2008

Modeling and analysis to improve the quality of healthcare services

Sharmin Shahriar Akhter
University of Windsor

Modeling and Analysis to Improve the Quality of Healthcare Services

By

Sharmin Shahriar Akhter

A Thesis

Submitted to the Faculty of Graduate Studies and Research

through the Department of Industrial and Manufacturing Systems Engineering
in Partial Fulfillment of the Requirements for

the Degree of Master of Applied Science at the

University of Windsor

Windsor, Ontario, Canada
2008

Your file Votre référence
ISBN: 978-0-494-47017-6
Our file Notre référence
ISBN: 978-0-494-47017-6

AUTHOR'S DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

For many healthcare services or medical procedures, patients have extensive risk of complication or face death when treatment is delayed. When a queue is formed in such a situation, it is very important to assess the suffering and risk faced by patients in queue and plan sufficient medical capabilities in advance to address the concerns. As the diversity of care settings increases, congestion in facilities causes many patients to unnecessarily spend extra days in intensive care facilities. Performance evaluation of current healthcare service systems using queueing theory gains more and more importance because of patient flows and systems complexity. Queueing models have been used in handsome number of healthcare studies, but the incorporation of blocking is still limited. In this research work, we study an efficient two-stage multi-class queueing network system with blocking and phase-type service time distribution to analyze such congestion processes. We also consider parallel servers at each station and first-come-first-serve non-preemptive service discipline are used to improve the performance of healthcare service systems.

DEDICATION

This thesis is dedicated to my husband, **Mesbah Khan** who has supported me all the way since the beginning of my studies. I also dedicate this thesis to my beloved son, **Samin Khan** and my beloved daughter, **Maleeha Khan** who offered me unconditional love and support throughout the course of this thesis.

I wish to dedicate this thesis to my late mother, **Lily Ahmed** and my late father, **A. N. Minhaj Uddin Ahmed**. They were a constant source of inspiration to my life. Although they are not here to give me strength and support I always feel their company, which supports me to struggle to achieve my goals in life. They taught me to endure and prepared me to face challenges with devotion and modesty.

Finally, I would like to dedicate this thesis to all those who believe in the affluence of learning.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my advisor Dr. Guoqing Zhang for his constant encouragement and support throughout my years at University of Windsor. This thesis work would not have been possible without his dedication for quality research and his guidance in finding solutions to research problems.

Sincere gratitude and many thanks to my thesis committee members, Dr. Mohammed A. S. Khalid, Dr. Myron Hlynka and Dr. Mike Hunglin Wang for their helpful comments and suggestions. I am deeply grateful to Dr. Waguih ElMaraghy, Professor & Head, Department of Industrial and Manufacturing Systems Engineering (IMSE), University of Windsor, for his cooperation and giving me mental support.

I like to thank Ms. Jacquie Mummery, Ms. Brenda M. Schreiber, Mr. Ramadan Barakat and Mr. David Mckenzie for their continuous support in the creation of this thesis and also my fellow colleagues for their cooperation.

Last but not least, I am grateful for the continuous love and support of my family. My husband Mesbah Khan helps me generously with almost every aspect of my life. So special thanks to my husband for brightening my life and for lifting up my spirits. Also my special thanks go to my son Samin Khan and daughter Maleeha Khan for their help and patience.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

ACRONYMS:

BAS	Blocking After Service
BBS	Blocking Before Service
CPU	Central Processing Unit
CTMP	Continuous Time Markov Process
FCFS	First Come First Serve
ICU	Intensive Care Unit
MTBF	Mean Time between Failure
MTTR	Mean Time to Repair
PH	Phase type probability distribution
QBD	Quasi Birth and Death process
QN	Queueing Networks
P.M.	Performance Measures

CAPITAL LETTER:

A	Arrival process
A_0	Sub-Block matrices of matrix Q
A_1	Sub-Block matrices of matrix Q
A_2	Sub-Block matrices of matrix Q

B	Capacity of the buffer/queue
C_i	Stationary probability matrices
D_i	Matrices that are on the diagonal of the generator matrix
E_0	Sub-Block matrices of matrix Q
F_1	Sub-Block matrices of matrix Q
F_2	Sub-Block matrices of matrix Q
$E(L)$	Expected mean queue length
$E(S)$	Expected mean sojourn time
I	Identity matrix of order $v_1 x v_2$
K	Population size
L_i	Matrices that are under the diagonal of the generator matrix
Q	Transition or infinitesimal matrix
SD	Service discipline
S_1	PH-distribution service time for patient type A at station 1 with order v_1
S_2	PH-distribution service time for patient type B at station 1 with order v_1
T_1	PH-distribution service time for patient type A at station 2 with order v_2
T_2	PH-distribution service time for patient type B at station 2 with order v_2
U_i	Matrices that are above the diagonal of the generator matrix

SMALL LETTER:

n	Number of servers
e	Row or column vector of one's
l	Mean queue length
n_1	Number of patients at station 1
n_2	Number of patients at station 2
m	Total patients at station 2 (including patients in service)
s_1	Number of states of station 1
s_2	Number of states of station 2
v	Number of phases of stations
v_1	Number of phases to complete service at station 1
v_2	Number of phases to complete service at station 2
w_q	Mean waiting time in the queue

GREEK LETTERS:

α	Initial vector for any station with order v
α_1	Initial vector for patient type A at station 1 with order v_1
α_2	Initial vector for patient type B at station 1 with order v_1
β_1	Initial vector for patient type A at station 2 with order v_2
β_2	Initial vector for patient type B at station 2 with order v_2
λ	Patient's arrival rate

λ_1	Patient arrival rate for patient type A at system 1
λ_2	Patient arrival rate for patient type B at system 1
μ	Mean processing or service time of server
ρ	Server utilization factor
θ	Vacation rate of servers at station
θ_1	Vacation rate of servers at station 1
θ_2	Vacation rate of servers at station 2
γ	Service resume rate of servers
γ_1	Service resume rate of servers at station 1
γ_2	Service resume rate of servers at station 2

CHAPTER 1

INTRODUCTION

1.1 Background:

Health care resources are becoming increasingly limited and expensive, thereby placing greater attention on the productivity of the resources and the corresponding level of service provided to patients. As a consequence, one of the important operational issues in health care involves capacity with the goals of high resource deployment and congregation of high quality services. Hospital waiting lists are also a major problem. It is very persuasive to imagine of them simply as queues of patients waiting for health services which can be lowered by simply providing more service. In simple terms, waiting lists will reduce if the throughput at each of these phases exceeds the demand. Unfortunately, one important feature of these waiting lists is the way in which the demand for service often increases when waiting list size or waiting times decrease.

There are some factors that are used to measure the performance of health care services. Some of those important performance measures of patients flow are cycle time, throughput, mean queue length and mean waiting time in the queue. A great amount of research has been conducted to develop analytical and numerical models of these performance measures in manufacturing systems and communication systems, but mathematical literature in healthcare performance is rare, though there is a significant amount of literature on using simulation to study health care clinics. Many opportunities

currently exist to apply the quantitative tools of operations management or management science to support health care decision making. As is the case with any successful application of these tools, an understanding of the underlying environment is essential. The use of queueing network techniques allows us to capture the stochastic nature of arrivals and service times that is typical in the health care system.

Nobody likes to wait in a queue – especially when the queue is formed in the emergency department of a hospital. There are two ways to improve system performance of health care services: one way is providing queues (Buzacott, 1971) to allow the other stations to continue working when one of them is down; the other way is providing some identical parallel servers, since during the vacation time of a server, another can then take over its function. A health care service system may consist of only one server or several servers each of which performs the same operation.

1.2 Objective:

The administrators of health care facilities are continually faced with the task of balancing the delivery of quality healthcare with the appropriate allocation of their resources. Although considerable research has been expended on health care services, the techniques and models developed to date have concentrated on simple systems where restricted aspects of the problem are considered. A majority of the previous work on the subject is concerned with health care service models having balanced, single server work stations with no vacation time. In systems where server vacation times are possible, the

blocking of the stations is not considered and the service time is restricted to the exponential distribution. "Blocking" denotes situations where patients are turned away from accommodations to which they are referred, are thus forced to remain in their present facilities until space becomes available (Koizumi et al., 2005). Queueing models have been used in handsome number of healthcare studies, but the incorporation of blocking is still limited as mentioned by Koizumi et al. (2005). This blocking phenomenon often generates some financial loss because the blocking clients are, in many cases, ready to move to a less intensive and hence less expensive facility. Thus, controlling the congestion in the existing healthcare service system is important not only from the clinical and human rights perspectives but also from the budgetary perspective for health policy makers.

In this thesis, a model is used to conduct a detailed study of how the length of server blockings, server vacation times, and inter-stage storage capacity can affect the performance of healthcare services with two stations. This thesis presents a numerical method for the performance evaluation of health care service systems with phase type service time distribution, vacation time and blocking. In factories, the server at any workstation (due to repair) can be affected by different kinds of vacation time which occur with different frequencies and require different lengths of time (Gershwin, Dallery, Papadapoulos and Smith, 2000). Exponentially distributed vacation times are considered in this research work.

Recently, many researchers have contributed towards the understanding of complexities present in healthcare service systems through the use of a variety of modeling tools such as simulation, Markov chains, Petri nets and queueing. Queueing models have been used in healthcare services systems. Analytical models analyze the system using mathematical or symbolic relationships. These are then used to develop a formula or to define an algorithm or computational procedure by which the performance measures of the system can be calculated. Analytical models can also be used to demonstrate properties of various operating rules and control strategies. The development of analytically tractable models to determine their performance is vitally important. Matrix analytic methods play the same role. "Matrix analytic methods constitute a success story, illustrating the enrichment of a science, applied probability, by a technology, that of digital computer" (Latouche and Ramaswami, 1999). Neuts (1973) also wrote: "To do work in computational mathematics is ... a commitment to a more demanding definition of what constitutes the solution to a mathematical problem. When done properly, it conforms to the highest standard of scientific research." The continuous time Markovian arrival process has been widely used to model the source behaviour of data traffic, while phase type (*PH*) distribution has been extensively applied to model the service time. This thesis focuses on the computation of the *M/PH/1* and *M/PH/n* queueing systems, in which the arrival process is considered to be Markovian and the service time obeys a continuous *PH* distribution. Such a queueing model has potential in performance evaluation of healthcare service systems. Based on matrix-analytic methods, computer programs have been developed for obtaining the stationary distribution of the system

numbers and further deriving the key performance indices of the *M/PH/1* and *M/PH/n* queueing systems.

The objectives of this research are given below:

1. To perform an efficient study of the effect of server failures, blocking, finite buffer capacities in front of the second station, service times with phase type distribution at each stage;
2. To construct a model of a healthcare service system using Queueing Theory, and to use this model in the analysis of the healthcare services system or medical procedures;
3. To provide the mathematical procedures, associated with the justification of convergence and accuracy.
4. To provide the concept and the analysis of the *M/PH/1*, *M/PH/n* queueing systems with several additional factors using a numerical method;
5. To provide the concept of the first-come-first-serve service discipline to handle multiple classes of patients.
6. To implement a computer program to demonstrate the analysis and calculations.
7. To demonstrate the applicability of this type of queueing system in life.

It is postulated that the results from this research will help to develop a better understanding of the effect of the above design factors on the efficient operation of healthcare systems.

1.3 Organization of This Thesis:

This research attempts to overcome some of the shortcomings of earlier work using an analytical approach. This thesis is organized in the following manner:

Chapter 1: Introduction - This chapter consists of three sections. These are – (i) Background of this research work; (ii) Objective of this thesis; and (iii) Organization of this thesis.

Chapter 2: Review of Literature – This chapter presents previous work related to this research according to different modeling techniques. Also the summary of previous work related to this research is included at the end of this chapter. This chapter is divided into three sections. In the first section, previous work related to performance evaluation of healthcare service systems is described. Then in the second section, summary of previous work is categorized according to the used parameters. Finally, conclusions are presented in the last section.

Chapter 3: Model Formulation – This chapter is divided into six sections. An open queueing model is developed. To develop the model, associated terms are explained in different sections.

Chapter 4: Analysis of Patients' Flow Line Systems with Server Vacation Times and Blocking – In this chapter, a two nodes *M/PH/1 and M/PH/n* queueing system subject to blocking and unreliable server is analyzed by using the state model method and phase type service time distribution.

Chapter 5: Conclusions and Future Research – In this chapter, conclusions, contributions and future research on this research topic are described.

CHAPTER 2

REVIEW OF LITERATURE

Performance evaluation of current healthcare service systems using queueing theory gains more and more importance because of patient flows and system complexity. Usually it is very difficult or even impossible to evaluate the performance by system measurement. A great number of researchers proposed many models and methods to analyze healthcare service systems with unreliable servers and finite waiting area. Previous work related to this research work is described in the following sections.

2.1 Review Of The Models Developed Previously Related To Healthcare Service Systems:

Systems models are used to evaluate the performance of healthcare service systems. The following figure 2.1.1 represents the modeling techniques that were used in previous research to solve the stochastic models:

2.1.1 Related Works Developed on Queueing Models:

Previous research relevant to performance analysis of the healthcare service systems using queueing network modeling techniques are described below:

Queueing networks with blocking are typically very difficult to analyze, and closed form results at steady state (or otherwise) are usually not available. Hunt (1956)

began the theory of tandem queues with blocking and its application to practical problems. Although a rigorous Markovian analysis was performed early on by Hunt (1956), to date noticeably few results have been obtained, with the bulk of the work focusing on continuous-time models. In this paper, Hunt studied the blocking influence in sequential arrays of waiting lines with Poisson arrivals and exponential service times. The author did not discuss the reliability of the system in this paper.

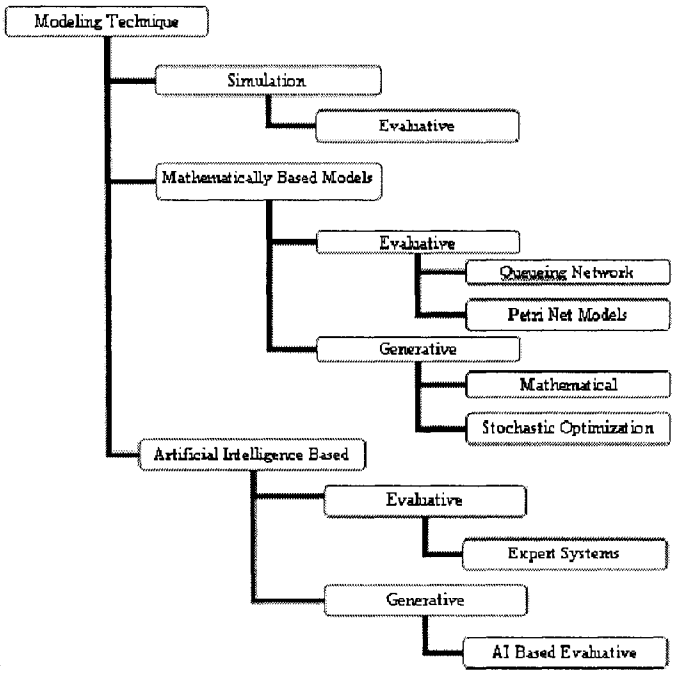

Figure 2.1.1: Modeling techniques for queueing network from model developer's point of view

Evaluating the performance of unreliable systems is important. There are two ways to improve performance of unreliable systems: one way is providing queues (Buzacott, 1971) to allow the other stations to continue working when one of them is

down; the other way is providing some identical parallel servers, since during a vacation an available server can then take over its function. In Buzacott (1967), the author studied both with and without inter-stage storage queues.

George et al. (1983) established a method of analyzing the general surgery waiting list problem at hospital and district level. In a public health system, one of the problems is the size of the waiting list for admission to hospital. This paper concentrated on a linear programming model to plan the aggregate throughput of the general surgical department. Since the largest number of patients and the longest waiting lists are for general surgery, this study concentrated (although not exclusively) on this specialty.

Shimatonis (1983) studied a simulation model for a medical organization, which is an application to a cardiological outpatient department. A medical organization is usually a complex multilevel dynamic system with interacting elements. The complexity is particularly increased by the stochastic nature of the flux of incoming patients and by the probability characteristics of the main and associated diseases. In this paper, a multi-channel queueing system is simulated.

Panayiotopoulos and Vassilacopoulos (1984) studied a queueing system with the following characteristics: (i) General independent inter-arrival times; (ii) General Service-time distribution; (iii) Limited waiting room; (iv) Patient priorities increase up to a certain number; (v) Time dependent number of servers (doctors); (vi) Infinite patient population; (vii) Each server meets the system only once within a certain period of time,

while the total number of the available servers is known. This research considered a hospital emergency department.

Worthington (1987) used some research on queueing models to investigate some of the management implications of feedback in waiting lists. The authors estimated fairly precisely the implications of certain management actions for waiting lists in which arrival rates decrease linearly as the waiting-list size increases. The queueing model used in this paper was denoted by *M/G/n*.

Gün, and Makowski (1989) considered a two node tandem queueing system with phase type servers and Bernoulli arrivals where the servers operate in discrete-time and subject to blocking and failure. In this paper, the algorithm combines a decomposition/ aggregation technique with exact analytical results for two node systems.

