Queueing Network Models of Ambulance Offload Delays

by

Eman Almehdawe

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

Waterloo, Ontario, Canada, 2012

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.

Abstract

Although healthcare operations management has been an active and popular research direction over the past few years, there is a lack of formal quantitative models to analyze the ambulance offload delay problem. Offload delays occur when an ambulance arriving at a hospital Emergency Department (ED) is forced to remain in front of the ED until a bed is available for the patient. Thus, the ambulance and the paramedic team are responsible to care for the patient until a bed becomes available inside the ED. But it is not as simple as waiting for a bed, as EDs also admit patients based on acuity levels. While the main cause of this problem is the lack of capacity to treat patients inside the EDs, Emergency Medical Services (EMS) coverage and availability are significantly affected. In this thesis, we develop three network queueing models to analyze the offload delay problem. In order to capture the main cause of those delays, we construct queueing network models that include both EMS and EDs. In addition, we consider patients arriving to the EDs by themselves (walk-in patients) since they consume ED capacity as well.

In the first model, ED capacity is modeled as the combination of bed, nurse, and doctor. If a patient with higher acuity level arrives to the ED, the current patient's service is interrupted. Thus, the service discipline at the EDs is preemptive resume. We also assume that the time the ambulance needs to reach the patient, upload him into the ambulance, and transfer him to the ED (transit time) is negligible. We develop efficient algorithms to construct the model Markov chain and solve for its steady state probability distribution using Matrix Analytic Methods. Moreover, we derive different performance measures to evaluate the system performance under different settings in terms of the number of beds at each ED, Length Of Stay (LOS) of patients at an ED, and the number of ambulances available to serve a region. Although capacity limitations and increasing demand are the main drivers for this problem, our computational analysis show that ambulance dispatching decisions have a substantial impact on the total offload delays incurred.

In the second model, the number of beds at each ED is used to model the service capacity. As a result of this modeling approach, the service discipline of patients is assumed to be nonpreemptive priority. We also assume that transit times of ambulances are negligible. To analyze the queueing network, we develop a novel algorithm to construct the system Markov chain by defining a layer for each ED in a region. We combine the Markov chain layers based on the fact that regional EDs are only connected by the number of available ambulances to serve the region. Using Matrix Analytic Methods, we find the limiting probabilities and use the results to derive different system performance measures. Since each ED's patients are included in the model simultaneously, we solve only for small instances with our current computational resources.

In the third model, we decompose the regional network into multiple single EDs. We also assume that patients arriving by ambulances have higher nonpreemptive priority discipline over walk-in patients. Unlike the first two models, we assume that transit times of ambulances are exponentially distributed. To analyze the decomposed queueing network performance, we construct a Markov chain and solve for its limiting probabilities using Matrix Analytic Methods. While the main objective for the first two models is performance evaluation, in this model we optimize the steady state dispatching decisions for ambulance patients. To achieve this goal, we develop an approximation scheme for the expected offload delays and expected waiting times of patients. Computational analysis conducted suggest that larger EDs should be loaded more heavily in order to keep the total offload delays at minimal levels.

Acknowledgements

I am most grateful to my supervisors Prof. Elizabeth Jewkes and Prof. Qi-Ming He for their guidance, support and inspiration throughout my Ph.D. studies. Thank you Prof. Jewkes for your encouragement and support during my Masters degree as well.

I would like to thank the members of my examination committee, Prof. Armann Ingolfsson, Prof. Steve Drekic, Prof. Samir Elhedhli and Prof. Fatma Gzara. I am very grateful for their insightful comments and helpful suggestions on my thesis that improved this research.

I wish to express my gratitude to all my friends and colleagues at the Department of Management Sciences who were always there when I needed them. And for making this an enjoyable experience. Special thanks to Bissan, Hossa, Mina and Tiffany.

I am greatly indebted to my husband Islam Saleh who has provided me with unconditional love and support throughout my long journey. Without him, I wouldn't have achieved this success. Also, I am indebted for my sweet little kids Noor and Dania for their patience and understanding when I was busy.

Finally, I am forever indebted to my parents, brothers and sister for their constant care, support, and encouragements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is a growing interest among Operations Research (OR) practitioners to apply OR methodologies to the healthcare sector. Healthcare systems present many complex problems that could benefit from operations research type analysis and applications [1]. The long waiting times of patients, the lack of resources and recently, the offload delay problem faced by the Emergency Medical Services (EMS) are among those issues. On the other hand, working in the healthcare brings up complicated non-technical challenges due to the fact that ”Patients Aren’t Widgets” [2].

