A Queueing Model To Study Ambulance Offload Delays

by

Mohammad Majedi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Mohammad Majedi

Abstract

The ambulance offload delay problem is a well-known result of overcrowding and congestion in emergency departments. Offload delay refers to the situation where area hospitals are unable to accept patients from regional ambulances in a timely manner due to lack of staff and bed capacity. The problem of offload delays is not a simple issue to resolve and has caused severe problems to the emergency medical services (EMS) providers, emergency department (ED) staff, and most importantly patients that are transferred to hospitals by ambulance. Except for several reports on the problem, not much research has been done on the subject. Almost all research to date has focused on either EMS or ED planning and operation and as far as we are aware there are no models which have considered the coordination of these units. We propose an analytical model which will allow us to analyze and explore the ambulance offload delay problem. We use queuing theory to construct a system representing the interaction of EMS and ED, and model the behavior of the system as a continuous time Markov chain. The matrix geometric method will be used to numerically compute various system performance measures under different conditions.

We analyze the effect of adding more emergency beds in the ED, adding more ambulances, and reducing the ED patient length of stay, on various system performance measures such as the average number of ambulances in offload delay, average time in offload delay, and ambulance and bed utilization. We will show that adding more beds to the ED or reducing ED patient length of stay will have a positive impact on system performance and in particular will decrease the average number of ambulances experiencing offload delay and the average time in offload delay. Also, it will be shown that increasing the number of ambulances will have a negative impact on offload delays and increases the average number of ambulances in offload delay. However, other system performance measures are improved by adding more

ambulances to the system. Finally, we will show the tradeoffs between adding more emergency beds, adding more ambulances, and reducing ED patient length of stay. We conclude that the hospital is the bottleneck in the system and in order to reduce ambulance offload delays, either hospital capacity has to be increased or ED patient length of stay is to be reduced.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisor, Professor Elizabeth Jewkes, for her guidance, encouragement, enthusiasm and patience throughout this research. One simply could not wish for a better or friendlier supervisor.

I would also like to thank my readers, Professor Mantin and Professor Bookbinder, for their valuable comments, insights and feedback. Further, I wish to thank my family and friends who were there when I needed them, offering me endless support and love.

Dedication

To my lovely and beautiful Parnian and my dearest family.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Healthcare is a large component of the Canadian economy that affects each and every citizen of this country. According to the latest OECD (Organization for Economic Co-operative and Development) health data, health spending accounted for 10% of the GDP in 2006 and is projected to reach 160 billion or 10.6% of the GDP in 2007 [2]. Canada has long been known for its good healthcare system. The ability to provide high quality care is one of the main reasons why Canada is among the best countries to live in, according to the United Nation's Quality of Life Survey. However, there have been some concerns in recent years regarding the quality of healthcare being provided to the Canadian public. Long waiting times due to overcrowded hospitals and shortage of medical practitioners are among the major complaints about the Canadian healthcare system. Studies conducted in 2007 have found that 57% of Canadians waited more than 4 weeks to see a specialist and 24% waited for more than 4 hours in the emergency room [12]. Also, Canada is well below the OECD average of 3 doctors per thousand population with only 2.2 doctors per thousand population [2].

In order to ensure continuous high quality healthcare is being provided to the Canadian public, healthcare issues such as those mentioned above must be dealt

with effectively and quickly. One challenge is that our healthcare system is complex and there are many different decision making units involved with any one healthcare issue. Therefore, coordinated decision making is of utmost importance, since each unit acting on its own may not produce the best overall solution.

One example of a system where co-ordination is necessary is in the provision of Emergency Medical Services (EMS). Over two million patients are treated by Canadian EMS every year. Traditionally, EMS have focused on emergency transport and inter-facility transfers for both emergency and non-emergency situations. However, due to demographic and health care trends in recent decades, EMS now integrate aspects of both health care and public safety services. Perhaps the most common and organized type of EMS is an ambulance organization. In general, an EMS provider such as an ambulance organization should be able to:

1. Detect and report an emergency incident
2. Identify the severity of the incident and its degree of urgency (call screening)
3. Respond to the incident as quickly as possible by dispatching appropriate number of ambulances to the scene
4. provide necessary care on emergency scene and while transferring the patient to the emergency department (ED) of the hospital

Unfortunately, long waiting times and congestion in the ED of the hospitals have caused serious problems with transfer of care for the EMS providers in recent years. When a patient is transferred to the hospital by ambulance, in order for the transfer of care to occur there has to be an emergency bed available. However with EDs becoming more and more crowded, many hospitals are experiencing staff and bed shortages. As a result, ambulance paramedics must spend time waiting for an emergency bed to become available since they cannot legally leave the patient

without the hospital accepting transfer of care of the patient. This situation is commonly known as "ambulance offload delay" and has a significant impact on the EMS response times [1] since it affects the availability of ambulances for emergency calls.

1.1 The Problem of Ambulance Offload Delay

Offload delays in Ontario have become more and more common and they are not a simple issue to resolve. In some regions such as Ottawa, the severity of the problem has reached the point where ambulances must leave patients unattended on stretchers at the hospital since offload delays have seriously impeded the delivery of EMS to the community. In other areas, ambulances from other regions have been called in to assist with offload delays; however, they have become more reluctant to respond since they also experience offload delays in their home regions and might not have enough ambulances to respond to their own emergency calls.

The region of Waterloo is among the regions suffering from the problem of offload delays. According to the latest statistics provided by the Region of Waterloo Public Health [34] in December of 2007, the region incurred as many as 22 offload delays in a single day and the number of ambulances lost to offload delays totaled to as many as 13.25 ambulance days per month in 2005. In 2006, Waterloo region incurred more than 6000 hours of offload delays and lost 12.36 ambulance days per month to offload delays. Cities such as Toronto and Ottawa have been experiencing even more severe offload delays. In order to prevent offload delays from increasing response times, EMS managers must employ additional resources to service emergency calls. Longer response times could be life-threatening when the patient requesting EMS service

[1] Response time is the time elapsed from notification of emergency until an ambulance arrives at the scene

is in a critical condition. In the region of Waterloo, the average response time was 13.43 minutes in 2005 [35] which was significantly higher than the provincially established response time of 10 minutes and 30 seconds for the Region of Waterloo.

Offload delays not only increase the EMS response times, they financially cost EMS providers as well. They cost, because EMS has to apply extra resources in order to keep response times low. When an offload delay occurs, EMS staff has to work overtime which is very costly to the EMS provider. In 2006, the city of Toronto spent $3,906,700 in overtime expenditures [46]. Overtime costs are not the only ones, when ambulances are operating for more hours than they are supposed to, EMS incurs ambulance operation costs as well. As mentioned before, the region of Waterloo incurred more than 6000 hours of offload delays in 2006 which, according to the EMS of the region of Waterloo, translates to a financial loss of approximately $840,000 in ambulance operations [34].

Perhaps the most negative impact of offload delays is on patients, paramedics, and emergency department (ED) staff. There have been some occasions where a patient died while waiting with paramedics for an emergency bed to become available [1]. ED staff have to work extremely hard to provide care to patients already in beds and at the same time manage to service the patients who arrive with an ambulance as well as the ones already waiting for service. Offload delays are frustrating to paramedics as well since in most cases they are unable to take a meal break and have to work overtime to provide necessary care to patients while waiting at the hospital.

As mentioned before the ambulance offload delay problem is not a simple issue to resolve. This is due to the fact that the offload delay problem is a component of a much larger problem stemming from ED overcrowding. The ED overcrowding problem is further a product of several internal and external factors not attributable to a single factor.

One of the factors contributing to ED overcrowding is a lack of availability of in-patient beds. When an ED patient's health condition becomes stable, he/she is usually transferred to an in-patient bed (if not sent home) where further care will take place. In this case, an ED bed can be used to treat another patient whose condition is not stable. However, lack of in-patient beds is itself another major problem in many Canadian hospitals. According to the study done by Estey et al. [14], it is commonly believed that lack of in-patient beds is one of the main causes of ED overcrowding and long ED waiting times. When there are no in-patient beds available patients have to stay in ED beds and as a result, the ED becomes congested. There are many other factors contributing to the ED overcrowding problem [15], some of them are:

- Use of an ED bed for non-emergency cases
- Staff shortages
- Aging population and increasing patient acuity

This discussion of ambulance offload delays clearly shows the severity and complexity of the problem. There have been various reports on offload delays indicating the need for long term plans to prevent the problem. There are also various analytical models, not specifically on offload delays, but on EMS operations that are focused on improving the efficiency of the EMS system. Most of these models are focused on ambulance location and relocation, status management systems, and staff optimization that are aimed to improve EMS performance. There are also several analytical and simulation models on ED planning and operation that are focused on improving waiting times and reducing ED overcrowding. However, what is important in situation of ambulance offload delay is the coordination between the ED of a hospital and the EMS provider. Each unit has its own performance

measures and each unit acting on its own may not produce the best overall solution to the problem. Therefore, coordinated decision making between these units is necessary. How does one tradeoff resource allocation to each unit in order to provide the best quality care for patients? We currently do not have models that allow us to evaluate tradeoffs in this specific context. Such a model is the main contribution of this thesis.

1.2 Contributions of this Work

The main focus of this thesis is on the interaction between the EMS provider and the cumulative effect of congestion in the ED of a hospital on ambulance offload delays. As mentioned before, most of the work done by researchers focuses on either ED or EMS operations and there are no analytical models specifically dealing with their interactions. We have developed an analytical model that allows us to analyze and explore the interaction between the two units by evaluating various system performance measures such as:

- The distribution of the number of patients waiting for ambulance
- The distribution of the number of ED beds occupied
- The distribution of the number of ambulances in offload delay
- Expected waiting time for an ambulance and the mean time in offload delay
- Expected ambulance and ED bed utilization

In our work we show that adding more emergency beds or reducing the ED patient treatment time have a similar effect on system performance and they both improve the overall system performance. Also, we analyze the effect of varying the number

of ambulances on the system performance and show that, although adding more ambulances would improve some of the system performance measures, it increases the average number of ambulances in offload delay. We also consider the tradeoffs between the number of ED beds, the number of ambulances, and the ED patient length of stay and show that how one can substitute a resource with another to achieve a similar system performance improvement. For example, instead of adding an extra bed in the ED to improve system performance, is it possible to add more ambulances or reduce ED treatment time (if possible) to achieve a similar gain in system performance measures.

1.3 Outline of the Report

The remainder of this thesis is organized as follows: In Chapter 2, we review models on EMS and ED planning and operations mainly focused on the areas of ambulance location and relocation, bed management and ED occupancy, staff optimization and scheduling. In Chapter 3, we introduce the ambulance offload delay model. We use queuing theory and continuous time Markov Chain concepts to model the situation. The matrix geometric method is then used in Chapter 4 to obtain the steady state probability distribution of the system. The results of Chapter 4 are used in Chapter 5 to compute the distribution of various system performance measures which allow us to compute more aggregate system performance measures. Chapter 6 contains sensitivity analysis and finally, conclusions and future research directions are discussed in chapter 7.

Chapter 2

Literature Review

There is a long history of research in both EMS and ED planning and operation. In this chapter, we provide an overview of various models developed in these areas. First, we will focus on ambulance location and relocation models used in the area of EMS planning and operation and next we will review various analytical and simulation models on staff scheduling and bed management in the area of ED operation. Finally we discuss the fact that, although there are similar analytical models in other areas of research such as manufacturing flow-lines that are useful when modeling the situation of offload delays, they are not directly applicable since design issues and performance measures are different.

2.1 EMS Planning and Operation

Two of the most important EMS operations that emergency managers are concerned with are call screening, determining the type and number of ambulances to dispatch to the incident, and ensuring response times are adequate. In emergency situations response time is vital and ambulances must be located in a way to ensure adequate coverage for fast response times. Most of the research done in EMS planning and

operation is focused on the problem of ambulance location and relocation. Most of the models in this area are based on the mathematical programming techniques and they are divided into three main categories: static and deterministic, probabilistic, and dynamic. Brotcorne, Laporte, and Semet [5] have done an extensive review of various ambulance location and relocation models in each of the above mentioned categories.

Static ambulance location models are among the early models that are meant to be used at the planing stage. One of the early static ambulance location models, known as Location Set Covering Model (LSCM), was introduced in 1971 [48] and was aimed to minimize the number of ambulances needed to service all demand points in the service area. The LSCM model did not take into account the fact that once an ambulance is dispatched, some demand points are no longer covered. Also, the model ignores the cost of the system since the optimal solution of the model usually requires many ambulances in order to provide complete coverage. To counter some of the shortcomings of LSCM models, Church and Revelle proposed Maximal Covering Location Problem (MCLP) model (1974) which aims to maximize population coverage subject to limited ambulance availability [9]. Although the MCLP model is more practical than the LSCM model, it has two major shortcomings: first, it assumes that response times are known and second that an ambulance close to the demand point is always available. Both models were useful in their own ways with LSCM determining the appropriate number of ambulances to cover all demand points and MCLP making the best possible use of limited ambulance resources.

When an emergency incident is identified, often two types of units with different capabilities are dispatched to the scene: Basic Life Support (BLS) and Advance Life Support (ALS) units. BLS is typically provided by firemen who are trained as paramedics and they are often the first to arrive on scene. ALS is covered by ambu-

lances. Both LSCM and MCLP models ignore the fact that, on occasion, different types of vehicles may be dispatched to the incident. A number of deterministic models were proposed to deal with this issue. Schilling et al. [42] was among the first to develop a model, known as Tandem Equipment Allocation Model (TEAM), to serve for this purpose. The model is capable of handling two types of vehicles and is a direct extension of MCLP, therefore the objective is still to maximize the demand covered. One problem with the TEAM is inadequate coverage when ambulances become busy. Researchers Daskin and Stern [11], and Hogan and Revelle [39], proposed extensions to the TEAM by introducing a second objective that would provide better ambulance coverage compared to the original model.

Deterministic models ignore the stochastic nature of the ambulance location problem and the fact that ambulances operate as servers in a queuing system and are sometimes unavailable. Hence, probabilistic models were then developed to overcome these shortcomings of deterministic models. Daskin [10] proposed a probabilistic model known as the Maximum Expected Covering Location Problem (MEXCLP) in which the same probability of q (called busy fraction) is assigned to each ambulance, where q is the fraction of time that an ambulance is unavailable. The model further assumes that all ambulances operate independently of each other. The objective of the MEXCLP is to maximize the expected demand covered and can only handle one type of vehicle. The MEXCLP model was later applied to the city of Bangkok by Fujiwara et al. [18] in which they considered 59 demand points and 46 ambulance location sites. Also, Repede and Bernardo [38] developed and applied an extension of MEXCLP model known as TIMEXCLP, to the city of Louisville, Kentucky. Authors considered variations in ambulance travel speed throughout the day in the TIMEXCLP formulation and used simulation to validate the proposed solution.

Revelle and Hogan [40] proposed two other probabilistic models with an ob-

jective of maximizing the demand covered with a given probability α. In the first model, they assumed the same busy fraction of q for all potential location sites as opposed to the same busy fraction of q for all ambulances that Daskin assumed. In their second model, they relaxed this assumption and assumed that different location sites have different busy fraction probabilities.

Erdogan, Erkut, and Ingolfsson [13] analyzed the effect of incorporating a survival function, which maps the response time to survival probability, into existing probabilistic location covering models. The authors generated 4 models (from existing models) with an objective of maximizing the expected number of patients who survive. The models were applied to the city of Edmonton, Canada where 180 demand points and 16 possible locations for EMS stations were considered. By comparing the results of the developed models with the existing location covering models such as the MEXCLP and MCLP, the authors were able to show that introducing a survival function can result in significantly better EMS unit locations with respect to the probability of survival.

Both deterministic and probabilistic ambulance location models did not consider the dynamic nature of the location problem and the need to repeatedly relocate ambulances in the same day. Kolesar and Walker [24] were among the first to recognize this and proposed a dynamic model for fire departments that was also applicable to ambulance systems. However, the model was not suitable enough for ambulance systems due to the fact that ambulances need to be relocated in short period of times, meaning that the model needed to be solved repeatedly. With an advancement of computer technology and development of faster heuristics, it is now possible to quickly solve the ambulance location problem in real time. Gendreau et al. [20] developed a model that uses the available information at any time t to recompute a new ambulance redeployment strategy. This model solves the ambulance relocation problem at each instant time t when a call is registered.

Gendreau et al's model also considers a number of practical considerations, such as:

- Long trips are avoided if they are between the first and final location sites,
- Repeated round trips are avoided if they are between the same two location sites,
- Ambulances moved in successive relocations cannot be always the same.

Another similar model known as the System Status Management (SSM) model was proposed by consultant Jack Stout [45]. The model uses historical data to predict when and where emergency calls will come in, and how to locate sufficient resources close enough to those anticipated calls to provide reasonable response times. Therefore, the SSM model aims to optimize response times while maximizing the use of personnel and vehicles. Many EMS systems are now employing models based on the concept of SSM. In general, dynamic models are becoming more popular these days and their advancement depends on sophisticated system technologies, and the availability of fast and accurate search heuristics.

One of the earliest models to incorporate queuing theory is the hypercube model developed by Larson [25]. The model was aimed at analyzing the problem of vehicle location and response district design in urban emergency services. Given a region with N response units such as ambulance stations that are spatially distributed throughout the region, and a certain spatial distribution of demands for service, the model is aimed to analyze the problem of:

1. How to partition the region into several districts in order to achieve certain levels of service
2. How to locate or position the N response units in different districts

Later, Larson proposed an approximate procedure to his queuing model for a faster and easier computation of system performance characteristics [26]. However, the approximate model assumes that there is only one server (ambulance) at each station and an ambulance is unavailable to respond to new emergency calls while providing service to a specific call. Further, the average service time is assumed to be independent of the location of the call and the location of ambulance's home station. Budge, Ingolfsson, and Erkut [6] proposed an extension to Larson's approximate model that allows for having multiple ambulances at each station and average service times that are dependent on call location and vehicle location. The model computes station-specific dispatch probabilities (the probability that a particular station will respond to an emergency call at a particular location) which allow for easy calculation of many system performance measures.

Taylor and Templeton [47] proposed and applied a queuing model to determine the optimal number of ambulances required in an urban fleet which serves two types of customers (low priority vs. high priority). They considered a priority queuing system with N servers and a cutoff service discipline where low priority customer arrivals are cutoff and placed in a queue whenever there are more than $0 < N_1 < N$ servers busy. Two models were proposed, one for the situation where high priority customers are lost if all N servers are busy; and one for the case where high priority customers join a queue for service if all servers are occupied.

In general, most of the models in the area of EMS planning and operation have focused on ambulance location and relocation as well as determining the optimal number of ambulances to optimize system performance measures such as response times and demand coverage. However, none of the models have looked at the link between EMS and ED. Next, we will provide a brief review of some of the analytical and simulation models used in the area of ED planning and operation.

2.2 Emergency Department Planning and Operation

As mentioned in the introduction, the problem of overcrowded emergency departments due to increased demand for ED services has reached a critical state [14]. Overcrowding has led to a number of problems including prolonged waiting times, ambulance diversions, offload delays, patient dissatisfaction, and many more [14]. Staff shortages and inefficient staff schedules are identified among causes of emergency overcrowding. Therefore, one of the areas in healthcare research that has been of interest to many scholars and it is becoming more crucial to assist with the problem of emergency overcrowding is nurse scheduling and rostering. Models in this area are mainly focused on determining the number of nursing personnel and their shifts to meet service demands while (in most cases) minimizing costs and other constraints such as nurse preferences, skill classes, and etc. Siferd and Benton [43] have done an extensive review on nurse scheduling models in which they have provided classifications of nurse rostering systems and review of methods for solving different classes of problems. Another similar literature review was done by Bradley and Martin on continuous nurse scheduling algorithms [3]. In general, most of the scholars in this area have used mathematical programming techniques to tackle the problem of nurse scheduling. As an example, Warner and Prawda [51] proposed a large scale mixed integer quadratic programming for allocating a fixed number of nursing staff in a number of skill classes to a number of units and shifts.

Miller and Pierskalla [31] recognized the fact that it is necessary to consider nurse preferences and proposed a mathematical programming model with an objective of minimizing the total penalties associated with a failure to provide minimum required coverage and nursing staff preferences for schedules. One of the benefits of this model is the fact that it is flexible enough to include a large number

of constraints, and it can also be used to solve part-time and full-time employee schedules. Another similar mathematical programming model which considered nursing staff preferences, rotation patterns, and requests for days off was proposed by Warner [50]. The author also incorporated multiple classes of nursing staff, a cyclical weekend coverage policy, and a four or six week planning horizon.

Beside models that used mathematical programming techniques, perhaps the most widely used models in the area of ED operations are simulation models. ED operations can easily be represented as a queuing system where a simulation approach can be used to model all the processes in the system. One of the main advantages of the simulation approach is its ability to model large systems with many processes. When developing a simulation model, the most important and time consuming step is data collection and validation. Although in some cases data has to be collected manually; with the advancement of technology and information systems, this is done automatically by computer systems nowadays. Therefore, easy availability of system data and the fact that many simulation software packages are now available are the main reasons why the simulation approach is becoming more and more popular. Most emergency departments are now using simulation to model their emergency system to shorten waiting times, and to optimize staff requirements.

McGuire [30] proposed a simulation model to reduce the length of stay of patients in the emergency department of a medium to large sized hospital in the southeast of Charlotte. With the help of the simulation model, Mcguire suggested five alternatives to the hospital's executive management to help reduce the average patient waiting time of 157 minutes. Similarly, Rossetti et al. [41] used simulation to determine the optimal staff schedules based on the emergency department at the University of Virginia Medical Center in Charlottesville, Virginia. They considered eighteen different alternatives for ED staff schedules and analyzed the impact of

each alternative on patient throughput and resource utilization.

Several researchers proposed analytical and simulation models in the area of bed management and ED occupancy which are found to be effective ways to control ED overcrowding. Forster et al. [17] analyzed the effect of hospital occupancy, measured by dividing the number of patients to the number of beds, on ED patient length of stay. Their analysis was based on an observational study design using administrative data and all patients presented in the ED between the years of 1993 to 1999. The authors showed that there is a positive correlation between hospital occupancy and the length of stay of patients in ED, and increasing hospital bed availability might reduce ED overcrowding.

Among researchers whom used queuing theory in the area of bed management, Gorunescu et al. [21] proposed a queuing model to optimize the allocation and use of hospital beds to improve patient care. The authors assumed Poisson patient arrivals, phase-type hospital service times, and c beds ($M/PH/c$). The queuing model was used to determine system performance characteristics such as mean bed occupancy and the probability of a patient being lost when there are no beds available. The authors also demonstrated a method to optimize the number of beds, c, to minimize the cost for a given arrival rate and average length of stay. Further, the model allows hospital managers to minimize the average cost per day by balancing costs of empty beds against costs of delayed patients.

In general, models in the area of ED planning and operation are mostly focused on optimizing the number of staff or ED beds while minimizing costs and improving ED performance measures such as ED patient length of stay. Again, there are no models which have specifically considered the interaction between EMS and ED.

As we discussed, the problem of offload delays is a two sided problem that affects both EMS and ED operations. Each unit has its own performance measures,

however each unit acting on its own may not produce the best overall solution. Unfortunately there are no models analyzing the interaction of these units in a situation such as ambulance offload delays. Therefore, our work is focused on developing a model to represent the situation of offload delays which allows for analysis of the tradeoffs between the two units. Albeit the fact that there are no analytical models specifically dealing with ambulance offload delays, there are similar analytical models in the area of manufacturing flow-lines.

2.3 Manufacturing Flow-lines

Manufacturing flow-line is one of the most studied class of manufacturing systems characterized by the pattern of material flow. A flow-line system is composed of m production stages in which materials or parts flow from stage 1 to stage m in order of $1, 2, ..., m$. Each stage consists of a number of functionally identical resources such as machines that process a variety of materials. There is a buffer between two consecutive stages where materials queue-up if machines at the next stage are not available. Therefore in order to produce a product, which consists of several tasks that need to be performed in the flow-line, materials has to flow from stage 1 to stage m where at each stage a certain task is being performed. In order for the performance of the system to be optimized, the following design issues need to be considered:

- The number of production stages
- The number of machines at each stage $1, 2, ..., m$
- The workload allocated to each stage
- The production capacity of each stage

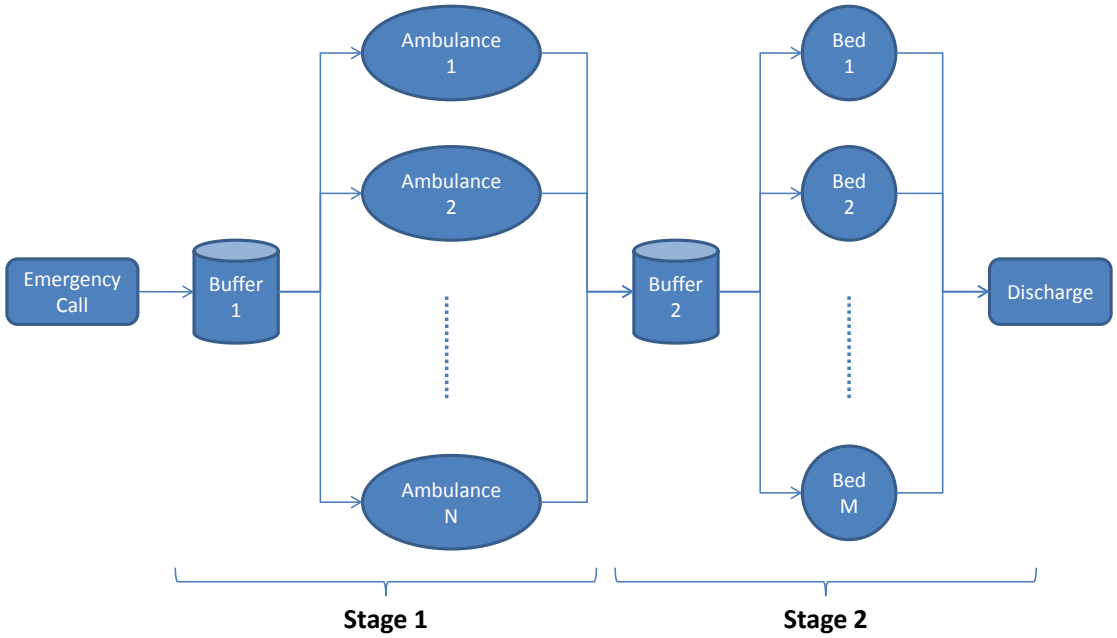

Figure 2.1: 2 Stage Flow-line Representation of Offload Delays

Buzacott and Shanthikumar [7] have done an excellent job in addressing the above mentioned design issues and reviewed various approaches using queuing models to resolve these issues. Another similar work is done by Govil and Fu [22] in which authors review the contributions and applications of queuing theory in the field of manufacturing systems including flow-lines.

The situation of ambulance offload delays has some similarities to a manufacturing flow-line system. It is possible to represent it as a two stage manufacturing flow-line shown in Figure 2.1, where patients have a similar role as parts. In this system, stage 1 consists of N ambulances and stage 2 consists of M emergency beds. Similarly there are two buffers, one in stage 1 with unlimited capacity for patients waiting for ambulance and the other one in stage 2 with a capacity of N for ambulances experiencing offload delay.