Albin et al. (1990) showed how a queueing network model helped to uncover causes of delay in a health center appointment clinic. An open network of queues was used to model the system. Single-server, infinite-capacity nodes were defined for the front desk, the screening area, and each of two physicians and three nurse practitioners. They collected data on service times, external arrival times and routing. QNA, a general queueing network software package, was used to analyze the system.

Worthington (1991) described work carried out with hospital consultants to develop and use computerized waiting list management models. Simple stocks-and-flows

models have been found to be capable of providing useful 'What if?' models of hospital waiting list systems. Worthington (1991) used a spreadsheet development tool to analyze the queueing model. The author concluded that while spreadsheets have been a good development tool, it has been necessary to reprogram the models in a more general computer language in order to achieve the flexibility required to apply them efficiently to other waiting list problems.

Liyanage and Gale (1995) studied a queueing model to design an appropriate service facility for the Campbelltown hospital emergency service in order to minimize patients waiting time while controlling the associated running cost. The authors developed a computer program to model the distributions of arrival time, waiting time and service time of the system and to estimate their parameters. Since the main concern of Liyanage and Gale (1995) is on the non urgent queue, the authors first designed this particular queue. An *M/M/n* queue is considered in this paper. These service times were mutually independent of the arrival process. The waiting space for the customers had an unlimited capacity. Customers were serviced on a first come first served basis.

Bretthauer and Côté (1998) presented a general model and solution methodology for planning resource requirements (i.e. capacity) in a health care organization. To illustrate the general model, the authors considered two specific applications: a blood bank and a health maintenance organization (HMO). To determine resource requirements, the authors developed a queueing network model that minimizes capacity costs. In this queueing network, six work stations are considered and stations 2, 4 and 6

have unlimited capacity. General inter-arrival time and general service time distributions are considered in this model.

Jun et al. (1999) presented a survey of the literature on applying discrete event simulation to understand the operation of health care facilities. This survey showed a large amount of research has been conducted in the area of patient flow as well as resource allocation. The numerous simulation studies reported in the literature have the common theme that they each attempted to understand the relationship that may exist between various inputs into a health care system (for example, patient scheduling and admission rules, patient routing and flow schemes, facility and staff resources) and various output performance measures from the system (for example, patient throughput, patient waiting times, physician utilization, staff and facility utilization). The breadth and scope of units within hospitals and clinics makes it impossible to undertake one single comprehensive study that addresses all these issues simultaneously.

Kim et al. (1999) analyzed the admission-and-discharge processes of one particular intensive care unit (ICU) within a classical steady-state queueing framework, by using computer simulation models built with actual data from the ICU. The results provided insights into the operations management issues of an ICU facility to help improve both the unit's capacity utilization and the quality of care provided to its patients. In this paper, exponentially distributed inter-arrival times and service times are considered.

Green and Nguyen (2001) developed insights on the impact of size, average length of stay, variability, and organization of clinical services on the relationship between occupancy rates and delays for beds. Most of the analyses of Green and Nguyen (2001) used an *M/M/n* queueing model to estimate delays. The model arrivals (patient demands for beds) occur according to a Poison process and that the service duration has an exponential distribution.

Gorunescu et al. (2002) described the movement of patients through a hospital department by using classical queueing theory and presented a way of optimizing the use of hospital resources in order to improve hospital care. A queueing model was used to determine the main characteristics of the access of patients to hospital, such as mean bed occupancy and the probability that a demand for hospital care is lost, when all beds are occupied. This paper used results from queueing theory to determine optimal bed number for a hospital system, where they described patient arrivals by a Poisson process, hospital beds assumed as servers and lengths of stay were modeled using phase type distributions. This paper considered an *M/PH/n* queue in which the number of beds was fixed and no inter-stage queue was allowed.

Isotupa and Stanford (2002) considered a single server queue that handles arrivals from N classes of customers on a non-preemptive priority basis. Each of the N classes of customers features arrivals from a Poisson process at rate λ_i and class-dependent phase type service. To analyze the queue length and waiting time processes of this queue, the

authors derived a matrix geometric solution for the stationary distribution of the underlying Markov chain.

Most applications of queueing with blocking have been done in the engineering fields (Koizumi, 2002).This model provided guidance for congestion management in the provision of public services. The work presented in this dissertation applied to the simplest blocking model and a simplified structure of the real world mental health system. This research considered an *M/M/n* queueing model to analyze the health care system.

The purpose of Preater (2002) was to advertise the bibliography contained in Preater (2001). Preater (2001) presented a bibliography of 166 papers that illustrated the wide variety of applications of queueing theory to health care and medicine. This bibliography included studies on: specific hospital departments or functions (out-patients, radiology, accident and emergency, intensive care, pharmacy, dentistry, etc.); ambulance location and dispatch problems; the management of elective and emergency surgery; compartmental modeling of pharmacokinetics and of patient flow; hospital campus design; and health service manpower planning.

Using queueing theory, McManus and Long (2004) constructed a mathematical model of patient flow, compared predictions from the model to observed performance of the unit and explored the sensitivity of the model to changes in unit size. A computer simulation model of Intensive Care Unit (ICU) flow was constructed using spreadsheet

software (Excel 2000®; Microsoft Corporation, Redmond, WA) and standard queueing formulae. The ICU was modeled as a multi-channel, single-stage system of identical parallel servers that process randomly patterned arrivals according to exponentially distributed service times. Each ICU bed was treated as one server and a "first come, first served" queueing discipline was assumed. It was further understood that no waiting line was possible for these critically ill patients and, therefore, the probability of waiting equals the probability of rejection. In the queueing literature, this model is denoted as *M/M/n/c* (Shorthand notation for a system with a Poison arrival process, exponential service times, c servers, and c spaces in the system).

In Wang (2004), the author developed a patient queue model that considers the condition and its changes over time for a patient in a queue. The risk faced by a patient is characterized under this model as a function of the arrival rate, the service capacity and the hazard rate of the disease. The objective of this study was to model and analyze a patient queue where patients wait for a medical treatment to treat an illness or to alleviate conditions that make them suffer. In this paper, patient arrival followed a Poisson distribution and service time followed an exponential distribution.

Gonzalez-Rubio (2005) presented a discrete-event, macro model of a geographical region where a population has access to general practitioners (GP) and to a hospital emergency room (ER). This paper designed a queuing model to simulate a pool of primary care clinics of a region where patients interact with doctors if available otherwise they may end in a hospital emergency room. This simulation model provided an insight on the expected results of different scenarios in terms of number of patients

treated in clinic, waiting to be treated, or treated in the emergency room. Each scenario is associated with a cost for evaluation purposes. The paper presented a first-model as a feasibility proof of the concept and included a critical view of the input data requirement and computing simulation costs of a comprehensive system. This work represented a first step of a larger project that is called Soleil. The aim of Soleil is to propose a decision aid system for health care management.

Although the diversity of care settings increased, congestion in facilities caused many patients to unnecessarily spend extra days in intensive facilities. Koizumi et al. (2005) studied a queueing network system with blocking to analyze such congestion processes. Two performance indicators, the number of patients waiting to enter each type of facility in the system and associated waiting time at the steady state, are derived in the steady-state analysis. For mathematical convenience, total arrival rates are often assumed to follow a Poison distribution with mean λ (so that inter-arrival times follow an exponential distribution with mean $1/\lambda$). Similarly, mean service times are typically assumed in this paper to follow an exponential distribution with mean $1/\mu$.

Atkinson et al. (2006) considered a queueing model and described the deployment of emergency medical service along a highway. The authors proposed two heuristic methods for the approximate evaluation of stationary loss probability and utilization of ambulance cars. This paper considered c medical service stations placed along a highway. The developed model generalized the well-known *M/M/n* system with loses.

Au-Yeung et al. (2006) formulated a (simplified) hierarchical Markovian queueing network model of patient flow in the Accident and Emergency department of a major London hospital. This paper developed the simplified multi-class queueing network model of patient flow. This model took the form of a hierarchical network of *M/M/n* queues. This queueing model had four customer classes. In this paper, the author used a simulation model to provide some insights into the effects of prioritizing different classes of patients in a real Accident and Emergency unit.

Chaussalet (2006) studied patient flow in health care systems with bed capacity constraints in order to provide a useful decision aid for health service managers. The author modeled the patient flow of health care systems using a closed queueing network framework with the assumption that the system is always full. Key performance measures of the health core system are derived. Chaussalet (2006) demonstrated that the model could help to improve decision-making by allowing managers to explore different options and evaluate their impacts on performance. The findings of this paper highlighted the importance of policy makers taking into account the interactions between different phases of care. Chaussalet (2006) developed a novel approach to modeling the flow of patients through health care systems with constrained bed capacity.

Cochran and Bharti (2006a) proposed a four-stage stochastic methodology for bed planning in hospitals using queueing networks and discrete event simulation models. This paper wanted to balance bed unit utilizations across an obstetrics hospital and minimize the blocking of beds from upstream units within given constraints on bed reallocation.

This paper used queueing networks first to assess the flows between units, and established target utilizations of bed units. Discrete-event simulation was used to maximize the flow through the balanced system including non-exponential lengths of stay, and blocking behaviour. Since the patients could be admitted to several of the units and could also leave the system from several of the units, the system was modeled as an Open Jackson Network. In this paper, the arrival patterns and unit lengths of stay are assumed to be exponentially distributed. The simulation software ARENA was used to analyze the simulation model.

Cochran and Bharti (2006b) proposed a multi-stage stochastic methodology to balance inpatient bed unit utilisations in an entire hospital. It minimised blocking of beds from upstream units, within given constraints on bed reallocation, while considering multiple patient classes. Queuing network analysis and optimization were used to achieve balanced targets of bed unit utilisation while building hospital staff involvement. Discrete event simulation was then used to maximize flow through the system including non-homogeneous effects of daily and hourly peak loading. Since patients can be admitted to several of the units and can also leave the system from several of the units, the system was modeled as an Open Jackson Network like Cochran and Bharti (2006a). In this paper, the arrival patterns and unit lengths of stay are assumed to be exponentially distributed. The simulation software ARENA was used to analyze the simulation model. Altmel and Ulas (1996) also analysed the capacity of an emergency department (ED). They report that 'quick fixes' in the units obscure the real bottleneck during times of overload and lead to faulty data collection.

The performance of healthcare systems in terms of patient flow times and utilization of critical resources could be assessed through queueing and simulation models (Creemers and Lambrecht, 2006). This paper modeled the orthopedic department of the Middelheim hospital (Antwerpen, Belgium) focusing on the impact of outages (preemptive and nonpreemptive outages) on the effective utilization of resources and on the flow time of patients. For this purpose they constructed an open re-entry queueing network *G/G/n* using a decomposition approach as well as a Brownian queueing model. Simulation was used as a validation tool.

In Griffiths and Price-Lloyd (2006), activities in an intensive care unit (ICU) at a major teaching hospital are modeled by means of a queue-theoretic approach. The model was utilized to investigate several "what if?" scenarios. Reference was made to a simulation model developed in conjunction with the queueing model. Griffiths and Price-Lloyd (2006) developed a mathematical model of the ICU environment. This paper considered a multi-server queue, a Poison's arrival process and hyper-exponential service time distribution, but no inter-stage storage or queue is considered in this paper.

Singh (2006) analyzed the theory and instances of use of queueing theory in healthcare organizations around the world and the benefits accrued from the same. In this paper, the author first discussed the fundamentals of queueing theory, then analyzed the use of queueing theory in the health care organization and finally, demonstrated a simplistic queueing model using POM-QM Software. The author described all of the four types of queueing systems, which are Single Channel Single Phase Systems, Multiple

Channel Single Phase Systems, Single Channel Multiple Phase Systems and Multiple Channel Multiple Phase Systems. With the increasing cost pressure, changing reimbursement mechanisms and affiliations, pressure for quality control, and awareness and demands of the patients, Singh (2006) suggested to tap into the benefits of engineering techniques such as queueing theory to provide smooth, safe and efficient healthcare services to customers, internal and external customer satisfaction and optimization of resources.

Huang and Liou (2007) proposed and implemented a medical record exchange model. According to this study, exchange interface servers (EISs) were designed for hospitals to manage the information communication through the intra and inter-hospital networks linked with a medical records database. Then queueing theory was applied to evaluate the performance of the proposed model. Huang and Liou (2007) estimated the service time for each queue of the CPU, database, and network, and measured the response time and possible bottlenecks of the model. This paper applied queueing theory to the model system, simulated the response time, and identified bottlenecks. The authors assumed that the arrival time of an outpatient at a hospital and the EMR request follows a Poisson distribution. The service time was assumed as exponentially distributed.

The length of stay in hospital of geriatric patients might be modelled using the Coxian phase type distribution (Marshall et al., 2007). This paper examined previous methods which had been used to model health-care costs and presented a new methodology to estimate the costs for a group of patients for their duration of stay in

hospital, assuming there were no further admissions. In this paper, a Coxian phase type distribution is used and the different phases of the distribution represented the different stages of care undergone by a patient while in hospital. This distribution provided the basis for an extended model which could accommodate patients who returned to previously visited states of the distribution, hence feedback, or where admissions into the ward were treated as a Poisson arrivals.

Xie et al. (2007) presented a modeling framework for patient flow in a healthcare system using semi-open queueing network models, which introduced a total bed constraint, above which new patients would be refused admission. In this paper, a service node of the queueing network is one of the phases and the servers at a service node are all hospital beds providing the corresponding phase of care at the time. The number of servers at each service node varies with time depending on the dynamic use of each bed, as patients move through care phases. The authors considered an *M/M/n* queueing model in their study.

2.1.2 The Works That Used Queueing Theory With The Following Considering Factors:

(i) Blocking:

Two types of blocking exist in tandem QN models: blocking-after-service (BAS) and blocking-before-service (BBS). BAS is the widespread blocking type and this

blocking type is considered in this research. Some studies of open queueing networks with blocking can be found in Hunt (1956), Hillier and Boling (1967), Altiok (1985), Perros and Altiok (1986), Gershwin (1987), Altiok (1989a), Altiok and Ranjan (1989b), Gün and Makowski (1989), Hillier and So (1991), Dallery and Gershwin (1992), Bracht (1995), Yamashita and Altiok (1998), Bihan and Dallery (2000), Tempelmeier and Bürger (2001), Tolio, Matta and Gershwin (2002), Koizumi et al. (2005), Vuuren, Adan and Resing-Sassen (2005), Gómez-Corral and Martos (2006).

(ii) **Single Server/ Multiple Servers:**

The literature concerning the single server case is extremely vast. Some of the papers with a single server are Hunt (1956), Hillier and Boling (1967), George et al. (1983), Panayiotopoulos and Vassilacopoulos (1984), Altiok (1985), Perros and Altiok (1986), Gershwin (1987), Altiok (1989a), Altiok and Ranjan (1989b), Altiok (1989c), Gün and Makowski (1989), Albin et al. (1990), Hillier and So (1991), Zhao and Grassmann (1991), Dallery and Gershwin (1992), Shiue and Altiok (1993), Altiok and Shiue (1994), Altiok and Shiue (1995), Bracht (1995), Amin and Altiok (1997), Bretthauer and Côtê (1998), Yamashita and Altiok (1998), Altiok and Shiue (2000), Bihan and Dallery (2000), Tempelmeier and Bürger (2001), Isotupa and Stanford (2002), Tolio, Matta and Gershwin (2002), Vandaele, Boeck and Callewier (2002), Wang (2004), Gómez-Corral and Martos (2006), Zhao et al. (2006), Marshall et al. (2007). For the multi-server case i.e. more than one server at each station, the number of published papers is more moderate, for instance, Shimatonis (1983), Elsayed and Hwang (1986),

Worthington (1987), Hillier and So (1989), Worthington (1991), Hillier and So (1995), Liyanage and Gale (1995), Magazine and Steck (1996), Green and Nguyen (2001), Hountum et al. (2001), Patchong and Willaeys (2001), Gorunescu et al. (2002), Koizumi, (2002), Vandaele et al. (2002), Yang (2003), McManus and Long (2004), Artalejo, Economou and Lopez-Herrero (2005), Van Nyen et al. (2005), Sleptchenko et al. (2005), Vuuren, Adan and Resing-Sassen (2005), Atkinson et al. (2006), Au-Yeung et al. (2006), Creemers and Lambrecht, (2006), Griffiths and Price-Lloyd (2006), Xie et al. (2007).

(iii) Vacation Time / Breakdown / Failure:

Regardless of an extensive research literature on the design and analysis of queueing network systems, analytic models and results for studying the effect of server Vacation Time/breakdowns and inter-stage storage remain limited. Some of the researchers who considered server vacation time, breakdown or failure in their work are Altiok (1985), Gershwin (1987), Gün and Makowski (1989), Hillier and So (1991), Dallery and Gershwin (1992), Bracht (1995), Bihan and Dallery (2000), Tempelmeier and Bürger (2001), Tolio, Matta and Gershwin (2002), Vandaele, Boeck and Callewier (2002), Yang (2003), Artalejo, Economou and Lopez-Herrero (2005), Li and Huang (2005).

