customer with ($N = 2$) or without ($N = 1$) the server having to serve at a lower rate is not affected by the fraction of Type 1 customers opting to serve Type 2 customers. This, initially might look somewhat surprising, but noticing that going from ($L = 1, N = 1$) to ($L = 1, N = 2$) the server serves at a lower rate only about 5% of the times (this is the maximum percentage when looking at all the combinations considered and the values for all the combinations are not displayed here).

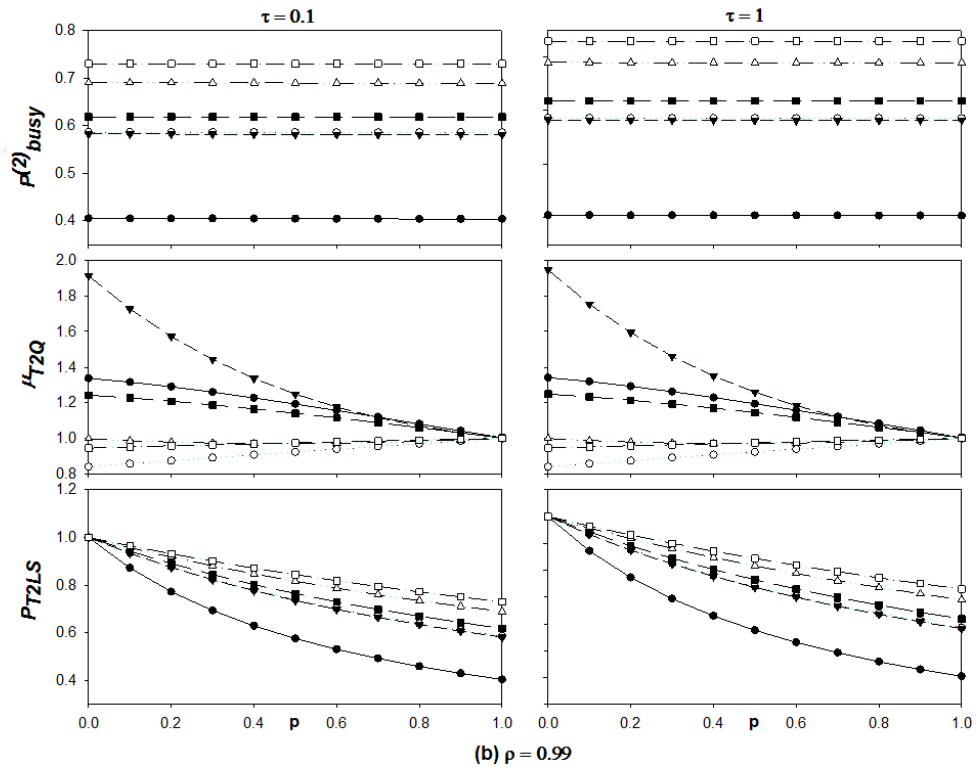
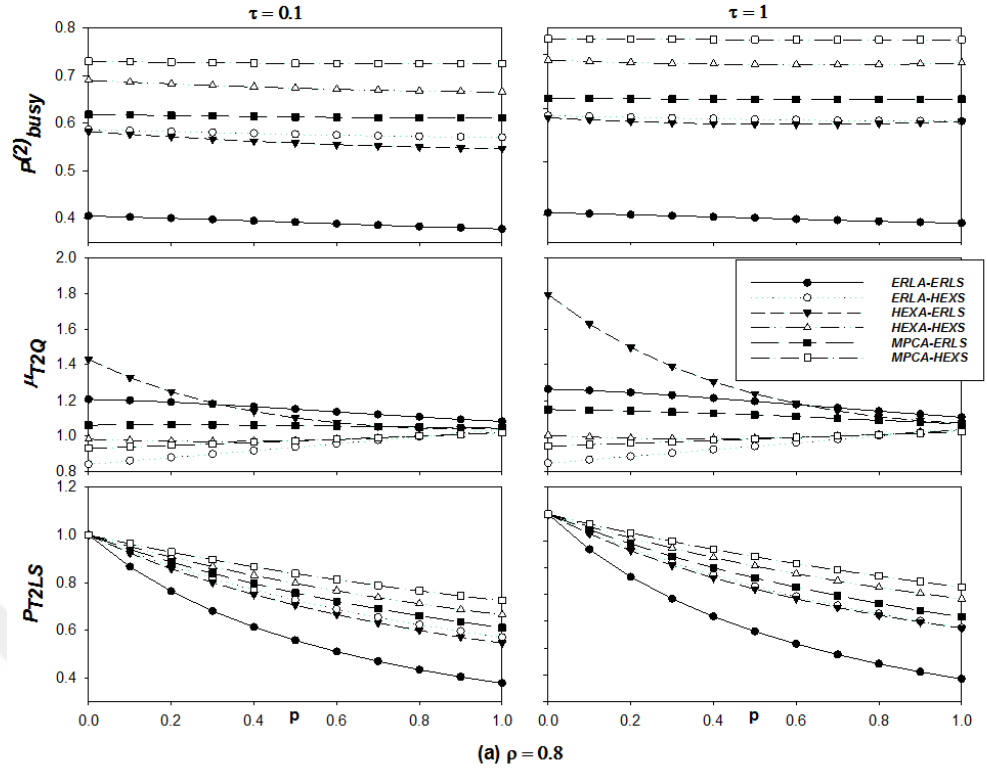