Although there are similarities between both system, there is one important difference and that is the fact that performance measures used are completely different. In the offload delay situation we are interested in evaluating performance measures such as the average time in offload delay, number of ambulances in offload delay, number of hospital beds occupied, and the number of patients waiting for ambulance. Whereas, in a flow-line system the two common performance measures used are the number of jobs in the system and the total time a job spends in the system. Also, most of the design issues are different between the two systems. As a result, the manufacturing flow-line models are not applicable to the situation of ambulance offload delays; however, knowledge and understanding of these models is helpful when designing a model for the ambulance offload delay problem.

In this chapter we reviewed several analytical and simulation models in the area of EMS and ED planning and operation and showed that although there are no models analyzing the interaction of these units, similar models exist in other areas of research such as the manufacturing flow-lines. In the next chapter, we use queuing theory to construct a system representing the problem of ambulance offload delay and model the behavior of the system as a continuous time Markov chain.

Chapter 3

The Ambulance Offload Delay Model

In this chapter, we propose a queuing system to represent the interaction between the EMS provider and the ED of a hospital. In section 3.1, we describe the queuing system and its components used to represent the situation of ambulance offload delay. In section 3.2, a continuous time Markov Chain (MC) representation of our queuing system is discussed in detail. The model state variables, transition rates between states, and the infinitesimal generator matrix and its components are all discussed in section 3.2. At last, a detailed example to illustrate the ideas presented and discussed in the chapter is presented in section 3.3.

3.1 Model Description

We consider a system with a single hospital that has a total of $M \geq 1$ emergency beds and an EMS center with a total of $N \geq 1$ ambulances to provide service to emergency calls. Both N and M are assumed to be integers. Calls arrive at the EMS center according to a Poisson process with rate λ and are served in a first come

first serve basis. When an emergency call arrives and an ambulance is available, it is dispatched to the incident scene. We define the "ambulance transit time" to be the time it takes for an ambulance to reach the emergency scene, performs on scene care, travels back to the hospital, offloads the patient and returns to its base location. The ambulance transit time is assumed to follow an Exponential distribution with rate τ. If there are no ambulances available to serve the emergency call, the patient will join a queue to wait until an ambulance becomes available in order to be served. An ambulance is assumed to be unavailable when it is either in transit (transferring a patient to the hospital) or experiencing an offload delay at the hospital.

It is important to note the difference between offload delay and offload time used in the ambulance transit time. As we defined in chapter 1, offload delay refers to the situation where an ambulance is unable to transfer care of a patient to the ED due to unavailability of an emergency bed. On the other hand, offload time is the time that it normally takes for an ambulance to offload a patient and to transfer him/her to the hospital ED when a bed is available. According to [35], the EMS in the region of Waterloo has set an offload time of twenty minutes, which means that if an ambulance takes forty five minutes to offload a patient; the first twenty minutes is considered as offload time and the extra twenty five minutes is considered as an offload delay. Another differentiation is based the fact that offload time is a part of the ambulance transit time which is an input parameter, whereas offload delay is an output of the system.

Once an ambulance transfers a patient to the hospital's ED, the time it takes for the hospital to treat the patient (the hospital service time) is assumed to be Exponentially distributed with rate μ_1. In the situation where there are no beds available at the ED and the ambulance is experiencing offload delay, we assume that it is possible to provide treatment in the ambulance and to directly transfer the patient to an in-patient bed without going through ED. If an emergency bed

becomes available while patient is being treated in an ambulance, we assume that the patient is transferred to the bed immediately and the treatment time is reset to the ED treatment time ($exp(\mu_1)$). The treatment time in an ambulance is assumed to follow an exponential distribution with rate μ_2. Basically, we assume that ambulances can be used as extra beds in the ED but with a treatment time of much longer than a regular ED bed due to the fact that there are not as many resources available in an ambulance as compared to the ED of a hospital. It is important to note that the ambulance treatment time parameter adds flexibility to our modeling. By appropriately setting the value of μ_2 we can model different policies with respect to offload delays. For example, suppose that the EMS policy with respect to offload delays is that ambulances must wait at the ED until a bed becomes available This policy can easily be adopted by setting the value of μ_2 to 0 which simply means that the rate at which ambulances treat patients is 0 (patients/time).

Patients arriving via ambulance are not the only arrivals to the ED of the hospital. In fact, according to [35] small percentage of ED's volume arrive by ambulance. In our model, we use the term "outside patients" to refer to those majority of patients that arrive at the hospital's ED without using an ambulance resource. There is no priority between the outside patients and the patients who arrive by ambulance in our model. That is, all patients that arrive to the ED are served in a FCFS basis. Further, we assume that outside patients are lost if the hospital's ED has reached its maximum capacity - i.e. no emergency beds are available. Simply put, if an outside patient arrives at the hospital when there are no emergency beds available, he/she will not wait to receive care and simply goes to another hospital. We assume that outside patients arrive according to a Poisson process with rate δ. These patients have the same service rate μ_1 as the patients arriving by an ambulance.

The Above mentioned queuing system is shown in Figure 3.1. The dashed line from the "Ambulance in offload delay" queue to the "inpatient/discharge" end point represents the situation where $\mu_2 > 0$. That is, the patient can either be transfer to a ED bed or can be treated in an ambulance and transferred to an inpatient bed or discharged. If a bed becomes available, then the patient is transferred to the ED bed and the treatment time is set to ED's treatment time with rate μ_1. However, if an ED bed does not become available and ambulance finishes its treatment with rate μ_2, the patient is either transferred to an in-patient bed or discharged from the hospital. For the case where $\mu_2 = 0$, the dash line is simply ignored since ambulances must wait at the ED until a bed becomes available.

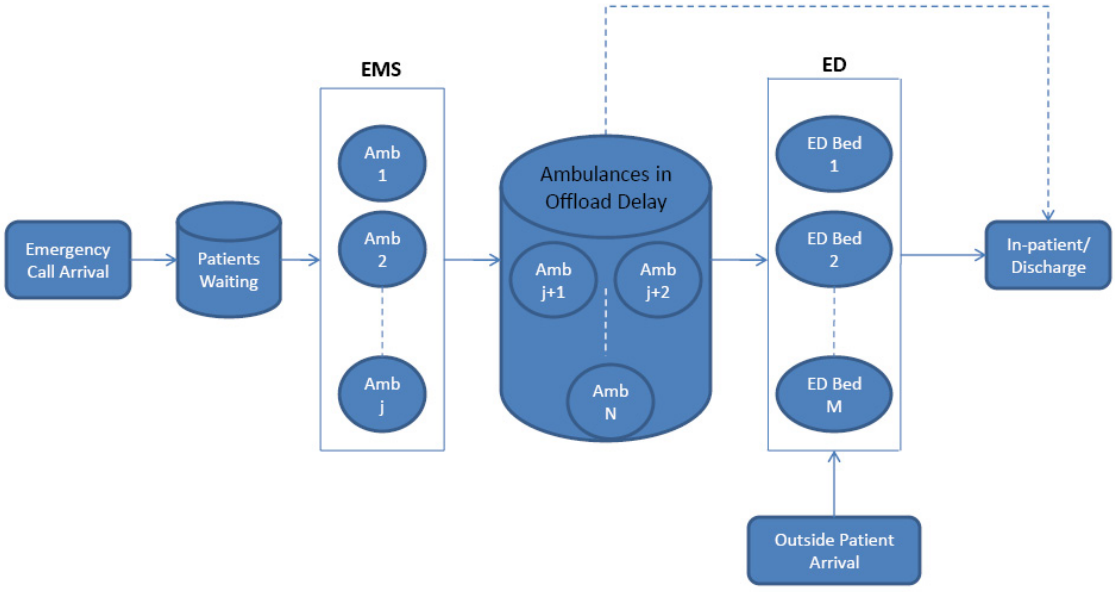

Figure 3.1: The Ambulance Offload Delay System

The following table summarizes the model parameters discussed above:

Parameter	Parameter Description
M	Total number of hospital beds
N	Total number of ambulances
λ	Patient arrival rate
τ	Ambulance transit time rate
δ	outside patient arrival rate
μ_1	Hospital service rate
μ_2	Ambulance service rate

Table 3.1: Model Parameters

The model assumptions can be summarized in the following points:

- Emergency calls arrive according to a Poisson process with rate λ and are served in a FCFS basis
- If an ambulance is not available when an emergency call arrives, the patient joins a queue
- The ambulance transit time, hospital treatment time, and ambulance treatment times are assumed to follow an Exponential distribution with rates τ, μ_1, and μ_2 respectively.
- When $\mu_2 = 0$ and there are no beds available at the ED of the hospital, ambulances must wait at the hospital until a bed becomes available (offload delay).
- When $\mu_2 > 0$ and an ED bed becomes available while a patient is receiving care in an ambulance, he/ she is transferred to the ED bed immediately and the treatment time is reset to ED's treatment time with rate μ_1.

- Outside patients are assumed to arrive according to a Poisson process with rate δ and have a service rate of μ_1.
- Patients arriving from outside and those arriving by ambulance have equal priority.
- Outside patients are lost in the case where the hospital's ED has reached its maximum capacity.

In the next section, we will show how this system can be modeled in a continuous time Markov Chain framework.

3.2 The Markov Chain

The system described in the previous section can be modeled by a continuous time Markov Chain (MC) with a discrete state space. In order to describe the system precisely at time $t : t > 0$ we define the following set of state variables:

1. $\{N_A(t) = 0, 1, ...\}$: Number of patients waiting for an ambulance at time t
2. $\{N_T(t) = 0, 1, ..., N\}$: Number of ambulances in transit (on the way to the hospital)
3. $\{N_D(t) = 0, 1, ..., N\}$: Number of ambulances experiencing offload delay at time t
4. $\{N_B(t) = 0, 1, ..., M\}$ Number of emergency beds occupied at time t

N is the total number of ambulances and M is the total number of emergency beds. Although the state of the system can be represented by the above 4 variables, we were able to simplify this representation by combining and reducing the above variables into the following 2 variables:

1. $\{I(t) = 0, 1, ..., N, ...\}$: The number of patients waiting for an ambulance at time t, $N_A(t)$, plus the number of ambulances in transit at time t, $N_T(t)$.

$$I(t) = N_A(t) + N_T(t) \tag{3.1}$$

2. $\{J(t) = 0, 1, ..., M, ..., M+N\}$: The number of ambulances in offload delay at time t, $N_D(t)$ plus the number of emergency beds occupied at time t, $N_B(t)$.

$$J(t) = N_D(t) + N_B(t) \tag{3.2}$$

This way of system representation has some benefits. It allows us to precisely describe the state of the system at any given point in time with only two variables, and more importantly it will dramatically reduce the amount of computations and calculations required. However, this representation comes with a price. It will complicate system performance measure calculations as we will see in Chapter 5.

We use the notation $(I(t), J(t))$ to represent the state of the system at time t. Table 3.2 shows how to properly interpret the state of the system, $(I(t), J(t)) = (i, j)$ at time t for $i = 0, 1, 2, ...$ and $j = 0, 1, ..., M + N$.

To show how Table 3.2 can be used, suppose the hospital under consideration has four emergency beds, and there are 3 ambulances in the system ($N = 3$ and $M = 4$). Say we want to interpret state $(4, 5)$ where $i = 4$ and $j = 5$. Since $i = 4 > N = 3$ and $j = 5 > M = 4$, we are looking at the case number four in Table 3.2 which indicates that the system is in a state where four emergency beds are occupied; one ambulance is experiencing offload delay at the hospital; two ambulances in transit; and two patients are waiting for ambulance. Another example would be state $(2, 6)$ which falls under the second case since $i = 2 < N = 3$ and $j = 6 > M = 4$. This state indicates that four beds are occupied; two ambulances are in offload delay;

Case	Condition	Interpretation
1	$i \leq N$, $j \leq M$	i patients in transit, j beds occupied, no ambulances stock at the hospital, and no patients waiting for an ambulance
2	$i \leq N$, $j > M$	M beds occupied, $j - M$ ambulances stock at the hospital, $N - (j - M)$ ambulances are in transit, and $i - (N - (j - M))$ patients waiting for an ambulance
3	$i > N$, $j \leq M$	N ambulances in transit, $i - N$ patients waiting for an ambulance, and j beds occupied
4	$i > N$, $j > M$	M beds occupied, $j - M$ ambulances stock at the hospital, $N - (j - M)$ ambulances in transit, and $i - (N - (j - M))$ patients waiting for an ambulance

Table 3.2: How to interpret state descriptor $(I(t), J(t)) = (i, j)$

one patient is waiting for an ambulance; and one ambulance is in transit. Other states can be interpreted in a similar way.

There are 5 events in our system that can cause the state of the system to be changed at any given point in time, they are: an emergency call arrival, patient transfer completion to the hospital, hospital service completion, ambulance service completion when $\mu_2 > 0$, and an outside patient arrival. Suppose the system is in state $(I(t), J(t)) = (i, j)$ at time $t > 0$, Table 3.3 summarizes all the possible transitions that can occur from this state to all other possible states when one of the above mentioned events occur.

Case	Event	Condition	Next State	Transition Rate
1	Call Arrival	None	$(i+1, j)$	λ
2	Ambulance Transfer Completion	$0 < i \leq N$, $0 \leq j \leq M$	$(i-1, j+1)$	$i\tau$
3	Ambulance Transfer Completion	$i > N$, $0 \leq j \leq M$	$(i-1, j+1)$	$N\tau$
4	Ambulance Transfer Completion	$i < (N-(j-M))$, $M < j \leq M+N$	$(i-1, j+1)$	$i-(N-(j-M))\tau$
5	Ambulance Transfer Completion	$i \geq (N-(j-M))$, $M < j \leq M+N$	$(i-1, j+1)$	$(N-(j-M))\tau$
6	Hospital Service Completion	$j \leq M$	$(i, j-1)$	$j\mu_1$
7	Hospital Service Completion	$M < j \leq M+N$	$(i, j-1)$	$M\mu_1$
8	Ambulance Service Completion	$M < j \leq M+N$	$(i, j-1)$	$M\mu_1 + (j-M)\mu_2$
9	Outside Patient Arrival	$0 \leq j < M$	$(i, j+1)$	δ

Table 3.3: Model Transition Rates from State (i,j)

We will leave our discussion of Table 3.3 until section 3.3 where we extensively analyze and explain it in a context of a numerical example. Now that we have all the required transition rates, it is possible to construct the infinitesimal generator matrix Q which shows all the transition rates between the states of the MC for our queuing system. The Q matrix and its components are shown below:

$$
\begin{array}{c}
\\
0 \\
1 \\
2 \\
3 \\
\vdots \\
N-2 \\
N-1 \\
N \\
N+1 \\
N+2 \\
N+3 \\
\vdots
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc}
0 & 1 & 2 & 3 & \dots & N-2 & N-1 & N & N+1 & N+2 & N+3 & \dots
\end{array} \\
\begin{pmatrix}
B_0 & A_0 & & & & & & & & & & \\
C_1 & B_1 & A_0 & & & & & & & & & \\
 & C_2 & B_2 & A_0 & & & & & & & & \\
 & & C_3 & B_3 & \ddots & & & & & & & \\
 & & & \ddots & \ddots & \ddots & & & & & & \\
 & & & & \ddots & B_{N-2} & A_0 & & & & & \\
 & & & & & C_{N-1} & B_{N-1} & A_0 & & & & \\
 & & & & & & A_2 & A_1 & A_0 & & & \\
 & & & & & & & A_2 & A_1 & A_0 & & \\
 & & & & & & & & A_2 & A_1 & A_0 & \\
 & & & & & & & & & A_2 & A_1 & A_0 \\
 & & & & & & & & & & \ddots & \ddots
\end{pmatrix}
\end{array}
$$

Figure 3.2: The Infinitesimal Generator Matrix Q

where A_0, A_1, A_2, $\{C_n : n = 1, 2, ..., N-1\}$, and $\{B_n : n = 0, 1, ..., N-1\}$ are of size $(M + N + 1) * (M + N + 1)$, and

$$
\begin{aligned}
B_0 &= -A_0 \\
B_n &= -(C_n + A_0), \quad n = 1, 2, ..., N-1 \\
A_1 &= -(A_0 + A_2)
\end{aligned}
$$

In order to properly interpret the infinitesimal generator matrix Q and its component matrices, we need to define the *state level* of the system. We define the *level* of the MC as the subset of all states that have the same $I(t)$, and the *phase* of the MC as the subset of all states that have the same $J(t)$.

First, we note that the generator matrix has a repetitive structure after the N^{th} column (N^{th} column included). The j^{th} column for $j \geq N+1$ is the same as the N^{th} column except that it is shifted down by $j-N$ steps. We call this portion of the generator matrix, the *repeating portion* since it has a repetitive structure. In the repeating portion, A_0 represents the transition rate matrix at which the system moves up one level, A_1 is the transition rate matrix at which the system returns to the same level, and A_2 is the transition rate matrix at which the system moves down one level. Note that the transition rates within the component matrices correspond to the movement along phases of the MC. The A_0, A_1, and A_2 matrices have the following form:

$$
\begin{array}{l}
\textit{phase} \\
0 \\
1 \\
2 \\
\vdots \\
M-1 \\
M \\
M+1 \\
\vdots \\
M+N-1 \\
M+N
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
0 & 1 & 2 & \dots & M-1 & M & M+1 & \dots & M+N-1 & M+N
\end{array} \\
\begin{pmatrix}
\lambda & & & & & & & & & \\
& \lambda & & & & & & & & \\
& & \lambda & & & & & & & \\
& & & \ddots & & & & & & \\
& & & & \lambda & & & & & \\
& & & & & \lambda & & & & \\
& & & & & & \lambda & & & \\
& & & & & & & \ddots & & \\
& & & & & & & & \lambda & \\
& & & & & & & & & \lambda
\end{pmatrix}
\end{array}
$$

Figure 3.3: Transition Rate Matrix A_0

$$
\begin{array}{l}
\textit{phase} \\
0 \\
1 \\
2 \\
\vdots \\
M-1 \\
M \\
M+1 \\
M+2 \\
\vdots \\
M+N-2 \\
M+N-1 \\
M+N
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc}
0 & 1 & \dots & M-1 & M & M+1 & \dots & M+N-1 & M+N
\end{array} \\
\begin{pmatrix}
\beta_{N,0,0,1} & \delta & & & & & & & \\
\mu_1 & \beta_{N,1,0,1} & & & & & & & \\
& 2\mu_1 & & & & & & & \\
& & \ddots & & & & & & \\
& & & \beta_{N,M-1,0,1} & \delta & & & & \\
& & & M\mu_1 & \beta_{N,M,0,0} & 0 & & & \\
& & & & M\mu_1+\mu_2 & \beta_{N-1,M,1,0} & & & \\
& & & & & M\mu_1+2\mu_2 & & & \\
& & & & & & \ddots & & \\
& & & & & & & 0 & \\
& & & & & & & \beta_{1,M,N-1,0} & 0 \\
& & & & & & & M\mu_1+N\mu_2 & \beta_{0,M,N,0}
\end{pmatrix}
\end{array}
$$

Figure 3.4: Transition Rate Matrix A_1

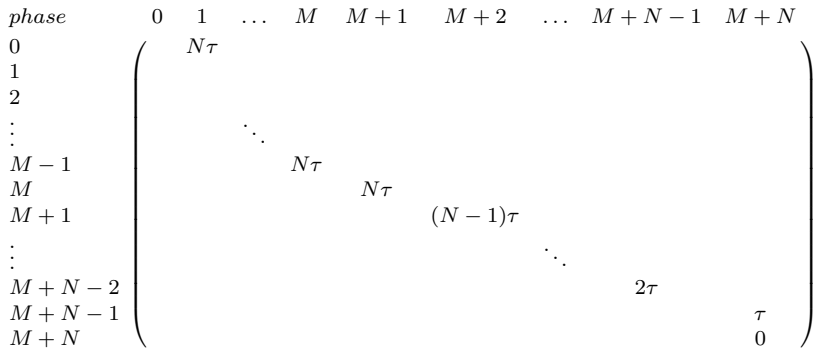

Figure 3.5: Transition Rate Matrix A_2

$$\beta_{n,i,j,k} = -(\lambda + n\tau + i\mu_1 + j\mu_2 + k\delta) \tag{3.3}$$

where $\beta_{n,i,j,k}$ is the function that makes the sum of the elements along each row of the Q matrix zero.

Next, we focus on the matrices prior the N^{th} column. We call this portion of the generator matrix the "boundary portion". The A_0 matrix in this portion is the same as the A_0 matrix in the repeating portion and corresponds to transitions from level n to $n+1$ for $n < N$. Matrices B_n and C_n have a similar interpretation to the A_1 and A_2 matrices. $\{B_n : n = 0, 1, ..., N-1\}$ is the transition rate matrix where the the system returns to the same level n whereas $\{C_n : n = 1, 2, ..., N-1\}$ is the transition rate matrix at which the system moves down from level n to $n-1$ given that the process started at level $n < N$. The B_n and C_n matrices have the following form:

$$
\begin{array}{c|c}
\textit{phase} & \begin{array}{ccccccccccccc} 0 & 1 & 2 & 3 & \dots & M & M+1 & \dots & M+N-(n-1) & \dots & M+N-1 & M+N \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ M-1 \\ M \\ \vdots \\ M+N-n \\ \vdots \\ M+N-2 \\ M+N-1 \\ M+N \end{array} &
\begin{pmatrix}
n\tau & & & & & & & & & & & \\
 & n\tau & & & & & & & & & & \\
 & & n\tau & & & & & & & & & \\
 & & & \ddots & & & & & & & & \\
 & & & & n\tau & & & & & & & \\
 & & & & & n\tau & & & & & & \\
 & & & & & & \ddots & & & & & \\
 & & & & & & & (n-1)\tau & & & & \\
 & & & & & & & & \ddots & & & \\
 & & & & & & & & & 2\tau & & \\
 & & & & & & & & & & \tau & \\
 & & & & & & & & & & 0 &
\end{pmatrix}
\end{array}
$$

Figure 3.6: Transition Rate Matrix C_n

$$
\begin{array}{c|c}
\textit{phase} & \begin{array}{ccccccccccc} 0 & 1 & \dots & M & M+1 & \dots & M+N-n & \dots & M+N-1 & M+N \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ M-1 \\ M \\ \vdots \\ M+N-n \\ \vdots \\ M+N-2 \\ M+N-1 \\ M+N \end{array} &
\begin{pmatrix}
\beta_{n,0,0,1} & \delta & & & & & & & & \\
\mu_1 & \beta_{n,1,0,1} & & & & & & & & \\
 & 2\mu_1 & & & & & & & & \\
 & & \ddots & & & & & & & \\
 & & & \delta & & & & & & \\
 & & & \beta_{n,M,0,0} & 0 & & & & & \\
 & & & & & \ddots & & & & \\
 & & & & & & \beta_{n,M,N-n,0} & & & \\
 & & & & & & & \ddots & & \\
 & & & & & & & & 0 & \\
 & & & & & & & & \beta_{n,M,N-1,1} & 0 \\
 & & & & & & & & M\mu_1 + N\mu_2 & \beta_{n,M,N,1}
\end{pmatrix}
\end{array}
$$

Figure 3.7: Transition Rate Matrix B_n

Although matrices B_n and C_n have a similar interpretation in the boundary portion as the A_1 and A_2 in the repeating portion, there is major difference between these matrices. Matrices B_n and C_n have different entries depending on the level of the MC. For example, the transition rate matrix with which the process moves from the second level to the first level, C_2 is not the same as the transition rate matrix corresponding to the movement from the third level to the second level, C_3. We refer to these matrices as *level dependent* matrices. On the other hand, the entries of matrices A_0, A_1 and A_2 are always fixed and they do not depend on the level of MC. These matrices are referred to as *level independent* matrices. Simply put, the process we are considering is a mixture of what is so called *level dependent quasi-birth-death process* and *level independent quasi-birth-death process*.

The process is level independent when the MC is in a level $I(t) \geq N$ and it is level dependent for $I(t) < N$.

Let $\pi(i,j)$ denote steady state probability or the long run (mean) fraction of time the system spends in state (i,j). For example, given that there are four beds and two ambulances in the system, $\pi(3,5)$ denotes the mean fraction of time the system spends in state $(3,5)$ in which four beds are occupied, one ambulance is in offload delay, one ambulance is in transit, and two patients are waiting for an ambulance. Now, let

$$\begin{aligned} \underline{\pi_i} &= \{\pi(i,0), \pi(i,1), ..., \pi(i,M), ..., \pi(i,M+N)\} \\ \underline{\pi} &= \{\pi_0, \pi_1, ..., \pi_N, ...\} \end{aligned}$$

We are interested in calculating the steady state probability vector $\underline{\pi}$, since having these long run probabilities will allow us to compute various system performance measures as we will see in Chapter 5.

The essential problem is in determining the steady state probability vector $\underline{\pi}$. This requires the solution to a set of linear flow balance equations, where there is an equation associated with each level of the MC. For a continuous time Markov chain process with an infinitesimal generator matrix Q, the balance equations are given by the following system of equations:

$$\begin{aligned} \underline{\pi} Q &= 0 \\ \underline{\pi} e &= 1 \\ \underline{\pi} &\geq 0 \end{aligned} \tag{3.4}$$

where e is a column vector of ones and the equation $\underline{\pi} e = 1$ is known as the *normalization equation*. In general, for a continuous time MC process we have that

the rate out of state i must be equal to the rate into state i for $i > 0$, and that is why we have $\underline{\pi}Q = 0$. The following table expands the $\underline{\pi}Q = 0$ and shows the balance equation for each specific level in our case:

Level	**Balance Equation**
0	$\pi_0 B_0 + \pi_1 C_1 = 0$
1	$\pi_0 A_0 + \pi_1 B_1 + \pi_2 C_2 = 0$
$\vdots$	$\vdots$
i	$\pi_{i-1}A_0 + \pi_i B_i + \pi_{i+2}C_{i+1} = 0$
$\vdots$	$\vdots$
$N-1$	$\pi_{N-2}A_0 + \pi_{N-1}B_{N-1} + \pi_N A_2 = 0$
N	$\pi_{N-1}A_0 + \pi_N A_1 + \pi_{N+1}A_2 = 0$
$N+1$	$\pi_N A_0 + \pi_{N+1}A_1 + \pi_{N+2}A_2 = 0$
$\vdots$	$\vdots$
j	$\pi_{j-1}A_0 + \pi_j A_1 + \pi_{j+1}A_2 = 0$
$\vdots$	$\vdots$

Table 3.4: System Balance Equations

Therefore, the general form of the balance equations in the repeating portion is given by:

$$\pi_{i-1}A_0 + \pi_i A_1 + \pi_{i+1}A_2 = 0, \quad i = N, N+1, ... \tag{3.5}$$

For the boundary portion, we have:

$$\pi_0 B_0 + \pi_1 C_1 = 0 \tag{3.6}$$

$$\pi_{i-1}A_0 + \pi_i B_i + \pi_{i+2}C_{i+1} = 0, \quad i = 1, 2, ..., N-2 \tag{3.7}$$

$$\pi_{N-2}A_0 + \pi_{N-1}B_{N-1} + \pi_N A_2 = 0 \tag{3.8}$$

Before we proceed to the next chapter and discuss our approach to solve for the steady state probability distribution $\underline{\pi}$, we present a numerical example to illustrate the ideas presented in this chapter and to discuss Table 3.3.