(iv) Phase Type Distribution:

Phase type distributions are important in queueing theory because their structure can give rise to a Markovian state-description. The phase (or stage) concept was first introduced by Erlang (1917), later generalized by Cox (1955), and then generalized by Neuts (1981). Phase type distribution is considered as the service time or processing time by Altiok (1985), Gün and Makowski (1989), Hillier and So (1991), Yamashita and Altiok (1998), Gorunescu et al. (2002), Yang (2003), Gómez-Corral and Martos (2006), Zhao et al. (2006), Marshall et al. (2007).

(v) Finite Queue/Buffer:

Queueing networks with finite queues/buffers have been studied extensively in the literature: e.g. Panayiotopoulos and Vassilacopoulos (1984), Altiok (1985), Perros and Altiok (1986), Gershwin (1987), Gün and Makowski (1989), Hillier and So (1989), Hillier and So (1991), Dallery and Gershwin (1992), Shiue and Altiok (1993), Perros (1994), Hillier and So (1995), Magazine and Steck (1996), Amin and Altiok (1997), Yamashita and Altiok (1998), Bihan and Dallery (2000), Tempelmeier and Bürger (2001), Yang (2003), Vuuren, Adan and Resing-Sassen (2005), Zhao et al. (2006).

The interested readers are referred to the books Neuts (1981), Askin (1993), Papadopoulos et al. (1993), Perros (1994) and Altiok (1996).

2.2 Conclusion:

The literature above focuses on the queueing model with different features. From the above literature, the application of queueing theory could be divided into two categories, patient flow and allocation. Systems with parallel servers have become more popular, and parallelism has the effect of reducing the variability of total throughput (Alves, 1990). Though much research have been done to estimate the performance of two stage or multi-stage unreliable healthcare service systems, there is still need for research to evaluate the performance of two-stage unreliable healthcare service systems with parallel servers with different configuration. To our best knowledge, there is no other research available in the current literature that studies the performance of an unreliable series-parallel healthcare service system subject to blocking and with phase type service time distribution for different classes of patients' flow system using analytical methods. So the objective of this research paper is to evaluate the performance of two-stage unreliable series-parallel healthcare service systems with blocking and phase type service time distribution. A two-stage tandem queueing system will also be analyzed before reaching the goal and to compare the performance between the healthcare service systems in series and the series-parallel healthcare service systems. Non-preemptive priority control policy will be used to serve multiple-classes of patients. An exact analytic method will be used to analyze there two node unreliable healthcare service systems with parallel servers. In both cases, the systems will be considered subject to failure and blocking. Computer programs will be implemented in MATLAB to apply the matrix analytic method. In order to design an effective and productive unreliable series-parallel

healthcare service system with different factors/parameters, quick and accurate analysis of the performance of such systems is still necessary and important. Such a queueing model has potential in performance evaluation of healthcare service systems. The analysis of service system schedules is oriented toward the evaluation of trade-offs between high facility utilization and patient waiting requirements since these operating characteristics are generally inversely related (Davis, 1981). The modeling approach presented in this research work is flexible and it could be applied to a variety of health care areas such as blood banks, outpatient clinics, emergency rooms, or diagnosis systems. Any type of healthcare service systems related to queues would benefit from this research. Model formulation will be described in chapter 3.

CHAPTER 3

MODEL FORMULATION

3.1 Background of the Problem:

Queueing theory is used widely in engineering and industry for analysis and modeling of processes that involve waiting lines. Queueing analysis is dependent on accurate measurement of three variables: arrival rate, service time, and the number of servers in the system (McManus and Long, 2004).

Queueing with blocking is a relatively new topic in health care systems and most applications have been done in the engineering fields. According to Koizumi et al. (2005), there have been no applications in the field of mental health and also the authors believe that Koizumi et al. (2005) is the first study that introduces a queueing model as a tool to plan and manage resource allocation in mental health systems. While there are numerous published healthcare studies that apply queueing theory, those analyzing a blocking mechanism are still rare. Few health care publications were found that involve analyses of blocking. El-Darzi, et al. (1998) and Cohen et al. (1980) analyzed the congestion in generic patient flows using simulation. Patient flow lines consist of several stages. Where queueing systems with unreliable servers are concerned, most research that has been done focuses on single-server systems or systems with a Poisson arrival process and exponential service time. However, in some situations we need to consider non-exponential service time or service rate changes with the number of available servers.

The analysis of queueing systems of the type discussed in this thesis has been done very rarely.

In asynchronous Patient flow lines, a patient enters the system from the first station, passes in order through all stations and the intermediate queue locations and exits the line from the last station. The flow of patients works as follows: if a patient has completed its service at a station and the next queue has space available, the served patient is moved on. Then the station starts to give service to a new patient that is taken from its input queue. If the queue next to the server does not have space, then the station holds the patient (who already received service) without service beginning for the next patient until there is room in the following queue. So the station is *blocked.* (This type of blocking is referred to in the literature by various names: blocking after service, type 1 blocking, transfer blocking, production blocking and non-immediate blocking). On the other hand, if the queue has no patient to give service, the station remains empty until a new patient is moved in the queue. In this situation, this station is said to be *starved.* Also sometimes the server could be unavailable due to vacation time. Blocking, starving of patients and vacation time of the servers are the main causes of inefficiency in patient flow lines. So it is assumed here that the first station is never starved and the last station is never blocked.

Singh (2006) analyzed the theory and instances of use of queueing theory in healthcare organizations around the world and the benefits accrued from the same. The author described all of the four types of queueing systems, which are Single Channel

Single Phase Systems, Multiple Channel Single Phase Systems, Single Channel Multiple Phase Systems and Multiple Channel Multiple Phase Systems. With increasing cost pressure, changing reimbursement mechanisms and affiliations, pressure for quality control, and awareness and demands of the patients, Singh (2006) suggested to tap into the benefits of engineering techniques such as queueing theory to provide smooth, safe and efficient healthcare services to our customers, internal and external customer satisfaction and for optimization of resources. In Singh (2006), the last one is the Multiple Channel Multiple Phase Systems. This type of system has numerous queues and a complex network of multiple phases of services involved. This type of service is typically seen in a hospital setting, ER, multi-specialty outpatient clinics, etc. For example in an hospital outpatient clinic, a patient first joins a queue for registration, then he/she is entered for assessment, then for diagnostics, review, treatment, intervention or prescription and finally exits from the system. Multiple Channel Multiple Phase Systems with phase type service distribution is used in Gorunescu et al. (2002), but an inter-stage queue was not allowed in this paper. So, this research could be extended by adding an inter-stage storage queue to the network. Koizumi (2002) studied Multiple Channel Multiple Phase Systems with blocking, but the author considered exponential service time distribution in the paper. So, a phase type service distribution could be considered as the service time distribution instead of considering an exponential service time distribution.

Another relevant paper to our research is Isotupa and Stanford (2002). This paper considers a single server queue that handles arrivals from N classes of customers on a non-preemptive priority basis. Each of the N classes of customers feature arrivals from a

Poisson process at rate λ_i and class-dependent phase type service. To analyze the queue length and waiting time processes of this queue, the authors derived a matrix geometric solution for the stationary distribution of the underlying Markov chain. Following the standard convention, each time a customer is selected, a customer of class j is chosen if there are no waiting customers from any class i, $i<j$. Each of the N classes of customers features arrivals from a Poisson process at rate λ_i, $i=1, 2$ and class-dependent phase type service, i.e. Class 1 receives priority treatment over class 2 according to a non-preemptive priority service discipline. This paper uses the notation $\sum_{i=1}^{N} M_i / PH_i / 1$ to refer the queue. This paper extended the research to manage different classes of customers. The queueing network considered is relevant to our research in that it concerns a single server queue with phase type service time distribution. Different types of queueing network features could be added to this research work.

After reviewing the above models and also the literature that are reviewed in chapter two, to our best knowledge, there is no other research available in the current literature that studies the performance of the multi-class unreliable series-parallel patients' flow line subject to blocking with phase type service time distribution using analytical methods. So the objective of this research is to evaluate the two-stage multi-class series-parallel patient flow lines with vacation time and blocking. Non-preemptive priority control policy will be used to serve multiple-class patients' classes. In this research work, we assume operating time or service time distribution is phase type. Phase type distributions are important in queueing theory because of this generality and because their structure can generate a Markovian state-description. The phase concept is based on

the representation of general distributions by mixtures of convolutions of exponential distributions. Because of the memoryless property of the exponential distribution, the stochastic processes formed by phase type distributions are of Markovian type and can be analyzed using well-known methods (Altiok, 1985). Patients get service on a non-preemptive priority basis. An infinite queue is allowed in front of the first station; and a finite queue is required between the work stations. The following model in the next section is considered to evaluate the performance of two-stage multi-class series-parallel patient flow line systems and the system is subject to vacation time and blocking.

3.2 Notation:

The following notation is used in small letters for the basic parameters of isolated servers. These basic parameters are used to formulate the problem:

n → Number of parallel servers at each station

n_1 → Number of patients at station 1

n_2 → Number of patients at station 2

s_1 → Number of states of station 1

s_2 → Number of states of station 2

v_1 → Number of phases of station 1

v_2 → Number of phases of station 2

S_1 → PH-distribution service time for patient class A at station 1 with order v_1

S_2 → PH-distribution service time for patient class B at station 1 with order v_1

T_1 → PH-distribution service time for patient class A at station 2 with order v_2

T_2 → PH-distribution service time for patient class B at station 2 with order v_2

α_1 → Initial vector for patient class A at station 1 with order v_1

α_2 → Initial vector for patient class B at station 1 with order v_1

β_1 → Initial vector for patient class A at station 2 with order v_2

β_2 → Initial vector for patient class B at station 2 with order v_2

(α_1, S_1) → Represents PH- type service time distribution for patient class A at station 1

(α_2, S_2) → Represents PH- type service time distribution for patient class B at station 1

(β_1, T_1) → Represents PH- type service time distribution for patient class A at station 2

(β_2, T_2) → Represents PH- type service time distribution for patient class B at station 2

λ_1 → Patient arrival rate for patient class A at system 1

λ_2 → Patient arrival rate for patient class B at system 1

θ_2 → Vacation rate of servers at station 2

γ_1 → Service resume rate of servers at station 1

γ_2 → Service resume rate of servers at station 2

m → Total number of patients at station 2 (including patients in service)

l → Mean queue length

w_q → Mean waiting time in the queue

q → Capacity of the queue in between station 1 and station 2

3.3 Assumptions for the Performance Model:

According to the Kendall notation, the following assumptions are used in our queueing model:

(i) The arrivals follow a Poisson process, i.e. the inter-arrival times between successive arrivals are exponentially distributed,

(ii) The service times are independent and identically distributed and the phase type distribution is assumed for service time.

(iii) There are parallel servers (identical) at each workstation.

(iv) Two classes of patients are considered in the system.

(v) A server can give service to only one patient at a time.

(vi) Time between vacation or failure (MTBF) and time to service resume (MTTR) are independent random variables and exponentially distributed.

(vii) When a workstation consists of parallel servers, patient is transferred to the first available parallel server.

(viii) All servers at Station 1 are always busy until blocked.

(ix) The inter-stage queue capacity is finite.

(x) Moving time between stations is zero.

(xi) No set-up time is considered during switch over to a new patient class.

(xii) Non-preemptive priority control policy is used to serve multiple-classes of patients.

(xiii) A server can not give service to any patient when it is starved, blocked or under vacation time.

(xiv) The patient flow line is assumed to be operating under saturation, i.e. the first station is never starved and the last station is never blocked, and so the line is operated at maximum capacity.

3.4 An Open Queueing Network Model:

Under the above assumptions, let us consider a patient flow line *M/PH/n* consisting of two stations in series at station M_i, $i = 1, 2$ i.e. at each station there are n parallel servers to perform the specified service and Q_1 is an intermediate queue of finite capacity in between station 1 and station 2 (figure 3.4.1). The capacity of this finite queue is q. These are illustrated in figure 3.4.1. Patient flow line serves two classes of patient A and B. Here a queueing system with two-servers at each station is considered in which patients arrive according to a Poisson process with rate λ_1 for patient class A and λ_2 for patient class B. Each station M_i has servers M_{ij} in parallel, where $j = 1, \ldots, n$. Patients arrive at the service system to be served by any one of the available service facilities at the first work station. Each patient requires several services performed on different servers. Patients, who received service, leave the system through station 2. The service times on the servers depend on the patient and follow a phase type distribution. For patient class A, PH-distribution processing time at station 1 is represented by (α_1, S_1) with order v_1 and it is also represented by (β_1, T_1) with order v_2 for station 2. For patient class B, PH-distribution processing time at station 1 is represented by (α_2, S_2) with order v_1 and it is also represented by (β_2, T_2) with order v_2 for station 2. Each patient class has a deterministic routing throughout the patient flow line system. The first station is considered blocked, when the last station is full, i.e. the queue in front of the last station has maximum number of patients and patients are receiving service from the servers at the last station. The main performance measure of the patient flow line is the mean queue length. The patients' waiting time is approximated by using Little's formula. The

service discipline of this patient flow line system is first come first serve (FCFS). Server vacation time significantly affects the efficiency of a patients' flow line. Assume all servers at station 1 have service resume rate γ_1 and all servers at station 2 have the same vacation time rate θ_2 and service resume rate γ_2 in this patients' flow line.

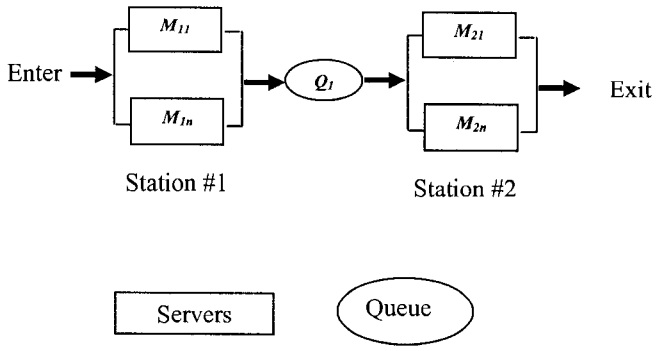

Figure 3.4.1: Configuration of the series-parallel HealthCare Service System

While our main focus in this paper is on *M/PH/n* (here, *n* is the number of parallel servers) approximations for open queueing networks of series-parallel topologies, we also analyze the *M/PH/1* systems with the same factors to compare the performance of a series parallel healthcare service system with the performance of a healthcare service system in series. Some researchers used Jackson's Theorem in their research work to analyze the system. Jackson's Theorem was the first significant development in the theory of networks of queues. This research will focus on *M/PH/n* open queueing networks, and Jackson's theorem can not be used to analyze this type of queueing system. The reason, according to Willig (1999) is that Jackson's theorem needs some conditions, such as,

(i) The number of units in the network is not limited;

(ii) Every node in the network can have Poisson arrivals from outside the network;

(iii) Any unit can leave the system from any node (or subset);

(iv) Service times are exponentially distributed;

(v) In every node the service discipline is FCFS;

(vi) The ith service facility consists of M_i identical servers, each with service rate μ_i (as a generalization the service rate μ_i may depend on the number of units in system i). The Matrix Geometric Procedure will be used to solve for the stationary distribution.

(vii) The utilization of all of the subsystems is less than one.

3.5 Performance Measures:

The objective of the present work is to evaluate the performance of an asynchronous, unreliable healthcare service system with the assumptions given above. The objective is also to figure out the different performance measures of an unreliable healthcare service system by providing parallel servers at different work-stations. The following basic performance measures in the analysis of healthcare service systems to be considered in this work are:

(i) Mean queue length;

(ii) Mean waiting time;

3.6 Conclusions:

An open network queueing model to measure the performance of an unreliable multi-class series-parallel patient flow line by using an exact analytical approach will be described in the following chapters. Such models are important because, (1) exact solutions are better than simulations or approximations, and (2) they provide useful qualitative insight into the behaviour of systems (Dallery and Gershwin, 1992). For different types of what-if scenarios, the open queueing network model is a very efficient approach, computationally fast and leads quickly to effective answers. Queueing theory will be used to evaluate the performance of two-stage multi-class patient flow lines with parallel servers subject to blocking and vacation time and non-preemptive priority control policy will be used to serve multiple class patient. The matrix analytic method will be used to determine the transition matrix. The size of the transition matrix depends on the complexity of the system. The size of the transition matrix will increase as much as the complexity of the system increases. Since the system is complex, the size of the transition matrix is big, too big to show as a single matrix. Being a numerically based technique, the system is dependent on the efficiency of numerical and storage procedures and on the computational power of the computer. Computer programs will be implemented in MATLAB to analyze this system. Analysis of series and series-parallel patient flow line system will be presented in the next chapter.