Figure 5.12 Comparison of the R_1 under different scenarios for $\rho = 0.8, 0.99$

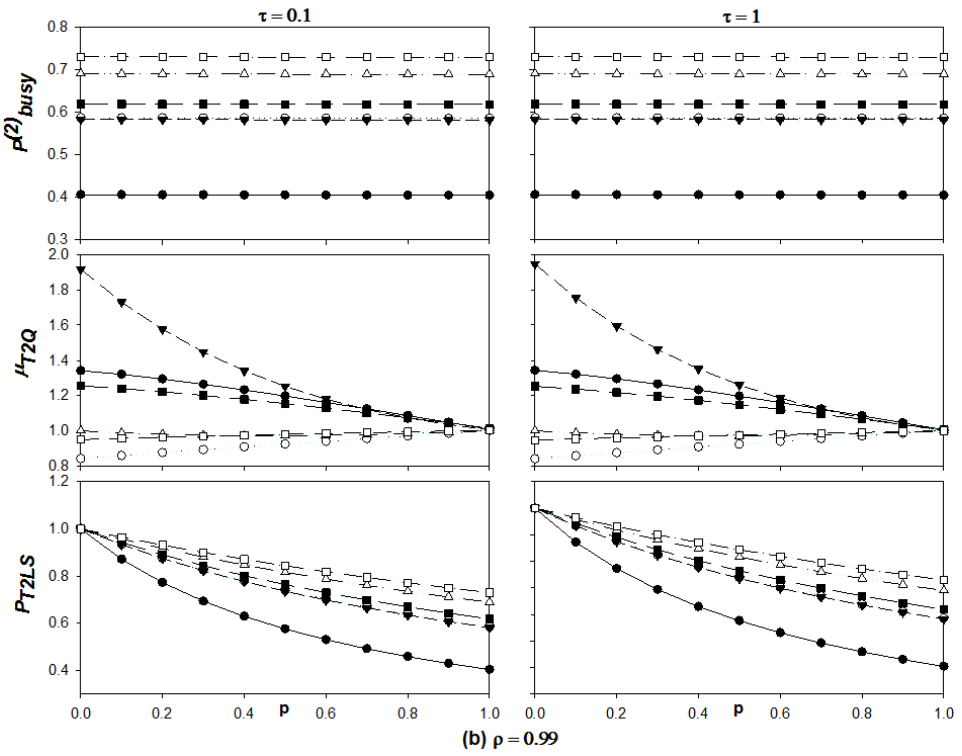
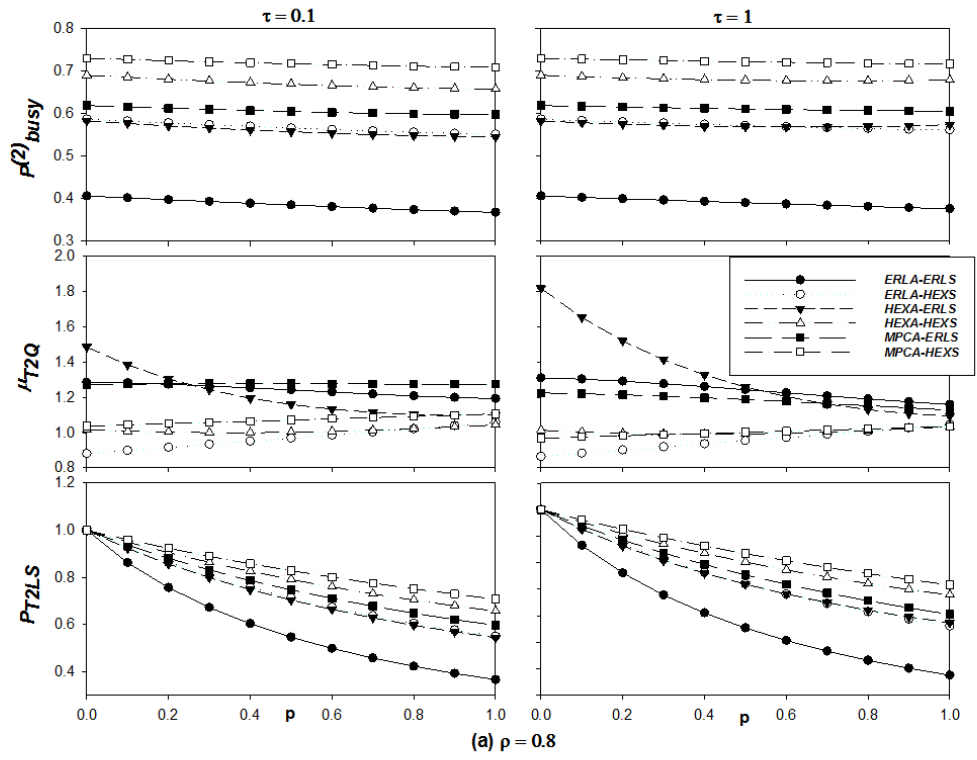