Ambulance offload delays are increasingly becoming a concern for healthcare providers across Canada and the United States. Offload time is the time taken to transfer a patient from an ambulance stretcher into an Emergency Department (ED); this time usually is around 30 minutes. This transfer includes unloading the patient from the ambulance, moving the ambulance stretcher inside the ED, giving the report to the triage nurse, transferring the patient to an ED gurney, completing paperwork, and preparing the equipment for the next run [3]. If an Emergency Department is overcrowded and cannot accept transfer of care for an incoming patient, an offload delay results, and the paramedic crews are delayed in emergency departments for extended periods of time, caring for their patients while waiting for an available hospital bed. Moreover, it is not possible to use ambulance crews and vehicles for other jobs whilst they are waiting to offload the patient. In some countries, such as the United States, an ED can declare ”diversion” status if they are overcrowded [4]. For EMS management, "diversion" means that patients should be routed to other less crowded EDs. Diversion, or reallocating patients to a regional hospital can be a key to minimize overall offload delays experienced by ambulances.

From an EMS perspective, offload delays decrease the EMS coverage in the community and increase their response times and costs due to the increase in the actual utilization of ambulances [5]. This leads to a waste of a scarce resource (ambulance). From a patient perspective, the delay of patient admission to the ED has serious consequences on his medical condition. Those clinical consequences are greater for the sicker patients than for those who are less ill. Difficulty offloading patients who need urgent resuscitation results in delayed ED care and worse patient outcomes, especially so in time-critical conditions such as stroke [6]. Delayed admission of patients leads also to patient discomfort and inconvenience, and in some cases, it might lead to poor patient outcomes [7].

The ambulance offload delay problem is a direct consequence of a much bigger problem, which is the lack of capacity in the healthcare system [8]. The principal cause of ambulance offload delay is the lack of capacity to treat hospital inpatients, leading to prolonged Emergency Department Length of Stay and ED overcrowding. Over the years, patients who would have been better cared for in alternate settings remained in acute care beds. As a result of hospital restructuring and financial constraints, acute care beds were reduced without the necessary community support. This led to the care of inpatients in EDs, followed by ED overcrowding and consequent offload delays leading to delayed ambulance response to emergency calls from the community [9].

Ambulance offload delay is a complex problem that happens due to the interaction between an EMS provider and regional EDs served by that EMS provider. So far, there are no quantitative models developed to analyze the interaction between EDs and EMS that capture the effect of ED crowding on ambulance operations and offload delays. In addition, the research available on EMS operations ignores the effect of offload delays on ambulance utilization and consequently, ambulance coverage for a region. To capture the effect of offload delays on EMS performance in general, and on EMS utilization in specific, quantitative models should be developed to quantify offload delays experienced. In the

	Grand River Hospital	St. Mary's Hospital	Cambridge Memorial Hospital
ED (visit/yr)	56,000	45,000	40,000
ED capacity (beds)	39	34	23
EMS (arrival/yr)	8900	5700	5200
Offload delays (hr/mo)	200	81	16

Table 1.1: Region of Waterloo data

context of long term strategic planning, EMS management needs to estimate the response time based on the EMS' available capacity, as well as the regional hospitals' capacity and arrival rates.

1.1 Motivation

Different regions in Canada have started to report on offload delays, in this section we highlight some of the statistics from EMS reports to show the significance of the problem. For example, in 2006 the Provincial government of Ontario invested $96 million in its comprehensive action plan to reduce the length of time paramedics wait to offload patients in front of hospital EDs. Offload delays cost Toronto EMS approximately 180 ambulance hours per day in December 2007 [10]. The average offload delay in the Toronto area in the same period was reported to be 3 to 8 hours. That's 3 to 8 hours a fully equipped ambulance and its trained paramedics wait for an ED bed while they could be available to respond to other emergency calls in the community.

The EMS in the Region of Waterloo, Ontario, own 18 ambulances that serve the region's three hospitals. According to the EMS 2008 Master Plan [11], the region reported a maximum of 22 offload delay incidents in a single day in December of 2007. In 2006, the Region of Waterloo incurred more than 6000 hours of offload delays and lost 12.36 ambulance days per month. In 2005, the number of ambulances lost to offload delays totaled to as many as 13.25 ambulance days per month. Table 1.1 illustrates some statistics about the offload delays in the region with respect to the three hospitals in 2007.

Middlesex-London, Ontario EMS have had at least one Cardiac Arrest in the hall-way at University Hospital. Table 1.2 summarizes some of the Middlesex-London region statistics for the offload delay problem in the last two years. In Peel, Ontario the average growth in offload delay since 2001 is approximately 25 per cent per year. The growth in offload delay hours for 2007-2008 is 36 per cent. In York region, the time spent in hospitals by paramedics was reported to be 12,946 hours in 2000, and 27,238 in 2004. All these data shows that the problem is serious and it is getting worse.