CHAPTER 4

ANALYSIS OF PATIENT FLOW LINE SYSTEMS WITH SERVER VACATION TIMES AND BLOCKING

4.1 Introduction:

Phase type distribution has been used extensively in healthcare service systems. But Phase type service time distribution with blocking is rare in healthcare service. The phase type distribution is convenient to use due to its convenience in representing various types of distribution. Its effectiveness stems from the fact that it possesses the Markovian property. In this chapter, first *M/PH/1* and then *M/PH/n* queueing systems with Poisson arrivals and phase type service times at all stations are considered to analyze the performance.

The following numerical approach (i.e. the state model method) is adopted for this study. A computer program will be written to generate the transition states systematically of the underlying Markov chain and also to determine the performance measures of the considered patient flow lines systems. The performance measures are determined from the infinitesimal matrix or the transition matrix, the stationary probability vector and the boundary conditions using the state model method.

4.2 PH - Distribution:

This section is provided to give a short description of PH-distribution according to Neuts (1981). We consider a Continuous Time Markov process (CTMP) on the states {1, ..., $m + 1$} with infinitesimal generator or rate matrix

$$Q = \begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix} \quad \text{..... (4.2.1)}$$

where, the $m \times m$ square matrix T satisfies $T_{ii} < 0$, for $1 \leq i \leq m$, and $T_{ij} \geq 0$, for $i \neq j$, i.e. here T is a square matrix with size m; the diagonal elements of T are negative; the off-diagonal elements are non-negative and T is assumed stable. Here $Te+T^0=0$, and the initial probability vector is given by (α, α_{m+1}), with $\alpha e + \alpha_{m+1} = 1$. We assume that the states 1, ..., m are all transient, so that absorption into the state $m+1$, from any initial state, is certain. In $Te+T^0=0$, T^0 is the probability matrix at the absorbing state, e is a column vector with all entries 1 and 0 is a column vector with all entries 0.

A probability distribution $F(\cdot)$ on $(0, \alpha)$ is a distribution of phase type (PH-distribution) if and only if it is the distribution of the time until absorption in a finite Markov process of the type defined in equation (4.2.1). The pair (α, T) is called a representation of $F(\cdot)$.

4.3 Quasi-Birth-and-Death Process (QBD Process):

This section is provided to give a description of Quasi-Birth-and Death process (QBD process) according to Neuts (1981). It has long been appreciated that the simple birth-and-death Markov process habitually provides an adequate initial model for the performance of service systems. QBD process is a Markov process on the state space $E = \{(i, j),\ i \geq 0,\ 1 \leq j \leq m\}$, with infinitesimal generator Q, is given by

$$Q = \begin{pmatrix} B_0 & A_0 & & & & & & \\ B_1 & A_1 & A_0 & & & & & \\ & A_2 & A_1 & A_0 & & & & \\ & & A_2 & A_1 & A_0 & \cdot & \cdot & \cdot \\ & & & A_2 & A_1 & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & & \end{pmatrix} \quad \text{..... (4.3.1)}$$

where $B_0e + A_0e = B_1e + A_1e + A_0e = (A_0 + A_1 + A_2)\ e = 0$. Here Q is also assumed to be irreducible. More information about QBD is given in Appendix C.

4.4 Analysis of Patient Flow Line Systems in Series with Vacation Time and Blocking:

The goal of this research is to analyze a series-parallel patient flow line. However, before reaching this goal, in this section first series and then in the next section series-parallel patient flow line systems with server vacation time and blocking are considered.

Server vacation time and blocking are considered to analyze the queueing systems. Two types of blocking exist in open queueing network models: blocking after service (BAS) and blocking before service (BBS). Blocking after service (BAS) is the widespread blocking type in queueing network systems and so this type of blocking is assumed here.

4.4.1 Description of Patient Flow Line Systems in Series with Vacation Time and Blocking:

This section is presented to analyze the *M/PH/1* queueing system. The transition matrix of the system is formed by using the state model method, which is described in this section. The stationary probability vector and different performance measures are determined in the following sections.

The patient flow line is represented by a two stage *M/PH/1* queueing system with queues in series (Figure 4.4.1). The queue in front of the first station is infinite and the queue in front of the last station is finite. This system with server vacation time and blocking consists of two single-server service facilities (stations) corresponding to the two work stations of the patient flow line, where every patient must be served by these stations in the same fixed sequence. Patients enter the system through the first station and after getting the service, patients leave the system through the last station. Each station is a single server that serves one patient at a time. Let λ_1 be the arrival rate of patient class A and λ_2 be the arrival rate of patient class B at station 1. For patient class A, (α_1, S_1) be the phase type service distribution with order v_1 at station 1 and (β_1, T_1) be the phase type

service distribution with order v_2 at station 2. For patient class B, (α_2, S_2) be the phase type service distribution with order v_1 at station 1 and (β_2, T_2) be the phase type service distribution with order v_2 at station 2. θ_2 be the server vacation time rate until next vacation/breakdown and γ_2 be the server resume/return time rate at station 2.

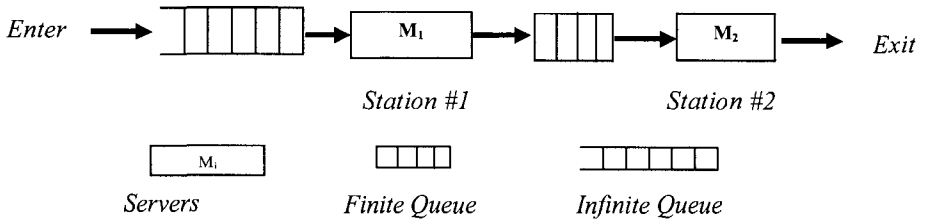

Figure 4.4.1: Configuration of the series connected two node patients flow lines

Assume one server at each station. Assume buffer at second station has size m (including patient in service). In order to describe the patient flow line as a Markov chain, the state vector has to keep track of the number of patients in each station and the status of the servers, i.e. vacation, busy or idle. It also has to show whether the first server is blocked or not. Let us define the state vector of this system, which is $E = \{(s_1, n_2, s_2, v_1, v_2); s_1 = 1, 2; 0 \leq n_2 \leq m; s_2 = 0, 1, 2\}$, where $s_1 = 1$ denotes the server at the 1st station is busy and $s_1 = 2$ represents that the server at station #1 is blocked by station 2 and n_2 denotes the number of patients at the 2nd station (including taking service) and the capacity of the queue in front of the second server has to be equal to or more than the

number of the server (here, assumed m=2); s_2 represents the availability of the server at station 2 (the server is idle when s_2=0; the server is busy when s_2=1; and the server is on vacation when s_2=2); and v_1 *and* v_2 denote the phases of the servers at the both station 1 and station 2 i.e. the number of phases of service time distribution at each station is 2, which is P_1 and P_2 (Figure 4.4.2). For feasibility assume here $v_1 = v_2 = 2$, but the implemented program in MATLAB, will work for any number of phases of operation at any work station (Figure 4.4.3). The number of phases to complete the service assumed can be increased as long as the CPU memory space allows. Saturation condition is assumed here, i.e. $n_1 > 0$ the first station never becomes starved or the last station never becomes blocked. Also $Se+S^0$ and $Te+T^0=0$. . Here in $Se+S^0$ and $Te+T^0=0$, S and T are the phase type service time distribution at station 1 and station 2 respectively, S^0 and T^0 are the phase type service time distribution at absorption state for station 1 and station 2 respectively, e is a column vector with all entries 1 and 0 is a column vector with all entries 0.

So, S^0=-Se and T^0=-Te.

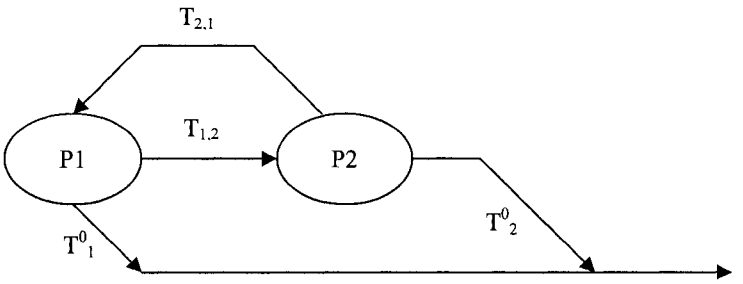

Figure 4.4.2: Representation of the distribution to complete service with two phases P1 and P2

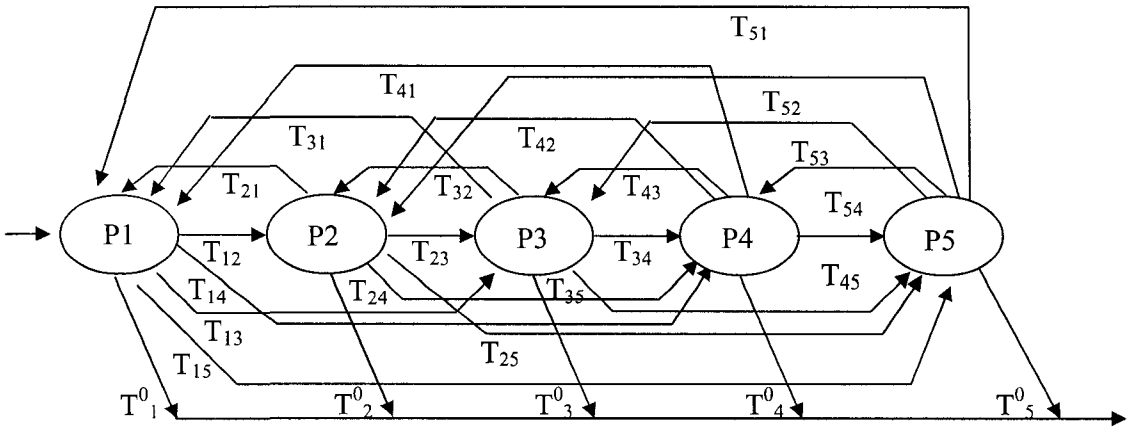

Figure 4.4.3: Representation of the phase type service distribution with five phases

Table 4.4.1 represents the algorithm to analyze the *M/PH/1* and *M/PH/n* patient flow queueing systems. This algorithm will be followed to implement the program to determine the performance measures of this queueing system. In Table 4.4.1, required parameters will be initialized in Step 0, transition matrix will be determined using the state model method in Step 1, stationary probability matrices and vector will be determined in Step 2 using Gaver's recursive method. Mean queue length will be determined in Step 4 and by using Little's formula, mean waiting time in inter-stage patient's queue will be determined in Step 5.

Table 4.4.1: Algorithm to perform the analysis of M/PH/1 and M/PH/n queueing system

Step 0	Initialization of the used parameters
Step 1	Forming or developing the infinitesimal matrix or the transition matrix using state model method (section 4.4 for single and section 4.5 for parallel server).
Step 2	Determining the stationary probability matrix and stationary probability vector using Gaver's recursive method with the boundary conditions (section 4.6).
Step 3	Determining the mean queue length from the stationary probability vector (section 4.6).
Step 4	Determining the mean waiting time of the patients in the queue in front of the second server from the mean queue length using Little's formula (section 4.6).

The matrix analytic method is used to analyze the transition or rate matrix. The size of the rate matrix depends on the complexity of the system. The size of the rate matrix will increase as the complexity of the system increases. Since, our system is complex; the size of the rate matrix is big, which is not possible to show in a single page. So, the block sub-matrices are again divided into small matrices. The rate matrix or transition matrix for this model is:

$$Q = \begin{pmatrix} A_0 & E_0 & \\ F_1 & A_1 & E_0 \\ & F_2 & A_2 \end{pmatrix} \qquad \text{..... (4.4.1)}$$

Definitions of the block sub-matrices of the transition matrix are given below:

A_0, A_1 and A_2 matrices are those matrices, which are on the diagonal of the transition matrix. These matrices are generated from the transitions of patients flow. These matrices keep track of the number of patients arrive at station #2, server resume and server's vacation related information of station #2. Also this function calculates the diagonal negative elements of the rate matrix. The size of the following sub-block matrices A_0, A_1 and A_2 depends on the identity matrix I of order $v_1 x v_2$:

$$A_0 = \begin{pmatrix} S-\gamma_2 I & \gamma_2 I & \\ \theta_2 I & S-\theta_2 I & \\ & \theta_2 I & S-\theta_2 I \end{pmatrix} \quad \text{..... (4.4.2)}$$

$$A_1 = \begin{pmatrix} S-\gamma_2 I & \gamma_2 I & \\ \theta_2 I & S+T-\theta_2 I & \\ & \theta_2 I & S-\theta_2 I \end{pmatrix} \quad \text{..... (4.4.3)}$$

$$A_2 = \begin{pmatrix} -\gamma_2 I & \gamma_2 I & \\ \theta_2 I & T-\theta_2 I & \\ & \theta_2 I & -\theta_2 I \end{pmatrix} \quad \text{..... (4.4.4)}$$

Matrix E_0 is that matrix which is above the diagonal of the transition matrix. This matrix is generated from the transitions of patients flow. This matrix keeps track of the number of patients who receive service from station #1 and enter into the second station. The size of the following sub-block matrix E_0 depends on $v_1 x v_2$:

$$E_0 = \begin{pmatrix} S^0\alpha & & \\ & S^0\alpha & \\ & & S^0\alpha \end{pmatrix} \quad \text{..... (4.4.5)}$$

F_1 and F_2 matrices are those matrices which are under the diagonal of the transition matrix. This matrix is generated from the transitions of patients flow. These matrices keep track the number of patients, who received service from station #2. The size of the following sub-block matrices F_1 and F_2 depend on $v_1 x v_2$:

$$F_1 = \begin{pmatrix} diag(T^0) \end{pmatrix} \quad \text{..... (4.4.6)}$$

$$F_2 = \begin{pmatrix} T^0\beta \end{pmatrix} \quad \text{..... (4.4.7)}$$

Single server patient flow line system is analyzed in this section. The method developed by Gaver et al. (1984) to calculate the probability distribution matrix and probability distribution vector for this patient flow line system will be described in section 4.6. In section 4.7, implemented MATLAB program for this queueing system will be described. Also the numerical example and sensitivity analysis for this queueing system will be described in section 4.8 and section 4.9 respectively. Parallel server patient flow line system is analyzed in the next section.

4.5 Analysis of Patient Flow Line for Series-Parallel Queueing Systems with Vacation Time and Blocking:

Systems with parallel servers have become more popular, and parallelism has the effect of reducing the variability of total throughput (Alves, 1990). Though much research has been done to estimate the performance of two stage or multi-stage unreliable patients' flow lines, there is still need for research to evaluate the performance of multi-stage unreliable patients' flow lines with parallel servers. So the objective of this section is to evaluate the performance of two-stage unreliable patients' flow line with parallel servers and blocking. The goal of this research is to analyze Series-Parallel patient's flow Line Systems. So, *M/PH/n* queueing systems with vacation time and blocking are considered in this section to reach this goal. Also, vacation time or failure of the servers at the second station and blocking of the servers at the first station are considered in this type of queueing systems. As mentioned in the previous section, two types of blocking

exist in open queueing network models: blocking after service (BAS) and blocking before service (BBS). Blocking after service (BAS) is the widespread blocking type in queueing systems and so this type of blocking is assumed here. The state model method or the matrix geometric method is adopted for this study. A computer program has been written to generate the transition states systematically of the underlying Markov chain and also to determine the performance measures of the considered patient's flow line systems. The performance measures are determined from the rate matrix, the stationary probability vector and the boundary conditions by using the state model method.

4.5.1 Description of Patient Flow Line Systems in Series-Parallel with Vacation Time and Blocking:

To analyze two stage *M/PH/n* queueing systems with server vacation time and blocking, parallel servers are considered at each work station (Figure 4.5.1). Let us consider, λ be the arrival rate at the 1st station, (α, S) be the phase type service distribution with order v_1 at station #1 and (β, T) be the phase type service distribution with order v_2 at station #2, θ be the server vacation time rate and γ be the vacation resume time rate. The same number of phases to complete the service is assumed here at both stations.

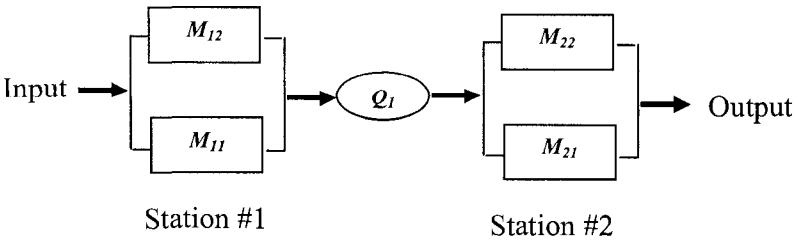

Figure 4.5.1: Configuration of the series-parallel two node patients' flow line

Let us define the state space of this system, which is $E = \{s_1, n_2, s_2, v_1, v_2); s_1=1,2; 0 \le n_2 \le m; s_2 = 0, \ldots, n, n+1\}$, where n denotes the number of parallel servers at the each station; $s_1=1$ means the parallel servers at station 1 are busy and $s_1=2$ represents that the servers at station 1 is blocked by station 2 and station 2 has m number of patients (including patients in service); n_2 denotes the number of patients at the 2nd station (number patients waiting in the queue and in service); s_2 represents the availability of the server at station 2; and v denotes the phase of the servers at the both stations. Assumed, both servers have the same numbers of phases, i.e. $v = v_1 = v_2$ and m is the capacity of station 2. Saturation condition is considered here, i.e. the first station never becomes starved or the last station never becomes blocked. Also $Se+S^0=0$ at station 1 *and* $Te+T^0=0$ at station 2. Here in $Se+S^0=0$ and $Te+T^0=0$, S and T are the phase type service time distribution at station 1 and station 2 respectively, S^0 and T^0 are the phase type service time distribution at absorption state for station 1 and station 2 respectively, e is a column vector with all entries 1 and 0 is a column vector with all entries 0. So, $S^0 = -Se$ and $T^0 = -Te$.