Figure 5.13 Comparison of the R_2 under different scenarios for $\rho = 0.8, 0.99$

Example 5.9: In this example we will discuss an optimization problem of interest. Suppose that r, c_1, c_2 and c_3 denote, respectively, the revenue per processed customer, the costs of holding a Type 1 customer and a Type 2 customer in the queue, and the cost using a Type 1 customer to serve a Type 2 customer, on a per unit of time basis. Without loss of generality we will fix $r = 1$. The goal of the optimization problem is to search for optimum values of p, L and N for the working vacation model and of p and L for the pure vacation model. The objective function, ETP , giving the expected total profit per unit of time, is to be maximized and the expression for ETP is as follows.

$$ETP = \lambda_1(1 - P_{lost}) + \lambda_2 - c_1\mu_{T1Q} - c_2\mu_{T2Q} - c_3\lambda_2P_{T2L1}.$$

Towards finding the optimum values, we fix $\lambda_1 = 1, \mu = 1.1, \theta = 0.2$ (for WV-model), $\tau = 0.1$ and $\rho = 0.99$. In Table 5.9 we display the optimum, (L^*, p^*, N^*) (for WV-model) and (L^*, p^*) (for V-model) under various scenarios. In the following we use the following notation for the cost vectors: A = (0.05, 0.001, 0.1); B = (0.05, 0.001, 0.2); C = (0.10, 0.001, 0.1); D = (0.10, 0.001, 0.2).

Table 5.9 Optimum values of (L^*, p^*, N^*) for WV-model, (L^*, p^*) for PV-model

<i>PH</i>	<i>Cost</i>	<i>ERLA</i>	<i>EXPA</i>	<i>HEXA</i>	<i>MNCA</i>	<i>MPCA</i>
<i>ERLS</i>	A	(5,1,1)-(5,1)	(5,1,1)-(5,1)	(6,1,1)-(6,1)	(6,1,1)-(6,1)	(10,1,1)-(10,1)
	B	(4,1,1)-(4,1)	(5,1,1)-(5,1)	(5,1,1)-(5,1)	(5,1,1)-(5,1)	(9,1,1)-(9,1)
	C	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(4,1,1)-(4,1)	(5,1,1)-(5,1)
	D	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(4,1,1)-(4,1)	(5,1,1)-(4,1)
<i>HEXS</i>	A	(6,1,1)-(6,1)	(7,1,1)-(6,1)	(7,1,1)-(7,1)	(7,1,1)-(7,1)	(10,1,1)-(9,1)
	B	(6,1,1)-(6,1)	(6,1,1)-(6,1)	(6,1,1)-(6,1)	(6,1,1)-(6,1)	(9,1,1)-(8,1)
	C	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(4,1,1)-(4,1)	(4,1,1)-(4,1)
	D	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)	(3,1,1)-(3,1)

All optimum occurs when $p = 1$ and $N = 1$ for WV-model; and for the PV-model, all optimum occurs when $p = 1$. These are as expected. The values of L^* depend on the type of arrival and service processes. Also, note that the values of L^* in Table 5.9

are identical to those of $MAP/PH/1 - CSQ$ model.