		Victoria Hospital	University Hospital
ED visits/yr	2008	$12,186$	$11,552$
	2009	$11,674$	$12,445$
Offload delays, min/yr	2008	138,720	116,580
	2009	117,780	95,400

Table 1.2: Middlesex Region data

1.2 Objectives

Ambulance offload delay is a complex problem that happens due to the interaction between an EMS provider and EDs in a region. Motivated by a project conducted with the EMS in Waterloo, Canada, in this research we model this interaction using queueing networks. For that purpose we develop three models to analyze the problem. We use those models to suggest possible solutions for the problem. While all the models developed are stochastic, each is model is unique in the modeling approach and assumptions. Our focus is primarily on the application, and indeed, we use the Region of Waterloo project as a running example throughout this research.

Our models are capable of capturing the offload delay variability in terms of the hospitals emergency departments' capacities, the Length Of Stay (LOS) of patients, and the number of ambulances available to serve a region. Our methodology provides exact solutions for various performance measures. The aim is to provide the decision makers with a decision support tool that can be used to investigate different possibilities in terms of EMS number of ambulances, ED capacity, and population arrival rates. The operational research method used can be customized to any regional EMS-ED system through the use of its corresponding data elements.

In this research we address the following issues:

1. The impact of limited capacity of emergency departments and the LOS of patients on offload delays, at the hospital level and at the regional level.

2. The impact of ambulance dispatching decisions on offload delays and on crowding at both hospital level and regional level.

3. The impact of arrival rate of patients arriving to a hospital ED by themselves (later are called walk-in patients) on offload delays incurred.

4. The likelihood that the EMS cannot respond to an emergency call because all the ambulances are busy.

5. The effect of offload delays on the probability distribution of the number of busy ambulances in a region, and as a result, the total actual ambulance utilization.

6. The optimal allocation of ambulance patients to regional hospitals such that offload delays are minimal.

The queueing models developed are different in terms of modeling assumptions, solution methodologies and research objectives. The first model we develop in Chapter 3 is based on the idea of modeling the capacity in an ED as the combination of beds, nurses, and doctors. Due to this modeling approach, the service discipline at the ED can be represented by a preemptive resume system. This means that if the combination (bed, doctor, and nurse) is not available, then the patient service will be interrupted. This model is built on a regional level to achieve the above objectives.

In Chapter 4, we model the beds as servers. In this case, when a patient with a more acute condition arrives to the ED, he will be admitted to the ED before other patients with less acute conditions that arrived to the ED before him. For this model, we use the nonpreemptive priority discipline to model the admission of patients into hospital EDs. The model is built on a regional level.

Model	**Network size**	**priority**	**transit time**
Model 1 (Ch. 3)	multiple EDs	preemptive	negligible
Model 2 (Ch. 4)	multiple EDs	nonpreemptive	negligible
Model 3 (Ch. 5)	single ED	nonpreemptive	exponential

Table 1.3: Summary of modeling assumptions for the three models developed

The third queueing model in Chapter 5 uses similar modeling assumptions as the second model but has different research objectives. While the first two models focus on performance evaluation for the queueing network and quantifying offload delays of ambulances, this model aims to optimize the allocation of patients arriving by ambulances to regional EDs. Table 1.2 summarizes the main assumption differences of the queueing models developed in this thesis. More details on each model assumptions, methodologies, and objectives are provided in the corresponding chapters.

1.3 Outline of the Thesis

The thesis is organized as follows. In Chapter 2 we present an overview of related literature. In Chapter 3, we develop and analyze the first queueing model when the service priority at the EDs are preemptive and transit times are negligible. In Chapter 4, we develop the second model when the service priority is nonpreemptive and the transit times are negligible, while in Chapter 5 we develop the decomposed model with nonpreemptive priority discipline and Markovian transit time. Finally, conclusions and future research directions are discussed in Chapter 6.

Chapter 2

Related Background

In this chapter, we classify the related background into four broad categories: first, the literature related to the modeling and analysis of the ambulance offload delay problem or ambulance diversion which is the problem we investigate in this thesis. Second, the literature on the use of queueing theory to model congestion and delays in healthcare systems. Third, the literature on queueing networks with blocking which we use in Chapter 5 to analyze the resulting queueing model. Fourth, we introduce the Matrix analytic Methods which we utilize to derive the limiting probabilities of the resulting Markov chains of each model considered.

2.1 Ambulance offload delay and ambulance diversion

The increasing awareness in the delay ambulances experience when they offload patients to the Emergency Departments has urged decision makers to start analyzing this problem, yet, there has been little research performed from an OR perspective. This research is motivated by recent work conducted by Majedi [12] who models the ambulance offload delay using queueing theory. He models the one hospital interaction with the ED using a two-dimensional Markov chain and analyzes the system performance under different input parameters. His model does not capture the impact of dispatching decisions on offload delays, nor the effect of walk-in patient arrivals to the EDs.