The infinitesimal matrix or transition matrix for this model is:

$$Q = \begin{pmatrix} A_0 & E_0 & & & & \\ F_1 & A_1 & E_0 & & & \\ & F_2 & A_1 & E_0 & & \\ & & \cdots & \cdots & \cdots & \cdots \\ & & & & F_2 & A_2 \end{pmatrix}$$

..... (4.5.1)

Definitions of the block sub-matrices of the transition matrix are given below:

A_0, A_1 and A_2 matrices are those matrices, which are on the diagonal of the transition matrix. These matrices are generated from the transitions of patients flow. These matrices keep track of the number of patients arrive at station #2, server resume and server's vacation related information of station #2. Also this function calculates the diagonal negative elements of the rate matrix. The size of the following sub-block matrices A_0, A_1 and A_2 depends on the identity matrix I of order v_1xv_2 (here n denotes the number of parallel servers):

$$A_0 = \begin{pmatrix} S-n\gamma_2 I & n\gamma_2 I & & & & \\ \theta_2 I & S-\theta_2 I-(n-1)\gamma_2 I & (n-1)\gamma_2 I & & & \\ & 2\theta_2 I & S-2\theta_2 I-(n-2)\gamma_2 I & & & \\ & & & \ddots & & \\ & & & & \ddots & \\ & & & & n\theta_2 I & S-n\theta_2 I \end{pmatrix}$$

..... (4.5.2)

$$A_1 = \begin{pmatrix} S-n\gamma_2 I & n\gamma_2 I & & & & \\ \theta_2 I & S+T-\theta_2 I-(n-1)\gamma_2 I & (n-1)\gamma_2 I & & & \\ & 2\theta_2 I & S+2T-\theta_2 I-(n-2)\gamma_2 I & (n-2)\gamma_2 I & & \\ & & \cdots & & & \\ & & & \cdots & & \\ & & & (n-1)\theta_2 I & S+nT-(n-1)\theta_2 I-\gamma_2 I & \gamma_2 I \\ & & & & n\theta_2 I & S-n\theta_2 I \end{pmatrix}$$

..... (4.5.3)

$$A_2 = \begin{pmatrix} -n\gamma_2 I & n\gamma_2 I & & & & \\ \theta_2 I & T-\theta_2 I-(n-1)\gamma_2 I & (n-1)\gamma_2 I & & & \\ & 2\theta_2 I & 2T-\theta_2 I-(n-2)\gamma_2 I & (n-2)\gamma_2 I & & \\ & & & \cdots & & \\ & & & (n-1)\theta_2 I & nT-(n-1)\theta_2 I-\gamma_2 I & \gamma_2 I \\ & & & & n\theta_2 I & -n\theta_2 I \end{pmatrix}$$

..... (4.5.4)

Matrix E_0 is that matrix which is above the diagonal of the transition matrix. This matrix is generated from the transitions of patients flow. This matrix keeps track of arrival of the patients who receive service from station #1 and enter into the second station. The size of the following sub-block matrix E_0 depends on the identity matrix I of order $v_1 x v_2$ (here n denotes the number of parallel servers):

$$E_0 = \begin{pmatrix} S^0\alpha & & & & & \\ & S^0\alpha & & & & \\ & & S^0\alpha & & & \\ & & & \cdots & & \\ & & & & \cdots & \\ & & & & & S^0\alpha \end{pmatrix} \quad \text{..... (4.5.5)}$$

F_1 and F_2 matrices are those matrices which are under the diagonal of the transition matrix. This matrix is generated from the transitions of patients flow. These matrices keep track the number of patients, who received service from station #2. The size of the following sub-block matrices F_1 and F_2 depend on the identity matrix I of order $v_1 x v_2$ (here n denotes the number of parallel servers):

$$F_1 = \begin{pmatrix} diag(T^0) & & & \\ & 2diag(T^0) & & \\ & & \cdots & \\ & & & ndiag(T^0) \end{pmatrix}$$

..... (4.5.6)

$$F_2 = \begin{pmatrix} T^0\beta & & & \\ & 2T^0\beta & & \\ & & \cdots & \\ & & & nT^0\beta \end{pmatrix}$$

..... (4.5.7)

Parallel server patient flow line system is analyzed in this section. The method to calculate the probability distribution matrix and probability distribution vector for this type patient flow line system will be described in section 4.6. In section 4.7, implemented MATLAB program for this queueing system will be described. Also the numerical example and sensitivity analysis for this queueing system will be described in section 4.8 and section 4.9 respectively.

4.6 Determination of the Stationary Distribution Matrix for the Patient Flow System:

This section represents the step 2 of the algorithm to perform the analysis of M/PH/1 and M/PH/n queueing system. An efficient computational approach to the analysis of finite birth-and-death models in a Markovian environment is given in Gaver et al. (1984). The emphasis is upon obtaining numerical methods for evaluating stationary distributions. In order to determine the stationary probability distribution matrix, the method developed by Gaver et al. (1984) is followed in this research.

Assume, U_i represents the matrices that are above the diagonal of the generator matrix, $D_{i\ \text{represents}}$ the matrices that are on the diagonal of the generator matrix and L_i represents the matrices that are under the diagonal of the generator matrix.

Then,

$U_i = E_0$, when $0 \leq i < m$

$D_i = A_0$, when $i = 0$

$D_i = A_1$, when $1 \leq i < m$

$D_i = A_2$, when $i = m$

$L_i = F_1$, when $i = 1$

$L_i = F_2$, when $1 \leq i \leq m$

Lemma (Gaver et al., 1984):

$C_0 = D_0$

$C_i = D_i + L_i (- C^{-1}_{i-1}) U_{i-1}$ when $1 \leq i < m$

Using the above Lemma, we get following relations for single and series-parallel server healthcare queueing system to calculate the probability distribution matrix,

$C_0 = D_0$

i,e, $C_0 = A_0$

$C_i = D_i + L_i (- C^{-1}_{i-1}) U_{i-1}$ when $1 \leq i < m$

i,e, $C_i = A_1 + F_1 (- C^{-1}_{i-1}) E_0$ when $i = 1$

$C_i = A_1 + F_2 (- C^{-1}_{i-1}) E_0$ when $1 < i < m$

$C_i = A_2 + F_2 (- C^{-1}_{i-1}) E_0$ when $i = m$

The stationary probability vectors Π_i, $0 \leq i \leq m$, are determined by the following equations:

$$\Pi_m C_m = 0 \qquad \text{.....(4.6.1)}$$

$$\sum_{0}^{m} \Pi_i e = 1 \qquad \text{.....(4.6.2)}$$

Here,

$P_i = P_{i+1} F_1 (- C^{-1}_{ii})$ when $i = 1$

$P_i = P_{i+1} F_2 (- C^{-1}_{ii})$ when $1 < i \leq m$

According to Gaver et al. (1984), we will determine the stationary probability vectors by using the following algorithm described in Table 4.6.1. In Table 4.6.1, required parameters are initialized in Step 0, matrices C_i are determined using the above mentioned lemma in Step 1, stationary probability vector is determined in Step 2 and Step 3. Finally, stationary probability vector is normalized in Step 4.

Table 4.6.1: Algorithm for the probability distribution matrix of M/PH/1 and M/PH/n queueing system

Step 0	Initialization of the used parameters
Step 1	Determine recursively the matrices C_i, $0 \leq i < m$
Step 2	Solve the system $\Pi_m C_m = 0$, and $\sum_0^m \Pi_i e = 1$
Step 3	Compute recursively the vectors P_i, $i = m\text{-}1, m\text{-}2, \ldots, 0$, using Π_i instead of P_i
Step 4	Re-normalize the vector P so obtained

The different performance measures, such as mean queue length (l) and the mean waiting time (w) in the queue are determined from the calculated stationary probability matrix and row vector. The mean waiting time is obtained according to the Little's formula.

Mean waiting time = Mean Queue length / Arrival rate ..... (4.6.3)

4.7 Description of the Implemented Program in MATLAB:

To analyze the series and series-parallel patients' flow line systems, the program is implemented in MATLAB. One main file and three function files are implemented to conduct this analysis. Main file calculates the result of the analysis, i.e. the mean queue length and Patients' mean waiting time. Gaver et al. (1984)'s method is implemented in this program to determine the stationary probability matrices and stationary probability vector. Also Little's formula is used to calculate the performance measure of this system. This is an interactive program. This program allows inputs from user or defaults values and then calculates the performance measures of the patients' flow line system.

There are three functions to calculate the sub-matrices A_i, E_i and F_i. Main program calls these functions to calculate the performance measures. A_i matrices are those matrices which are on the diagonal of the generator matrix. This function keeps track of all server resume and servers vacation related information of station #2. Also this function calculates the diagonal negative elements of the rate matrix. The inputs or variables for this function are total number of states of station #2, number of parallel servers, servers' resume rate per hour, server vacation rate per hour, phase type service distribution matrix for station #1, phase type service distribution matrix for station #1, number of phases of service, and a counter to keep track the transition.

There is another function for the sub-matrices E_i. E_i matrices are those matrices which are above the diagonal of the generator matrix. This function calculates matrices

E_i. This function keeps track of the number of patients who receive service from station #1 and enter into the second station. The inputs or variables for this function are total number of states of station #2, number of parallel servers, phase type service distribution matrix for station #1, initial probability vector for station #1, and number of phases of service.

The last function is for the sub-matrices F_i. F_i matrices are those matrices which are under the diagonal of the generator matrix. This function calculates matrices F_i. This function keeps track the number of patients, who received service from station #2. The inputs or variables for this function are total number of states of station #2; number of parallel servers, phase type service distribution matrix for station #2, initial probability vector for station #2, number of phases of service; and a counter to keep track the transition.

4.8 Numerical Example and Test Run Results for *M/PH/1* and *M/PH/n* Queueing Systems:

Consider a patients' flow line consisting of two stations in series. Before completing service, each patient must go through two stages. Phase type service distribution is considered at each station. An average of 4 patients per hour of class A and an average of 7 patients per hour of class B arrives at station 1. Patients' death rate is not considered here. So, all patients completing service at station 1 immediately move on to station 2. The servers at station 2 may take off at any operating time. The vacation rate of the servers is 2 per hour and the server's vacation resume rate is 5 per hour. The queue in front of station 1 is infinity and the queue in front of station 2 is finite. Assume each station has the same phase type service distribution for each patient's class. For patient class A, the initial probability vector is [0.4, 0.6] and phase type service time distribution is [-0.7, 0.4; 0.8, -0.6]. For patient class B, the initial probability vector is [0.2, 0.6, 0.2] and phase type service time distribution is [-0.5, 0.2, 0.3; 0.02, -0.5, 0.3; 0.03, 0.2, -0.23];

a. Determine the mean queue length for patients' class A.

b. Determine the mean waiting time for patients' class A.

c. Determine the mean queue length for patients' class B.

d. Determine the mean waiting time for patients' class B.

The program to analyze the considering system is implemented in MATLAB. This problem is analyzed by both Single Server Queueing System and also by Parallel-Server Queueing System. The solution of the above problem is given below as the test run result:

4.8.1 Test Run Result Using Single Server:

```
**********************************************************************
**********************  SERVERS IN SERIES  ************************
**********************************************************************

First station or node is always saturated;
i.e. First station is never starved.
Total number of states of station #1 = 2
Server(s) at station #1 is busy, when s1 = 1
Server(s) at station #1 are blocked, when s1 = 2

Please, input the number of the parallel servers at the first Station or node:
Program is considering the default value since no value is entered at the prompt.
Number of server(s) at the first Station or node: 1

Please, input the capacity of the queue in front of the second Station or node.
Program is considering the default value since no value is entered at the prompt.
```

Capacity of the queue in front of the second Station or node = 1

Assume Station #1 and Station #2 have the same number of parallel servers

Number of patients at the second Station (including Patients in service) = 2

The number of states of the server(s) at the station #2 = 4

Server(s) on vacation, when s2=0;

Server(s) at station #2 is busy, when 1<=s2=2;

Servers(s) at station #2 are idle, when s2=3;

Enter the vacation rate per hour at station 2 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Server vacation rate per hour at station 2 is 2.00

Enter the Service Resume Rate per hour at station 2 [Example to enter: 5]:

Program is considering the default value since no value is entered at the prompt.

Service Resume Rate per hour at station 2 is 5.00

****** LAMDA_A, ALPHA_A & S_A FOR for Patient Class A at STATION 1 *****

Input the arrival rate per hour for Patient Class A [Example to enter: 4]:

Program is considering the default value since no value is entered at the prompt.

Arrival rate at station #1 for Patient Class A is 4 per hour.

Enter the number of phases of service at station 1 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 1 is 2

Enter the value of the vector alpha [Example to enter: [0.4 0.6]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.4 0.6

Size of the alpha vector is 2.

Value of the phase type distribution

[Example to enter: [-0.7 0.4; 0.8 -0.6]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix is 2

Value of the phase type distribution:

-0.7000 0.4000

0.8000 -0.6000

****** LAMDA_B, ALPHA_B & S_B FOR for Patient Class B at STATION 1 *****

Input the arrival rate per hour for Patient Class B [Example to enter: 4]:

Program is considering the default value since no value is entered at the prompt.

Arrival rate at station #1 for Patient Class B is 7 per hour.

Enter the number of phases of service at station 1 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 1 is 3

Enter the value of the vector alpha [Example to enter: [0.2, 0.6, 0.2]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.2 0.6 0.2

Size of the alpha vector is 3.

Value of the phase type distribution

[Example to enter: [-0.5 0.2 0.3; 0.02 -0.5 0.3; 0.03 0.2 -0.23]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix is 3

Value of the phase type distribution:

-0.5000 0.2000 0.3000

0.0200 -0.5000 0.3000

0.0300 0.2000 -0.2300

********** BETA_A & T_A for Patient Class A at STATION 2 **********

Enter the number of phases of service at station 2 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 2 is 2

Enter the value of the vector beta [Example to enter: [0.4 0.6]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.4 0.6

Size of the alpha vector is 2.

Value of the phase type distribution

[Example to enter: [-0.7 0.4; 0.8 -0.6]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix at station 2 is 2

Value of the phase type distribution:

-0.7000 0.4000

0.8000 -0.6000

********** BETA_B & T_B for Patient Class B at STATION 2 **********

Enter the number of phases of service at station 2 [Example to enter: 3]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 2 is 3

Enter the value of the vector beta [Example to enter: [0.2, 0.6, 0.2]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.2 0.6 0.2

Size of the alpha vector is 3.

Value of the phase type distribution

[Example to enter: [-0.5 0.2 0.3; 0.02 -0.5 0.3; 0.03 0.2 -0.23]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix at station 2 is 3

Value of the phase type distribution:

-0.5000 0.2000 0.3000

0.0200 -0.5000 0.3000

0.0300 0.2000 -0.2300

PROGRAM RUN RESULT:

FOR PATIENT CLASS A:

Mean length of the queue is 0.64825

Mean waiting time in the queue is 9.7237 minutes

FOR PATIENT CLASS B:

Mean length of the queue is 0.64847

Mean waiting time in the queue is 5.5583 minutes

4.8.2 Test Run Result Using Series-Parallel Servers:

****

********** SERIES-PARALLEL SERVERS **********

****

First station or node is always saturated;

i.e. First station is never starved.

Total number of states of station #1 = 2

Server(s) at station #1 is busy, when s1 = 1

Server(s) at station #1 are blocked, when s1 = 2

Please, input the number of the parallel servers at the first Station or node: 5

Number of server(s) at the first Station or node: 5

Please, input the capacity of the queue in front of the second Station or node.

Program is considering the default value since no value is entered at the prompt.

Capacity of the queue in front of the second Station or node = 5

Assume Station #1 and Station #2 have the same number of parallel servers

Number of patients at the second Station (including Patients in service) = 10

The number of states of the server(s) at the station #2 = 12

Server(s) on vacation, when s2=0;

Server(s) at station #2 is busy, when 1<=s2=10;

Servers(s) at station #2 are idle, when s2=11;

Enter the vacation rate per hour at station 2 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Server vacation rate per hour at station 2 is 2.00

Enter the Service Resume Rate per hour at station 2 [Example to enter: 5]:

Program is considering the default value since no value is entered at the prompt.

Service Resume Rate per hour at station 2 is 5.00

****** LAMDA_A, ALPHA_A & S_A FOR for Patient Class A at STATION 1 *****

Input the arrival rate per hour for Patient Class A [Example to enter: 4]:

Program is considering the default value since no value is entered at the prompt.