Next, we fix $p = 0.5$. That is, now we look at the case where only 50% of Type 1 customers opt to serve. In Table 5.10 we display the reduction in the percentages of ETP when going from the optimal point when $p = 1$ to $p = 0.5$ for the WV as well as PV models. While the reduction percentages for WV model are somewhat closer to those of the $MAP/PH/1 - CSQ$ model (see Section), the reduction percentages are significantly smaller for the PV model case as compared to WV model.

Table 5.10 Percentages of reduction in ETP at optimum values

PH	$Cost$	$CSQ - WV$			$CSQ - PV$		
		$ERLA$	$HEXA$	$MPCA$	$ERLA$	$HEXA$	$MPCA$
$ERLS$	A	24.65	23.01	17.55	24.50	22.88	17.40
	B	23.07	21.27	16.20	22.93	21.15	16.02
	C	24.09	21.12	17.25	23.95	21.01	17.06
	D	22.45	19.53	15.91	22.29	19.41	15.73
$HEXS$	A	23.54	23.09	19.75	23.31	22.92	19.51
	B	21.22	21.01	17.70	20.96	20.79	17.46
	C	21.15	19.98	15.91	20.97	19.84	15.63
	D	18.88	18.03	13.82	18.63	17.80	13.59

In Table 5.11, we also display the reduction in the percentages of ETP when going from the optimal point when $p = 1$ to $p = 0.5$ for various N , $2 \leq N \leq 5$ in the $CSQ - WV$ model. We see that the reduction (in the percentages) increases as value of N is increased.

Table 5.11 Percentages of reduction in *ETP* at optimum values for various *N*

<i>Cost</i>	<i>N</i>	<i>ERLS</i>			<i>HEXS</i>		
		<i>ERLA</i>	<i>HEXA</i>	<i>MPCA</i>	<i>ERLA</i>	<i>HEXA</i>	<i>MPCA</i>
A	2	24.54	22.90	17.26	23.30	22.94	19.42
	3	24.55	22.91	17.35	23.33	22.95	19.56
	4	24.56	22.92	17.43	23.46	22.96	20.26
	5	24.94	23.28	17.47	24.26	23.51	21.14
B	2	22.99	21.19	15.90	20.97	20.81	17.38
	3	23.01	21.19	15.99	21.00	20.82	17.52
	4	23.02	21.20	16.06	21.41	20.90	18.44
	5	23.51	21.74	16.14	22.44	21.69	19.49
C	2	24.01	21.07	17.09	21.06	19.92	15.74
	3	24.02	21.07	17.14	21.09	19.93	17.05
	4	24.96	22.30	17.25	23.19	21.74	19.10
	5	26.17	24.03	17.42	25.64	24.32	21.15
D	2	22.35	19.46	15.74	18.65	17.85	13.71
	3	22.37	19.47	15.81	19.14	18.07	15.41
	4	23.57	21.01	15.98	21.67	20.32	17.74
	5	24.78	22.73	16.17	24.14	22.90	19.88

CHAPTER SIX

CONCLUSION

In this thesis, the queueing systems with phase-type service have been studied. The systems have been analyzed by using *matrix geometric method* due to the all systems have the structure of *quasi-birth-and-death process*. The contribution this study can be summarized as follows.

The exact analytical solution to the $M/G/c/K$ model is only possible for special cases, such as for exponential service, a single server, or no waiting room. On the other hand, there are some approximations for the multi-server, finite-buffer and general service time assumptions. So, we have firstly studied the $M/C_2/c/K$ queueing system (**Model 1**).

Model 1: This research has developed a methodology for a solution to the multi-server and finite-buffer queueing system where the service time is modeled by the two-phase Cox distribution. The queueing system is characterized by the QBD process. The definition of the state variable of the system includes the number of customers in the system and the phase of the process. The solution to the multi-server queueing system with the defined Cox service process has been found with the matrix geometric method, and a closed-form expression is proposed for stationary probabilities. Two algorithms are proposed for the calculation of the level-dependent rate matrices for the boundary and repetitive states, respectively. When the balance equation for server number c and the normalizing condition are solved together, the probability of presence of c customers in the system in steady state π_c can be computed. After finding π_c , the other components of the invariant probability vector can be calculated by using the given theorem. Furthermore, an analytic expression is obtained for the mean number in the queue.

Finally, the proposed solution approach has been applied to the bed occupancy management problem in healthcare systems. Geriatric patients' length of stay is considered as the service time and the service time is modeled by the two-phase Cox

distribution. In this way, the problem is analyzed as a queuing system in which beds correspond to servers. Because the defined system allows patients to wait, it abolishes the constraint that the patients are lost when the capacity is full, which was assumed in the previous studies. Total cost is formulated as a combination of the patient's waiting cost and the cost of the unoccupied bed. Moreover, optimization is performed for different arrival rates and for different buffer sizes. Additionally, a simulation study is conducted over three different data sets. An extensive comparison between the methods included in the existing literature and the proposed method is presented. The comparison has been provided in terms of both such performance measures as loss probability and the MBU and cost.