3.3 Numerical Example

In order to keep the example simple but at the same time be illustrative we consider a system with $N = 3$ and $M = 4$ - i.e. four emergency beds and three ambulances. For now we do not assign any numerical values to the rest of the model parameters *i.e* arrival and service rates. By the definition of the state variables we know that $I(t) \geq 0$ and $0 \leq J(t) \leq 7$. The transition rates for this example are given in Table 3.5. Now we will take some time to explain Table 3.5.

When an emergency call arrives, regardless of the system state, the MC level increases by one with rate λ. Suppose the system is currently in state $(5, 7)$ where four hospital beds are occupied, three ambulances are in offload delay, and five patients are waiting for an ambulance. In an event of call arrival, the system will make a transition to state $(6, 7)$, where now six patients are waiting for an ambulance instead of 5.

The situation is more complicated in the case of ambulance transfer completion event (transferring a patient to the hospial). In order to explain cases 2, 3, 4, and 5 we use the following fact: *if* $\{T_i : i = 1, 2, ..., N\}$ *are* N *i.i.d exponential random variables with rate* τ*, then* $T_{min} = min(T_1, T_2, ..., T_N)$ *is exponentially distributed with rate* $N\tau$. We know that the ambulance transit time is exponentially distributed with rate τ. Therefore, if there are $n > 0$ ambulances in transit, the rate at which a patient is transferred to the hospital is $n\tau$ since we are looking at the the minimum of n Exponential distributions with rate τ.

Case	**Event**	**Condition**	**Next State**	**Transition Rate**
1	Call Arrival	None	$(i+1, j)$	λ
2	Ambulance Transfer Completion	$0 < i \leq 3$, $0 \leq j \leq 4$	$(i-1, j+1)$	$i\tau$
3	Ambulance Transfer Completion	$i > 3$, $0 \leq j \leq 4$	$(i-1, j+1)$	3τ
4	Ambulance Transfer Completion	$i < (3-(j-4))$, $4 < j \leq 7$	$(i-1, j+1)$	$i-(3-(j-4))\tau$
5	Ambulance Transfer Completion	$i \geq (3-(j-4))$, $4 < j \leq 7$	$(i-1, j+1)$	$(3-(j-4))\tau$
6	Hospital Service Completion	$j \leq 4$	$(i, j-1)$	$j\mu_1$
7	Hospital Service Completion	$4 < j \leq 7$	$(i, j-1)$	$4\mu_1$
8	Ambulance Service Completion	$4 < j \leq 7$	$(i, j-1)$	$4\mu_1 + (j-4)\mu_2$
9	Outside Patient Arrival	$0 \leq j < 4$	$(i, j+1)$	δ

Table 3.5: Example 3.3 Transition Rates from State (i,j)

According to case 2, if there are no ambulances in offload delay, $0 \leq j \leq 4$, and no patients waiting for an ambulance, $0 < i \leq 3$, the rate at which patients are transferred to the hospital depends on the number of ambulances in transit. The rate is given by the "number of ambulances in transit $\times\tau$". An ambulance transfer completion event would cause the MC level to decrease by one and the phase of the system to increase by one To illustrate this, suppose the system is in state $(3, 3)$, that is 3 beds are occupied and 3 ambulances are in transit. When a patient is successfully transferred to the hospital, the system will make a transition into state $(2, 4)$ with rate 3τ since at the same time a patient is transferred to the hospital and an emergency bed is occupied. As a result, the level of MC is decreased by one and the phase of the MC is increased by one.

Case 3 is similar to case 2 except that there are $\{i - 3 : i > 3\}$ patients waiting for an ambulance, 3 ambulances in transit, and no ambulances in offload delay. In this situation, the rate at which an ambulance transfer completion occurs is always 3τ since there are 3 ambulances in transit. As an example, consider state $(10, 3)$. The only difference between this state and state $(3, 3)$ discussed for case 2 is the fact that there are 7 patients waiting for ambulance but the number of ambulances in transit is still 3. Therefore the rate at which an ambulance transfer completion occurs is still 3τ and in this case the system makes a transition into state $(9, 4)$.

Cases 4 and 5 are similar to cases 2 and 3 except that there are $(j-4)$ ambulances experiencing offload delay at the hospital. This results in $(3 - (j - 4))$ ambulances available that are either in transit or idle depending on the value of i. Again, the transition rates depend on the number of ambulances in transit only.

The hospital service completion event is similar to ambulance transfer completion event in the sense that the rate at which a patient is discharged from the hospital depends on the number of beds occupied. Again, we know that the hospital service rate is exponentially distributed with rate μ_1, therefore if there are

$n > 0$ beds occupied, the rate at which a bed becomes available is $n\mu_1$. Unlike an ambulance transfer completion event, a hospital service completion only affects the MC phase by increasing its value by one. An example of a hospital service completion for case 6 would be transition from state $(4, 3)$, with 3 beds occupied, to the state $(4, 2)$ where a patient is discharged from the hospital. In this case, the rate of transition is $3\mu_1$ since there were 3 beds occupied. Now, if there are ambulances experiencing offload delay at the hospital (case 7), the rate at which a patient is discharged from the hospital is $M\mu_1$ since all beds are occupied; that is why an ambulance is in offload delay.

In case 8, all the emergency beds are occupied which means that there are ambulances experiencing offload delay. Suppose the system is in the state $(2, 6)$ where all beds are occupied, 2 ambulances are in offload delay, 1 ambulance is in transit, and 1 patient is waiting for an ambulance. Given that $\mu_2 > 0$, a patient can be discharged either through the hospital or an ambulance depending on whichever finishes its service first. If the patient is discharged through the hospital the rate is $4\mu_1$ since there are 4 beds occupied; and if the patient is discharged through an ambulance the rate is $2\mu_2$ since there are 2 ambulances in offload delay. Therefore the overall rate would be $4\mu_1 + 2\mu_2$ since we are looking at the minimum of two exponential random variables with corresponding rates $4\mu_1$ and $2\mu_2$. Hence, the rate at which the system makes a transition from state $(2, 5)$ to state $(2, 4)$ is given by $4\mu_1 + 2\mu_2$. Now if we assume that $\mu_2 = 0$, that is ambulances do not have the ability to provide treatment to patients within ambulance, case 8 is the exact same as case 7.

Finally, outside patient arrivals (case 9) can only occur when there is an emergency bed available i.e $0 \leq j < 4$, and this would cause the MC phase to increase by one. The rate at which this transition occurs is simply δ. As an example, suppose an outside patient arrives when the system is in the state $(1, 3)$ where 3 beds are

occupied but one bed is still available. This arrival would cause the system to make a transition into state $(1, 4)$ where all emergency beds are now occupied.

Using the transition rates from Table 3.5, we can construct the infinitesimal generator matrix Q and its components for this example:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & \dots \end{array} \\ \begin{pmatrix} B_0 & A_0 & & & & & \\ C_1 & B_1 & A_0 & & & & \\ & C_2 & B_2 & A_0 & & & \\ & & A_2 & A_1 & A_0 & & \\ & & & A_2 & A_1 & A_0 & \\ & & & & A_2 & A_1 & \ddots \\ & & & & & \ddots & \ddots \end{pmatrix} \end{array}$$

$$B_0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} \beta_{0,0,0,1} & \delta & & & & & & \\ \mu_1 & \beta_{0,1,0,1} & \delta & & & & & \\ & 2\mu_1 & \beta_{0,2,0,1} & \delta & & & & \\ & & 3\mu_1 & \beta_{0,3,0,1} & \delta & & & \\ & & & 4\mu_1 & \beta_{0,4,0,0} & 0 & & \\ & & & & 4\mu_1 + \mu_2 & \beta_{0,4,1,0} & 0 & \\ & & & & & 4\mu_1 + 2\mu_2 & \beta_{0,4,2,0} & 0 \\ & & & & & & 4\mu_1 + 3\mu_2 & \beta_{0,4,3,0} \end{pmatrix} \end{array}$$

$$B_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} \beta_{1,0,0,1} & \delta & & & & & & \\ \mu_1 & \beta_{1,1,0,1} & \delta & & & & & \\ & 2\mu_1 & \beta_{1,2,0,1} & \delta & & & & \\ & & 3\mu_1 & \beta_{1,3,0,1} & \delta & & & \\ & & & 4\mu_1 & \beta_{1,4,0,0} & 0 & & \\ & & & & 4\mu_1+\mu_2 & \beta_{1,4,1,0} & 0 & \\ & & & & & 4\mu_1+2\mu_2 & \beta_{1,4,2,0} & 0 \\ & & & & & & 4\mu_1+3\mu_2 & \beta_{1,4,3,0} \end{pmatrix} \end{array}$$

$$B_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} \beta_{2,0,0,1} & \delta & & & & & & \\ \mu_1 & \beta_{2,1,0,1} & \delta & & & & & \\ & 2\mu_1 & \beta_{2,2,0,1} & \delta & & & & \\ & & 3\mu_1 & \beta_{2,3,0,1} & \delta & & & \\ & & & 4\mu_1 & \beta_{2,4,0,0} & 0 & & \\ & & & & 4\mu_1+\mu_2 & \beta_{2,4,1,0} & 0 & \\ & & & & & 4\mu_1+2\mu_2 & \beta_{2,4,2,0} & 0 \\ & & & & & & 4\mu_1+3\mu_2 & \beta_{1,4,3,0} \end{pmatrix} \end{array}$$

$$C_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & \tau & & & & & & \\ & & \tau & & & & & \\ & & & \tau & & & & \\ & & & & \tau & & & \\ & & & & & \tau & & \\ & & & & & & \tau & \\ & & & & & & & \tau \\ & & & & & & & \end{pmatrix} \end{array}$$

$$C_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & 2\tau & & & & & & \\ & & 2\tau & & & & & \\ & & & 2\tau & & & & \\ & & & & 2\tau & & & \\ & & & & & 2\tau & & \\ & & & & & & 2\tau & \\ & & & & & & & \tau \\ & & & & & & & \end{pmatrix} \end{array}$$

$$A_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} \beta_{3,0,0,1} & \delta & & & & & & \\ \mu_1 & \beta_{3,1,0,1} & \delta & & & & & \\ & 2\mu_1 & \beta_{3,2,0,1} & \delta & & & & \\ & & 3\mu_1 & \beta_{3,3,0,1} & \delta & & & \\ & & & 4\mu_1 & \beta_{3,4,0,0} & 0 & & \\ & & & & 4\mu_1 + \mu_2 & \beta_{3,4,1,0} & 0 & \\ & & & & & 4\mu_1 + 2\mu_2 & \beta_{2,4,2,0} & 0 \\ & & & & & & 4\mu_1 + 3\mu_2 & \beta_{1,4,3,0} \end{pmatrix} \end{array}$$

$$A_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & 3\tau & & & & & & \\ & & 3\tau & & & & & \\ & & & 3\tau & & & & \\ & & & & 3\tau & & & \\ & & & & & 3\tau & & \\ & & & & & & 2\tau & \\ & & & & & & & \tau \\ & & & & & & & \end{pmatrix} \end{array}$$

$$A_0 = \lambda \cdot I_8$$

where I_8 is an identity matrix of size 8 and $\beta_{n,i,j,k}$ is given by equation 3.3. The balance equations for this example have the general form given by equations 3.5-3.8.

This concludes the presentation of our MC model. In Chapter 4, we continue our discussion of steady state probabilities and present a method to numerically compute the steady state probability distribution. We also extend example 3.3 discussed above by assigning numerical values to the rest of the model parameters and solve for its steady state probability distribution.

Chapter 4

Calculating Steady State Probabilities

In this chapter we present a computational procedure to allow us to numerically compute the steady state probability distribution $\underline{\pi}$ for the Markov Chain model developed in chapter 3. Recall the system of balance equations (3.4) presented in Section 3.2:

$$\begin{aligned} \underline{\pi}Q &= 0 \\ \underline{\pi}e &= 1 \\ \underline{\pi} &\geq 0 \end{aligned} \tag{4.1}$$

Various methods can be used to solve the above system of equations with each one having its own strengths and weaknesses. A simple (but not efficient) approach to solve 4.1 is the substitution method. This is possible since the number of equations in 4.1 is always one more than the number of variables due to the normalization equation $\underline{\pi}e = 1$. However, this method gets extremely tedious as the number of

variables increases. Fortunately, several researchers have been able to recognize possible structural features of these equations, and were able to develope efficient and more sophisticated algorithms compared to the simple substitution method. Snyder and Stewart [44] reviewed two commonly used computational approaches for solving the above set of equations. The first approach, called the explicit approach, was developed by Marie and Pellaumail in scalar form [29] and Carroll et al. in matrix form [8]. In these approaches, an expression for the steady state probabilities is explicitly written as the function of the server parameters. Snyder and Stewart show how this approach applies to a queuing system with phase type distribution service times where (n, j) represents the state of the system with n number of customers in a service facility and j being the current stage of service. The explicit approach constructs an extra set of balance equations that would reduce the normal balance equations from second order difference equations to first order difference equations. This allows for defining a coefficient matrix H_n that relates the steady state probabilities in the form $\pi_{i-1} = H_n \cdot \pi_i$, where H_n elements are given explicitly as a function of server parameters.

The second approach discussed by Snyder and Stewart is based on Neuts' work [33]. Neuts was able to show that if one can group the states of the Markov Chain (MC) into vectors which possess a certain repetitive structure, there exists a recursive relationship between the steady state probabilities that can be expressed in the following form:

$$\pi_{i+1} = R \cdot \pi_i$$

where R is a constant matrix of an appropriate dimension. This approach is called the matrix geometric method and it was proposed by Evans [16] and Wallace [49] and extensively developed by Neuts later. There are a wide range of

processes that have a regular structure which lead to Markov models with a repetitive structure that could fit within the matrix geometric framework. Examples of such processes are various types of queueing systems, computer performance models, and telecommunication models. For many problems it is possible to use either the explicit or the matrix geometric approach, but some of the problems are more amenable to solution via the matrix geometric method. In our work, we will use Neuts' matrix geometric approach to numerically compute the steady state probability distribution.

4.1 The Matrix Geometric Approach

Recall the infinitesimal generator matrix, Q, from Chapter 3:

$$
\begin{array}{c}
 \\
0 \\
1 \\
2 \\
3 \\
\vdots \\
N-2 \\
N-1 \\
N \\
N+1 \\
N+2 \\
N+3 \\
\vdots
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc}
0 & 1 & 2 & 3 & \dots & N-2 & N-1 & N & N+1 & N+2 & N+3 & \dots
\end{array} \\
\begin{pmatrix}
B_0 & A_0 & & & & & & & & & & \\
C_1 & B_1 & A_0 & & & & & & & & & \\
 & C_2 & B_2 & A_0 & & & & & & & & \\
 & & C_3 & B_3 & \ddots & & & & & & & \\
 & & & \ddots & \ddots & \ddots & & & & & & \\
 & & & & \ddots & B_{N-2} & A_0 & & & & & \\
 & & & & & C_{N-1} & B_{N-1} & A_0 & & & & \\
 & & & & & & A_2 & A_1 & A_0 & & & \\
 & & & & & & & A_2 & A_1 & A_0 & & \\
 & & & & & & & & A_2 & A_1 & A_0 & \\
 & & & & & & & & & A_2 & A_1 & A_0 \\
 & & & & & & & & & & \ddots & \ddots
\end{pmatrix}
\end{array}
$$

We know that the process is in the repeating or level independent portion whenever the MC is in the level $I(t) \geq N$ and it is in the boundary, or level dependent

portion whenever $I(t) < N$. For now, let us focus on the level independent portion. From chapter 3, the set of balance equations for the repeating portion have the following form:

$$\pi_{i-1}A_0 + \pi_i A_1 + \pi_{i+1}A_2 = 0, \quad i = N, N+1, ... \tag{4.2}$$

First we note that the transition rates between the adjacent levels in the repeating portion do not depend upon the level. For example, the rate at which the process moves up from level i to $i+1$ is A_0 for all $i \geq N$. Similarly, the rates at which the process moves down from level i to $i-1$ or returns to i are A_2 and A_1 respectively and are the same for all $i \geq N$. Therefore it is not surprising to see that the value of $\{\pi_i,\, i \geq N\}$ is a function of the transition rates between the adjacent levels. Hence, this suggests that there is some constant matrix R such that

$$\pi_{i+1} = R \cdot \pi_i \quad i = N-1, N, ... \tag{4.3}$$

which is what Neuts discovered. The matrix R is usually called the "Rate Matrix" (not to be confused with the transition rate matrices) and in our case has a dimension of $(M+N+1) \times (M+N+1)$. According to Ramaswami and Taylor [37], the rate matrix, R, has the following physical interpretation: given that the process starts in state (k,i) for $k \geq N$ and $0 \leq i \leq M+N$, the $(i,j)^{th}$ entry of the matrix R is the expected sojourn time in state $(k+1,j)$ for $0 \leq j \leq M+N$, before returning to level k. Note that the expected sojourn time in state i is the expected number of visits to state i multiplied by the expected time spent in state i per visit.

Rewriting equation 4.3 results in the following *matrix geometric* form:

$$\pi_k = R^{k-N+1} \cdot \pi_{N-1} \quad k \geq N \tag{4.4}$$

Now we substitute this equation into equation 4.2:

$$R^{i-N}\,\pi_{N-1}A_0 + R^{i-N+1}\,\pi_{N-1}A_1 + R^{i-N+2}\pi_{N-1}A_2 = 0. \quad i = N, N+1, ... \quad (4.5)$$

Multiplying both sides by R^{N-i} and simplifying yields:

$$A_0 + RA_1 + R^2A_2 = 0 \quad (4.6)$$

This is a quadratic equation that is normally solved numerically. Computing the rate matrix R in 4.6 will allow us to compute the repeating portion steady state probability distribution $(\pi_N, \pi_{N+1}, ...)$.

Computing steady state probabilities $(\pi_0, \pi_1, ..., \pi_{N-1})$ for the level dependent portion is not as simple as the level independent portion due to the fact that transition rate matrices depend on the level of MC. Because of this, relationship 4.3 is no longer applicable to the level dependent portion of the generator matrix. Although the rate at which the process moves up a level (A_0) is still the same, the rates at which the process moves down a level or returns to the same level are not the same and they depend on the level $I(t)$. Neuts also considered such processes and showed that in the level dependent process the following relationship holds among the steady steady probabilities:

$$\pi_i = R_i \cdot \pi_{i-1} \quad i = 0, 1, ..., N-1 \quad (4.7)$$

The only difference between relationship 4.3 and 4.7 is in the fact that in 4.7 the rate matrix R_i is also level dependent. The rate matrix R_i has a similar interpretation as the rate matrix R in the level independent portion. The $(j,k)^{th}$ entry of the matrix R_i is the expected sojourn time in state $(i+1, k)$, before returning to level i given that the process started in state (i, j) for $i < N$.

Similar to 4.4 rewriting 4.7 yields:

$$\pi_i = \pi_0 \prod_{j=0}^{i-1} R_j \quad i = 1, 2, ..., N-1 \tag{4.8}$$

Now recall the boundary balance equations from chapter 3:

$$\pi_0 B_0 + \pi_1 C_1 = 0 \tag{4.9}$$

$$\pi_i A_0 + \pi_{i+1} B_i + \pi_{i+2} C_{i+1} = 0, \quad i = 1, 2, ..., N-2 \tag{4.10}$$

$$\pi_{N-2} A_0 + \pi_{N-1} B_{N-1} + \pi_N A_2 = 0 \tag{4.11}$$

Substituting 4.8 in 4.10-4.11 and simplifying yields:

$$A_0 + R_i B_i + R_i R_{i+1} C_{i+1} = 0, \quad i = 0, 1, 2, ..., N-3 \tag{4.12}$$

$$A_0 + R_{N-2} B_{N-1} + R_{N-2} R_{N-1} A_2 = 0 \tag{4.13}$$

Note that equations 4.12-4.13 and 4.6 are very similar. Simply, the rate matrix R_i is replaced with R and matrices A_1 and A_2 are replaced with B_i and C_i respectively. Therefore, by first solving the family of matrices $\{R_i, i = 0, 1, ..., N-2\}$, we can compute the steady state probabilities associated with the boundary portion, $(\pi_0, \pi_1, ..., \pi_{N-1})$ through equation 4.7.

As you can see, the rate matrix R is the heart of the matrix geometric method. Several scholars such as Neuts [33], Lucantoni and Ramaswami [36], and Latouche and Ramaswami [28] have proposed iterative algorithms to solve for the rate matrix R in equation 4.6. Researchers such as Gaver, Jacobs and Latouche [19], Bright and Taylor [4], and Ye and Li [52] are among the ones who have developed algorithms for calculating the rate matrices R_i in the level dependent case. However, as Bright and Taylor pointed out in [4] it is not necessary to solve the system of equations

4.12 and 4.13 in our case. Let us rewrite equations 4.12 and 4.13 in the following form:

$$R_i = -A_0(B_i + R_{i+1}C_{i+1})^{-1}, \quad i = 0, 1, 2, ..., N-3 \tag{4.14}$$

$$R_{N-2} = -A_0(B_{N-1} + R_{N-1}C_N)^{-1} \tag{4.15}$$

Note that now we can recursively calculate $\{R_i, i = 0, 1, ..., N-2\}$ if we know the value of R_{N-1}. However, R_{N-1} is basically the rate matrix R for the repeating portion. Therefore, if we calculate the rate matrix R in 4.6, we can recursively calculate the rest of rate matrices $\{R_i, i = 0, 1, ..., N-2\}$ by using equations 4.14-4.15. Now that we have all the required rate matrices, we can use $\{R_i, i = 0, 1, ..., N-2\}$ in equation 4.7 to compute $(\pi_0, \pi_1, ..., \pi_{N-1})$ and then use R in equation 4.3 to calculate $(\pi_N, \pi_{N+1}, ...)$.

First we need to compute the value of π_0 in order to use equations 4.7 and 4.3 to calculate the rest of the steady state probabilities. This is where the normalization equation $\underline{\pi}e = 1$ becomes handy.

The balance equation for state 0 is given by equation 4.9. Rewriting this equation using equation 4.8 gives:

$$\pi_0(B_0 + R_0C_1) = 0 \tag{4.16}$$

Also, we can use equations 4.3 and 4.7 to simplify the normalization equation and get the following result (see Appendix B for the complete proof):

$$\pi_0[\sum_{i=0}^{N-1}\prod_{j=0}^{i-1} R_j + \prod_{j=0}^{N-1} R_j(I-R)^{-1}]e = 1 \tag{4.17}$$

Therefore, to solve for π_0 we need to solve the following system:

$$\pi_0(B_0 + R_0C_1) = 0 \quad s.t. \tag{4.18}$$

$$\pi_0[\sum_{i=0}^{N-1}\prod_{j=0}^{i-1} R_j + \prod_{j=0}^{N-1} R_j(I-R)^{-1}]e = 1$$

We have all the necessary tools to calculate the steady state probabilities but first we have to solve for the rate matrix R in 4.6 in order to compute $\{R_i, i = 0, 1, ..., N-2\}$ and hence the steady state probability distribution. As we mentioned before there are several iterative algorithms for solving equation 4.6. Below, we will discuss and present two of these algorithms.

4.1.1 Simple Iterative Algorithm

The algorithm that we are about to present is very simple and performs well for small size problems. This algorithm has also been presented by Neuts [33] and has been used by many researchers. We will start from equation 4.6 and rewrite it in the following form:

$$R = -(A_0 + R^2A_2){A_1}^{-1} \tag{4.19}$$

Hence the rate matrix R can be computed through the following iterative procedure:

$$R_{n+1} = -(A_0 + {R_n}^2A_2){A_1}^{-1}, \quad R_0 = 0 \tag{4.20}$$

until $|R_{n+1} - R_n| < \epsilon$, where ϵ represents the accuracy of the calculation. The following figure shows this algorithm in more detail:

Note that smaller values of ϵ represent more accurate calculations. As Nelson [32] also points out, the number of iterations required for convergence in this method (and most of other methods) increases as the spectral radius of the rate matrix R

$R_0 = 0$

$i = 0$

Repeat

$R_{i+1} = -(A_0 + {R_i}^2 A_2){A_1}^{-1}$

$diff = \max |R_{i+1} - R_i|_{(i,j)}$

$R_{temp} = R_{i+1}$

$i = i + 1$

$Until \quad diff < \epsilon$

$R = R_{temp}$

Figure 4.1: Simple Iterative Algorithm

increases. If we let λ_i be the eigenvalues of the rate matrix R, then its spectral radius $\rho(R)$ is defined by:

$$\rho(R) = \max(|\lambda_i|) \tag{4.21}$$

The spectral radius of R can be interpreted as the measure of the utilization of the system. As the system becomes more utilized, equation 4.20 needs more iterations in order to converge. Therefore, a highly utilized system or a spectral radius close to 1 results in computationally intensive calculations depending on the size of the problem. In general this approach is very easy to implement and it produces quite similar results in comparison with other approaches.

4.1.2 The Logarithmic Reduction Algorithm

The second approach is based on the logarithmic reduction algorithm developed by Latouche and Ramaswami [28] for the level independent QBDs. This approach has also been extended by Bright and Taylor [4] to handle the case of level dependent QBDs. We will not get into the detail of how this algorithm is derived since it is not

the focus of our work. Refer to [28] for detailed implementation of the algorithm. Figure 4.2 shows the algorithm in detail, where I is an identity matrix of size $M + N + 1$ and e is a column vector of ones:

$i = 0$

$B_0 = (-A_1)^{-1}A_0$

$B_2 = (-A_1)^{-1}A_2$

$S_1 = B_2$

$S_2 = B_0$

Repeat

$i = i + 1$

$(A_1)' = B_0B_2 + B_2B_0$

$(A_0)' = (B_0)^2$

$(A_2)' = (B_2)^2$

$B_0 = (I - (A_1)')^{-1}(A_0)'$

$B_2 = (I - (A_1)')^{-1}(A_2)'$

$S_1 = S_1 + S_2B_2$

$S_2 = S_2B_0$

$Until \quad |e - S_1e| < \epsilon$

$G = S_1$

$U = A_1 + A_0G$

$R = A_0(-U^{-1})$

Figure 4.2: Logarithmic Reduction Algorithm

This algorithm is based on calculating the matrices G and U which are analogous to the rate matrix R, and are used to define most of the characteristics of the Markov Chain. The $(i, j)^{th}$ entry of matrix G represents the probability that a process starting in phase i of level $k+1$ will first enter level k in phase j for $k \geq N$. The $(i, j)^{th}$ entry of the matrix U is the probability that starting in phase i of level $k + 1$ the process eventually returns to phase j of the same level $k + 1$ without

visiting any state in level k for $k \geq N$. The algorithm presented in Figure 4.2 uses an iterative procedure to calculate the G matrix and then uses the following relationships derived by Hajek [23] and Latouche [27] to compute the U matrix and hence the R matrix:

$$\begin{aligned} U &= A_1 + A_0 G \\ R &= A_0(-U)^{-1} \end{aligned}$$

According to Latouche and Ramaswami [28], the logarithmic reduction algorithm is numerically very stable, converges much faster on large problems compared to other similar algorithms, and performs very well on simpler problems.

The simple iterative algorithm or the logarithmic reduction algorithm allow us to compute the R matrix for the repeating portion. Setting $R_{N-1} = R$ in equations 4.14 and 4.15, we can compute the family of matrices $\{R_i : i = 0, 1, ..., N-2\}$ for the boundary portion of the process. Finally, equations 4.7 and 4.3 allow us to compute the steady state probabilities for the boundary and repeating portion respectively after π_0 is calculated through 4.18. Before we discuss how the above algorithms are implemented, we have to check whether the system is stable in the long run or not since equations 4.3 and 4.7 will not produce correct results if the system is not stable in the long run. Therefore, we will discuss and present the stability conditions next.