Arrival rate at station #1 for Patient Class A is 4 per hour.

Enter the number of phases of service at station 1 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 1 is 2

Enter the value of the vector alpha [Example to enter: [0.4 0.6]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.4 0.6

Size of the alpha vector is 2.

Value of the phase type distribution

[Example to enter: [-0.7 0.4; 0.8 -0.6]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix is 2

Value of the phase type distribution:

-0.7000 0.4000

0.8000 -0.6000

****** LAMDA_B, ALPHA_B & S_B FOR for Patient Class B at STATION 1 ******

Input the arrival rate per hour for Patient Class B [Example to enter: 4]:

Program is considering the default value since no value is entered at the prompt.

Arrival rate at station #1 for Patient Class B is 7 per hour.

Enter the number of phases of service at station 1 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 1 is 3

Enter the value of the vector alpha [Example to enter: [0.2, 0.6, 0.2]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.2 0.6 0.2

Size of the alpha vector is 3.

Value of the phase type distribution

[Example to enter: [-0.5 0.2 0.3; 0.02 -0.5 0.3; 0.03 0.2 -0.23]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix is 3

Value of the phase type distribution:

-0.5000 0.2000 0.3000

0.0200 -0.5000 0.3000

0.0300 0.2000 -0.2300

********** BETA_A & T_A for Patient Class A at STATION 2 **********

Enter the number of phases of service at station 2 [Example to enter: 2]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 2 is 2

Enter the value of the vector beta [Example to enter: [0.4 0.6]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.4 0.6

Size of the alpha vector is 2.

Value of the phase type distribution

[Example to enter: [-0.7 0.4; 0.8 -0.6]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix at station 2 is 2

Value of the phase type distribution:

-0.7000 0.4000

0.8000 -0.6000

********** BETA_B & T_B for Patient Class B at STATION 2 **********

Enter the number of phases of service at station 2 [Example to enter: 3]:

Program is considering the default value since no value is entered at the prompt.

Number of phases of service at station 2 is 3

Enter the value of the vector beta [Example to enter: [0.2, 0.6, 0.2]]:

Program is considering the default value since no value is entered at the prompt.

Value of the vector alpha is: 0.2 0.6 0.2

Size of the alpha vector is 3.

Value of the phase type distribution

[Example to enter: [-0.5 0.2 0.3; 0.02 -0.5 0.3; 0.03 0.2 -0.23]]:

Program is considering the default value since no value is entered at the prompt.

Size of the PH-distribution matrix at station 2 is 3

Value of the phase type distribution:

-0.5000 0.2000 0.3000

0.0200 -0.5000 0.3000

0.0300 0.2000 -0.2300

PROGRAM RUN RESULT:

FOR PATIENT CLASS A:

Mean length of the queue is 0.11111

Mean waiting time in the queue is 1.6667 minutes

FOR PATIENT CLASS B:

Mean length of the queue is 0.11111

Mean waiting time in the queue is 0.9524 minutes

4.9 Sensitivity Analysis:

In order to study this model as a Continuous Parameter Markov Process, large and complex matrices need to be manipulated. Also, careful implementation of the computer coding for this complex system and sufficient computer memory space is essential. Computer programs are implemented in the MATLAB software. The number of the transition states depends on the number of parallel identical servers at each station, the size of the used intermediate finite queue placed in between the two stations and also on the number of phases of service at each station. Different types of performance measures, such as, mean queue length, mean waiting time in the queue are calculated by using the implemented MATLAB program. These performance measures depend on some variables, which are described below:

Table 4.9.1: Relation between No. of Servers and Mean Waiting Time when n=q

No. of Servers	Mean Waiting Time (sec)	
	Patient Class A	Patient Class B
1	583.422	333.498
2	299.616	171.252
3	180.018	102.87
4	128.574	73.470
5	100.002	57.144
6	81.816	46.752
7	69.228	39.558
8	60.000	34.284
9	52.944	30.252
10	47.37	27.066

Table 4.9.1 shows the relation between the numbers of servers and mean waiting time for patients class A and also the relation between the number of servers and mean waiting time for patients class B when number of server is equal to the capacity of the queue in front of the second station (i.e. $n=q$). For both patients' class, the value of mean waiting time is decreasing significantly with the increase of the number of identical parallel servers at each station. Figure 4.9.1 shows the relation between the numbers of servers and mean waiting time for patients class A and also the relation between the number of servers and mean waiting time for patients class B when number of server is equal to the capacity of the queue in front of the second station (i.e. $n=q$).

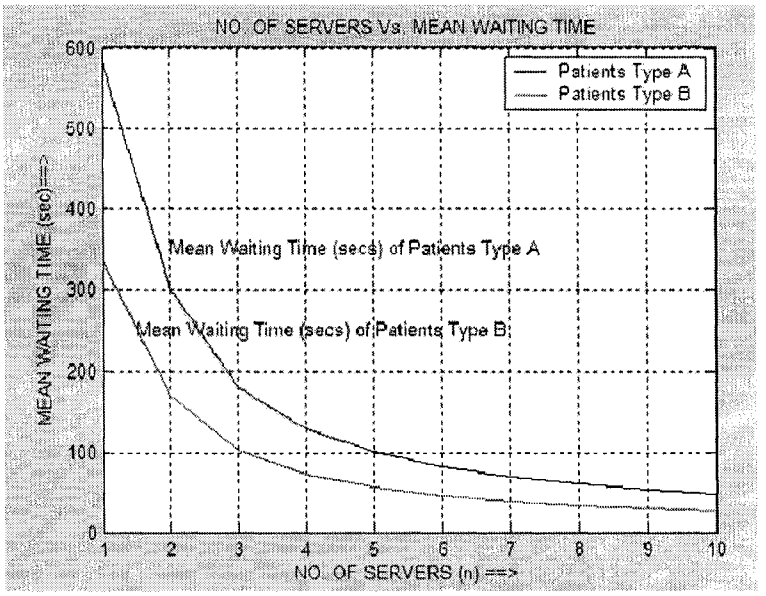

Figure 4.9.1: Relation between No. of Servers and Mean Waiting Time when n=q

Table 4.9.2 shows the relation between the numbers of servers and mean waiting time for patients class A and also the relation between the number of servers and mean waiting time for patients class B when the capacity of the queue in front of the second station is 2 (i.e. $q=2$). For both patients' class, the value of mean waiting time is decreasing significantly with the increase of the number of identical parallel servers at each station. Figure 4.9.2 shows the relation between the numbers of servers and mean waiting time for patients class A and also the relation between the number of servers and mean waiting time for patients class B when the capacity of the queue in front of the second station is 2 (i.e. $q=2$).

Table 4.9.2: Relation between No. of Servers and Mean Waiting Time when q=2

No. of Servers	Mean Waiting Time (sec)	
	Patient Class A	Patient Class B
1	420.9221	244.7138
2	287.2907	171.2503
3	225.7947	128.6485
4	179.1256	102.8688
5	150.0577	85.7154
6	128.5765	73.4695
7	112.4995	64.2857
8	100.001	57.1429
9	90.0001	51.4286
10	81.8182	46.7532

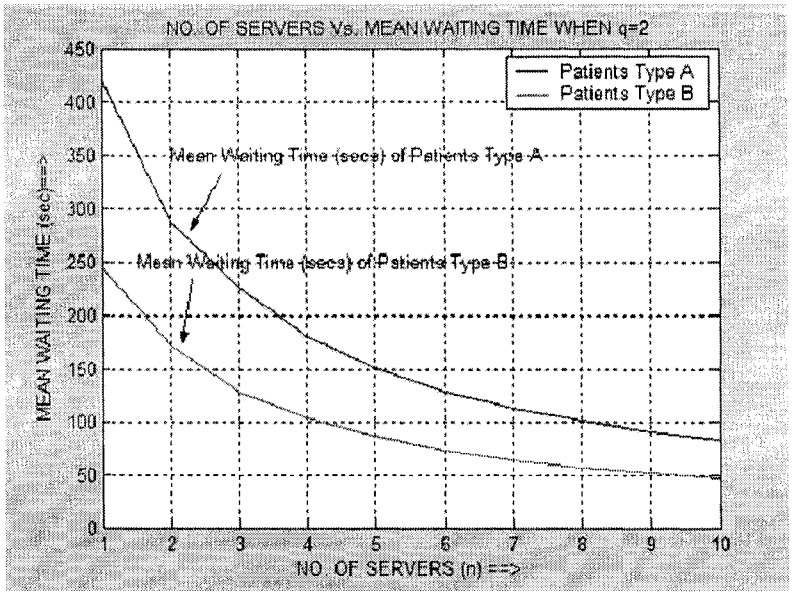

Figure 4.9.2: Relation between No. of Servers and Mean Waiting Time when q=2

Table 4.9.3: Relation between Mean Queue Length and Program Run Time with No. of Servers

For Both Patients Class			
No. of Servers	**Mean Queue Length**	**No. of Servers**	**Program Run Time (sec)**
1	0.46769	1	0.063
2	0.31921	5	0.11
3	0.25088	10	0.187
4	0.19903	15	0.188
5	0.16673	20	0.312
6	0.14286	25	0.468
7	0.1250	30	0.781
8	0.11111	35	1.109
9	0.10000	40	1.64
10	0.090909	45	2.265

Table 4.9.3 shows the relation between the numbers of servers and means queue length for both patients class when the capacity of the queue in front of the second station is 2 (i.e. $q=2$). Also the relation between the number of servers and program run time for patients class A and B when the capacity of the queue in front of the second station is 2 (i.e. $q=2$). For both patients' class, the value of mean queue length is decreasing significantly with the increase of the number of identical parallel servers at each station. Figure 4.8.3 shows the relation between the numbers of servers and mean queue length for both patients class when the capacity of the queue in front of the second station is 2 (i.e. $q=2$). Also Figure 4.9.4 shows the relation between the number of servers and program run time for both patients' class when the capacity of the queue in front of the second station is 2 (i.e. $q=2$). The shape of the curve is parabolic when program run time increases with the increase of the number of identical parallel servers at each station.

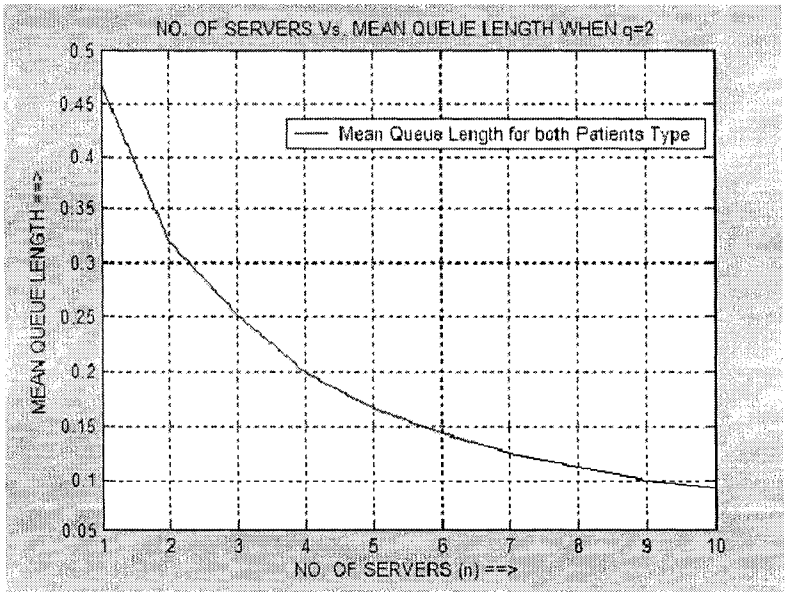

Figure 4.9.3: Relation between No. of Servers and Mean Queue Length when q=2

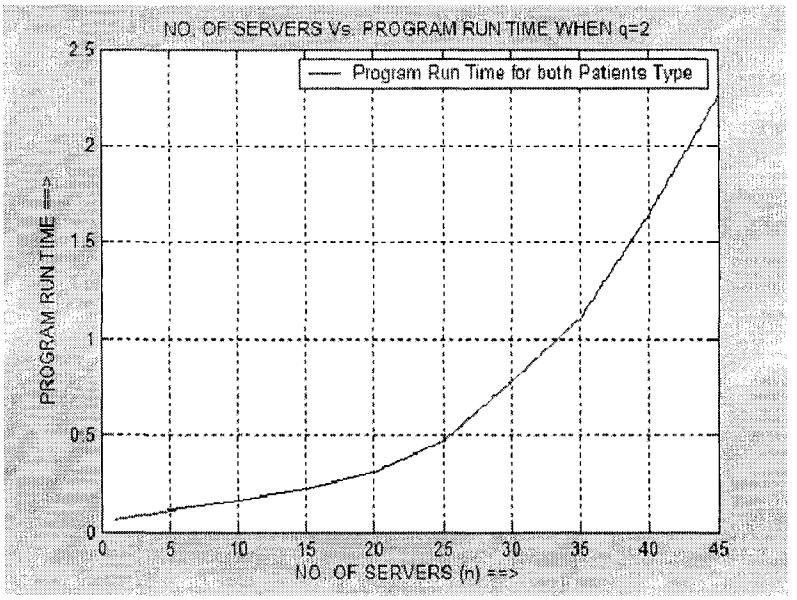

Figure 4.9.4: Relation between No. of Servers and Program Run Time when q=2

Table 4.9.4: Mean Waiting Time with different value of server vacation rate θ

No. of Servers (n)	Patient Class A		Patient Class B	
	Mean Waiting Time when θ=2 per hour	Mean Waiting Time when θ=0.25 per hour	Mean Waiting Time when θ=2 per hour	Mean Waiting Time when θ=0.25 per hour
1	420.9221	408.2074	244.7138	233.0716
2	287.2907	283.0685	171.2503	169.1947
3	225.7947	222.6809	128.6485	128.4057
4	179.1256	169.0558	102.8688	102.847
5	150.0577	150.0393	85.7154	85.7138

Mean waiting time depends on server vacation rate. Table 4.9.4 shows the relation between the numbers of servers and mean waiting time for both patients' class when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server vacation rate changes. For both patients' class, the value of mean waiting time is decreasing significantly with the increase of the number of identical parallel servers at each station. Figure 4.9.5 shows the relation between the numbers of servers and mean waiting time for both patients, class A when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server vacation rate changes. Also figure 4.9.6 shows the relation between the number of servers and program run time for both patients' class B when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server vacation rate changes. So, it is clear from figure 4.9.5 and figure 4.9.6 that patients' mean waiting time in the queue depends on server vacation rate.

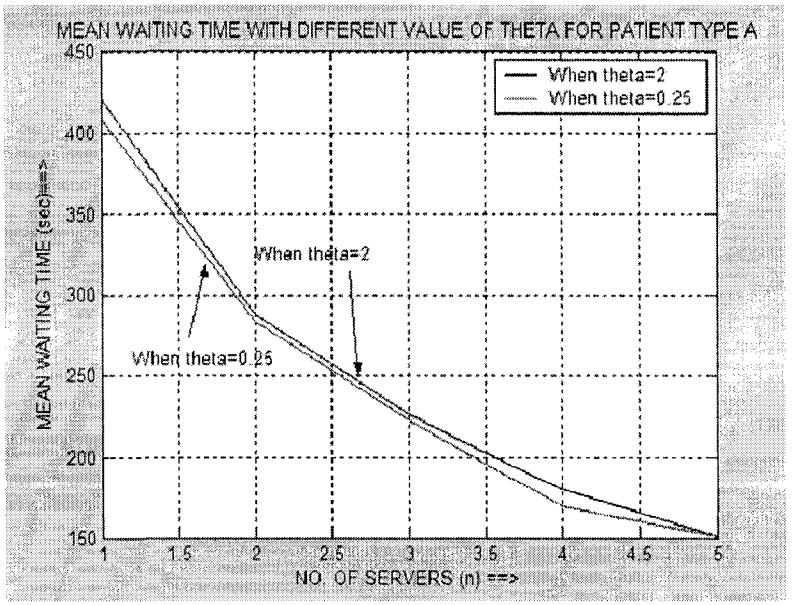

Figure 4.9.5: Mean Waiting Time with different value of θ for Patients Class A

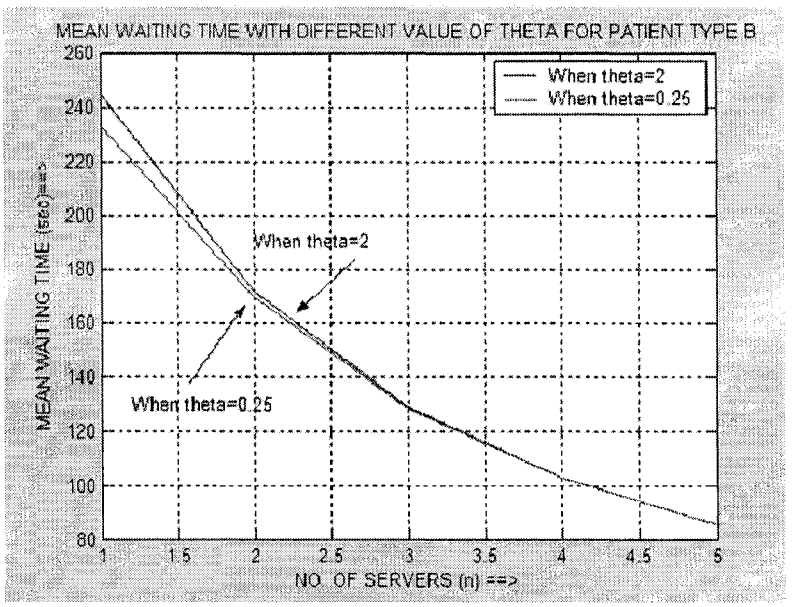

Figure 4.9.6: Mean Waiting Time with different value of θ for Patients Class B

Table 4.9.5: Mean Waiting Time with different value of server resume rate γ

No. of Servers (n)	Patient Class A		Patient Class B	
	Mean Waiting Time when γ=5 per hour	Mean Waiting Time when γ=0.25 per hour	Mean Waiting Time when γ=5 per hour	Mean Waiting Time when γ=0.25 per hour
1	420.9221	1441.9044	244.7138	733.4933
2	287.2907	1311.2805	171.2503	394.7278
3	225.7947	703.0342	128.6485	171.7468
4	179.1256	201.7443	102.8688	107.7921
5	150.0577	157.0877	85.7154	86.1167

Mean waiting time depends on server resume rate. Table 4.9.5 shows the relation between the numbers of servers and mean waiting time for both patients' class when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server resume rate changes. For both patients' class, the value of mean waiting time is decreasing significantly with the increase of the number of identical parallel servers at each station. Figure 4.9.7 shows the relation between the numbers of servers and mean waiting time for both patients, class A when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server resume rate changes. Also figure 4.9.8 shows the relation between the number of servers and program run time for both patients' class B when the capacity of the queue in front of the second station is 2 (i.e. $q=2$) and server resume rate changes. So, it is clear from figure 4.9.7 and figure 4.9.8 that patients' mean waiting time in the queue depends on server resume rate.