It is concluded that the proposed method can be used for the solution to the $M/C_2/c/K$ queuing system with any given parameter. In addition, the proposed solution method may be applied to many other areas apart from healthcare systems, such as telecommunication or manufacturing systems.

For the $M/PH/1$ queue with working vacations, the performance measures in the stochastic decomposition structure were given conditionally. The condition is that the service times in the normal period have exponential distribution instead of phase-type distribution. Secondly, we have studied $M/C_k/1$ queue model with working vacation (**Model 2**) in which the performance measures have been given by the stochastic decomposition structure.

Model 2: In this study, the steady-state probability distribution for the $M/C_k/1$ queuing model with multiple working vacations has been obtained by using MGM . Performance measures such as the expected number in the system and the probability of being in the busy/vacation period are obtained. Besides, the stochastic decomposition structure is examined and the average number of customers in the system is expressed by some of the mean number of customers in the system without any vacation and the mean additional number of customers due to vacations. Finally, a sensitivity analysis of the model parameters is examined through a numerical study.

As a result of the numerical study, it can be concluded that μ_b has more effects on the performance measures than μ_v , since it determines the traffic intensity. The second factor which affects the performance measures is the number of phases and CV_b^2 of the service time distribution accordingly. By using stochastic decomposition, the average number in the system without any vacation and the mean additional number due to vacations can be calculated separately. This result provides an inference about the vacation period and can be used for the optimization of the industrial system.

Thirdly, we considered queueing systems useful in crowdsourcing (**Model 3**, **Model 4** and **Model 5**). Specifically, we considered the queueing models in which one type of customers may be available to serve another type of customers leading to more efficiency as well as to help the management to increase their productivity and hence revenues. Through illustrative numerical examples, it is shown that the significant benefits in introducing this type of variants to the classical queueing models.

Model 3: The queueing model incorporating the crowdsourcing concept in the sense of using the customers as possible servers after these customers are served by one of the system servers was first studied in Chakravarthy & Dudin (2017). The authors studied the $M/M/c$ model in which used exponential distribution to model arrival and service processes. We have generalized the model studied in Chakravarthy & Dudin (2017) by considering a more versatile point process, namely, Markovian arrival process (MAP) for Type 1 arrivals and use phase type distribution for services in the context of a single server. Thus, the model is of $MAP/PH/1$ -type with crowdsourcing. This model have been studied analytically using matrix-analytic methods.

Even for the small value of the probability introduced for in-store customers to opt for servicing the other type, we point out the significant advantage in considering crowdsourcing by offering more traffic load (through increased the rate of online customers without violating the stability condition) to the system resulting in more customers served (which in turn increasing the revenues when cost/profits are incorporated). Also, the role of correlation, especially a positive one, present in the

inter-arrival times in the system performance measures is highlighted.

Model 4: We have developed our **Model 3** by considering in the context of a multi-server. Thus, the model is of $MAP/PH/c$ -type with crowdsourcing and it has been studied by using simulation because of huge state space.

Model 5: We have again improved our **Model 3** with vacation/working vacation policies. In this model, we considered a queueing system useful in crowdsourcing and with server taking (multiple) vacations. Further, the server can offer services to a special group of customers during vacationing but at a lower rate. Using a versatile class of point process for arrivals and phase type services, we showed the significant benefits in introducing this type of variants to the classical queueing models through illustrative numerical examples.

REFERENCES

- Akar, N., Oguz, N., & Sohraby, K. (2000). A novel computational method for solving finite QBD processes. *Stochastic Models*, 16(2), 273-311.
- Akar, N., & Sohraby, K. (1997). Finite and infinite QBD chains: A simple and unifying algorithmic approach. In *INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution., Proceedings IEEE*, 3, 1105-1113.
- Artalejo, J. R., Gomez-Correl, A., & Qi-Ming, H. (2010). Markovian arrivals in stochastic modelling: a survey and some new results. *SORT: Statistics and Operations Research Transactions*, 34(2), 101-156.
- Asmussen, S. (2008). *Applied probability and queues*. Springer Science and Business Media.
- Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4), 419-441.
- Augustin, R., & Büsher, K. J. (1982). Characteristics of the Cox-distribution. *ACM Sigmetrics Performance Evaluation Review*, 12(1), 22-32.
- Baba, Y. (2005). Analysis of a GI/M/1 queue with multiple working vacations. *Operations Research Letters*, 33(2), 201-209.
- Baba, Y. (2010). The M/PH/1 queue with working vacations and vacation interruption. *Journal of Systems Science and Systems Engineering*, 19(4), 496-503.
- Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2003). *Bayesian modeling of hospital bed occupancy times using a mixed generalized Erlang distribution*. USA: In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting* (p. 443). Oxford University Press.
- Bernstein, M. S., Karger, D. R., Miller, R. C., & Brandt, J. (2012). Analytic methods for optimizing realtime crowdsourcing. *arXiv preprint arXiv:1204.2995*.