4.2 Stability Conditions

The queuing system we are considering is said to be stable in the long run if the expected drift towards lower states exceeds that towards the higher states [33]. That is, in the long run, the system tends to move to lower levels rather than

higher level states, or the rate at which the system moves down a level exceeds the rate at which the system moves up a level. According to Neuts [33] the process is stable when the MC is positive recurrent.

One way to check for the stability condition is to compute the spectral radius of the R matrix. If it is less than 1 we know that the system is stable [33]. However, this requires calculating the R matrix. Another method, perhaps a more efficient one, is to compute the stationary distribution $\underline{f} = (f_0, f_1, ..., f_M + N)$ for the following set of linear equations:

$$\begin{aligned} \underline{f}A &= 0 \\ \underline{f}e &= 1 \end{aligned} \tag{4.22}$$

where e is a column vector of ones and $A = A_0 + A_1 + A_2$. Then, the stability condition is given by

$$\underline{f}A_0e < \underline{f}A_2e \tag{4.23}$$

If this condition holds, we know that the system is stable. This method is discussed in more detail by Neuts in [33]. We know that A_2 is the matrix corresponding to moving down a level and A_0 corresponds to moving up a level. Therefore, equation 4.23 simply states that in the long run we want the rate at which the process moves down a level to exceed the rate at which it moves up a level. For our model, the above stability condition can be explicitly given by:

$$\begin{bmatrix} f_0 & \cdots & f_M & f_M+1 & \cdots & f_{M+N} \end{bmatrix} \cdot \begin{bmatrix} N\tau \\ \vdots \\ N\tau \\ (N-1)\tau \\ \vdots \\ \tau \end{bmatrix} > \begin{bmatrix} f_0 & \cdots & f_M & f_M+1 & \cdots & f_{M+N} \end{bmatrix} \cdot \begin{bmatrix} \lambda \\ \vdots \\ \lambda \\ \lambda \\ \vdots \\ \lambda \end{bmatrix}$$

which reduces to:

$$\left[N\tau \cdot \sum_{i=0}^{M} f_i + \sum_{i=M+1}^{M+N} (N-i+M+1)\tau f_i\right] > \left[\lambda \cdot \sum_{i=0}^{M+N} f_i\right]$$

Now if we let:

$$\alpha_1 = \sum_{i=0}^{M+N} f_i, \qquad \alpha_2 = \sum_{i=0}^{M} f_i, \qquad \alpha_3 = \sum_{i=M+1}^{M+N} (N-i+M+1)\tau f_i$$

The stability condition for our case is simplified to the following condition:

$$\frac{\lambda\alpha_1}{N\tau\alpha_2 + \alpha_3} < 1$$

In some systems it is possible to derive an explicit expression for f_i in terms of the arrival and service rates, which makes it very easy to check for the stability condition without the need to compute the stationary distribution $\underline{f}$. In our case, it is possible to derive an explicit expression for $\{f_i, i = 0, 1, ..., M+N\}$; however, this expression gets very complicated as the number of ambulances (N) and beds (M) increases. To illustrate this, for the case where there is only a single ambulance and bed ($N = 1, M = 1$) the stationary distribution $\underline{f} = (f_0, f_1, f_2)$ is given by:

$$f_0 = \frac{\mu_1(\mu_1+\mu_2)}{\tau^2+\tau\delta+\tau\mu_1+\tau\mu_2+\delta\mu_1+\delta\mu_2+\mu_1^2+\mu_1\mu_2} \tag{4.24}$$

$$f_1 = \frac{(\delta+\tau)(\mu_1+\mu_2)}{\tau^2+\tau\delta+\tau\mu_1+\tau\mu_2+\delta\mu_1+\delta\mu_2+\mu_1^2+\mu_1\mu_2} \tag{4.25}$$

$$f_2 = \frac{\tau(\tau+\delta)}{\tau^2+\tau\delta+\tau\mu_1+\tau\mu_2+\delta\mu_1+\delta\mu_2+\mu_1^2+\mu_1\mu_2} \tag{4.26}$$

As you can see, with a single ambulance and emergency bed, the explicit expressions are already complicated and hard to interpret. Increasing the values of M and N would result in even more complicated expressions.

Finally, having the stability condition and all the required tools to calculate the steady state probabilities, we can show how all the above ideas are implemented in the mathematical software, Matlab. The following points summarize the step by step algorithm for calculating the steady state probability distribution $\underline{\pi}$:

1. Calculate the rate matrix R using either the simple iterative algorithm or the logarithmic reduction algorithm presented in Figures 4.1 and 4.2.

2. Set $R_{N-1} = R$ and recursively calculate $R_{N-2}, R_{N-3}, ..., R_1, R_0$ using equations 4.14 and 4.15.

3. Compute π_0 by solving 4.18.

4. Use equation 4.7 to recursively compute $\pi_1, \pi_2, ..., \pi_{N-1}$.

5. Use π_{N-1} in equation 4.3 to recursively compute $\pi_N, \pi_{N+1}, ...$

4.3 Implementation

We used Matlab 7.1 to implement the matrix geometric method described in the previous section. All of the Matlab codes for calculating the steady state probabilities can be found in Appendix A. The calculations are done through "The Main Execution File" (found in Appendix A.1) which manages all the functions written in Matlab. In this file, the model parameters presented in Figure 3.2 are first initialized. Next, the boundary portion matrices B_n and C_n and the repeating portion matrices A_0, A_1, and A_2 are constructed. The Matlab code for constructing these matrices are all included in Appendix A.2. After the model has been setup, the stability conditions are checked. This is done through the "stabilitycond(A,A_0,A_2)" function found in Appendix A.3. This function returns 1 if the model is stable and returns 0 otherwise. If the model is stable, the rate matrix R for the repeating

portion and $\{R_i : i = 0, 1, ..., N-2\}$ for the boundary portion are calculated (refer to Appendix A.4). The output of the "Main Execution File" is the steady state probability matrix in the following form:

$$\pi = \begin{pmatrix} \pi_{0,0} & \pi_{0,1} & \cdots & \pi_{0,(M+N)} \\ \pi_{1,0} & \pi_{1,1} & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \pi_{L,0} & \cdots & \cdots & \pi_{L,(M+N)} \end{pmatrix}$$

The $(i,j)^{th}$ entry of the π matrix above can be interpreted as the long run fraction of time that the process spends in state $[I(t), J(t)] = (i,j)$. For computational purposes, in calculating the steady state probabilities we need to truncate the state space $I(t) = i$ at some point, which we call "L". The value of L must be chosen in a way that $\pi_i \simeq 0$ for $i > L$. That is, the probability of having more than a certain number of patients waiting for an ambulance ($i > L$) is approximately zero. One way to do this is to set an arbitrary value for L. However, this might cause a few problems. If the value is set too high it might lead to unnecessary calculations for models that are small and have low utilization. On the other hand, if the value is set too low, we might end up not calculating some of steady state probabilities that are needed. To prevent these problems, we will truncate the state space at a point where:

$$1 - \epsilon \;<\; \sum_{i=0}^{L} \pi_i \;<\; 1$$

where ϵ is a number close to zero. With the above condition, we will make sure that the state space is truncated at a point where the sum of $\{\pi_i : i = 0, 1, ..., L\}$ is close to 1.

We will extend the numerical example 3.3 and compute the steady state probability distribution matrix $\underline{\pi}$ for it in the section section.

4.4 Numerical Example (3.3) Continued

In this section, we continue the numerical example that we presented in section 3.3 by calculating the steady state probability distribution. In order to do so, we will assign a numerical value to the input parameters presented in Figure 3.2 in chapter 3. The assigned values are presented in the following table:

Parameter	λ	τ	δ	μ_1	μ_2	M	N
Value (avg/hour)	2	2	4	2	0.25	4	3

We are assuming that the time unit that we are working with is in "hours". According to the above set of input, on average, 2 patients arrive at the hospital by ambulance every hour, and 4 patients arrive by means other than ambulance every hour. Also, it takes an ambulance 30 minutes to transfer a patient to the hospital from the time it is dispatched to the scene. $\mu_1 = 2$ indicates that the hospital discharges a patient every 30 minutes on average. Further, we assume that the ambulance treatment time is on average 4 hours ($\frac{1}{\mu_2}$) for each patient. That is, if there are no beds available at the ED and an ambulance is experiencing offload delay, it is possible to treat the patient in the ambulance (takes on average 4 hours) and transfer him/her to an in-patient bed directly. As we assumed in chapter 3, if an emergency bed becomes available, the patient is transferred to the bed right away. Note that the above assigned values are for illustration purposes only; they are not intended to be realistic.

In section 3.3 we presented the general form of the boundary and repeating portion matrices for the numerical example in terms of the input parameters. Below we restate these matrices, now in terms of the assigned values:

$$B_0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} -6 & 4 & & & & & & \\ 2 & -8 & 4 & & & & & \\ & 4 & -10 & 4 & & & & \\ & & 6 & -12 & 4 & & & \\ & & & 8 & -10 & 0 & & \\ & & & & 8.25 & -10.25 & 0 & \\ & & & & & 8.5 & -10.5 & 0 \\ & & & & & & 8.75 & -10.75 \end{pmatrix} \end{array}$$

$$B_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} -8 & 4 & & & & & & \\ 2 & -10 & 4 & & & & & \\ & 4 & -12 & 4 & & & & \\ & & 6 & -14 & 4 & & & \\ & & & 8 & -12 & 0 & & \\ & & & & 8.25 & -12.25 & 0 & \\ & & & & & 8.5 & -12.5 & 0 \\ & & & & & & 8.75 & -10.75 \end{pmatrix} \end{array}$$

$$B_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} -10 & 4 & & & & & & \\ 2 & -12 & 4 & & & & & \\ & 4 & -14 & 4 & & & & \\ & & 6 & -16 & 4 & & & \\ & & & 8 & -14 & 0 & & \\ & & & & 8.25 & -14.25 & 0 & \\ & & & & & 8.5 & -12.5 & 0 \\ & & & & & & 8.75 & -10.75 \end{pmatrix} \end{array}$$

$$C_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & 2 & & & & & & \\ & & 2 & & & & & \\ & & & 2 & & & & \\ & & & & 2 & & & \\ & & & & & 2 & & \\ & & & & & & 2 & \\ & & & & & & & 2 \\ & & & & & & & \end{pmatrix} \end{array}$$

$$C_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & 4 & & & & & & \\ & & 4 & & & & & \\ & & & 4 & & & & \\ & & & & 4 & & & \\ & & & & & 4 & & \\ & & & & & & 4 & \\ & & & & & & & 2 \\ & & & & & & & \end{pmatrix} \end{array}$$

$$A_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} -12 & 4 & & & & & & \\ 2 & -14 & 4 & & & & & \\ & 4 & -16 & 4 & & & & \\ & & 6 & -18 & 4 & & & \\ & & & 8 & -16 & 0 & & \\ & & & & 8.25 & -14.25 & 0 & \\ & & & & & 8.5 & -12.5 & 0 \\ & & & & & & 8.75 & -10.75 \end{pmatrix} \end{array}$$

$$A_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} & 6 & & & & & & \\ & & 6 & & & & & \\ & & & 6 & & & & \\ & & & & 6 & & & \\ & & & & & 6 & & \\ & & & & & & 4 & \\ & & & & & & & 2 \\ & & & & & & & \end{pmatrix} \end{array}$$

$$A_0 = 2 \cdot I_8$$

where I_8 is an identity matrix of size 8.

Now we can check to see whether the system is stable in the long run or not. In order to do this, first we need to compute the A matrix which is the sum of A_0, A_1, and A_2 matrices and then solve for the stationary distribution $\underline{f}$ in 4.22. Using Matlab, we get the following result for $\underline{f}$:

$$\underline{f} = \begin{pmatrix} 0.0105 & 0.0525 & 0.1312 & 0.2187 & 0.2734 & 0.1988 & 0.0936 & 0.0214 \end{pmatrix}$$

Now checking inequality 4.23 for the stability condition yields the following results:

$$\underline{f} A_0 e = 2 < \underline{f} A_2 e = 5.1$$

Therefore, we conclude that the system is stable. Using either algorithm presented in Figures 4.1 and 4.2 we get the following rate matrix R:

$$R = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0.1796 & 0.0774 & 0.0388 & 0.0215 & 0.0129 & 0.0044 & 0.0010 & 0.0001 \\ 0.0294 & 0.1764 & 0.0711 & 0.0337 & 0.0185 & 0.0057 & 0.0012 & 0.0001 \\ 0.0089 & 0.0534 & 0.1705 & 0.0645 & 0.0303 & 0.0079 & 0.0016 & 0.0002 \\ 0.0038 & 0.0229 & 0.0721 & 0.1642 & 0.0618 & 0.0117 & 0.0021 & 0.0002 \\ 0.0022 & 0.0133 & 0.0416 & 0.0926 & 0.1707 & 0.0179 & 0.0027 & 0.0003 \\ 0.0016 & 0.0096 & 0.0295 & 0.0642 & 0.1144 & 0.1656 & 0.0108 & 0.0008 \\ 0.0014 & 0.0086 & 0.0261 & 0.0554 & 0.0953 & 0.1299 & 0.1801 & 0.0066 \\ 0.0016 & 0.0094 & 0.0284 & 0.0590 & 0.0980 & 0.1254 & 0.1605 & 0.1977 \end{pmatrix} \end{array}$$

Setting $R_2 = R$ and using equations 4.14 and 4.15, we get the following results for the rate matrices R_0 and R_1:

$$R_0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0.3080 & 0.2321 & 0.1874 & 0.1428 & 0.0976 & 0.0261 & 0.0059 & 0.0007 \\ 0.0872 & 0.3489 & 0.2430 & 0.1719 & 0.1135 & 0.0291 & 0.0065 & 0.0007 \\ 0.0442 & 0.1766 & 0.3701 & 0.2279 & 0.1407 & 0.0333 & 0.0072 & 0.0008 \\ 0.0296 & 0.1183 & 0.2454 & 0.3626 & 0.1967 & 0.0395 & 0.0081 & 0.0009 \\ 0.0238 & 0.0951 & 0.1962 & 0.2858 & 0.3426 & 0.0474 & 0.0091 & 0.0010 \\ 0.0196 & 0.0785 & 0.1612 & 0.2322 & 0.2724 & 0.2180 & 0.0181 & 0.0016 \\ 0.0168 & 0.0671 & 0.1374 & 0.1960 & 0.2254 & 0.1710 & 0.1862 & 0.0072 \\ 0.0177 & 0.0709 & 0.1446 & 0.2043 & 0.2303 & 0.1653 & 0.1669 & 0.1983 \end{pmatrix} \end{array}$$

$$R_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0.2243 & 0.1214 & 0.0729 & 0.0450 & 0.0278 & 0.0077 & 0.0017 & 0.0002 \\ 0.0457 & 0.2283 & 0.1148 & 0.0637 & 0.0370 & 0.0095 & 0.0020 & 0.0002 \\ 0.0170 & 0.0848 & 0.2254 & 0.1045 & 0.0547 & 0.0124 & 0.0024 & 0.0003 \\ 0.0087 & 0.0437 & 0.1149 & 0.2180 & 0.0960 & 0.0172 & 0.0030 & 0.0003 \\ 0.0058 & 0.0292 & 0.0763 & 0.1424 & 0.2200 & 0.0243 & 0.0038 & 0.0004 \\ 0.0041 & 0.0206 & 0.0535 & 0.0983 & 0.1478 & 0.1700 & 0.0115 & 0.0009 \\ 0.0036 & 0.0181 & 0.0467 & 0.0844 & 0.1234 & 0.1335 & 0.1807 & 0.0066 \\ 0.0039 & 0.0196 & 0.0502 & 0.0894 & 0.1271 & 0.1292 & 0.1611 & 0.1977 \end{pmatrix} \end{array}$$

At last, using equations 4.3, 4.7, and 4.18 we get the following steady state probability matrix:

$$\pi = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0.0208 & 0.0623 & 0.0934 & 0.0934 & 0.0700 & 0.0168 & 0.0038 & 0.0007 \\ 0.0208 & 0.0624 & 0.0936 & \underline{0.0936} & 0.0702 & \underline{0.0169} & 0.0037 & 0.0005 \\ 0.0104 & 0.0313 & 0.0470 & 0.0471 & 0.0354 & 0.0087 & 0.0019 & 0.0002 \\ 0.0035 & 0.0105 & 0.0158 & 0.0160 & 0.0123 & 0.0035 & 0.0008 & 0.0001 \\ 0.0012 & 0.0035 & 0.0054 & 0.0055 & \underline{0.0043} & 0.0013 & 0.0003 & 0.0000 \\ 0.0004 & 0.0012 & 0.0018 & 0.0019 & 0.0015 & 0.0005 & 0.0001 & 0.0000 \\ 0.0001 & 0.0004 & 0.0006 & 0.0007 & 0.0005 & 0.0002 & 0.0000 & 0.0000 \\ 0.0000 & 0.0001 & 0.0002 & 0.0002 & 0.0002 & 0.0001 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0001 & 0.0001 & 0.0001 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix} \end{array}$$

To interpret some of the above results, consider $\pi(1,3)$, $\pi(1,5)$, and $\pi(4,4)$ underlined in the above π matrix. According to the value of $\pi(1,3)$, in 9.36% of the time the system is in the state $(1,3)$ where 1 ambulance is in transit and 3 emergency beds are occupied. In 1.69% of the time the system is found to be in the state $(1,5)$ where 1 ambulance is in transit, all hospital beds are occupied, and 1 ambulance is experiencing offload delay. Finally, the probability of finding the system in the state $(4,4)$ where all ambulances are in transit, all hospital are occupied, and 1 patient is waiting for an ambulance is 0.43%. The rest of the values of the π matrix are interpreted in the similar way.

In the next chapter we will show how the steady state probability distribution $\underline{\pi}$ can be used to compute more aggregate system performance measures.

Chapter 5

System Performance Measures

In this chapter, first we will discuss some of the important performance measures for our model and demonstrate how they can be computed using the steady state probability distribution $\underline{\pi}$. Next, we will compute these performance measures for the numerical example presented in Section 3.3 and extended in Section 4.4. Finally, simulation is used to validate the basic model and performance measures derived from it.

5.1 Computing Various Performance Measures

Let us define the following set of variables for our model:

Random variable	Description
$P_w = 0, 1, 2, ...$	Number of patients waiting for ambulance
$H_b = 0, 1, ..., M$	Number of ED beds occupied
$A_t = 0, 1, ..., N$	Number of ambulances in transit
$A_d = 0, 1, ..., N$	Number of ambulances experiencing offload delay
$A_b = 0, 1, ..., N$	Number of ambulances busy

Table 5.1: System Random Variables

Recall that N is the total number of ambulances and M is the total number of emergency beds. The steady state probability distribution $\underline{\pi}$ can be used to compute the distribution (probability function) of each of the above random variables which will allow us to compute aggregate system performance measures. For example, if we know the distribution of the number of ambulances in offload delay, we can easily compute the expected number of ambulances in offload delay by computing $E[A_d]$. Below we will show how the distribution of the above mentioned random variables are derived from the steady state probability distribution $\underline{\pi}$.

5.1.1 Distribution of P_w

Consider a system with $N = 2$ ambulances and $M = 4$ emergency beds. Suppose we are interested in computing the probability that 1 patient is waiting for an ambulance. To compute this, we would have to sum over all states with a level of 3 and a phase of less than 5, since in those states all ambulances are in transit and 1 patient is waiting for an ambulance. However, we should also consider states $(2, 5)$ and $(1, 6)$ since they also correspond to the situation where there is 1 patient waiting for an ambulance. In state $(2, 5)$, there are 4 beds occupied, 1 ambulance in offload delay, 1 ambulance in transit, and 1 patient waiting for an ambulance. In state $(1, 6)$ we have all beds occupied, all ambulances in offload delay, and 1 patient waiting for an ambulance. Therefore we have:

$$Pr(P_w = 1) = \sum_{j=0}^{4} \pi(3, j) + \pi(2, 5) + \pi(1, 6)$$

Similarly, to compute the probability of having 2 patients waiting for ambulance, we would have to sum over states $(4, 0), ..., (4, 4)$, $(3, 5)$, and (2,6), since in all those states there are 2 patients waiting for ambulance. In general, for this example, the

probability that $i > 0$ patients are waiting for ambulance is given by:

$$Pr(P_w = i) = \sum_{j=0}^{4} \pi(2+i, j) + \pi(2+i-1, 5) + \pi(2+i-2, 6)$$

Now, to compute the case where $i = 0$, that is no patients are waiting for an ambulance, we can use the fact that the sum of $(P_w = i)$ for $i = 0, 1, ..., L$ is 1 and calculate the $(P_w = 0)$ in the following way:

$$Pr(P_w = 0) = 1 - \sum_{n=1}^{L} Pr(P_w = n)$$

Note that from chapter 4, "L" is the point where $\pi_i \simeq 0$ for $i > L$. For the general case where we have N ambulances and M patients, the distribution of P_w is given by:

$$Pr(P_w = n) = \begin{cases} 1 - \sum_{n=1}^{L} Pr(P_w = 0) & n = 0, \\ \sum_{j=0}^{M} \pi(N+n, j) + \sum_{i=1}^{N} \pi(N+n-i, M+i) & n = 1, 2, ..., L. \end{cases} \tag{5.1}$$

5.1.2 Distribution of H_b

To compute the distribution of the number of emergency beds occupied, first we define $\pi_{J(t)}(j)$, which is the marginal distribution of the phase of the MC, $J(t)$:

$$\pi_{J(t)}(j) = \sum_{i=0}^{L} \pi(i, j), \quad j = 0, 1, ..., M+N \tag{5.2}$$

Now, the p.f. of H_b is given by:

$$Pr(H_b = n) = \begin{cases} \pi_{J(t)}(n) & n = 0, 1, ..., M-1, \\ 1 - \sum_{i=0}^{M-1} \pi_{J(t)}(i) & n = M. \end{cases} \tag{5.3}$$

Basically, to compute the probability of having $n < M$ beds occupied, we would have to sum over all states with a phase n. However, for the case where $n = M$, we should note that all states with a phase greater than M represent the situation where M beds are occupied. Therefore, we would have to sum over all states with phase $M, M+1, ..., M+N$.

5.1.3 Distribution of A_t

Again we will use an example to show how the distribution of A_t can be calculated. Consider a system with $N = 2$ ambulances and $M = 4$ emergency beds. To compute the probability of having no ambulances in transit we would have to sum over all states with a level of 0 since they all correspond to the situation where none of the ambulances are in transit. However, we should also consider states with a phase of 6, $\{(1,6),(2,6),...,(L,6)\}$ since they represent the situation where all ambulances are in offload delay, which means that there are no ambulances in transit. Therefore, the probability of having no ambulances in transit is given by:

$$Pr(A_t = 0) = \sum_{j=0}^{6} \pi(0,j) + \sum_{i=1}^{L} \pi(i,6)$$

Similarly, to compute the probability of having 1 ambulance in transit, we have to consider states $(1,0),(1,1),...,(1,5)$ where there is one ambulance in transit, and states $(2,5),(3,5),...,(L,5)$ where there is one ambulance experiencing offload delay and one ambulance in transit. Therefore, we have:

$$Pr(A_t = 1) = \sum_{j=0}^{5} \pi(1,j) + \sum_{i=2}^{L} \pi(i,5)$$

To calculate the probability of having 2 ambulances in transit, we note that all the other states that were not considered before correspond to the situation

where there are 2 ambulances in transit. Basically, all states with a level $I(t) \geq 2$ and a phase $J(t) \leq 4$ refer to the situation where there are 2 ambulances in transit. Instead of summing over all those states we can calculate $Pr(A_t = 2)$ in the following way:

$$Pr(A_t = 2) = 1 - [Pr(A_t = 0) + Pr(A_t = 1)]$$

since $\sum_{n=0}^{2} Pr(A_t = n) = 1$. For the general case where there are N ambulances and M emergency beds, we have:

$$Pr(A_t = n) = \begin{cases} \sum_{j=0}^{N+M-n} \pi(n, j) + \sum_{i=n+1}^{L} \pi(i, N + M - n) & n = 0, 1, ..., N-1, \\ 1 - \sum_{i=0}^{N-1} Pr(A_t = i) & n = N. \end{cases} \tag{5.4}$$

5.1.4 Distribution of A_d

To compute the distribution of the number of ambulances in offload delay, we would have to consider all states with a phase greater than M since in those states all hospital beds are occupied and ambulances are experiencing offload delay. The p.f. of A_d is given by:

$$Pr(A_d = n) = \begin{cases} \sum_{j=0}^{M} \pi_{J(t)}(j) & n = 0, \\ \pi_{J(t)}(M + n) & n = 1, 2, ..., N. \end{cases} \tag{5.5}$$

where $\pi_{J(t)}(j)$ is defined in 5.2. For example, if we are interested in calculating the probability that there is one ambulance in offload delay in as system where $N = 2$ and $M = 4$, we have to sum over all states that have phase of 5. States with a phase of 5 are those states where all hospital beds are occupied and one ambulance

is in offload delay.

5.1.5 Distribution of A_b

At last, we will show how to compute the probability that $0 \leq n \leq N$ ambulances are busy in a context of an example. A busy ambulance refers to the situation where the ambulance is either in offload delay or it is transferring a patient to the hospital. Consider a system with 3 ambulances and 4 emergency beds, hence $N = 3$ and $M = 4$. The probability of having zero ambulances busy can be easily calculated by:

$$Pr(A_b = 0) = \sum_{j=0}^{4} \pi(0, j)$$

In the above formula we did not include states with level 0 and phase greater than 4 since they represent the situation where at least one ambulance is in offload delay. To calculate the probability of 1 ambulance busy, we have to consider states where there is either 1 ambulance in transit,$\{(1,0),(1,1),...,(1,4)\}$, or 1 ambulance in offload delay $\{(0,5)\}$. Therefore, the probability of having 1 ambulance busy is given by:

$$Pr(A_b = 1) = \sum_{j=0}^{4} \pi(1, j) + \pi(0, 5)$$

Similarly, in order to compute the probability of 2 ambulances being busy, we have to consider states $\{(2,0),(2,1),...,(2,4)\}$, $(1,5)$, and $(0,6)$ since they represent the situation where there are 2 ambulances in transit, 1 ambulance in transit and 1 in offload delay, and 2 ambulances in offload delay, respectively. Therefore, we have:

$$Pr(A_b = 2) = \sum_{j=0}^{4} \pi(2, j) + \pi(1, 5) + \pi(0, 6)$$

Again, we can use the fact that:

$$\sum_{n=0}^{3} Pr(A_b = n) = 1$$

to compute the probability of having 3 ambulances busy:

$$Pr(A_b = 3) = 1 - [Pr(A_b = 0) + Pr(A_b = 1) + Pr(A_b = 2)]$$

In general, for the case of N ambulances and M emergency beds we have:

$$Pr(A_b = n) = \begin{cases} \sum_{j=0}^{M} \pi(0, j) & n = 0 \\ \sum_{j=0}^{M} \pi(n, j) + \sum_{i=1}^{n} \pi(n - i, M + i) & n = 1, ..., N - 1, \\ 1 - \sum_{i=0}^{N-1} Pr(A_t = i) & n = N. \end{cases} \tag{5.6}$$

As can be seen, computing the distribution of some of the random variables defined in Table 5.1 is quite complicated. As we mentioned in chapter 3, this is due to the fact that we combined and reduced 4 system state variables into 2 state variables. However as you can see, computing various performance measures is still manageable and not problematic at all. Using the above distributions next we show how to compute more aggregate performance measures.