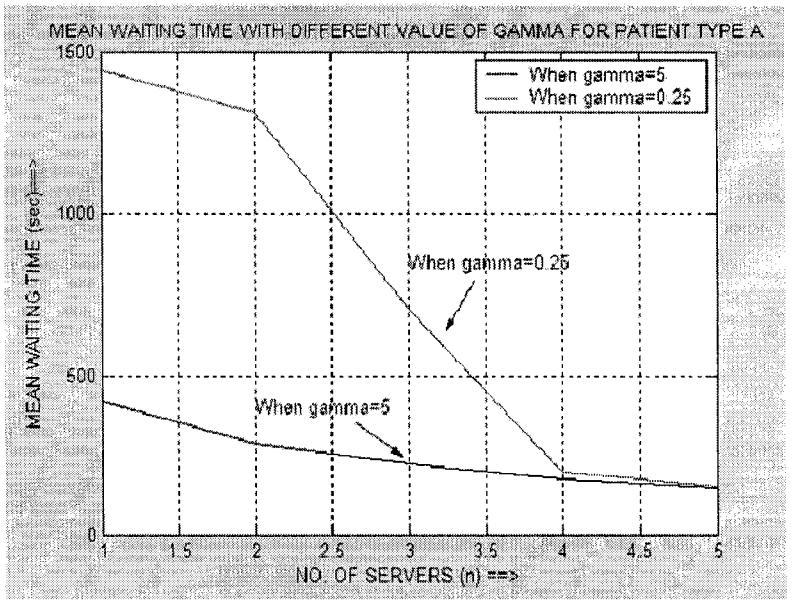

Figure 4.9.7: Mean Waiting Time with different value of γ for Patients Class A

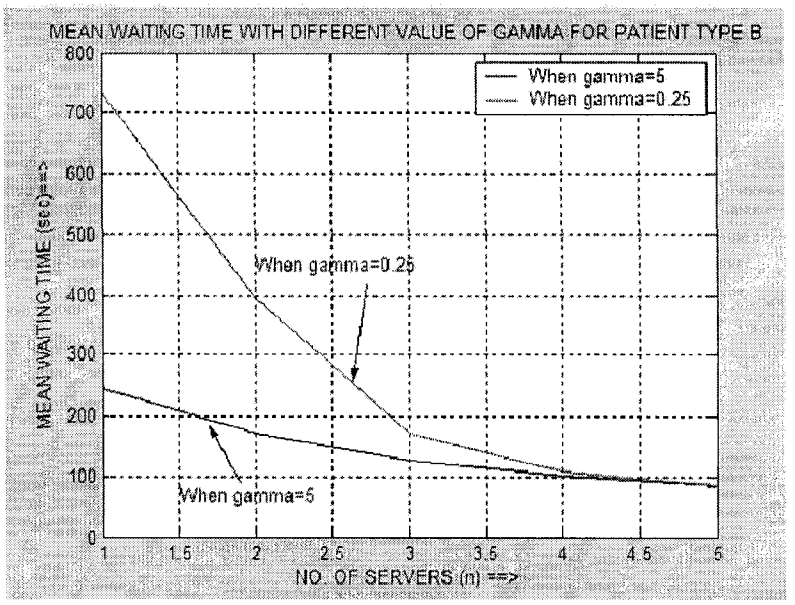

Figure 4.9.8: Mean Waiting Time with different value of γ for Patients Class B

4.10 Conclusions:

In this chapter, two node *M/PH/1* and *M/PH/n* queueing system is analyzed by using state model method by implementing MATLAB computer programs. The system is subject to blocking and vacation. Numerical examples are used to analyze this M/PH/1 and *M/PH/n* queueing system. MATLAB program is implemented using the above matrices. This implemented MATLAB program will determine different performance measures; such as, mean queue length and mean waiting time of the patients flow line system.

In this chapter, first a two node *M/PH/1* queueing system is analyzed by using state model method by implementing MATLAB computer programs in section 4. In section 5, a two node *M/PH/n* queueing system is analyzed by using state model method by implementing MATLAB computer programs. Both systems are subject to blocking and unreliable. Gaver's method to calculate the probability distribution is described in section 6. The purpose of the implemented program is described in section 7. In section 8, numerical examples are represented with the implemented program run results. Sensitivity analysis is described in section 9 from the implemented MATLAB program. From this analysis, it is apparent that the series-parallel patients' flow line systems are more efficient than the patients' flow line systems in series, which is described in this research.

CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

5.1 Conclusions:

For many healthcare services or medical procedures, patients have extensive risk of complication or face death when treatment is delayed. When a queue is formed in such a situation, it is very important to assess the suffering and risk faced by patients in queue and plan sufficient medical capabilities in advance to address the concerns. As the diversity of care settings increases, congestion in facilities causes many patients to unnecessarily spend extra days in intensive care facilities. Performance evaluation of current healthcare service systems using queueing theory gains more and more importance because of patient flows and system complexity. Although queueing models have been used in numerous healthcare studies, the inclusion of blocking is still rare. In this research work, we study an efficient two-stage multi-class queueing network system with blocking and phase type service time distribution to analyze such congestion processes. We also consider parallel servers at each station and first-come-first-serve non-preemptive service discipline are used to improve the performance of healthcare service systems. Such a queueing model has potential in performance evaluation of patients' flow line systems. From this research, it is apparent that the series-parallel patients' flow line systems are more efficient than the patients' flow line systems in series, which is described in this research. The modeling approach presented in this research work is flexible and it could be applied to a variety of health care areas, such as

blood banks, outpatient clinics, emergency rooms, or diagnosis systems. Any type of healthcare service systems related to queues would benefit from this research.

5.2 Contributions:

This research has resulted in the following contributions in the field of patients' flow line systems:

- Considering continuous time Markov chain with discrete states.

- The two-stage tandem queueing model is developed for patients' flow line systems with unreliable servers and finite queue in front of the second station. This model is easy to implement in real patients' flow line systems.

- To analyze the system the blocking is also considered in this research work. This type of blocking i.e. blocking-after-service is a widespread type of blocking in the patients' flow line systems.

- The phase type service distribution is applied to analyze the developed model using *M/PH/1* and *M/PH/n* queueing systems. To analyze the systems, computer programs are implemented using the software MATLAB.

- The series-parallel queueing model is developed for patients' flow line systems with unreliable servers and finite buffer. This model is easy to implement in real healthcare service systems.

- Multiple-class instead of considering only one class patients' class is considered to evaluate the performance of the patients' flow line systems in this research.

- This research studies mean queue length of the finite queue in front of the second station, and patients' mean waiting time in the queue simultaneously.

The ultimate goal of this research is the application of the queueing model to real-life problems with phase type service distribution.

5.3 Future Works:

A solution to a research problem always leads to many other interesting problems which remain to be solved. This research is no exception, and provides numerous opportunities for improvement and extension. Some of key future directions are pointed out below:

The following works could be conducted in future on this research:

- Optimization of the current research could be possible.

- Comparison with other computation methods has to be done in the future.

- A significant direction for future effort is the expansion to more than two stages i.e. multiple stages in patients' flow line systems with all the factors that are considered in this research.

- Patients' death rate is not considered here. So Patients' death rate or backorder could be considered in the future research work.

REFERENCES:

Albin, S. L., Barrett, J. Ito, D. and Mueller, J. E. 1990. A queueing network analysis of a health center. *Queueing Systems*. 7: 51 – 61.

Altmel, I.K. and Ulas, E. 1996. Simulation modeling for emergency bed requirement planning, *Annals of Operations Research*, 67: 183 – 210.

Altiok, T. and Shiue, G.A. 2000. Pull-type manufacturing systems with multiple product types. *IIE Transactions*. 32: 115-124.

Altiok, T. 1996. Performance analysis of manufacturing systems. *Springer Series in Operation Research. Springer, USA.*

Altiok, T. and Shiue, G.A. 1995. Single-stage, multi-product production/inventory systems with lost sales. *IIE Transactions*. 42: 889-913.

Altiok, T. and Shiue, G.A. 1994. Single-stage, multi-product production/inventory systems with backorders. *IIE Transactions*. 26: 52-61.

Altiok, T. 1989a. Approximate analysis of queues in series with phase type service times and blocking. *Operation Research. 37(4)*: 601-609.

Altiok, T. and Ranjan, R. 1989b. Analysis of production lines with general service times and finite buffers: a two-node decomposition approach. *Engineering Costs and Production Economics. 17*: 155-165.

Altiok, T. 1989c. (R, r) Production/inventory systems. *Operation Research. 37(2)*: 266-276.

Altiok, T. 1985. Production lines with phase type operation and repair times and finite buffers. *International Journal of Production Research*. 23(3): 489-498.

Alves, R. 1990. Performance evaluation of series-parallel systems. *Journal of Manufacturing Operation Management,* 3: 224 – 250.

Amin, M. and Altiok, T., 1997, Control policies for multi-product, multi-stage manufacturing systems: an experimental approach. *International Journal of Production Research*, 35(1): 201-223.

Artalejo, J. R. Economou, A. and Lopez-Herrero, M. J. 2005, Analysis of a multi-server queue with setup times. *Queueing Systems*, 52: 53-76.

Askin, R. (1993). Modeling and analysis of manufacturing systems. Wiley, New York.

Atkinson, J. B., Kovalenko, I. N., Kuznetsov, N. Y. and Mikhalevich, K. V. 2006, Heuristic methods for the analysis of a queueing system describing emergency medical service deployed along a highway. *Cybernetics and Systems Analysis*, 42(3): 379 - 391.

Au-Yeung, S. W. M., Harrison, P. G. and Knottenbelt, W. J. 2006. A queueing network model of patient flow in an accident and emergency department. *Department of Computing*, Imperial College, London, UK.

Bihan, H. L. and Dallery, Y. 2000. A robust decomposition method for the analysis of production lines with unreliable machines and finite buffers. *Annals of Operation Research*, 93: 265-297.

Bracht Erik Van. 1995. Performance Analysis of a serial production line with machine breakdowns. *IEEE.* 4: 417-424

Bretthauer, K. M. and Côté, M. J. 1998. A model for planning Resource Requirements in Health Care Organizations. *Decision Sciences*. 29(1): 243 – 270.

Buzacott, J. A. 1971. The role of inventory in flow-line production systems, *International Journal of Production Research*, 9(4): 425-436.

Buzacott, J. A. 1967. Automatic transfer lines with buffer stocks, *International Journal of Production Research*, 5(3): 183-199.

Chaussalet, T. J. 2006. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*. 45(5): 492 – 497.

Cochran, J. K., and Bharti, A. 2006a. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*. 9: 31 – 45.

Cochran, J. K., and Bharti, A. 2006b. A multi-stage stochastic methodology for ehole hospital bed planning under peak loading. *International Journal of Industrial and Systems Engineering*. 1(1): 08 – 36.

Cohen, M. A., Hershey, J. C. and Weiss, E. N. 1980. Analysis of capacity decisions for progressive patient care hospital facilities. *Health Services Research*. 15: 145 – 160.

Cox, D. R.1955. A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society,* 51: 313-319.

Creemers, S., and Lambrecht, M. R. 2006. Modeling a healthcare system as a queueing network: the case of a Belgian hospital. *Department of Decision Sciences & Information Management, Research Center for Operations Management, Katholieke University Leuven, Belgium.*

Dallery, Y. and Gershwin, S. B. 1992. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems*, 12: 3-94.

Davis, S. G. 1981.Analysis of the deployment of emergency medical services. *Omega*, 9(6):655 - 657.

El-Darzi et al. 1998. A simulation modeling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Management Science, 1:143-149.

Elsayed, E. A. and Hwang, C. C. 1986. Analysis of two-stage manufacturing systems with buffer storage and redundant machines. *International Journal of Production Research*, 24(1): 187-201.

Erlang, A. K. 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *P. O. Electrical Engineering Journal 10*: 189-197.

Gaver. D. P., Jacobs, P. A. and Latouche, G. 1984. Finite birth-and-death models in randomly changing environments. *Advances in Applied Probability.* 16 (4): 715 – 731.

George. J. A., Fox, D. R. and Canvin, R. W. 1983. A hospital throughput model in the context of long waiting lists. *Journal of Operational Research Society.* 34: 27 – 35.

Gershwin, S. B., Dallery, Y., Papadopoulos, C. T. and Smith, J.M. 2000. Efficient algorithm for buffers space allocation. *Annals of Operation Research*, 93:117-144.

Gershwin, S. B. 1987. An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operation Research*, 35(2): 291-305.

Gómez-Corral, A. and Martos, M.E. 2006. Performance of two-stage tandem queues with blocking: The impact of several flows of signals. *Performance Evaluation*, 63: 910 – 938.

Gómez-Corral, A. and Martos, M.E. 2002. A tandem queue with blocking and Markovian arrival process. *Queueing Systems*, 41: 343-370.

Gonzalez-Rubio, R. 2005. Soleil: modeling a health system. *3rd International Industrial Simulation Conference*. pp. 129 – 136.

Gorunescu, F. McClean, S. I. and Millard, P. H. 2002. A queueing model for bed-occupancy management and planning of hospitals. *Journal of Operational Research Society*. 53: 19 – 24.

Green, L. V. and Nguyen, V. 2001. Strategies for cutting hospital beds: the impact on Patient service. *Health services Research*. 36(2): 421 –442.

Griffiths, J. D. and Price-Lloyd, N. 2006. A queueing model of activities in an intensive care unit. *Journal of Management Mathematics*. 17: 277 – 288.

Gün, L. and Makowski, A. M.1989. Matrix-geometric solution for two node tandem queueing systems with phase type servers subject to blocking and failures. *Commun. Statist.-Stochastic Models,* 5(3): 431-457.

Hillier, F. and So, K. C. 1995. On the optimal design of tandem queueing systems with finite buffers. *Queueing Systems*, 21: 245-266.

Hillier, F. S. and So, K. C. 1991. The effect of machine breakdowns and interstage storage on the performance of production line systems. *International Journal of Production Research.* 29(10): 2043 – 2055.

Hillier, F. S. and So, K. C. 1989. The assignment of extra servers to stations in tandem

queueing systems with small or no buffers. *Performance evaluation.*10: 219 – 231.

Hillier, F. S. and Boling, R. W. 1967. Finite Queues in series with exponential or Erlang service times – a numerical approach. *Operations Research.* 15(2): 286 – 303.

Houtum, G. J. V., Adan, I. J. B. F., Wessels, J. and Zijm, W. H. M. 2001. Performance analysis of parallel identical machines with a generalized shortest queue arrival mechanism. OR Spektrum, 23: 411-427.

Huang, E. W. and Liou, D. M. 2007. Performance analysis of a medical record exchanges model. *IEEE Transactions on Information Technology in Biomedicine*, 11(2): 153 - 160.

Hunt, G. C. 1956. Sequential arrays of waiting lines. Operation Research, 4: 674-683.

Isotupa, K. P. S. and Stanford, D. A. 2002. An infinite-phase quasi-birth-and-death model for the non-preemptive priority M/PH/1 queue. *Stochastic Models.* 18(3): 387-424.