- Bhat, U. N. (2015). *An introduction to queueing theory (2nd ed.)*. New York: Springer.
- Bhattacharya, S., Celis, L. E., Chander, D., Dasgupta, K., Karanam, S., & Rajan, V. (2015). Crowds of crowds: Performance based modeling and optimization over multiple crowdsourcing platforms. *Human Computation*, 2(1), 105-131.
- Bhulai, S. (2006). On the value function of the M/Cox(r)/1 queue. *Journal of Applied Probability*, 43(2), 363-376.
- Bladt, M., & Nielsen, B. F. (2008). Lecture notes on phase type distributions.
- Chakravorthy, S. R. (2001). The batch Markovian arrival process: A review and future work. *Advances in Probability Theory and Stochastic Processes*. A. Krishnamoorthy et al. (ed) Notable Publications Inc., 21-39.
- Chakravorthy, S. R. (2010). Markovian arrival processes. *Wiley Encyclopedia of Operations Research and Management Science*. Published Online.
- Chakravorthy, S. R., & Dudin, A. N. (2017). A queueing model for crowdsourcing. *Journal of Operational Research Society*, 68(3), 221-236.
- De-Nitto Persone, V., & Grassi, V. (1996). Solution of finite QBD processes. *Journal of Applied Probability*, 33(4), 1003-1010.
- Elhafsi, E. H., & Molle, M. (2007). On the solution to QBD processes with finite state space. *Stochastic Analysis and Applications*, 25(4), 763-779.
- Evans, R. D., Gao, J. X., Mahdikhah, S., Messaadia, M., & Baudry, D. (2015). A review of crowdsourcing literature related to the manufacturing industry. *Journal of Advanced Management Science*, 4(3), 224-231.
- Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, 12(1), 11-26.
- Faddy, M. J. (1990). Compartmental models with phase-type residence-time distributions. *Applied Stochastic Models and Data Analysis*, 6(2), 121-127.

- Faddy, M. J. (1994). Examples of fitting structured phase-type distributions. *Applied Stochastic Models and Data Analysis*, 10(4), 247-255.
- Faddy, M. J., & McClean, S. I. (1999). Analyzing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15(4), 311-317.
- Faddy, M. J., & McClean, S. I. (2005). Markov chain modelling for geriatric patient care. *Methods Archive*, 44(3), 369-373.
- Fadıloğlu, M., & Yeralan, S. (2002). Models of production lines as quasi-birth-death processes. *Mathematical and Computer Modelling*, 35(7-8), 913-930.
- Faridani, S., Hartmann, B., & Ipeirotis, P. G. (2011). What is the right price? Pricing tasks for finishing on time. *Human Computation*, 11, 11.
- Forbes (2013). Retrieved March 26, 2016 from. <https://www.forbes.com/sites/erikamorpho-/2013/03/28/about-walmarts-idea-to-crowdsource-its-same-day-delivery-service/3a6ddbd05e5e>.
- Gorunescu, F., McClean, S. I., & Millard, P. H. (1999). Using a M/Ph/c queue to optimize hospital bed occupancy. *Proceedings of the Applied Stochastic Models and Data Analysis Conference*, 106-111.
- Gorunescu, F., McClean, S. I., & Millard, P. H. (2002a). A queuing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1), 19-24.
- Gorunescu, F., McClean, S. I., & Millard, P. H. (2002b). Using a queuing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4), 307-312.
- Graham, A. (1981). *Kronecker products and matrix calculus with applications*. Chichester, UK: Ellis Horwood.