5.1.6 System Performance Measures

The expected value of the random variables defined in Table 5.1 can easily be calculated using the definition of expectation. For example, the expected number of patients waiting for an ambulance is calculated in the following way:

$$E[P_w] = \sum_{n=1}^{L} n \, Pr(P_w = n) \tag{5.7}$$

Another performance measure of interest is the expected waiting time for an ambulance. It is important to note the expected waiting time for ambulance refers to the time from emergency call arrival until an ambulance is dispatched to the scene. The time it takes for an ambulance to reach the scene is not included in this performance measure since it is a part of the ambulance transit time. We can apply Little's law to compute the expected waiting time for ambulance. All we need is the average number of patients waiting for ambulance, $E(P_w)$, and the rate of emergency call arrival, λ. Given these, the average waiting time for an ambulance is given by:

$$E[W] = \frac{E[P_w]}{\lambda} \tag{5.8}$$

where W is the waiting time random variable. Similarly, we can compute the expected time in offload delay by applying Little's law. Given the average number of ambulances in offload delay, $E(A_d)$ and the rate at which ambulances arrive at the hospital, $N\tau$, the average time in offload delay is given by:

$$E(D) = \frac{E[A_d]}{N\tau} \cdot \Pr(\text{at least one ambulance in offload delay}) \tag{5.9}$$

where $E[A_d] = \sum_{n=1}^{N} n\, Pr(A_d = n)$ and D is the offload delay time random variable. Note that the average time in offload delay is calculated over the ambulances that actually experience offload delay, and they are the ones contributing to the average. That is why the $\frac{E[A_d]}{N\tau}$ factor is multiplied by the probability of having at least one ambulance in offload delay. Also in 5.9, we are using the fact that the minimum of N Exponential distributions with rate τ results in another exponential distribution with a rate $N\tau$.

Another important performance measure of interest is the expected ambulance utilization. It can be calculated by dividing the expected number of ambulances busy by the total number of ambulances:

$$U_A = \frac{E[A_b]}{N} \tag{5.10}$$

Similarly, the expected emergency bed utilization is given by:

$$U_B = \frac{E[H_b]}{M} \tag{5.11}$$

In section 4.4 we computed the steady state probability distribution $\underline{\pi}$ for the numerical example presented in section 3.3. Next, we will further extend example 4.4 and compute the above mentioned performance measures.

5.2 Numerical Example (4.4) Continued

In this section we use the steady state probability distribution computed in example 4.4 in chapter 4 and compute the above mentioned performance measures. Again, we have used Matlab 7.1 to numerically compute the performance measures (refer to Appendix A.6 for the performance measure calculation codes). Let us restate the assigned values for the model parameters from section 4.4:

Parameter	λ	τ	δ	μ_1	μ_2	M	N
Value (avg/hour)	2	2	4	2	0.25	4	3

Table 5.2 shows the distribution and the expected value of the random variables defined in Table 5.1 for this example. Table 5.3 shows the average waiting time for ambulance, the average time in offload delay, and the expected ambulance and ED bed utilization.

There are 4 emergency beds in the system and on average 4 outside patients and 2 patients via ambulance arrive at the hospital every hour. With these arrival rates

n	$Pr(A_t = n)$	$Pr(H_b = n)$	$Pr(A_t = n)$	$Pr(A_d = n)$	$Pr(A_b = n)$
0	0.9602	0.0572	0.3620	0.9399	0.3398
1	0.0257	0.1718	0.3643	0.0479	0.3574
2	0.0091	0.2580	0.1854	0.0107	0.1919
3	0.0032	0.2585	0.0883	0.0016	0.1109
4	0.0011	0.2546			
5	0.0004				
6	0.0001				
Mean	0.0614	2.4815	1	0.0730	1.0739

Table 5.2: Example 4.4 - Performance Measure Distribution Results

E(W) (minutes)	E(D) (minutes)	U_A (%)	U_B (%)
1.84	12.29	35	62

Table 5.3: Example 4.4 - Aggregate Performance Measure Results

to the hospital, we would expect the hospital to be somewhat busy considering the hospital's overall treatment rate of 8 patients per hour. According to Tables 5.2 and 5.3, in 25.46% ($Pr(H_b = 4)$) of the time all emergency beds are occupied and the expected ED bed utilization is 62%. Also, by looking at the value of $E[H_b]$ we can see that on average 2.5 out of 4 hospital beds are occupied all the time. In this case, if more than roughly two ambulances arrive at the hospital we would have an ambulance offload delay. However, the chances of having an offload delay is only 6% ($1 - Pr(A_d = 0)$) and the average time in offload delay is about 12.3 minutes.

From the input parameter values, we know that on average 2 emergency call arrive every hour and there are 3 ambulances that can transfer 2 patients per hour on average. We would expect to see almost no queue for patient waiting for ambulance since there enough ambulances to transfer patients to the hospital. From Table 5.2 we can see that there are no patients waiting for ambulance in 96% ($Pr(P_w = 0)$) of the time and if there is a patient waiting (4% of the time), the

average waiting time is very small (2 minutes). Tables 5.2 and 5.3 further show that in 34% ($Pr(A_b = 0)$) of the time all ambulances are available (idle) and the ambulance utilization is 35%.

5.3 Model Validation

In the previous sections we demonstrated how to compute various system performance measures for our model and computed them for the example 4.4 in chapter 4. In this section we will show that the results obtained for the performance measures using our model are accurate and valid. Simulation is chosen for our model validation. Developing a simulation model for the system that we described in section 3.1 is quite easy and simple. In general, designing a simulation model is always easier than an analytical model. However, one of the main advantages of the analytical approach (such as the matrix geometric method) over the simulation is the fact that the analytical approach produces more accurate results in general; especially when system is highly utilized as we will see below. For a low utilized system we would expect to achieve similar results using either approach.

We have used the simulation software Simul8 to model the system that we described in section 3.1. All of the assumptions that were made in section 3.1 hold for the simulation model as well. We will compare the results of the analytical approach with the simulation approach for a fairly large system (compared to our previous examples) with 7 ambulances $n = 7$ and 12 emergency beds $M = 12$. Three sets of input parameters, shown in Table 5.5 were chosen for the arrival and service rates. Again, assigned parameter values are for illustration purposes only and they are not intended to be realistic. The first set of input parameters correspond to a low utilized system where there is almost no queue for ambulance and on average most ambulances are available. The second set represents a system

with a medium utilization. The third set of the input parameters correspond to a high utilized system where almost all ambulances are busy most of the time and the probability of having at least one patient waiting for an ambulance is quite high.

	λ	τ	δ	μ_1	μ_2	M	N
Set 1-Low Utilization System	4	2	6	2	1	12	7
Set 2-Medium Utilization System	6.5	1.2	6	0.9	0.5	12	7
Set 3-High Utilization System	6.2	1.5	10	0.5	0.25	12	7

Table 5.4: Input Parameter Sets for Model Validation

Let us briefly discuss the simulation model shown in Appendix C. Emergency calls arrive through the "Call Arrival" entry point and outside patients arrive through the "Outside Patients" entry point. As you can see, there is no queue between the "hospital" work center and the "Outside Patients" entry point since we assumed that outside patients are lost when there are no emergency beds available. Note that the replication property of the "hospital" work center has been set to 12 since $M = 12$. Also we are assuming that $\mu_2 > 0$, that is, it is possible to treat patients in the ambulance without the need of occupying an emergency bed.

To validate our model we will compute the distribution of the random variables P_w, A_t, A_d, A_b, and H_b as well as their expected values for both the simulation model and the analytical model, and compare the results. For each set of the input parameters, we will conduct a simulation trial with 20 runs, each with a results collection period and warm-up period of 1000 and 100 hours, respectively.

We have used the PASTA (Poisson Arrival See Time Averages) property in our simulation model to compute the above mentioned distributions. For example, to compute the distribution of the number of patients waiting for ambulance (P_w), first we recorded the number of work items in the "Call Queue" each time an arrival occurred. Next, we counted the total number of times that there were n work

items in the "Call Queue" at the end of each run. The probability of n patients waiting for an ambulance was then calculated by dividing the value of n by the total number of arrivals. The validation results are summarized in Tables 5.4-5.6[1]. Note that we have abbreviated the simulation approach to "SA" and the analytical approach (Matrix Geometric Method) to "AA" in Tables 5.4-5.6.

[1]Note that thorough analysis would include constructing confidence intervals; however, we did not go into such detail since it is not the focus of our work

N	$\Pr(P_w = n)$		$\Pr(H_b = n)$		$\Pr(A_t = n)$		$\Pr(A_d = n)$		$\Pr(A_b = n)$	
	SA	AA	SA	AA	SA	AA	SA	AA	SA	AA
0	0.9985	0.9986	0.1345	0.1353	0.0074	0.0067	0.9993	0.9994	0.1344	0.1352
1	0.0010	0.0010	0.2696	0.2706	0.0340	0.0337	0.0006	0.0006	0.2696	0.2705
2	0.0003	0.0003	0.2693	0.2706	0.0852	0.0843	0.0001	0.0001	0.2691	0.2706
3	0.0001	0.0001	0.1796	0.1804	0.1402	0.1406	0.0000	0.0000	0.1797	0.1805
4	0.0000	0.0000	0.0914	0.0902	0.1784	0.1757	0.0000	0.0000	0.0915	0.0903
5	0.0000	0.0000	0.0378	0.0361	0.1740	0.1757	0.0000	0.0000	0.0378	0.0361
6	0.0000	0.0000	0.0127	0.0120	0.1450	0.1464	0.0000	0.0000	0.0128	0.0120
7	0.0000	0.0000	0.0051	0.0048	0.1042	0.1046				
8	0.0000	0.0000			0.0657	0.0654				
9	0.0000	0.0000			0.0358	0.0363				
10	0.0000	0.0000			0.0176	0.0182				
11	0.0000	0.0000			0.0084	0.0083				
12	0.0000	0.0000			0.0041	0.0041				
Mean	0.0021	0.0019	2.0133	1.9999	4.9728	4.9884	0.0008	0.0008	1.9786	1.9669

Table 5.5: Parameter Set 1-Performance Evaluation

N	$\Pr(P_w = n)$		$\Pr(H_b = n)$		$\Pr(A_t = n)$		$\Pr(A_d = n)$		$\Pr(A_b = n)$	
	SA	AA	SA	AA	SA	AA	SA	AA	SA	AA
0	0.5361	0.5307	0.0033	0.0030	0.0002	0.0000	0.7592	0.7583	0.0025	0.0023
1	0.0785	0.0795	0.0172	0.0167	0.0002	0.0000	0.1295	0.1289	0.0132	0.0128
2	0.0650	0.0666	0.0461	0.0465	0.0003	0.0002	0.0686	0.0684	0.0357	0.0359
3	0.0539	0.0555	0.0901	0.0882	0.0009	0.0008	0.0294	0.0304	0.0696	0.0678
4	0.0445	0.0461	0.1292	0.1296	0.0030	0.0027	0.0103	0.0107	0.0974	0.0967
5	0.0370	0.0382	0.1599	0.1590	0.0080	0.0076	0.0025	0.0028	0.1138	0.1118
6	0.0307	0.0317	0.1703	0.1700	0.0184	0.0177	0.0005	0.0005	0.1095	0.1094
7	0.0256	0.0262	0.3838	0.3870	0.0369	0.0357	0.0000	0.0000	0.5581	0.5633
8	0.0211	0.0217			0.0634	0.0628				
9	0.0177	0.0180			0.0991	0.0986				
10	0.0150	0.0149			0.1385	0.1395				
11	0.0125	0.0123			0.1779	0.1798				
12	0.0104	0.0102			0.4532	0.4547				
≥ 13	0.0519	0.0484								
Mean	2.8261	2.7248	5.4045	5.4167	10.6036	10.6267	0.4117	0.4167	5.8162	5.8333

Table 5.6: Parameter Set 2-Performance Evaluation

N	$\Pr(P_w = n)$		$\Pr(H_b = n)$		$\Pr(A_t = n)$		$\Pr(A_d = n)$		$\Pr(A_b = n)$	
	SA	AA	SA	AA	SA	AA	SA	AA	SA	AA
0	0.14433	0.1120	0.0111	0.0102	0.0000	0.0000	0.1495	0.1408	0.0009	0.0006
1	0.03229	0.0244	0.0507	0.0501	0.0001	0.0000	0.1486	0.1452	0.0038	0.0029
2	0.03152	0.0241	0.1233	0.1206	0.0001	0.0000	0.1882	0.1882	0.0099	0.0074
3	0.03098	0.0236	0.1888	0.1881	0.0001	0.0000	0.1987	0.2019	0.0165	0.0130
4	0.03022	0.0231	0.2118	0.2128	0.0001	0.0000	0.1659	0.1699	0.0237	0.0182
5	0.02878	0.0225	0.1847	0.1857	0.0001	0.0000	0.1016	0.1044	0.0276	0.0219
6	0.02813	0.0220	0.1285	0.1298	0.0001	0.0000	0.0393	0.0416	0.0304	0.0238
7	0.02652	0.0214	0.1011	0.1027	0.0002	0.0001	0.0081	0.0080	0.8873	0.9123
8	0.02457	0.0208			0.0006	0.0006				
9	0.02399	0.0202			0.0027	0.0024				
10	0.02338	0.0196			0.0095	0.009				
11	0.02317	0.0191			0.0344	0.0308				
12	0.02259	0.0185			0.9521	0.957				
≥ 13	0.5295	0.6287								
Mean	20.1980	24.5383	4.1137	4.1330	11.9308	11.9399	2.5856	2.6345	6.6992	6.7679

Table 5.7: Parameter Set 3-Performance Evaluation

As you can see, the results are fairly close for the parameter set 1. The same holds for the parameter set 2 except for the case of P_w where we can see some deviation between the two approaches. For the high utilized system (parameter set 3) the results deviate the most. As we mentioned before, producing inaccurate results as the system becomes more and more utilized is the major downfall of the simulation approach. On the other hand, in the analytical approach we always achieve consistent and accurate results no matter what the system utilization is.

In this chapter we demonstrated how to compute some of the important performance measures for our model and we validated the results of our model by simulation. In the next chapter we will perform sensitivity analysis by varying some of the input parameters and analyzing their effect on a number of system performance measures.

Chapter 6

Sensitivity Analysis

In this chapter we use sensitivity analysis to analyze the effect of varying some of the input parameters on a number of system performance measures. In section 6.1, we develop and present a base case for our sensitivity analysis. In subsequent sections, we examine the impact of changing some of the input parameters such as the number of beds, ED treatment time (or patient length of stay), and the number of ambulances, on system performance measures.

6.1 The Base Case Model

In this section, we construct a base case model for the purposes of analyzing the impact of parameter changes on system performance. In order to present a reasonable base case, the parameter values are based on those of the public health system of the region of Waterloo. For some parameters we had to make some approximating assumptions due to the complexity of the real system, but we have attempted to retain as much reality as possible. In this way, the model, though approximate, can still provide some useful insights into the ambulance offload time problem. Extensions to the model (future research) are necessary to make the model highly

accurate, but the current level of detail still provides good insight into the tradeoffs involved in capacity planning.

The region of Waterloo has a population of approximately 509,000 and geographic area of 1382 sq.kms. There are three community hospitals in the region: Grand River Hospital, St. Mary's Hospital, and Cambridge Memorial Hospital. The delivery of ambulance services is the responsibility of the EMS division of public health which operates a fleet of 29 vehicles including 18 ambulances through 8 stations. Among the above mentioned hospitals in the region, we have chosen the Grand River Hospital (GRH) to develop our base case on since it has the largest emergency department in the region.

When a patient arrives at the ED via ambulance, the hospital must accept transfer of care of the patient before the paramedic staff can remove the patient from the ambulance. If the patient is in a stable health condition, they can be moved into the waiting room of the ED for further care. If, however, their condition is not stable, the hospital will not accept transfer of care until there are sufficient facilities available (*i.e.* a bed). In these cases, the paramedics and ambulance must continue to provide care at the hospital until such time as a bed opens up.

In the base model, we capture only the most severely ill patients - those that are classified as CTAS level I and II. CTAS (The Canadian Triage and Acuity Scale) is a scheme that allows medical staff to rank the severity of a patients health condition to ensure that the sickest patients are seen first. Patients that arrive to emergency and are classified as CTAS level I or II are in the worst health condition and are in need of immediate care. They typically arrive by ambulance. Patients with a CTAS level IV and V are in a stable health condition and are the least urgent. These patients typically arrive to the ED via means other than ambulance. The base case focuses on the flow of patients delivered by ambulance i.e. the most severely ill. That said, the model also includes non-ambulance arrivals of CTAS

levels III to V patients, as some of them do become more ill and take up a hospital bed. This is necessary to capture the overall arrivals of patients to the hospital beds not only those delivered by ambulance. Next, we will discuss the input parameter values that we have chosen for our base case.

6.1.1 Input Parameter set values

The input parameter values used for our base case are as follows:

- We let M, the number of beds, to be equal to 12 since the GRH currently has 12 beds in its emergency department.
- We assume that 40% of emergency calls result in a patient being transferred to the GRH. Given this and using the data provided in the report prepared by the region of Waterloo's public health [34] the rate of emergency call arrival or λ is computed to be 1.2 emergency calls/hour. This rate only represents the patients with a CTAS level of I and II.
- Since we assumed that 40% of emergency calls result in a patient being transferred to the GRH, we also assume that 40% of EMS ambulances in the region of Waterloo (7 out of 18) are dedicated to serve those calls. The remaining 11 ambulances are used to service the other two hospitals. As a result we let N, the number of ambulances, to be equal to 7.
- The average transit time for an ambulance (τ) which includes the travel time to the emergency scene, on scene care time, travel time to the hospital and offload time is assumed to be 1 patient/hour based on the discussions with EMS staff in the region of Waterloo.

- Using the raw data that we were provided from the GRH on their ED operation[1], we were able to calculate the average time a patient spends on an emergency bed to be 6.75 hours ($\frac{1}{\mu_1}$) which gives the value of the hospital's service rate μ_1 in our model. Note that the value of 6.75 hours is only valid for the patients that are assigned a CTAS level of I or II.

- The ambulance service rate μ_2 is assumed to be zero in our sensitivity analysis. That is, we are assuming that it is not possible to treat a patient in an ambulance and transfer him/her to an in-patient bed without going through the ED of the hospital. Therefore ambulances must wait at the hospital until a bed at the ED becomes available.

- We used the given raw data from the GRH and computed the rate at which outside patients arrive at the ED of the GRH (δ) which is given by 0.97 patients/hour. Note that this value corresponds to all patients that occupy an ED bed no matter what their CTAS level is.

Table 6.1 below summarizes the input parameter values chosen for our base case model. Next we will compute various performance measures for our base case model.

Parameter	λ	τ	δ	μ_1	μ_2	M	N
Value (avg/hour)	1.2	1	0.97	0.148	0	12	7

Table 6.1: Input Parameter Values for the Base Case Model

6.1.2 Performance Measure Results

Figures 6.1-6.4 show the distribution of the number of patients waiting for ambulance (P_w), the number of ambulances in offload delay (A_b), the number of ambu-

[1]The raw data that we were provided with corresponds to 7 months worth of data that were collected between January 2007 and August 2007 by the Grand River Hospital

lances busy (A_b), and the number of beds occupied (H_b), respectively. Table 6.2 shows the average waiting time for an ambulance (W), the average time in offload delay (D), and ambulance and bed utilization (U_A and U_B), as well as the mean value of the distributions presented in Figures 6.1-6.4. By looking at the results of Table 6.2 and Figure 6.1 we can see that in 95% of the time there are no patients waiting for an ambulance, but if there is a patient waiting for an ambulance (5% of the time), he/she must wait for approximately 6.6 minutes. It is important to note the waiting time for an ambulance corresponds to the time from an arrival of an emergency call to the time when an ambulance is dispatched. Now looking at the distribution of A_d, we would expect to have no ambulances in offload delay in 62% of the time and at least one ambulance in offload delay in 38% of the time. The mean value of A_d however indicates that on average there is one ambulance in offload delay at the GRH and the average time in offload is estimated to be around 23 minutes per ambulance.

Table 6.2 shows a 32% utilization for ambulances and 91% utilization for the ED beds of the GRH. The distribution of H_b in Figure 6.2 further indicates that all ED beds are occupied in 57% of the time. It is interesting to note that the utilization results are very well matched with the actual results in the region. Based on the discussion we had with the director of the EMS in Waterloo and the data provided in the report prepared by the region of Waterloo's public health (reference), ambulance utilization is usually between 30% to 35% and our result of 32% is indeed in this range. The 91% ED bed utilization result is also quite reasonable based on the talk we had with the GRH staff and it also matches the results of the surrounding areas. According to the report on the hospital care in the Greater Toronto Area (GTA) (reference) the ED bed occupancy rate is around 94% for GTA hospitals.

Now that we have our base case setup we can start our sensitivity analysis. First we are interested to see what happens if the number of beds in the ED is

either increased or decreased. Increasing the number of beds would definitely help the ED of the GRH to treat more patients which in turn would increase the flow of patients through the hospital. Being able to treat more patients would also benefit the EMS of Waterloo since it decreases the probability that an ambulance experiences an offload delay. This would result in more ambulances to be available and provide service to emergency calls which in turn reduces patient waiting time for an ambulance. In the following section we will analyze the effect of adding more beds to the ED of the GRH on various system performance measures.

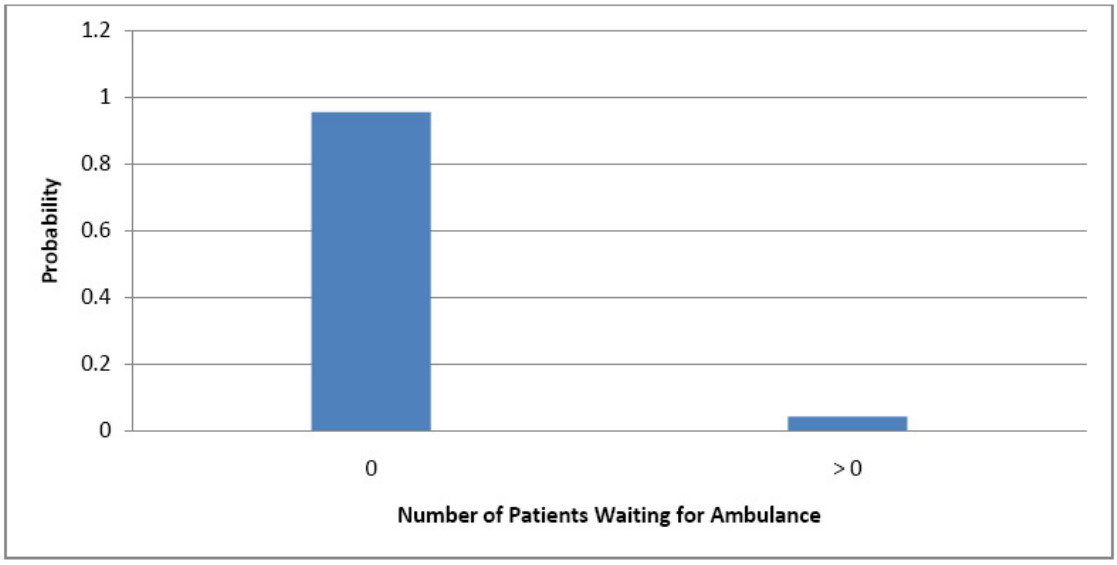

Figure 6.1: Distribution of the Number of Patients Waiting for Ambulance

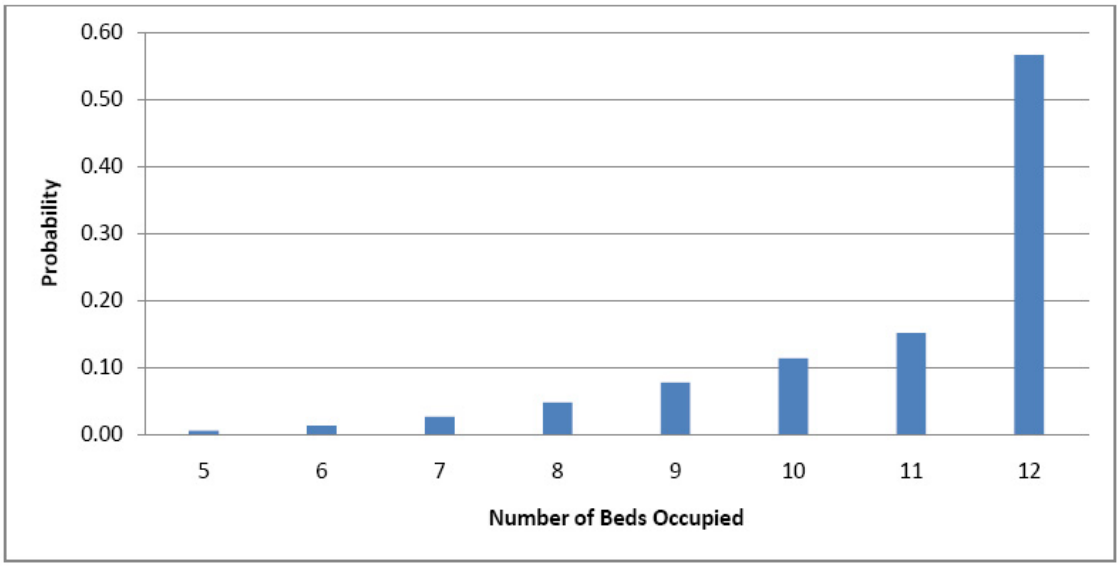

Figure 6.2: Distribution of the Number of Emergency Beds Occupied

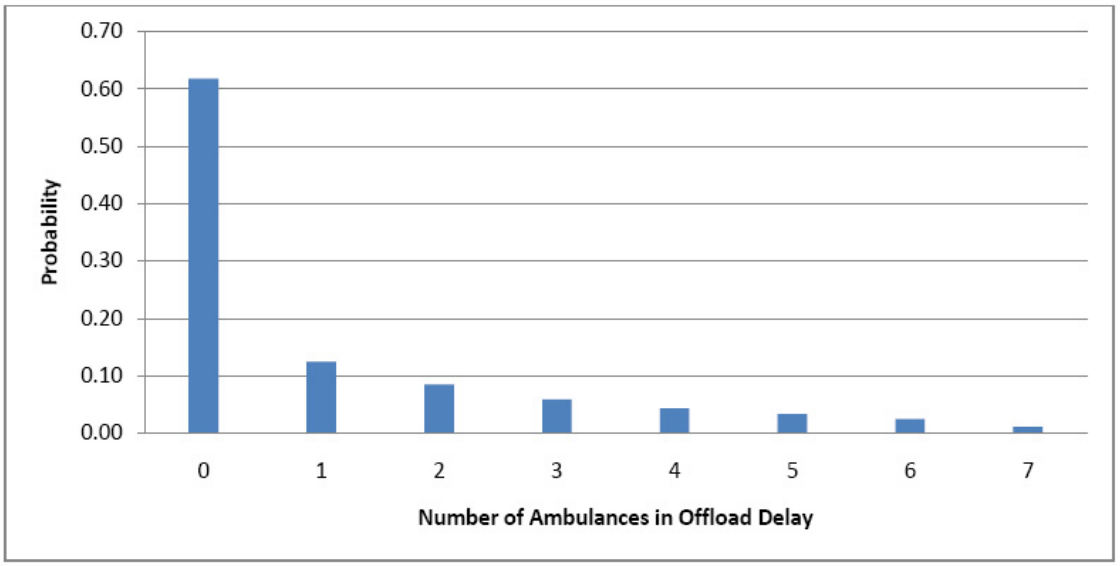

Figure 6.3: Distribution of the Number of Ambulances in Offload Delay

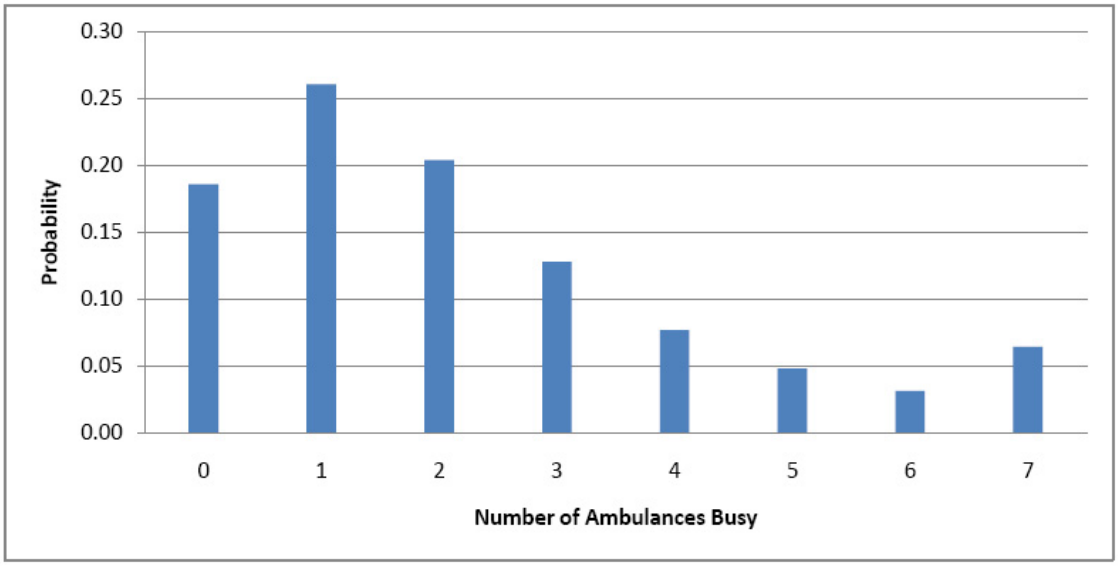

Figure 6.4: Distribution of the Number of Ambulances Busy

$E(W)$(minutes)	$E(D)$(minutes)	$U_A(\%)$	$U_B(\%)$	$E(P_w)$	$E(A_d)$	$E(A_b)$	$E(H_b)$
6.62	23.41	32	91	0.13	1.04	2.24	10.93