Jackson, J. R. 1957. Networks of waiting lines. Operation Research, 5: 518-521.

Jun, J. B., Jacobson, S. H. and Swisher, J. R. 1999. Application of discrete event simulation in health care clinics: A survey. *Journal of Operational Research Society.* 50(2): 109 – 123.

Kim, S., Horowitz, I., Young, K. K. and Buckley, T. A. 1999. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research.* 115: 36 – 46.

Koizumi, N. 2002. A queueing network Model with blocking: analysis of congested

patients flows in mental health systems. PhD Dissertation in Regional Science, University of Pennsylvania, USA.

Koizumi, N., Kuno, E. and Smith, T. E. 2005. Modeling patient flows using a queueing network with blocking. *Health Care Management Science*. 8: 49 – 60.

Latouche, G. and Ramaswami, V. 1999. Introduction to matrix analytic methods in stochastic modeling. *Society for Industrial and Applied Mathematics (SIAM) and American Statistical Association (ASA), USA.*

Li, J. and Huang, N. 2005. Modeling and analysis of a multiple product manufacturing system with split and merge. *International Journal of Production Research*; 43(19): 4049-4066.

Liyanage, L. and Gale, M. 1995. Quality improvement for the Campbelltown hospital emergency service.

Marshall, A. H. Shaw, B. and McClean, S. I. 2007. Estimating the costs for a group of geriatric patients using the Coxian phase type distribution. *Statistics in Medicine.* 26: 2716 – 2729.

Magazine, M. J. and Stecke, K. E. 1996. Throughput for production lines with serial work stations and parallel service facilities. *Performance Evaluation.* 25: 211 – 232.

McManus, M. L. and Long, M. C. 2004. Queueing theory accurately models the need for critical care resources. *Anesthesiology*. 100: 1271 – 1276.

Neuts, M. F., 1981, Matrix-Geometric solutions in Stochastic Models: An Algorithmic approach, The Johns Hopkins University Press, Baltimore, Maryland, London.

Neuts, M. F., 1973, Computer power or the liberation of applied probability. *Technical Report 312, Department of Statistics, Purdue University, West Lafayelle, IN, USA.*

Panayiotopoulos, J. C. and Vassilacopoulos, G. 1984. Simulating hospital emergency departments queueing systems: (GI/G/m(t)): (IHFF/N/∞). European *Journal of Operational Research*. 18: 250 – 258.

Papadopoulos, H. T., Heavey, C. and Browne, J. 1993. Queueing theory in manufacturing systems analysis and design. *CHAPMAN &HALL, Cambridge, Great Britain.*

Patchong, A. and Willaeys, D. 2001. Modeling and analysis of an unreliable flow line composed of parallel-machine stages. *IIE Transaction*; 33: 559 – 568.

Perros, H. G. 1994. Queueing networks with blocking. *Oxford university press, Oxford.*

Perros, H. G. and Altiok, T. 1986. Approximate analysis of open networks of queues with blocking: Tandem configurations. *IIE Transaction*. 17(2): 110-116.

Preater, J. 2002. Queues in health. *Health Care Management Science*. 5: 283 – 283.

Preater, J. 2001. A bibliography of queues in health and medicine. *Mathematics Department, Keele University, UK.*

Shimatonis, K. S. 1983. A simulation model for medial organization: Application to a cardiological outpatient department. *Translated from Medltslnskaya Tekhnika*. 5: 8 – 11.

Shiue, G.A. and Altiok, T. 1993. Two-stage, multi-product production/inventory systems. *Onvural and Akyildiz, (Eds.), Proc. Second International Workshop on Queueing Networks with Finite Capacity, (North-Holland, Amsterdam*, 213-223.

Singh, V. 2006. Use of queueing models in health care: Decision analysis. *Department of*

Health Policy and Management, University of Arkansas for Medical Sciences. pp. 1 - 34.

Sleptchenko, A. Harten, A. and Heijden, M. V. 2005. An exact solution for the probabilities of the multi-class, multi-server queue with preemptive priorities. *Queueing Systems*, 50: 81-107.

Tempelmeier, H. and Bürger, M. 2001. Performance evaluation of unbalanced flow lines with general distributed processing times, failures and imperfect production. *IIE Transactions*; 33: 293 – 302.

Tolio, T., Matta, A. and Gershwin, S. B. 2002. Analysis of two-machine lines with multiple failure modes. *IIE Transactions*, 34: 51-62.

Van Nyen, P. L. M., Bertrand, J. W. M., Van Ooijen, H. P. G. and Vandaele, N. J. 2005. A heuristic to control integrated multi-product multi-machine production-inventory systems with job shop routings and stochastic arrival, set-up and processing times. *Operation Research Spectrum (OR Spectrum)*, 27: 399-434.

Vandaele, N., Boeck, L. D. and Callewier, D. 2002. An open queueing network for lead time analysis. *IIE Transaction*. 34: 1-9.

Vuuren, M. V, Adan, I. J. B. F. and Resing-Sassen, S. A. E. 2005. Performance analysis of multi-server tandem queues with finite buffers and blocking. *Operation Research Spectrum (OR Spectrum)*, 27: 315-338.

Wang, Q. 2004. Modeling and analysis of high risk patient queue. *European Journal of Operational Research*. 155: 502 – 515.

Willig, Andreas. 1999. A short introduction to Queueing theory. *Telecommunication Networks Grou, Technical University Berlin, Berlin.*

Worthington, D. J. 1987. Hospital waiting list models. *The Journal of Operational Research Society*. 42(10): 833 – 843.

Worthington, D. J. 1987. Queueing models for hospital waiting lists. *The Journal of Operational Research Society*. 38(5): 413 – 422.

Xie, H., Chaussalet, T. and Rees, M. 2007. A semi-open queueing network approach to the analysis of patient flow in healthcare systems. *Computer-Based Medical Systems, 2007. CBMS '07. Twentieth IEEE International Symposium on 20-22 June*. pp. 719-724.

Yamashita, H. and Altiok, T. 1998. Buffer capacity allocation for a desired throughput in production lines. *IIE Transaction*. 30: 883-891.

Yang, X. 2003. A class of multi-server queueing systems with unreliable servers: models and application. *MASc. Thesis, University of Windsor, Windsor, Ontario, Canada*.

Zhao, J., Li, B., Cao, Xi. and Ahmed, I. 2006. A matrix-analytic solution for the DBMAP/PH/1 priority queue. *Queueing Systems*, 53: 127-145.

Zhao, Y. and Grassmann, W. K. 1991. A numerical stable algorithm for two server queue models. *Queueing Systems*, 8: 59-80.

APPENDIX A

QUEUEING MODELS

This appendix is provided to give brief description on queueing models:

Queueing models have been proved to be very useful in many practical applications in areas such as, e.g. production systems, inventory systems and communication systems. In many applications the variability in the arrival and service processes greatly affect the performance of the system. Queueing models help us to understand and quantify the effect of variability. This field is referred to as Queueing Theory.

Queue: a queue is a line of waiting customers who require service from one or more service providers.

Queueing System: The system that contains the combination of waiting lines/queues, customers and service providers is termed as a queueing system.

Kendall-Lee Notation (Kendall Notation):

The Kendall notation is used for a short characterization of queueing systems. A queueing system description looks as follows:

A/S/n/B/K/SD

where,

A → Arrival process

S → Service time distribution

n → Number of servers

B → Capacity of the buffer/queue

K → Population size

SD → Service discipline

Performance Measures: Relevant performance measures in the analysis of queueing models are:

i. Server utilization

ii. Length of waiting lines

iii. Delays of customers

Little's Law: Little's Law gives a very important relation between the mean number of customers in the system $E(L)$, the mean sojourn time or throughput time $E(S)$ and the average number of customers entering the system per unit time λ. Little's law states that:

Mean number in the system = Arrival rate x Mean sojourn time

$$i.e.\ E(L) = \lambda\, E(S) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(A.1)}$$

Queueing Networks:

(i) **Open Network:** In an open network new customers may arrive from outside the system (coming from a conceptually infinite population) and later leave the system.

(ii) **Closed Network:** In a closed queueing network the number of customers is fixed and no customer enters or leaves the system.

Some Common Types Of Queueing Systems:

M/M/1, M/M/1/K, M/M/n, M/G/n, M/D/n, D/D/1, G/M/1, M/PH/1 etc.

APPENDIX B

MARKOV PROCESS

This appendix is provided to give some idea about the Markov process:

The common characteristic of all Markovian Systems is that distributions, namely the distribution of the interarrival times and the distribution of the service times are exponential distributions and thus exhibit the Markov (memoryless) property. The Markov property means the system is memoryless, i.e. it does not remember the states it was in before, just knows its present state, and hence bases its decision to which future state it will transit, purely on the present, not considering the past.

A stochastic process $\{X(t) \mid t \in T\}$ is called a Markov process if for any $t_0 < t_1 < t_2 < \cdots < t_n < t$, the conditional distribution of $X(t)$ for given values of $X(t_0), X(t_1), \ldots, X(t_n)$ is independent of $X(t_0), X(t_1), \ldots, X(t_{n-1})$; that is:

$$P\left[X(t) \le x \middle| X(t_n) = x_n,\ X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0\right] = P\left[X(t) \le x \middle| X(t_n) = x_n\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(B.1)}$$

This definition applies to a Markov process with continuous state space.

The Markov process with discrete-state and discrete time is termed as Markov chain. A Markov chain is a sequence $X_1, X_2, X_3, \ldots$ of random variables with the property (Markov property): the conditional probability distribution of the next future state X_{n+1} given the present and past states is a function of the present state X_n alone, i.e.

$$P\left(X_{n+1}=x\middle|X_0=x_0,X_1=x_1,\ldots,X_n=x_n\right)=P\left(X_{n+1}=x\middle|X_n=x_n\right)$$

..... (B.2)

The range of the variables, i.e. the set of their possible values, is called the state space, the value of X_n being the state of the process at time n.

APPENDIX C

QUASI-BIRTH-AND-DEATH PROCESS

This appendix is provided to give a description of Quasi-Birth-and Death process (QBD process) according to Neuts (1981) and Papadopoulos et al. (1993). Birth-death process is another important Markovian process with regard to queueing theory. The transitions can only take place between neighbouring states. QBD process is a Markov process on the state space $E = \{(i, j),\ i \geq 0,\ 1 \leq j \leq m\}$, with infinitesimal generator Q that describes the model, has the following block tri-diagonal form:

$$Q = \begin{vmatrix} B_0 & A_0 & & & & & & \\ B_1 & A_1 & A_0 & & & & & \\ & A_2 & A_1 & A_0 & & & & \\ & & A_2 & A_1 & A_0 & \cdots & & \\ & & & A_2 & A_1 & \cdots & & \\ & & & & \cdots & & & \\ & & & & \cdots & & & \\ & & & & \cdots & & & \end{vmatrix}$$

where $B_0e + A_0e = B_1e + A_1e + A_0e = (A_0 + A_1 + A_2)\ e = 0$. Here Q is also assumed to be irreducible. The matrix $A = A_0 + A_1 + A_2$ is a finite generator. The equilibrium equations $\pi Q = 0$ can be expressed in the matrix-difference form as

$$\pi_k A_0 + \pi_{k+1} A_1 + \pi_{k+2} A_2 = 0 \qquad \text{for k} = 0, 1, 2, \ldots,$$

..... (C.1)

and

$$\pi_0 B_0 + \pi_1 B_1 = 0 \quad \text{..... (C.2)}$$

for the boundary equations, where,

(i) A_0 is an *(m x m)*-matrix describing the transitions in the subnetwork, which simultaneously produce inputs to the first queue;

(ii) A_1 is an *(m x m)*-matrix describing the transitions in the subnetwork, which produce neither inputs to nor outputs from the first queue assuming that the queue is not empty;

(iii) A_2 is an *(m x m)*-matrix describing the transitions in the subnetwork, which simultaneously produce outputs from the first queue; and

(iv) B_0 is an *(m x m)*-matrix describing the transitions in the subnetwork, which simultaneously produce neither inputs to nor outputs from the first queue, assuming that the queue is empty.

A Markov process whose equilibrium equations have the form of equations (C.1) and (C.2) was termed by Wallace (1973) a Quasi-Birth-and-Death (QBD) process.

APPENDIX D

PHASE TYPE DISTRIBUTION

This appendix is provided to give a description of PH-distribution according to Neuts (1981). Also, the references that used PH distribution as the service time distributions are provided in this appendix.

The beauty of the phase type distribution is that it can model almost all the probabilistic distribution we encounter in our practical life. Phase type distributions are important in queueing theory because their structure can give rise to a Markovian state-description. Phase type (PH) distributions are based on the method of stages technique. The phase (or stage) concept was first introduced by Erlang (1917), later generalized by Cox (1955), and then generalized by Neuts (1981). Phase type distribution was considered as a service time or processing time by Altiok (1985), Gün and Makowski (1989), Hillier and So (1991), Yamashita and Altiok (1998), Gorunescu et al. (2002), Yang (2003), Gómez-Corral and Martos (2006), Zhao et al. (2006), Marshall et al. (2007).

Latouche and Ramaswami (1999) wrote a chapter on PH distributions for several reasons. PH distributions provide a simple framework to demonstrate how one can extend many simple results on exponential distributions to more complex models without losing computational tractability. Finally, phase models provide convenient examples for the

paradigms for which matrix analytic methods apply. Continuous PH distributions are the natural generalization of the exponential and Erlang distributions.

We consider a Continuous Time Markov process on the states $\{1, \ldots, m + 1\}$ with infinitesimal generator

$$Q = \begin{vmatrix} T & T^0 \\ 0 & 0 \end{vmatrix} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(D.1)}$$

where, the $m \times m$ matrix T satisfies $T_{ii} < 0$, for $1 \leq i \leq m$, and $T_{ij} \geq 0$, for $i \neq j$. Also $Te+T^0=0$, and the initial probability vector of Q is given by (α, α_{m+1}), with $\alpha e + \alpha_{m+1} = 1$. We assume that the states $1, \ldots, m$ are all transient, so that absorption into the state $m+1$, from any initial state, is certain. Here in $Te+T^0=0$, T^0 is the probability matrix at the absorption state, e is a column vector with all entries 1 and 0 is a column vector with all entries 0.

A probability distribution $F(\cdot)$ on $(0, \alpha)$ is a *distribution of phase type (PH-distribution)* if and only if it is the distribution of the time until absorption in a finite Markov process of the type defined in equation

$$Q = \begin{vmatrix} T & T^0 \\ 0 & 0 \end{vmatrix}$$

The pair (α, T) is called a representation of $F(\cdot)$.

APPENDIX E

STATE MODEL METHOD

This appendix is provided to present a short description about the state model method, which is mainly collected from Papadopoulos et al. (1993). The idea behind the state model method is very simple and straightforward. There are three steps involved.

First, all the feasible states of the Markov chain describing the model are identified.

In the second step, the transition matrix or the infinitesimal matrix is generated from an analysis of the states of the model.

Third, once the transition matrix is obtained, the stationary equations together with the boundary condition can be used to solve for the stationary distribution. In the numerical procedures, the normalization of the stationary probability vector is handled by replacing one of the equations by the normalizing equation. Finally, each of these probabilities is divided by the sum of all the probabilities so that they all add up to 1.

As the number of states is very large for any realistically sized model, the state model method involves developing computer programs to automate the application of a solution procedure. This procedure can be used when the rate matrix has a particular block lower (or upper) Hessenberg structure. A special case of this general structure, which is commonly found in queueing systems, is the case where the rate matrix has the following block tri-diagonal structure

$$Q = \begin{bmatrix} A & B & & & \\ C & A & B & & \\ & C & A & B & \\ & & C & A & B \\ & & & & \cdots \end{bmatrix} \quad \text{.....(E.1)}$$

where each capital letter represents a block sub-matrix.

VITA AUCTORIS

NAME	Sharmin Shahriar Akhter
PLACE OF BIRTH	Jessore, Bangladesh
YEAR OF BIRTH	1968
EDUCATION	**1976**. Completed elementary level (Grade-V) from Chuadanga Govt. Girls High School, Jessore, Bangladesh and promoted to Grade VI at high school level
	1982. Completed 10-year schooling at Jessore Govt. Girls High School, Jessore, Bangladesh and passed the first compulsory public exam SSC (Secondary School Certificate) in Science
	1984. Completed 2-year college schooling at Jessore Cantonment College, Jessore, Bangladesh and passed the second compulsory public exam HSC (Higher Secondary Certificate) in Science
	1990. Received the Degree of Bachelor of Applied Science (B.A.Sc.) in Civil Engineering from Bangladesh Institute of Technology – BIT, Rajshahi, Bangladesh (renamed as Rajshahi University of Engineering and Technology – RUET)
	2004. Received the Degree of Bachelor of Science (B.Sc.) in Computer Science from the University of Windsor, Windsor, ON, Canada
CURRENT	**2008**. A candidate for the degree of Master of Applied Science (M.A.Sc.) in Industrial and Manufacturing Systems Engineering (IMSE) from the University of Windsor, Windsor, ON, Canada and expecting to graduate in Winter 2008