- Griffiths, J. D., Price-Lloyd, N., Smithies, M., & Williams, J. (2006). A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics*, 17(3), 277-288.
- Gun, L., & Makowski, A. M. (1987). Matrix-geometric solution for finite capacity queues with phase-type distributions. Technical report.
- Hajek, B. (1982). Birth-and-death processes on the integers with phases and general boundaries. *Journal of Applied Probability*, 19(3), 488-499.
- Hasteer, N., Bansal, A., & Murthy, B. K. (2015). Crowdsourced software development process: Investigation and modeling through markov decision theory. *International Journal of Software Engineering and Its Applications*, 9(9), 41-54.
- He, Q. M. (2010). Construction of continuous time markovian arrival processes. *Journal of Systems Science and Systems Engineering*, 19(3), 351-366.
- He, Q. M. (2014). *Fundamentals of matrix analytic methods*. New York: Springer Science Business Media.
- Hochrainer, S., Hochreiter, R., & G., P. (2014). An algorithm for calculating steady state probabilities of M/Er/c/K queueing systems. *arXiv preprint arXiv:1401.4691*.
- Hosseini, M., Shahri, A., Phalp, K., Taylor, J., & Ali, R. (2015). Crowdsourcing: A taxonomy and systematic mapping study. *Computer Science Review*, 17, 43-69.
- Howe, J. Crowdsourcing: A definition. Retrieved June 02, 2006 from [http://www.crowdsourcing.com/cs/2006/06/crowdsourcing a.html](http://www.crowdsourcing.com/cs/2006/06/crowdsourcing%20a.html).
- Ke, J. C., Wu, C. H., & Zhang, Z. G. (2010). Recent developments in vacation queueing models: a short survey. *International Journal of Operations Research*, 7(4), 3-8.
- Kelton, W., Sadowski, R., & Swets, N. (2010). *Simulation with ARENA. Fifth ed*. New York: McGraw-Hill.
- Kim, J., Choi, D., & Chae, K. (2003). Analysis of queue-length distribution of the M/G/1 queue with working vacations. *International Conference on Statistics and Related Fields, Hawaii*.

- Kimura, T. (1996). A transform-free approximation for the finite capacity M/G/s queue. *Operations Research*, 44(6), 984-988.
- Kimura, T. (2000). Equivalence relations in the approximations for the M/G/s/s+r queue. *Mathematical and Computer Modeling*, 31(10-12), 215-224.
- Koole, G. (2004). A formula for tail probabilities of Cox distributions. *Journal of Applied Probability*, 41, 935-938.
- Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. American Statistical Association and Society for Industrial and Applied Mathematics.
- Li, J., & Tian, N. (2007). The M/M/1 queue with working vacations and vacation interruptions. *Journal of Systems Science and Systems Engineering*, 16(1), 121-127.
- Li, S. Q. (1989). Overload control in a finite message storage buffer. *IEEE Transactions on Communications*, 37(12), 1330-1338.
- Liu, W. Y., Xu, X. L., & Tian, N. S. (2007). Stochastic decompositions in the M/M/1 queue with working vacations. *Operations Research Letters*, 35(5), 595-600.
- Lucantoni, D., Meier-Hellstern, K. S., & Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22(3), 676-705.
- Lucantoni, D. M. (1991). New results on the single server queue with a batch markovian arrival process. *Communications in Statistics. Stochastic Models*, 7(1), 1-46.
- Marcus, M., & Minc, H. (1964). *A survey of matrix theory and matrix inequalities*. Boston, MA: Allyn and Bacon.
- Marshall, A. H., & M., Z. (2009). Simulating Coxian phase-type distributions for patient survival. *International Transactions in Operational Research*, 16(2), 213-226.

- Marshall, A. H., & McClean, S. I. (2003). Conditional phase-type distributions for modelling length of stay in hospital. *International Transactions in Operational Research*, 10(6), 565-576.
- Marshall, A. H., & McClean, S. I. (2004). Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science*, 7(4), 285-289.
- Marshall, A. H., & Zenga, M. (2012). Experimenting with the Coxian phase-type distribution to uncover suitable fits. *Methodology and Computing in Applied Probability*, 14(1), 71-86.
- Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16(4), 764-779.
- Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Baltimore, MD: The Johns Hopkins University Press [1994 version is Dover Edition].
- Neuts, M. F. (1989). *Structured stochastic matrices of M/G/1 type and their applications*. New York: Marcel Dekker.
- Neuts, M. F. (1992). Models based on the Markovian arrival process. *IEICE Transactions on Communications*, E75-B(12), 1255-1265.
- Neuts, M. F. (1995). *Algorithmic Probability: A collection of problems*. New York: Chapman and Hall.
- Nozaki, S. A., & Ross, S. M. (1978). Approximations in finite-capacity multi-server queues with Poisson arrivals. *Journal of Applied Probability*, 15(4), 826-834.
- Payne, K., Marshall, A. H., & Cairns, K. J. (2012). Investigating the efficiency of fitting Coxian phase-type distributions to health care data. *IMA Journal of Management Mathematics*, 23(2), 133-145.
- Quora (2015). Retrieved March 26, 2016 from. <http://www.quora.com/What-are-the-best-examples-of-crowdsourcing>.