Table 6.2: Base Case Performance Measure Results

6.2 Varying the Number of Emergency Beds

As we discussed, increasing the number of beds in the ED of the GRH would improve many system performance measures. The performance measures that we are considering in our analysis are:

1. $E(P_w)$: Expected number of patients waiting for ambulance
2. $E(A_b)$: Expected number of ambulances experiencing offload delay
3. $E(D)$: Average time in offload delay
4. U_A: Expected ambulance utilization

5. U_B: Expected ED bed utilization

We will start by analyzing the effect of adding extra beds in the ED on the expected number of patients waiting for ambulance (Figure 6.5). Point "C" represents the ED's current operation point throughout the chapter. As can be seen when the number of beds increases, the expected number of patients waiting for ambulance approaches zero, which is what we had expected. Note that the same pattern also holds for the average waiting time for ambulance, since the average waiting time is calculated by dividing the average number of patients waiting for ambulance by the call arrival rate λ. With more beds in the ED, it is less likely for an ambulance to experience an offload delay. This is further proven by looking at Figures 6.6 and 6.7. As can be seen, the expected number of ambulances in offload delay and the average time in offload delay are similarly decreasing as the number of ED beds increases. Therefore, with more ambulances available less number of patients have to wait for an ambulance and they have to wait for a much shorter period of time.

Now let us look at the ambulance and the ED bed utilization results shown in Figures 6.8 and 6.9. Both ambulance and ED bed utilization decreases as the number of beds increases. With 20 ED beds, ambulance utilization reaches below 20% from 32% and similarly bed utilization reaches around 70% from 91% before. As can be seen all performance measures improved dramatically as the number of ED beds increased from 12 to 20.

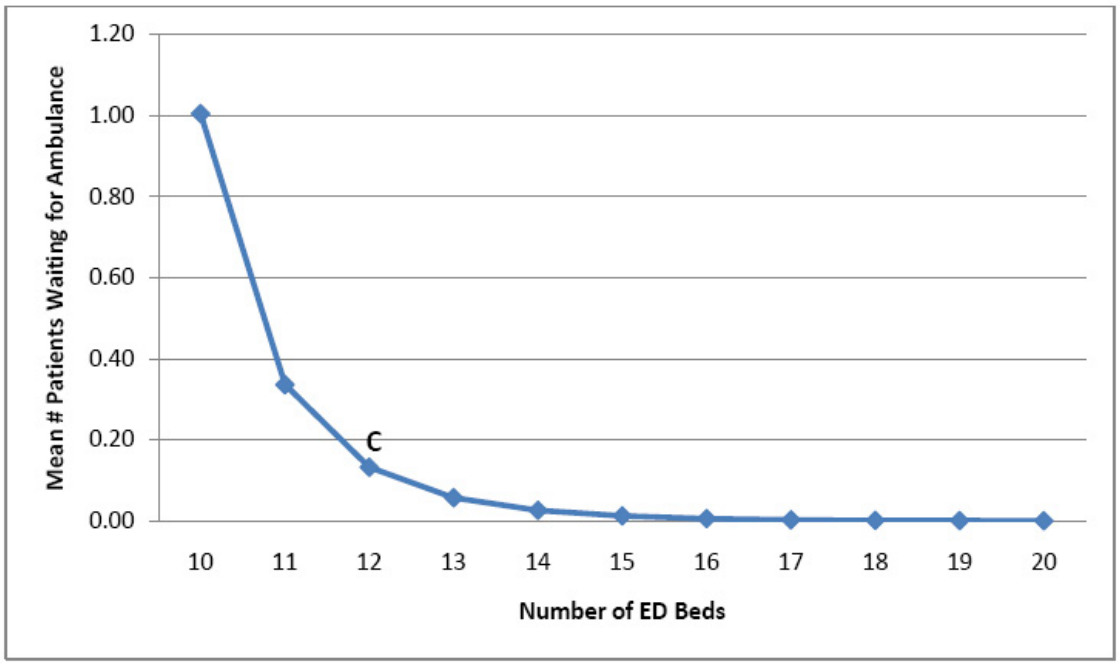

Figure 6.5: Number of Beds vs. Expected Number of Patients Waiting

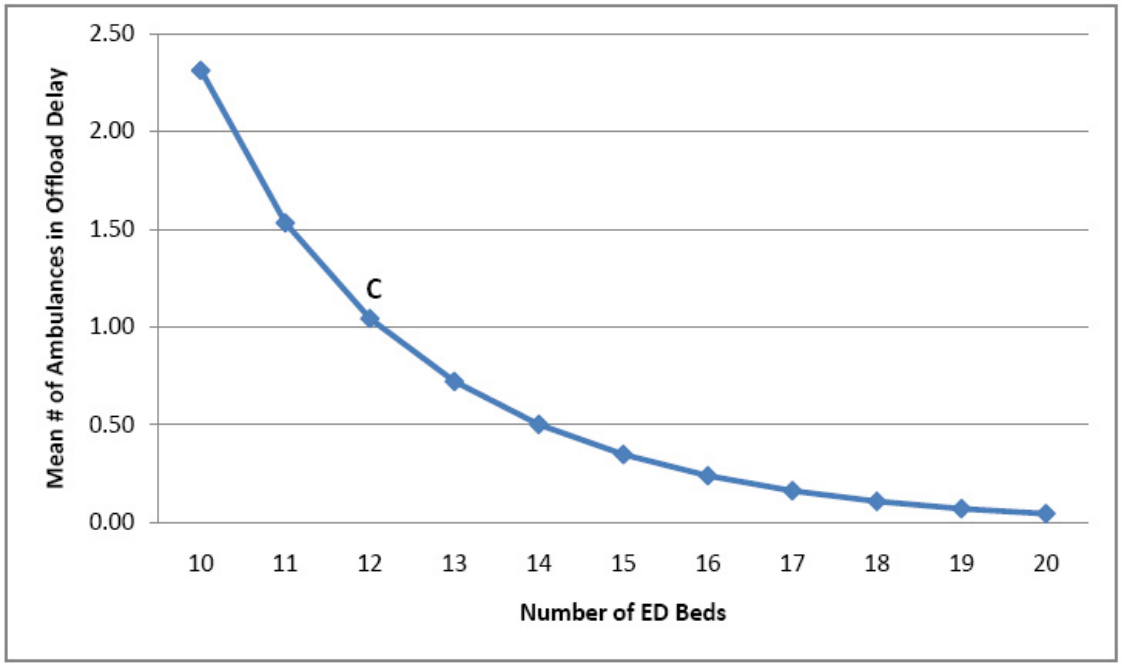

Figure 6.6: Number of Beds vs. Expected Number of Ambulances in Offload Delay

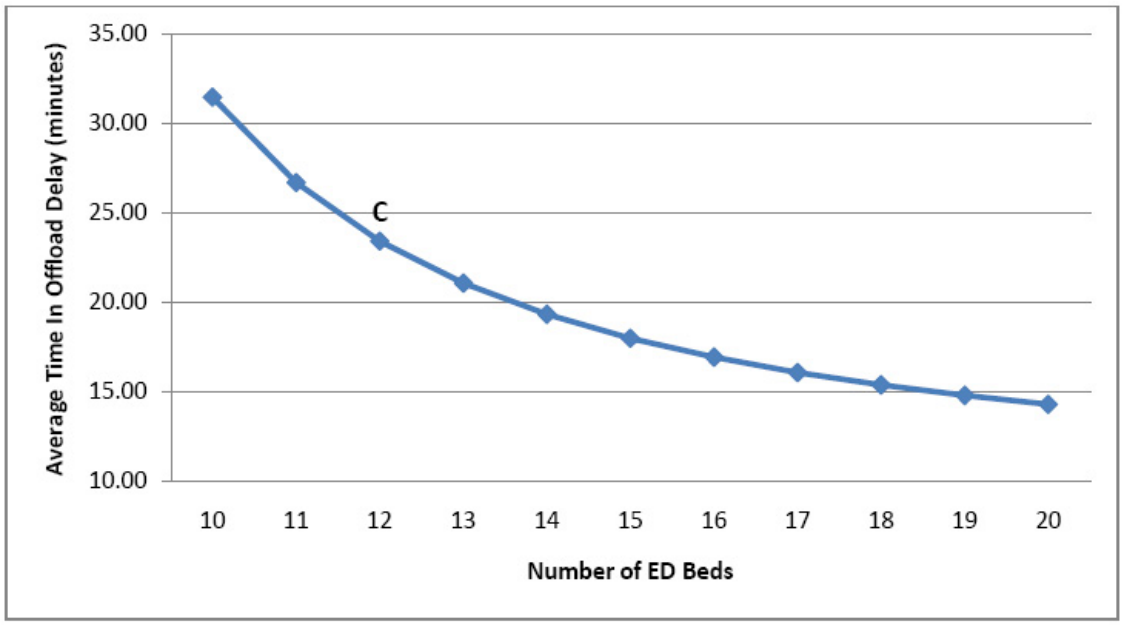

Figure 6.7: Number of Beds vs. Mean Time in Offload Delay

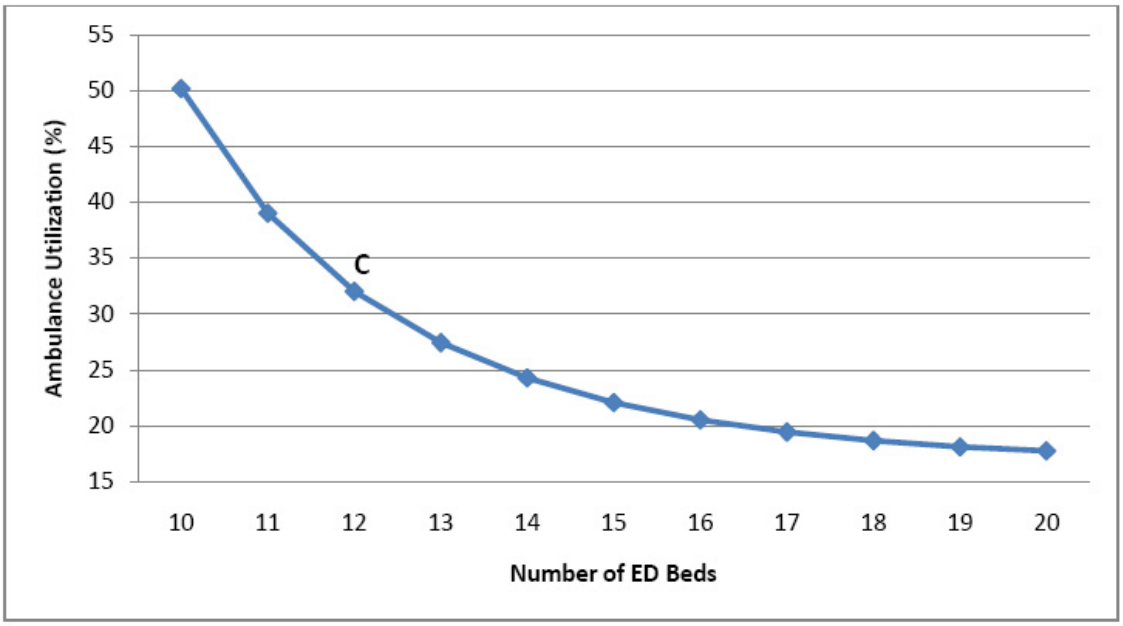

Figure 6.8: Number of Beds vs. Ambulance Utilization

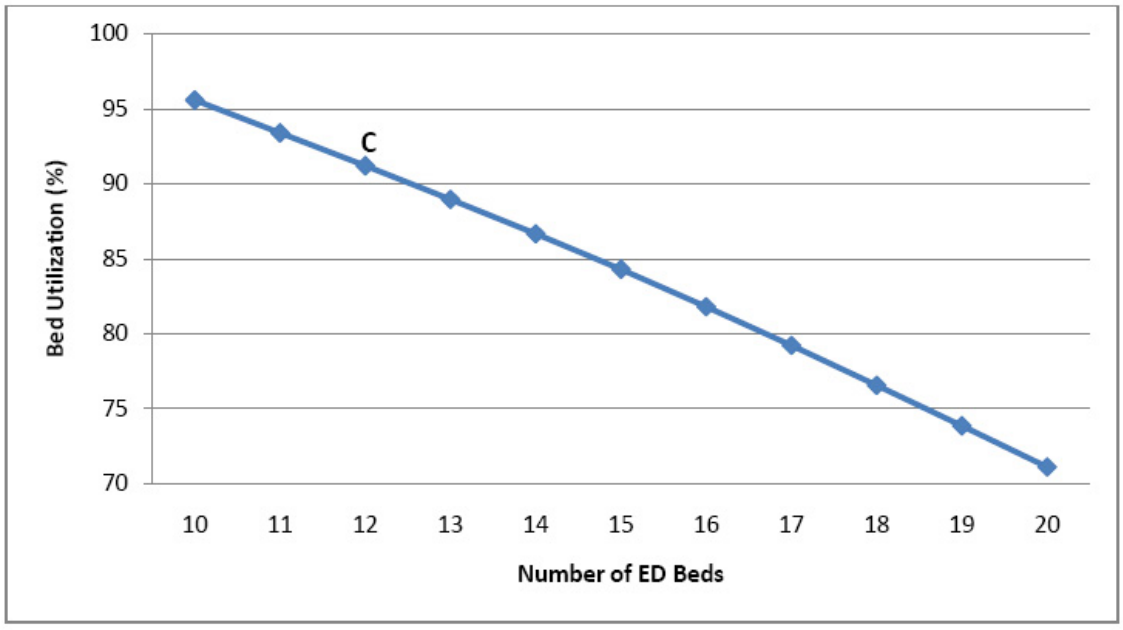

Figure 6.9: Number of Beds vs. ED Bed Utilization

By looking at the above figures we also see how all performance measures worsen as we decrease the number of ED beds to 10. In particular, the average number of ambulances in offload delay is increased to 2.3 and the average time in offload delay is now 31 minutes compared to 23 minutes before. It is interesting to note that how a small decrease in number of ED beds has a big impact on system performance. This clearly shows that the results are very sensitive to the number of beds.

As we saw in this section, increasing the number of beds would allow the ED of the GRH to provide treatment to more patients every hour and as a result we saw a successful improvement in system performance. Another way of achieving similar performance results is by shortening the ED patient treatment time (μ_1) which would have a similar effect as adding more beds.

6.3 Varying ED Treatment Time

According to our base case, it takes 6.75 hours for the ED of the GRH to treat a patient. In this section we will discuss the effect of reducing the ED treatment time by up to 1 hour in 15 minutes intervals on the same performance measures considered in the previous section. We will also show the effect of increasing the ED treatment time by up to 1 hour, since this situation is quite possible, considering the population and emergency call growth in the region.

The results for the mean number of patients waiting for ambulance, mean number of ambulances in offload delay, and average time in offload delay are shown in Figures 6.10-6.12. The average number of patients waiting for ambulance is approaching zero when the ED treatment time is reduced to 5.75 hours from 6.75 hours per patient. Similarly, the average number of ambulances in offload delay is reduced to 0.5 from 1, a 50% improvement, and the average time in offload delay is reduced to 19 minutes from roughly 23 minutes before. Again this is what we had expected due to the fact that having faster treatment times would increase the availability of ambulances to service more emergency calls. On the other hand, if treatment times worsen by up to 1 hour to 7.75 hours from 6.75 hours, The average number of ambulances in offload delay increases to 1.89 from 1.04 and the average time in offload delay rises to 29 minutes from 23 minutes. Other performance measures are similarly effected.

A similar pattern can be seen in ambulance and ED bed utilization (Figures 6.13 and 6.14). Reducing ED treatment time decreases ambulance and bed utilization to roughly 25% and 85% respectively. Whereas if treatment times worsen, ambulance and bed utilization rise to 95% and 44% from 91% and 32% respectively.

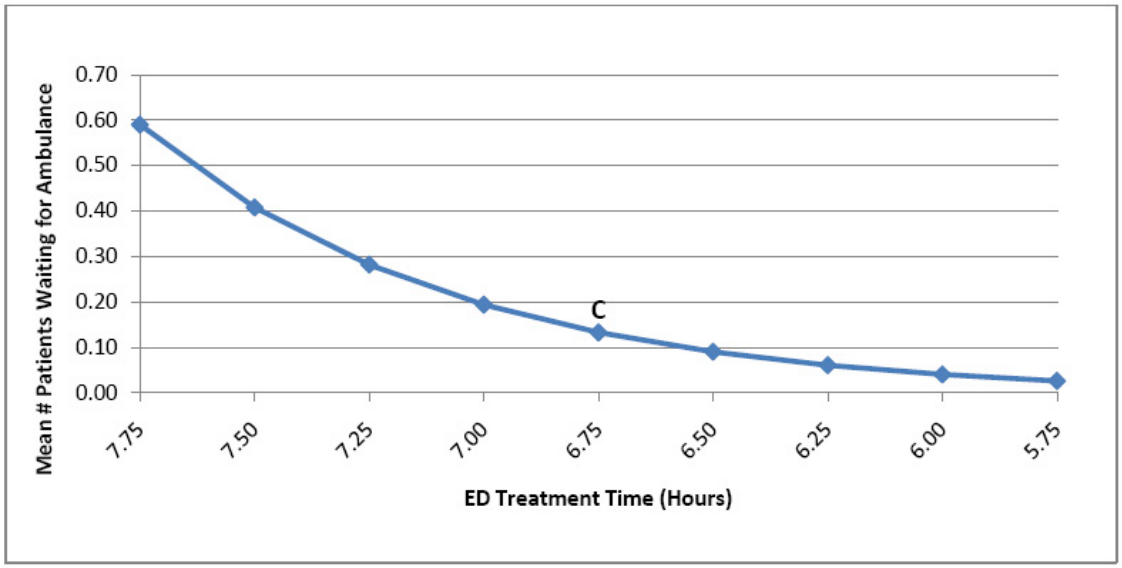

Figure 6.10: ED Treatment Time vs. Expected Number of Patients Waiting

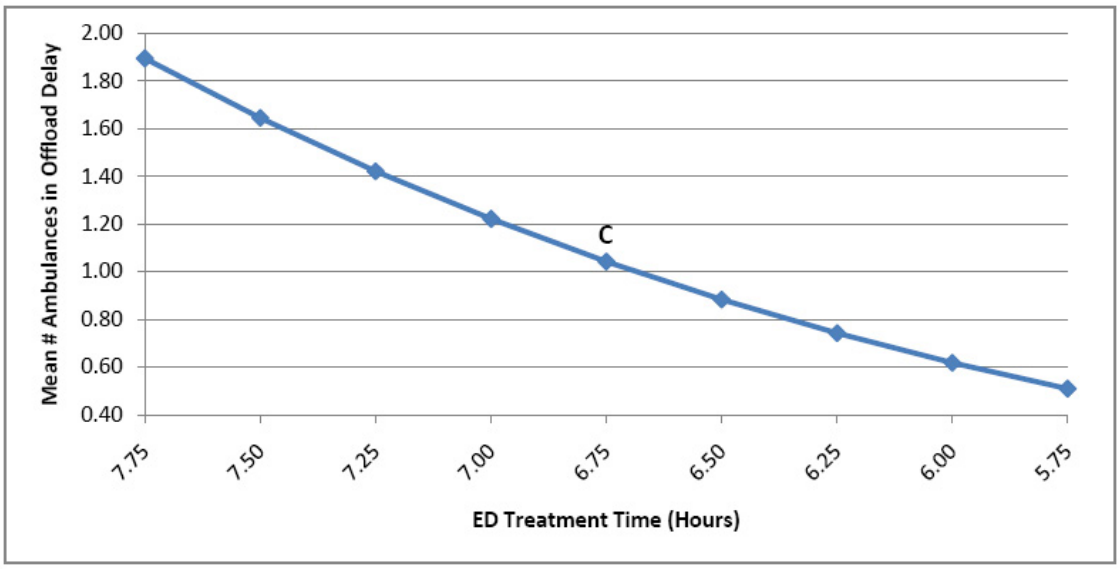

Figure 6.11: ED Treatment Time vs. Expected Number of Ambulances in Offload Delay

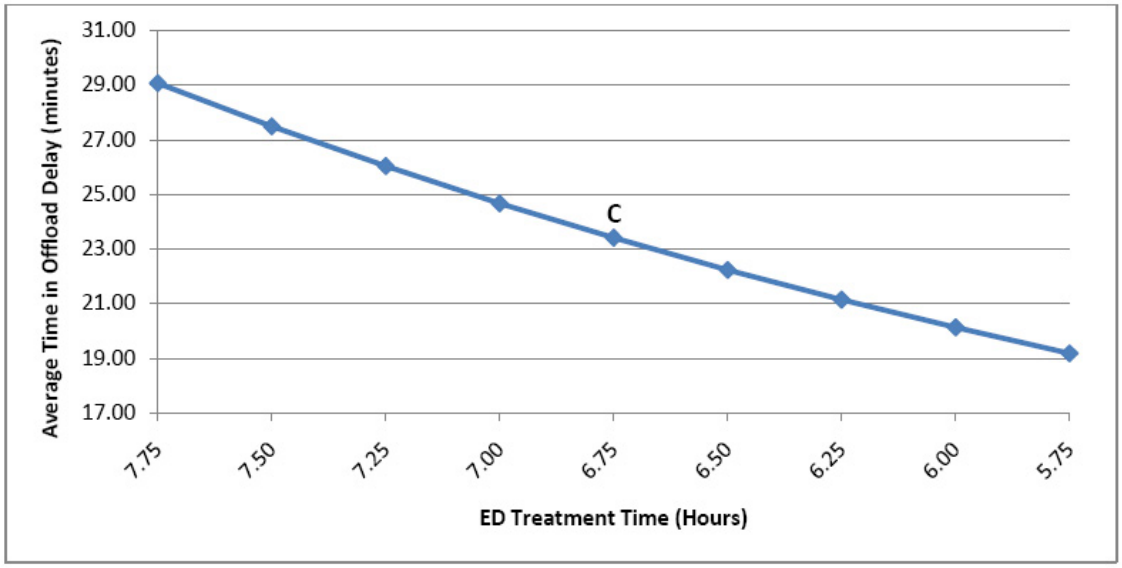

Figure 6.12: ED Treatment Time vs. Mean Time in Offload Delay

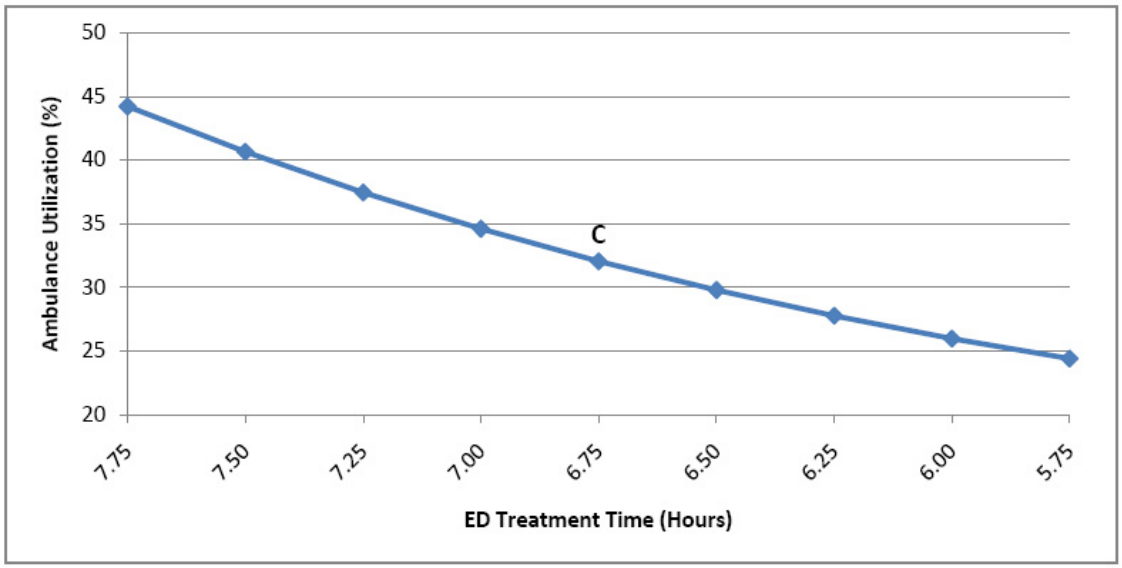

Figure 6.13: ED Treatment Time vs. Ambulance Utilization

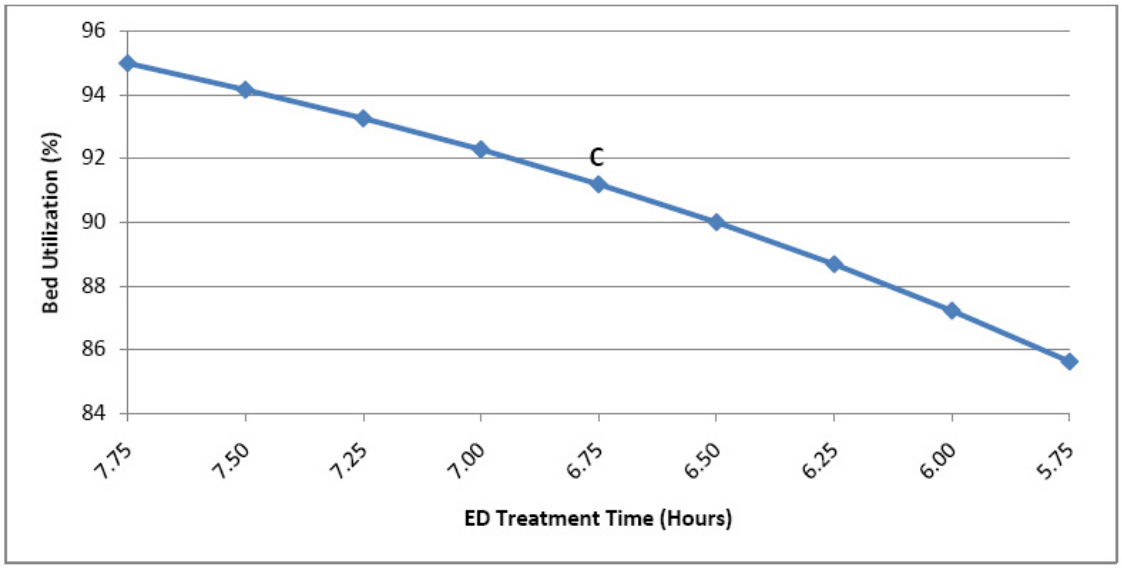

Figure 6.14: ED Treatment Time vs. ED Bed Utilization

As the results indicated, reducing ED treatment times produces similar results in comparison with adding extra beds in the ED. An interesting question to consider is by how much ED treatment times need to be reduced to, in order to achieve similar performance results as if we had added an extra bed. We will analyze this situation next.

6.4 ED expansion vs. ED Treatment Time Reduction

In the previous sections we showed how adding more beds in the ED or reducing the ED treatment time would improve the system performance. But we are interested to know which one is a better option in terms of costs and performance results. Several factors must be considered in this situation. Adding extra beds on a temporary basis or even permanently if possible seems to be a reasonable and easy way to improve system performance but it comes with a hefty cost. First, there has to be enough space available. Second and most importantly, more resources such as

equipment, nurses, and physicians are needed to accommodate this. On the other hand, reducing the ED treatment time is a possible alternative but it is limited and depends on the situation. It is possible that the reason why treatment times are high is because of the fact that there are no in-patient beds available in the hospital causing patients to stay in ED beds when they are ready to be transferred to in-patient beds. In this case it is possible to improve treatment times. Another possibility is technological advancements in medicine and medical equipment that could shorten service times. Also, if there are not enough staff resources such as nurses and physicians, staff addition could improve treatment times.

Perhaps a more cost effective way to achieve better performance results is to add certain number of beds while reducing treatment times at the same time. For example, instead of adding 2 extra beds in the ED, it could be more cost effective to add 1 extra bed and reduce the ED treatment times by say 15 minutes. This combination might achieve the same performance results as having 2 extra beds but might cost less. There are several possibilities that can be considered and our model can be used to analyze and compare these alternatives.

The first thing we are interested inn knowing by how much ED service times should be reduced, in order to achieve the same performance results as adding an extra bed. Given that the number of ED beds is N and the ED treatment time is μ_1, we want the overall ED treatment time with $N+1$ beds to be equal to the overall ED treatment time with N beds but improved service time. That is on average we want

$$(N+1)\mu_1 = N\mu_1\frac{1}{\alpha} \quad N > 0$$

where $0 < \alpha < 1$. Therefore, the new ED treatment rate $\mu_1^{'}$ corresponding to an extra bed is given by:

$$\mu_1^{'} = \frac{N\mu_1}{N+1} \quad N > 0 \tag{6.1}$$

Using the above relationship the following figure shows the tradeoff between the number of beds and the ED treatment time:

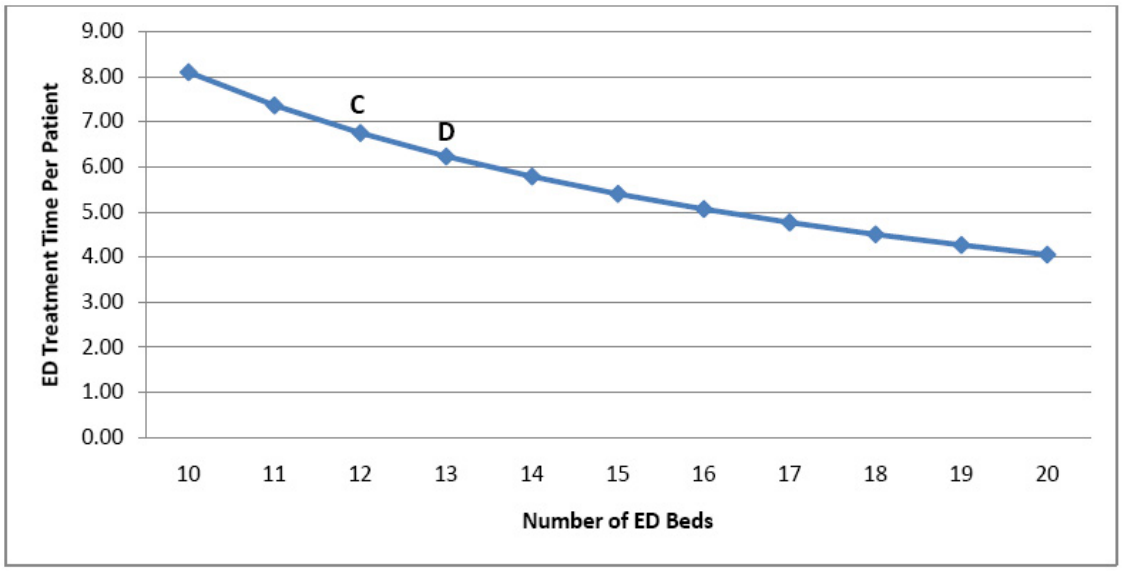

Figure 6.15: Tradeoff Curve

Point 'C' with 12 beds and 6.75 hours of treatment time per patient is the current ED operation point. The above figure shows that if the ED of the GRH can reduce its treatment time to roughly 6.25 hours per patient (point D) from 6.75 hours (point C) they can achieve the same performance results as if they had added an extra bed (13 beds). We used our model to confirm this and have constructed the following table:

	U_A (%)	U_B (%)	E(D) (minutes)	$E(P_w)$	$E(A_d)$
ED with 13 Beds	27.43	88.96	21.06	0.06	0.72
6.25 hours Treatment Time	27. 75	88.68	21.14	0.06	0. 74

As can be seen, the results are quite similar in both cases. A very simple cost analysis would indicate that if the cost of adding an extra bed is less than the cost

of reducing the ED treatment time to 6.25 hours, the GRH should consider adding an extra bed. So far we only analyzed the effect of hospital capacity and service times on system performance. In the next section we will focus on EMS capacity and analyze the effect of increasing and decreasing the number of ambulances on the same performance measures we have considered so far.

6.5 Varying the Number of Ambulances

In our base case we assumed that 7 out of 18 EMS ambulances in the region of Waterloo are dedicated to serve the emergency calls that result in a patient being transferred to the GRH. Let us see what happens to the system performance as we vary the number of ambulances. The first thing we expect to see is that when we increase the number of ambulances, the average number of patients waiting for ambulance and their average waiting time should decrease. The results shown in Figure 6.16 indeed shows this pattern. Again, the average waiting time for ambulance has the exact same shape as in Figure 6.16 since it is scaled by a factor of λ.

The results for the average number of ambulances in offload delay and the average time in offload delay are shown in Figures 6.17 and 6.18. It is interesting to note that the average number of ambulance in offload delay increases as we increase the number of ambulances. When the number of ambulances increase there will be more ambulances available to service emergency calls and hence more patients will be transferred to the hospital. But, the rate at which the ED of the GRH discharges patients is the same as before which causes more ambulances to experience offload delay. This clearly shows that hospital is the bottleneck in the system.

The average time in offload delay on the other hand decreases as we increase the number of ambulances since the offload delay time is spread over more number

of ambulances. Recall the average offload delay time formula, $E(D)$:

$$E(D) = \frac{E[A_d]}{N\tau} \cdot \Pr(\text{at least one ambulance in offload delay})$$

As we increase the number of ambulances, the $N\tau$ factor in the denominator of $E(D)$ increases faster than the $E[A_d]$ factor in the nominator causing the $E(D)$ to decrease.

The results for ambulance and ED bed utilization are shown in Figures 6.19 and 6.20. As we can see, increasing the number of ambulances will not have any effect on the ED bed utilization. It is important to note that this is not always the case and it depends on the input parameters such as the hospital's service rate and call arrival rate. For example, if the ED of a hospital has a utilization rate of 50% and the rate of call arrivals is high, increasing the number of ambulances would cause more patients to be transferred to the hospital and hence bed utilization rate is increased. But when the utilization rate is already high like in our case, increasing the number of ambulances will not have much effect on the ED bed utilization. However, the ambulance utilization rate always decreases as the number of ambulances increase as we can see in Figure 6.19. Recall the ambulance utilization formula, U_A:

$$U_A = \frac{E[A_b]}{N} \tag{6.2}$$

Again as we increase the number of ambulances, the denominator of U_A increases faster than the nominator and hence U_A decreases.

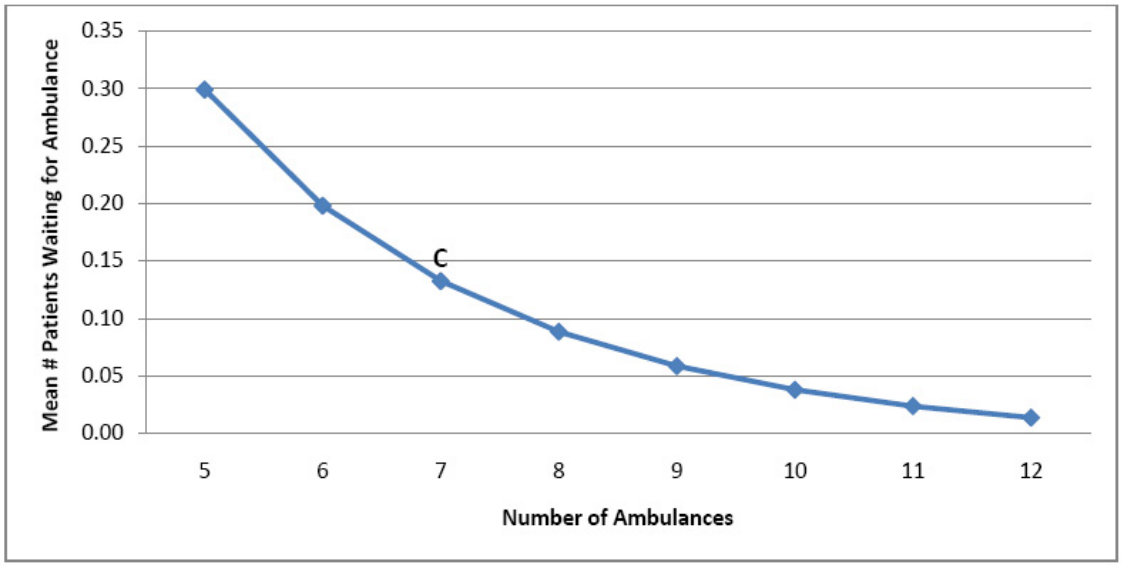

Figure 6.16: Number of Ambulances vs. Expected Number of Patients Waiting

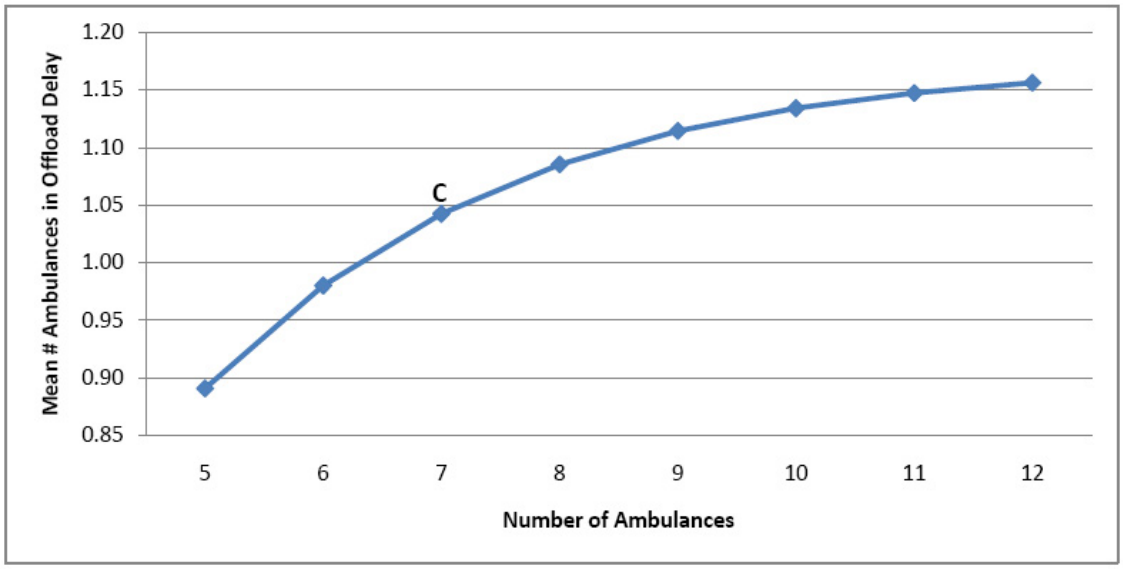

Figure 6.17: Number of Ambulances vs. Expected Number of Ambulances in Offload Delay

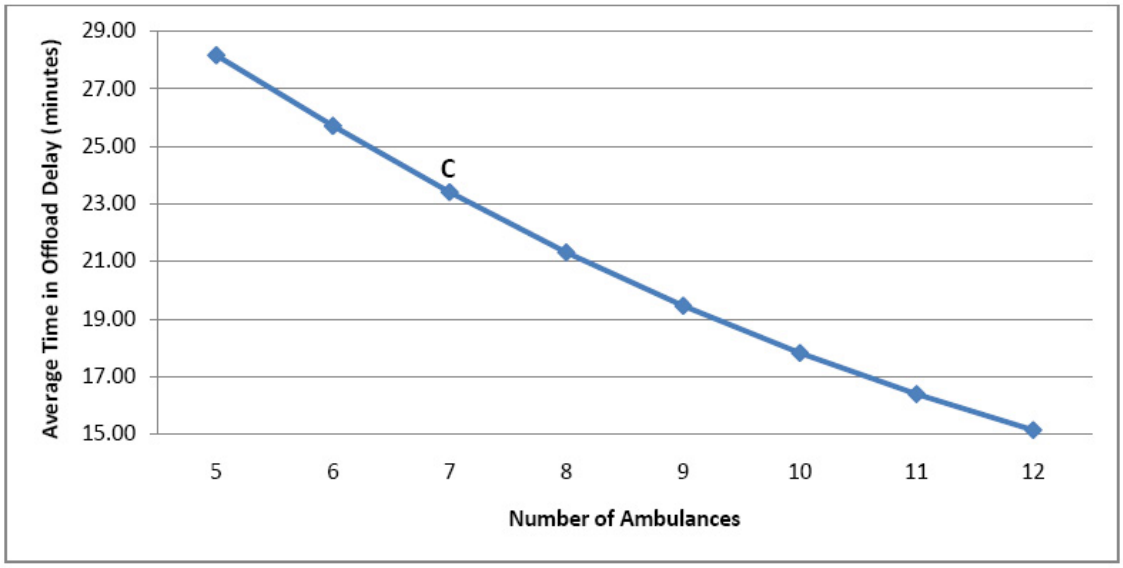

Figure 6.18: Number of Ambulances vs. Mean Time in Offload Delay

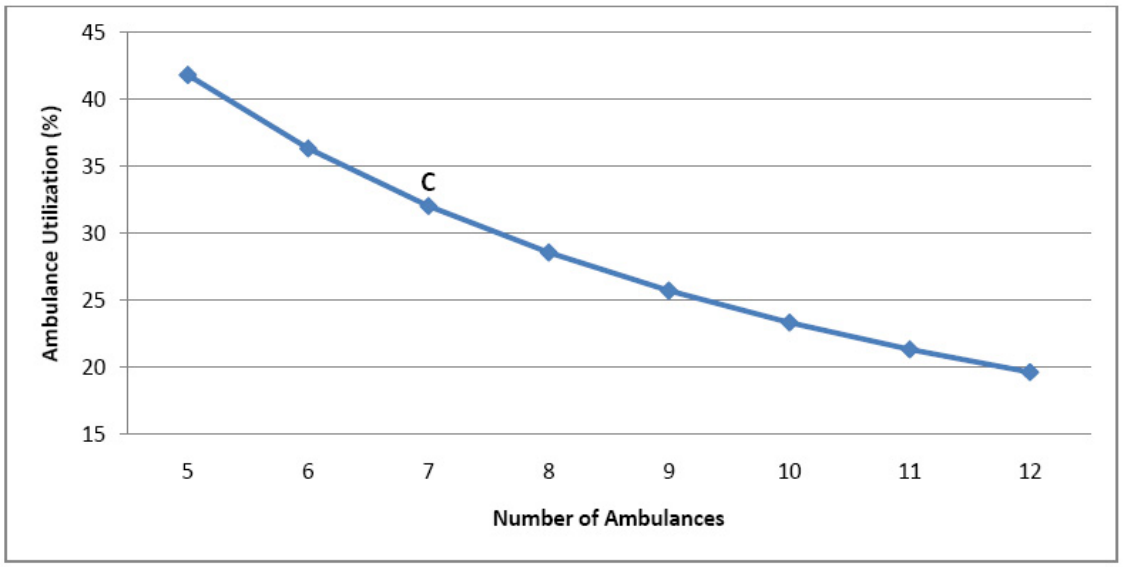

Figure 6.19: Number of Ambulances vs. Ambulance Utilization

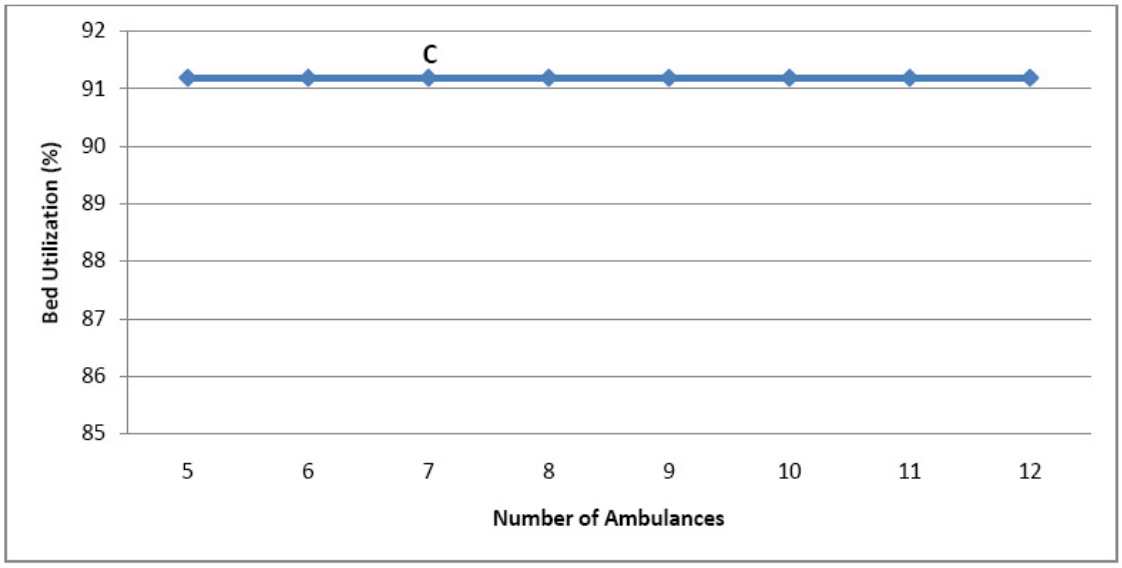

Figure 6.20: Number of Ambulances vs. ED Bed Utilization

In this section we analyzed the effect of the number of ambulances on various system performance measures and showed that hospital is the bottleneck in the system. In the next section, we will combine all of our analysis so far in this chapter and discuss the tradeoffs between the ED of the GRH and the EMS of Waterloo.

6.6 Tradeoffs

In Figure 6.15 we showed the tradeoffs between emergency beds and ED treatment time. It was shown that in order to achieve the same performance results as adding an extra bed, ED treatment time has to be reduced by a certain amount. In this section we will perform a similar analysis as in section 6.4, but now we will consider ambulances as well. That is, we will show the tradeoffs between emergency beds, EMS ambulances, and the ED treatment time. We have used our model to analyze this situation and have constructed the following table:

The above table shows the tradeoffs based on our base case between emergency

Number of Beds	Performance Measure	Number of Ambulances	ED Treatment Time (hours)
$12 \rightarrow 13$	P_w	$7 \rightarrow 9$	$6.75 \rightarrow 6.25$
	W	$7 \rightarrow 9$	$6.75 \rightarrow 6.25$
	D	$7 \rightarrow 9$	$6.75 \rightarrow 6.2$
	U_A	$7 \rightarrow 9$	$6.75 \rightarrow 6.2$
	U_B	Not Possible	$6.75 \rightarrow 6.25$

Table 6.3: Tradeoffs Between ED beds, EMS Ambulances, and the ED Treatment Time

beds, ambulances, and the ED treatment time. Let us demonstrate how the above table can be used. Consider the row which contains the P_w performance measure, the average number of patients waiting for ambulance. According to the table, instead of adding an extra bed in the ED, it is possible to add 2 more ambulances or reduce the ED treatment to 6.25 hours to achieve the same performance results in terms of $E(P_w)$. If the performance measure under consideration is the ambulance utilization (U_A), again we can add 2 ambulances or reduce the ED treatment time to 6.2 hours to achieve the same ambulance utilization result as if we had added an extra bed. Therefore, results depend on the performance measure under consideration. Note that it is not possible to achieve the same ED bed utilization result of adding an extra bed by adding more ambulances. This is due to the fact that in our case adding more ambulances did not have any effect on ED bed utilization, whereas adding an extra bed reduces the ED bed utilization. So by looking at the results of Table 6.3 we can see that adding an extra bed is equivalent to adding 2 more ambulances or the ED treatment time of 6.25 hours. That is, we can get a similar performance improvement by either adding an extra bed or adding two more ambulances or reduce the ED treatment time to 6.25 hours. Again we can see that hospital is the bottleneck in the system. It is important to note that the

result presented in table 6.3 only hold for the case where we are expanding the ED of the GRH from 12 beds to 13 beds. The results would have been different if were to add 2 or more extra beds. But all those scenarios can easily be analyzed by our model.

There are other sensitivity analyses that can be done in a similar way we did in this chapter. Those would include analyzing the effect of emergency call growth and changing the rate of outside patient arrivals on system performance. Another possible scenario that can be considered is when it is possible to have $\mu_2 > 0$. That is ambulances have the ability to treat patients in the ambulance and directly transfer them to in-patient beds. Again, our model can easily be used to analyze these possible scenarios. In the next chapter we conclude our work and set some future directions for our research.

Chapter 7

Conclusions and Future Research

Long waiting times and congestion in the emergency department of hospitals have caused serious problems in recent years in Canada and several other countries. Perhaps one of the most known causes of this problem is the inability of hospitals to accept patients from regional emergency medical services (EMS) ambulances in a timely manner. When a patient arrives at the ED via ambulance, the hospital must accept transfer of care of the patient before the paramedic staff can remove the patient from the ambulance. However, many hospital are experiencing bed and staff shortages due to increased demand of emergency services, and as a result ambulances must spend hours waiting at the hospital while providing care until an emergency bed becomes available. This situation is well-known as "ambulance offload delay" and it is not a simple issue to resolve.

Ambulance offload delays have a significant impact on EMS response times due to the fact that they affect ambulance availability to respond to emergency calls. As more ambulances experience offload delay, less ambulances are available to service emergency calls and hence patients have to wait longer to be taken care of by ambulance. Offload delays financially cost the EMS provider as well, since extra resources are required to provide quality service. Even though offload

delays considerably affect EMS performance and costs, but patients, paramedics and emergency department staff are the ones that are mostly hurt in this situation due to the mental and physical stress and pressure of the offload delays.

Although there is a long history of research in both EMS and ED planning and operation, except several reports on offload delays there are no models specifically dealing with the problem. In a situation such as offload delays one has to focus on the interaction between the EMS provider and the ED of the hospital, but most of the proposed models only focus on either EMS or ED. In this thesis we developed an analytical model that allowed us to extensively analyze and explore the situation of ambulance offload delays. We constructed a queuing system representing the interaction between these units and modeled the behavior of the system in a continuous time Markov chain framework. The matrix geometric method was used to numerically compute the steady state probability distribution for the Markov chain model developed. We computed the steady state probability distribution of our system and used it to compute the distribution of the following set of random variables defined on the system:

1. Number of patients waiting for ambulance
2. Number of ED beds occupied
3. Number of ambulances in transit
4. Number of ambulances experiencing offload delay
5. Number of ambulances busy

The distribution of the above random variables allowed us to compute more aggregate performance measures such as the expected number of ambulances in offload delay, mean number of beds occupied, expected ambulance and bed utilization,

average waiting time for ambulance, and average time in offload delay. The effect of varying some of the model input parameters such as the number of beds, ED treatment time, and the number of ambulances, on various system performance measures were analyzed extensively. We showed that adding more beds in the ED and reducing ED treatment time have a positive effect on system performance measures. We also showed that increasing the number of ambulances would increase the average number of ambulances in offload delay which indicates that the hospital is clearly the bottleneck in the system. Tradeoffs between adding extra emergency beds, adding more ambulances, and reducing ED treatment time were also discussed and analyzed.

The model we proposed and developed in this thesis can easily be extended to be more realistic. In particular we may consider the following enhancements in near future:

- A more complicated set of performance measures can be calculated by computing the distribution of offload delay time and patient waiting time. In our work we only computed the average time in offload delay and the average waiting time for ambulance. Calculating the distribution of offload delay time D and patient waiting time W would allow us to compute other useful performance measures such as the probability that an ambulance will be in offload delay for more than 'x' minutes before it becomes available, or the probability that a patient has to wait for more than 'y' minutes for an ambulance.

- In our work we assumed that outside patients are lost if there are no emergency beds available at the ED of the hospital. In future research we will relax this assumption and allow outside patients to wait for an emergency bed when there are no beds available. An additional state variable needs to be introduced in our model to keep track of the number of outside patients

waiting for a bed at any time t. This would also change the generator matrix and will likely introduce further complications in modeling and performance measure calculations.