- Ramaswami, V., & Lucantoni, D. M. (1985a). Algorithms for the multi-server queue with phase type service. *Communications in Statistics. Stochastic Models*, 1(3), 393-417.
- Ramaswami, V., & Lucantoni, D. M. (1985b). Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death processes. *Communications in Statistics. Stochastic Models*, 1(2), 125-136.
- Ramaswami, V., & Lucantoni, D. M. (1993). A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability*, 30(3), 650-674.
- Ross, S. M. (2014). *Introduction to probability models (11th ed.)*. Elsevier Inc.
- Schwartz, C., Borchert, K., Hirth, M., & Tran-Gia, P. (2015). Modeling crowdsourcing platforms to enable workforce dimensioning. *International Telecommunication Networks and Applications Conference (ITNAC)*, 30-37.
- Servi, L. D., & Finn, S. G. (2002). M/M/1 queue with working vacations (M/M/1/WV). *Performance Evaluation*, 50(1), 41-52.
- Sinclair, B. (2005). Coxian distributions. *OpenStax-CNX*.
- Smith, J. M. (2003). M/G/c/K blocking probability models and system performance. *Performance Evaluation*, 52(4), 237-267.
- Smith, J. M. (2007). Multi-server finite waiting room M/G/c/K optimization models. *Information Systems and Operational Research*, 45(4), 257-274.
- Sreenivasan, C., Chakravarthy, S. R., & Krishnamoorthy, A. (2013). MAP/PH/1 queue with working vacations, vacation interruptions and N policy. *Applied Mathematical Modelling*, 37(6), 3879-3893.
- Steeb, W. H., & Hardy, Y. (2011). *Matrix calculus and kronecker product*. Singapore: World Scientific Publishing.
- Stewart, W. J. (2009). *Probability, Markov chains, queues, and simulation- The mathematical basis of performance modeling*. New Jersey: Princeton University Press.

- Takahashi, Y. (1981). Asymptotic exponentiality of the tail of the waiting time distribution in a PH/PH/c queue. *Advances in Applied Probability*, 13(3), 619-630.
- Taylor, G. J., McClean, S. I., & Millard, P. H. (2000). Stochastic models of geriatric patient bed occupancy behavior. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1), 39-48.
- Tian, N., Zhao, X., & Wang, K. (2008). The M/M/1 queue with single working vacation. *International Journal of Information and Management Sciences*, 19(4), 621-634.
- Tian, N. S., Li, J. H., & Zhang, Z. G. (2009). Matrix analytic method and working vacation queues - a survey. *International Journal of Information and Management Sciences*, 20(4), 603-633.
- Tian, N. S., & Zhang, Z. G. (2006). *Vacation Queueing Models: Theory and applications*. New York: Springer Publishers.
- Wang, J., Faridani, S., & Ipeirotis, P. G. (2011). Estimating the completion time of crowdsourced tasks using survival analysis models. *Crowdsourcing for search and data mining (CSDM 2011)*, 31.
- Wolff, R. W. (1989). *Stochastic modeling and the theory of queues*. New Jersey: Prentice-Hall: Englewood.
- Wu, D., Rosen, D. W., & Schaefer, D. (2014). Modeling and analyzing the material flow of crowdsourcing processes in cloud-based manufacturing systems using stochastic petri nets. In *ASME 2014 International Manufacturing Science and Engineering Conference collocated with the JSME 2014 International Conference on Materials and Processing and the 42nd North American Manufacturing Research Conference*, (V001T04A011--V001T04A011). American Society of Mechanical Engineers.
- Wu, D. A., & Takagi, H. (2006). M/G/1 queue with multiple working vacations. *Performance Evaluation*, 63(7), 654-681.

- Xiu, C., Tian, N., & Liu, Y. (2010). 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.
- Yao, D. D. W., & Buzacott, J. A. (1985). Queueing models for a flexible machining station part II: The method of coxian phases. *European Journal of Operational Research*, 19(2), 241-252.
- Zhang, M., & Hou, Z. (2011). Performance analysis of MAP/G/1 queue with working vacations and vacation interruption. *Applied Mathematical Modelling*, 35(4), 1551-1560.
- Zhang, Z. J., & Xu, X. L. (2008). Analysis for the M/M/1 queue with multiple working vacations and N-policy. *International Journal of Information and Management Sciences*, 19(3), 495-506.