- The total number of ambulances was assumed to be fixed in our work and ambulances were dedicated to serve a specific hospital only. In future research we may consider the total number of ambulances as a variable with its initial value to be the total number of ambulances for the region and allow for the fact that ambulances can service other hospitals in the region as well. In this case, when an emergency call arrives which requires a patient to be transferred to the hospital under consideration, an ambulance is pulled and the number of ambulances is reduced by one. But at the same time we may introduce a rate that would cause the number of ambulances to decrease by one every time an ambulance is needed to service another hospital.

- In our work we assumed that all emergency calls result in patient being transferred to the hospital. In reality, there are situations where an ambulance is sent to the scene and provided care but there was no need to transfer the patient to the hospital. In such a system there are emergency calls and non-emergency calls. In this case we would likely to have another input parameter which corresponds to non-emergency call arrivals. An even more realistic system would include calls that require an ambulance to be sent to another region to provide coverage. This case can be handled in a similar way as non-emergency call as well.

In conclusion we believe that all this work will lead to a better understanding of the ambulance offload delay problem.

Appendix A

Matlab Code

A.1 The Main Execution File

```
%% Parameter initialization %%
% Total number of ED beds
global M
M = 1;
% Total number of ambulances
global N
N = 1;
global limit
% Emergency call arrival rate
global lambda
lambda = 2;
% Hospital service rate
global mu1
mu1 = 5;
% Ambulance service rate
```

```
global mu2
mu2 = 1;
% Ambulance transit rate
global theta
theta = 3;
% Outside patient arrival rate
global delta
delta = 2;

%% A0 matrix setup %%
A0 = Generate('A0');
%% A2 matrix function called%%
A2 = Generate('A2');
%% A1 matrix function called %%
A1 = Generate('A1');
A = A1+A2+A0;
%% Checking for the stability condition
if(stabilitycond(A,A0,A2) == 1)
    Rmatrix = RecR();
    P = Xmatrix(Rmatrix);
    L = limit;
    P1 = PnumQ(P);
    P2 = PnumTransit(P);
    P3 = PnumBeds(P);
    P4 = PnumStock(P);
    P5 = PnumAmbBusy(P);
%% The distribution of the random variables presented in Table 5.1
```

```
    P1
    P2
    P3
    P4
    P5
%% Computing the expected value of the above random variables P1-P5
    Ex = zeros(5,1);
    for i=1:size(P1,1)
    Ex(1,1)=(i-1)*P1(i,1)+Ex(1,1);
    end;
    for i=1:size(P2,1)
    Ex(2,1)=(i-1)*P2(i,1)+Ex(2,1);
    end;
    for i=1:size(P3,1)
    Ex(3,1)=(i-1)*P3(i,1)+Ex(3,1);
    end;
    for i=1:size(P4,1)
    Ex(4,1)=(i-1)*P4(i,1)+Ex(4,1);
    end;
    for i=1:size(P5,1)
    Ex(5,1)=(i-1)*P5(i,1)+Ex(5,1);
    end;
    Ex
%% Average waiting time for ambulance and average time in offload delay
    ((Ex(4,1)/(N*theta))*60)/(1-P4(1,1))
    (Ex(1,1)/lambda)*60
%% Ambulance and bed utilization
```

```
    (Ex(3,1)/M)
    (Ex(5,1)/N)
else
    message = 'The system is not stable we the given set of parameters';
    message
end;
```

A.2 The Infinitesimal Generator Component Matrices

```
%% B Matrix Setup %%
function B = Bmatrix(n)
global M
global N
global lambda
global mu1
global mu2
global theta
global delta
if(n > N-1)
    B = A1matrix();
else
    B1 = zeros(M+1,M+1);
    for k=0:M-1
        B1(k+1,k+1) = -(lambda + k*mu1 + delta);
    end;
```

```
    B1(M+1,M+1)= -(lambda + M*mu1);
    for h=1:M
        B1(h+1,h)=h*mu1;
    end;
    B2=zeros(M+1,N);
    B12 = horzcat(B1,B2);
    B3 = zeros(N,M+1);
    B3(1,M+1) = M*mu1+mu2;
    B4 = zeros(N,N);
    for s=1:N
        B4(s,s) = -(lambda + M*mu1 + s*mu2);
    end;
    for t=1:N-1
        B4(t+1,t)=M*mu1+(t+1)*mu2;
    end;
    B34 = horzcat(B3,B4);
    B = vertcat(B12,B34);
    B(M+N+1,M+N+1) = beta(0,M,N);
    for i = 0:n-1
        B(M+N-i,M+N-i)=B(M+N-i,M+N-i)-(i+1)*theta;
    end;
    for i = n:M+N-1
        B(M+N-i,M+N-i)=B(M+N-i,M+N-i)-n*theta;
    end;
end;
for h=1:M
    B(h,h+1)=delta;
```

```
end;



%% C Matrix Setup %%
function c = Cmatrix(n)
global M
global N
global theta
if(n >= N)
    c = A2matrix();
else
    a = (n*theta)*eye(M+N,M+N);
    c = vertcat(horzcat(zeros(M+N,1),a),zeros(1,M+N+1));
    for i = 0:n-1
        c(M+N-i,M+N+1-i)=(i+1)*theta;
    end;
end;



%% This function generates an A1, A2, and A0 matrix
function matrix = Generate(m)
if(m == 'A1')
    matrix = A1matrix();
elseif(m == 'A2')
    matrix = A2matrix();
elseif(m == 'A0')
    matrix = A0matrix();
```

```
end;


%% A0 Matrix Setup %%
function A0 = A0matrix()
global M
global N
global lambda
I = eye(M+N+1);
A0 = lambda * I;


%% A1 Matrix Setup %%
function A1 = A1matrix()
global M
global N
global lambda
global mu1
global mu2
global theta
global delta
B1 = zeros(M+1,M+1);
for k=0:M
    B1(k+1,k+1)=beta(N,k,0)-delta;
end;
for h=1:M
    B1(h+1,h)=h*mu1;
```

```
    B1(h,h+1)=delta;
end;
B1(M+1,M+1)=beta(N,M,0);
B2=zeros(M+1,N);
B12 = horzcat(B1,B2);
B3 = zeros(N,M+1);
B3(1,M+1) = M*mu1+mu2;
B4 = zeros(N,N);
for s=1:N
    B4(s,s)=beta(N-s,M,s);
end;
for t=1:N-1
    B4(t+1,t)=M*mu1+(t+1)*mu2;
end;
B34 = horzcat(B3,B4);
A1 = vertcat(B12,B34);



%% A2 Matrix Setup %%
function A2 = A2matrix()
global M
global N
global theta
b1 = (N*theta) * eye(M+1,M+1);
b2 = zeros(M+1,N-1);
b12 = horzcat(b1,b2);
b3 = zeros(N,M+1);
```

```
b4 = zeros(N,N-1);
for i=1:N-1
    b4(i,i)= (N-i)*theta;
end;
b34 = horzcat(b3,b4);
A2 = horzcat(zeros(M+N+1,1),vertcat(b12,b34));



%% Beta function used in A1 and B matrices %%
function b = beta(n,i,j)
global lambda
global mu1
global mu2
global theta
b=-(lambda + n*theta + i*mu1 + j*mu2);
```

A.3 Stability Condition Check

```
%% The following function returns 1 if the system is stable with the
%% given set of parameters and returns 0 otherwise.
function s = stabilitycond(A,A0,A2)
global M
global N
%% calculating the f vecto by solving the system of equations represented
%% by 53 and 54.
e = ones(M+N+1,1);
xx = A;
```

```
xx(:,M+N+1) = 1;
vec=zeros(1,M+N+1);
vec(M+N+1)=1;
s = vec*inv(xx);
%% left hand side of inequality 55
exp1 = s*A0*e;
%% Right hand side of inequality 55
exp2 = s*A2*e;
%%checking the inequality 55 and returning 1 if it holds and 0 otherwise.
if(exp1 < exp2)
    s = 1;
else
    s = 0;
end;
```

A.4 Computing the Rate Matrices

Simple Iterative Aprroach

```
% Recursive equation used to calculate the R matrices. See Equation 2.8 on
% page 501.
function R = RecR()
global N
global M
A0 = Generate('A0');
A1 = Generate('A1');
A2 = Generate('A2');
```

```
R0 = zeros(M+N+1,M+N+1);
eps1 = 1.0;
eps = 0.000001;
while (eps1 > eps)
    R1 = -(A0 + R0*R0*A2)*inv(A1);
    eps1 = max(max(abs(R1-R0)));
    R0 = R1;
end;
s = R1;
R(:,:,N) = s;
for i = N-1:-1:1
    R(:,:,i) = -A0*inv(Bmatrix(i)+R(:,:,i+1)*Cmatrix(i+1));
end;
```

The Logarithmic Reduction Algorithm

```
function R = R();
global M
global N
i = 0;
A0 = Generate('A0');
A1 = Generate('A1');
A2 = Generate('A2');
B0 = inv(-A1)*A0;
B2 = inv(-A1)*A2;
S = B2;
P = B0;
```

```
I = eye(M+N+1);
eps1 = 1.0;
eps = 0.000001;
while (eps1 > eps)
    i = i+1;
    A11 = B0*B2 + B2*B0;
    A00 = B0^2;
    A22 = B2^2;
    B0 = inv(I-A11)*A00;
    B2 = inv(I-A11)*A22;
    S = S+P*B2;
    P = P*B0;
    eps1 = max(abs(ones(M+N+1,1)-S*ones(M+N+1,1)));
end;
G = S;
U = A1 + A0*G;
R = A0*inv(-U);
R(:,:,N) = R;
for i = N-1:-1:1
    R(:,:,i) = -A0*inv(Bmatrix(i)+R(:,:,i+1)*Cmatrix(i+1));
end;
```

A.5 Calculating the Steady State Probability Distribution

```
%% This function calculates the steady state probabilities
function X = Xmatrix(R)
```

```
global N
global M
global limit
limit = M+N;
eps1 = 0;
eps = 0.99999;
while(eps > eps1)
    xx = Bmatrix(0) + R(:,:,1)* Cmatrix(1);
    xx(:,M+N+1) = 1;
    vec = zeros(1,M+N+1);
    vec(M+N+1)=1;
    X0 = vec*inv(xx);
    I = eye(M+N+1);
    NM = X0 *normcond(N,R)*ones(M+N+1,1);
    X0 = X0/NM;
    X = zeros(limit,M+N+1);
    for j = 1:M+N+1
        X(1,j)=X0(j);
    end;
    for i = 1:limit
        if(i <= N)
            X1 = X0 * R(:,:,i);
            for j = 1:M+N+1
                X(i+1,j) = X1(j);
            end;
            X0 = X1;
        else
```

```
            X1 = X0 * R(:,:,N);
            for j = 1:M+N+1
                X(i+1,j) = X1(j);
            end;
            X0 = X1;
        end;
    end;
    eps1 = sum(X*ones(M+N+1,1));
    limit = limit + 1;
end;


%% The following function computes the normalization condition 4.17
function NM = normcond(L,R)
global M
global N
if(L == 0)
    NM = R(:,:,1)*inv(I - R(:,:,N));
else
    I = eye(M+N+1);
    E = eye(M+N+1);
    NM1 = zeros(M+N+1,M+N+1);
    for k = 0:L
        for m = 0:k-1
            if(m <= N-1)
                R1 = R(:,:,m+1);
                E = E*R1;
```

```
            else
                R1 = R(:,:,N);
                E = E*R1;
            end;
        end;
        NM1 = E + NM1;
        E = eye(M+N+1);
    end;
    A = inv(I - R(:,:,N));
    NM2 = eye(M+N+1);
    for m = 0:L
        if(m <= N-1)
            R2 = R(:,:,m+1);
            NM2 = NM2*R2;
        else
            R2 = R(:,:,N);
            NM2 = NM2*R2;
        end;
    end;
    NM2 = NM2 * A;
end;
NM = NM1 + NM2;
```

A.6 Computing the Distribution of Performance Measures

```
%% The following function computes the distribution of the
```

```
number of patients waiting for ambulance
function P_Q = PnumQ(P)
global N
global M
global limit
L = limit;
if(L<=N)
    P_Q=1;
else
Pmod = zeros(L-N,M+N+1);
P_Q = zeros(L-N,1);
for i = 2:L-N
    for j = 1:M+1
        Pmod(i,j)=P(i+N,j);
    end;
end;
for i = 2:L-N
    k = 1;
    for j = M+2:M+N+1
        Pmod(i,j)=P(i+N-k,j);
        k = k + 1;
    end;
end;

for j = 1:M+1
    for k = 1:N+1
        Pmod(1,j) = Pmod(1,j) + P(k,j);
```

```
    end;
end;
for j = M+N:M+N+1
    K = N;
    for i = 1:N
        Pmod(1,j)= Pmod(1,j) + P(i,j);
    end;
    k = k - 1;
end;
for i = 2:L-N
    P_Q(i,1) = sum(Pmod(i,:));
end;
P_Q(1,1) = sum(sum(P)) - sum(P_Q);
end;



%% The following function computes the distribution of
the number of ED beds occupied
function P_B = PnumBeds(P)
global M
global N
P_B = zeros(M+1,1);
for i = 1:M
    P_B(i,1) = sum(P(:,i));
end;
P_B(M+1,1) = sum(sum(P)) - sum(P_B);
```

```
%% The following function computes the distribution of
the number of ambulances in transit
function P_transit = PnumTransit(P)
global M
global N
global limit
L = limit;
P_transit = zeros(N+1,1);
j = M+N+1;
for i = 1:N
    for r = j:-1:1
        P_transit(i,1) = P_transit(i,1) + P(i,r);
    end;
    for c = i+1:L
        P_transit(i,1) = P_transit(i,1) + P(c,j);
    end;
    j = j - 1;
end;
P_transit(N+1,1) = sum(sum(P)) - sum(P_transit);


%% The following function computes the distribution of
the number of ambulances in offload delay
function P_stock = PnumStock(P)
global M
global N
```

```
P_stock = zeros(N+1,1);
for i = 1:M+1
    P_stock(1,1) = P_stock(1,1) + sum(P(:,i));
end;
for j = 2:N+1
    P_stock(j,1) = sum(P(:,M+j));
end;

%% The following function computes the distribution of
the number of ambulances busy
function P_ambbusy = PnumAmbBusy(P)
global M
global N
global limit
L = limit;
P_ambbusy = zeros(N+1,1);
P_ambbusy(1,1) = sum(P(1,1:M+1));
k = M+2;
for i = 2:N
    P_ambbusy(i,1) = sum(P(i,1:M+1));
    for j = i-1:-1:1
        P_ambbusy(i,1) = P_ambbusy(i,1) + P(j,k);
        k = k + 1;
    end;
    k = M+2;
end;
P_ambbusy(N+1,1) = sum(sum(P)) - sum(P_ambbusy);
```

Appendix B

Proof of the Normalization Condition 4.17

In this appendix we will prove the relationship 4.17:

$$\underline{\pi}e = \pi_0[\sum_{i=0}^{N-1}\prod_{j=0}^{i-1} R_j + \prod_{j=0}^{N-1} R_j(I-R)^{-1}]e = 1 \tag{B.1}$$

We will start with the left hand side of B.1 and expand it:

$$\begin{aligned} \underline{\pi}e &= \sum_{i=0}^{limit} \pi_i\, e \\ &= \sum_{i=0}^{N-1} \pi_i\, e + \sum_{i=N}^{\infty} \pi_i\, e \end{aligned} \tag{B.2}$$

Using equations 4.4 and 4.8, we can further expand B.2:

$$\begin{aligned} B.2 &= \sum_{i=0}^{N-1}\pi_0 \prod_{j=0}^{i-1} R_j\, e + \sum_{i=N}^{\infty} R^{i-N+1}\pi_{N-1}\, e \\ &= \sum_{i=0}^{N-1}\pi_0 \prod_{j=0}^{i-1} R_j\, e + \sum_{i=N}^{\infty} R^{i-N+1}\pi_0 \prod_{j=0}^{N-2} R_j\, e \\ &= \pi_0 [\sum_{i=0}^{N-1} \prod_{j=0}^{i-1} R_j\, e + \sum_{i=N}^{\infty} R^{i-N+1} \prod_{j=0}^{N-2} R_j]\, e \end{aligned} \tag{B.3}$$

We can simplify B.4 using the fact that $\sum_{i=N}^{\infty} R^{i-N+1} = R(I-R)^{-1}$:

$$\begin{aligned} B.3 &= \pi_0 [\sum_{i=0}^{N-1} \prod_{j=0}^{i-1} R_j\, e + R(I-R)^{-1} \prod_{j=0}^{N-2} R_j]\, e \\ &= \pi_0 [\sum_{i=0}^{N-1} \prod_{j=0}^{i-1} R_j\, e + \prod_{j=0}^{N-1} R_j\, (I-R)^{-1}]\, e \\ &= 1 \end{aligned}$$

Appendix C

Simulation Model

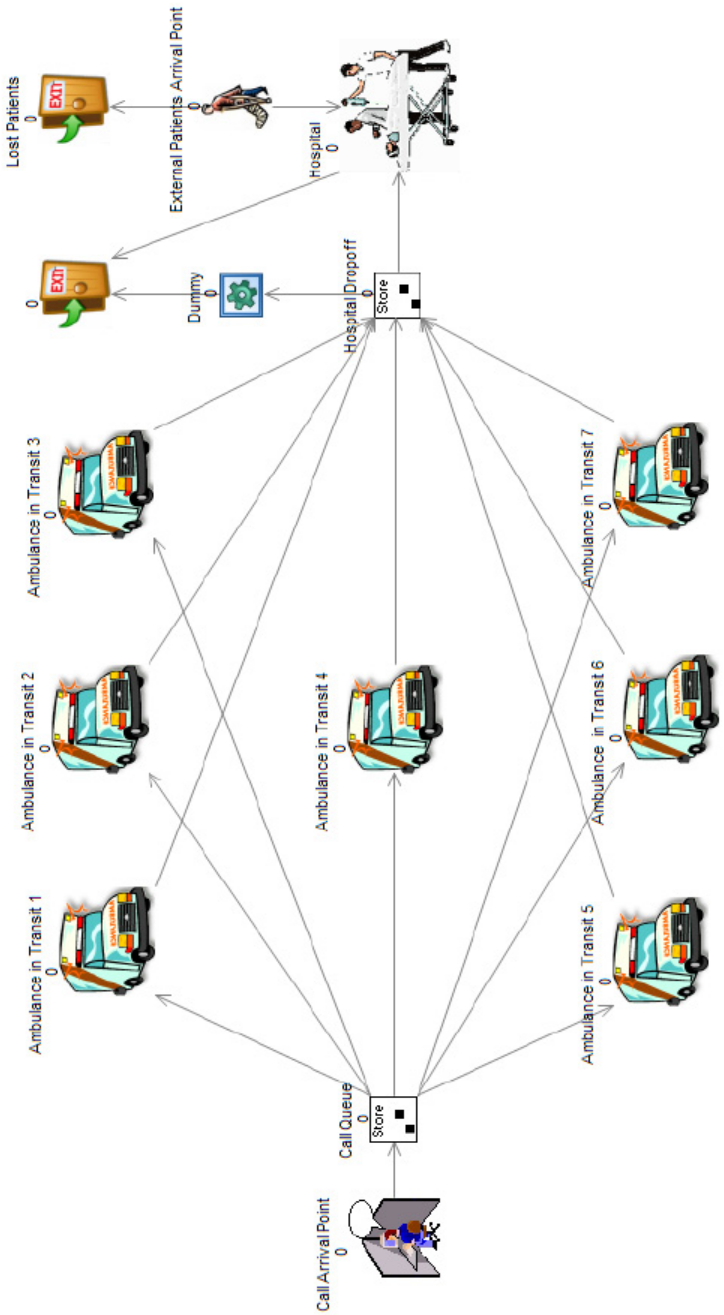

Figure C.1: Snapshot of the Simulation Model

References

[1] http://www.medindia.net/news/Toronto-Patients-Die-Waiting-for-Emergency-Care-32814-1.htm, February 2008. 4

[2] Oecd health data 2008: How does canada compare, 2008. 1

[3] D.J. Bradley and Martin J.B. Continuous personnel scheduling algorithms: A literature review. *Journal of the Society for Health Systems*, 2(2):8–23, 1991. 14

[4] L. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi birth and death processes. *Stochastic Models*, 11(3):497–525, 1995. 49, 52

[5] L. Brotcorne, L. Laporte, and Frederic. S. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003. 9

[6] S. Budge, E. Erkut, and A. Ingolfsson. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. 2007. manuscript, 12 pages. 13

[7] J.A. Buzacott and J.G. Shanthikumar. Design of manufacturing systems using queueing models. *Queuing Systems*, 12:135–214, 1992. 18

[8] J.L. Carroll, A. Van De Liefwoort, and L. Lipsky. Solution of m/g/1//n-type loops with extensions to m/g/1 and gi/m/1 queues. *Operations Research*, 30:490–514, 1982. 45

[9] R.L. Church and C.S. ReVelle. The maximal covering location problem. *Papers of Regional Science Association*, 32:101–118, 1974. 9

[10] M.S. Daskin. A maximum expected location model: Formulation, properties and heuristic solution. *Transportation Science*, 7:48–70, 1983. 10

[11] M.S. Daskin and E.H. Stern. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15:137–152, 1981. 10

[12] K. Davis. Mirror on the wall: An international update on the comparative performance of american health care. Technical report, Commonwealth Fund, 2007. 1

[13] G. Erdogan, E. Erkut, and A. Ingolfsson. Ambulance deployment for maximum survival. 2006. manuscript, 29 pages. 11

[14] A. Estey, K. Ness, L.D. Saunders, A. Alibhai, and R.A. Bear. Understanding the causes of overcrowding in emergency departments in the capital health region in alberta: A focus group study. *Canadian Journal of Emergency Medicine*, 5(2):81–94, 2003. 5, 14

[15] A. Estey, K. Ness, L.D. Saunders, A. Alibhai, and R.A. Bear. Understanding the causes of overcrowding in emergency departments in the capital health region in alberta: A focus group study. *Canadian Journal of Emergency Medicine*, 5(2):87–94, 2003. 5

[16] R.V. Evans. Geometric distribution in some two-dimensional queuing systems. *Operations Research*, 15:830–846, 1967. 45

[17] A.J. Forster, I. Stiell, G. Wells, A.J. Lee, and C.V. Walraven. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine*, 10(2):127–133, 2003. 16

[18] Makjamroen T. Fujiwara, O. Ambulance deployment analysis: A case study of bangkok. *European Journal of Operational Research*, 31:9–18, 1987. 10

[19] Jacobs P. A Gaver, D. P. and G. Latouche. Finite birth and death models in randomly changing environments. *Advances in Applied Probability*, 16(4):715–731, 1984. 49

[20] M. Gendreau and G. Laporte. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27:1641–1653, 2001. 11

[21] F. Gorunescu, S. I. McClean, and P. H. Millard. A queueing model for bed occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002. 16

[22] M.K. Govil and M.C. Fu. Queuing theory in manufacturing: A survey. *Journal of Manufacturing Systems*, 18(3):214–240, 1999. 18

[23] B. Hajek. Birth-and-death processes on the integers with phases and general boundaries. *Journal of Applied Probability*, 19:488–499, 1982. 54

[24] P. Kolesar and W.E. Walker. An algorithm for the dynamic relocation of fire companies. *Operations Research*, 22:249–274, 1974. 11

[25] R.C. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1:67–95, 1974. 12

[26] R.C Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, 1975. 13

[27] G. Latouche. A note on two matrices occurring in the solution of quasi birth and death processes. *Communications in Statistics*, 3:251–257, 1987. 54

[28] G. Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth date processes. *Journal of Applied Probability*, 30:650–674, 1993. 49, 52, 53, 54

[29] R.A. Marie and J.M. Pellaumail. Steady state probabilities for a queue with a general service distribution and state dependent arrivals. *IEEE Transactions on Software Engineering*, 9:109–113, 1983. 45

[30] F. McGuire. Using simulation to reduce length of stay in emergency departments. In *Proceedings of the 1994 Winter Simulation Conference*, 1994. 15

[31] H.E. Miller and W.P. Pierskalla. Nurse scheduling using mathematical programming. *Operations Research*, 24(5):857–870, 1976. 14

[32] R. Nelson. Matrix geometric solutions in markov models - a mathematical tutorial. IBM Research Division. T.J. Watson Research Center, 1991. 51

[33] M.F. Neuts. Matrix geometric solution in stochastic models - an algorithmic approach. John Hopkins University Press, 1981. 45, 49, 51, 54, 55

[34] Region of Waterloo Public Health. Emergency medical services master plan, December 2007. 3, 4, 86

[35] J. Prno. Ambulance offload delays at hospitals in waterloo region. Technical report, Region of Waterloo Public Health, 2005. 4, 21, 22

[36] V. Ramaswami and D.M. Lucantoni. Efficient algorithms for solving the non-linear matrix equations arising in phase type queues. *Communications in Statistics - Stochastic Models*, 1:29–52, 1985. 49

[37] V. Ramaswami and P.G. Taylor. Some properties of the rate operators in level dependent quasi birth and death processes with countable number of phases. *Stochastic Models*, 12(1):143–164, 1996. 47

[38] J.F. Repede and J.J. Bernardo. Developing and validating a decision support system for locating emergency medical vehicles in louisville, kentucky. *European Journal of Operational Research*, 75:567–581, 1994. 10

[39] C.S. ReVelle and K. Hogan. Concepts and applications of backup coverage. *Management Science*, 34:1434–1444, 1986. 10

[40] C.S. ReVelle and K. Hogan. The maximum availability location problem. *Transportation Science*, 23:192–200, 1989. 10

[41] M.D. Rosetti, G.F. Trzcinski, and S.A. Syverud. Emergency departments simulation and determination of optimal attending physician staffing schedules. In *Proceedings of the 1999 Winter Simulation Conference*, 1999. 15

[42] D.A. Schilling, D.J. Elzinga, J. Cohon, R.L. Church, and C.S. ReVelle. The team/fleet models for simultaneous facility and equipment sitting. *Transportation Science*, 13:163–175, 1979. 10

[43] S.P. Siferd and W.C. Benton. Workforce staffing and scheduling: Hospital nursing specific models. *European Journal of Operational Research*, 60:233–246, 1992. 14

[44] P.M. Snyder and W.J. Stewart. Explicit and iterative numerical approaches to solving queueing models. *Operations Research*, 33(1):183–202, 1985. 45

[45] J. Stout. System status management: The strategy of ambulance placement. *Journal of Emergency Medical Services*, 9(5), 1983. 12

[46] R. Suthons. 2007 budget briefing note: Hospital offload delay, 2007. 4

[47] I.D.S. Taylor and Templeton J.G.C. Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research*, 28(5):1168–1188, 1980. 13

[48] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971. 9

[49] V.L. Wallace. *The Solution of Quasi Birth and Death Processes Arising from Multiple Access Computer Systems*. PhD thesis, University of Michigan, 1969. 45

[50] D.M. Warner. Scheduling nursing personnel according to nursing preference: A mathematical programming approach. *Operations Research*, 24(6):842–856, 1976. 15

[51] D.M. Warner and J. Prawda. A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science*, 19(4):411–422, 1972. 14

[52] J. Ye and S. Li. Folding algorithm: A computational method for finite qbd processes with level-dependent transitions. *IEEE Transactions on Communications*, 42:625–639, 1994. 49