Other research done on the offload delay problem has been conducted by MD practitioners who try to shed some light on the importance of the problem and its implications. For example, Ting [6] investigates the causes of ambulance offload delay and the impact of delayed ED care for patients. Taylor et al. [13] conduct an observational study to determine the difference between documented ambulance arrival times and the actual arrival times of patients from the ambulance into the emergency department.

There is some work that investigates the impact of limited capacity in hospital beds on the EMS offload delay. Silvestri et al. [14] examine the effect of ED bed availability on offload delays experienced by conducting an observational study for 22 months in Orange County in Florida. The study suggests that ED bed availability has an impact on EMS unit offload delays. Later, Silvestri et al. [15] conduct an observational study to evaluate offload delay intervals and the association between out-of-hospital patient triage categorization (PTC, which is similar to CTAS in Canada) and admission. The study concludes that delayed EMS units have reduced the EMS response availability, and PTCs are not able to determine need for admission and should not be used to support offload delays.

Eckstein and Chan [3] investigate the effect of ED crowding on paramedic ambulance availability from April 2001 through March 2002 in Los Angeles, CA. Their empirical study suggests a direct relation between ED crowding and the ability of EMS to provide timely responses. Schull et al. [16] conduct a quantitative analysis to determine the relationship between physician, nursing, and patient factors on emergency department use of ambulance diversion.

The main cause of ambulance delays is ED overcrowding. Drummond [7] investigates and summarizes the causes of ED overcrowding to be:

1. Lack of beds for patients admitted to the hospital;

2. Shortage of nursing staff;

3. Increased volume, complexity and acuity of patients in the ED;

4. Delays in service provided by other departments, e.g. labs and consultants.

The delays inside the ED have a cascading effect which result in ambulance offload delays or ambulance diversion on the EMS side. As a result, ambulance offload delays are being proposed as a realistic measure of ED overcrowding [17]. According to a quantitative study by Schull et al. [16] on the determinants of ambulance diversion, they found that admitted patients in the ED is the main cause for ambulance diversion, whereas nurse hours and the volume of walk-in patient arrivals to the ED are minor contributors for ambulance diversion.

In Australia, the term "access block" is used to define the situation where patients are unable to gain access to hospital beds within a specified amount of time [18]. Forero et. al [18] survey access block studies and report their impact on patients' mortality and ambulance diversion. They conclude that the problem will remain unless the hospital capacity is addressed in an integrated approach. They suggest that this should be done at both national and state levels.

In order to solve this problem, some countries, e.g. the United States, allow the ED to declare "diversion" status if they are overcrowded [4]. For EMS management, "diversion" means that patients should be routed to other less crowded EDs. Recently, Deo and Gurvich [4] developed a queueing game model for two EDs that each try to minimize their waiting times. They show that decentralized diversion decisions result in depooling of the network resources. They also provide a near optimal solution for the ambulance diversion problem when a centralized dispatcher (social planner) coordinates diversion.

Our work is unique because we develop quantitative models to *analyze and minimize* the offload delays in terms of ED-specific parameters. While the literature above focuses on a single ED, our models combine the effect of overcrowding in multiple EDs for a region. This is because EMS are always provided on a regional basis. In that view, our models are more general and can give more insightful results for EMS decision makers and analysts.

2.2 Queueing Models for Healthcare Systems

Queueing theory has been used in literature to analyze systems that are characterized by limited resources and variable customer arrival and service times. Queueing network models have been used extensively to model production, telecommunication systems, and traffic flow to help determine capacity levels that are needed to fulfill demand within an acceptable time frame [19]. Although queueing theory is useful for analyzing systems faced with extended delays and resource shortages, as faced by healthcare systems, the use of queueing models in this field is limited.

Utley and Worthington [20] review the modeling methods available for healthcare organizations in terms of resources and service levels. They focus on the insights that can be drawn from queueing and simulation models. Formundam and Herrmann [21] and Green [22] provide extensive surveys on the contributions and applications of queueing theory in healthcare systems. Queueing model performance measures are available in the form of analytical, numerical, or approximate solutions.

2.2.1 Benefits of queueing models

We summarize the benefits of the queueing modeling and analysis approach for the healthcare systems as follows:

1. It can help determine levels of staffing, equipment and beds to achieve a service standard;

2. It can be used to assess the implications of decisions with respect to resource allocation and design of new services;

3. It is helpful in gaining insight on the degree of flexibility in organizing resources;

4. It can give simple formulae results for the system performance, e.g. expected delays, expected queue length and the probability of waiting, among other measures;

5. The performance measures derived can be used to develop optimization models.

Those models can be used to find optimal solutions efficiently and with minimum input data.

Simulation is another approach that has been used more frequently to model healthcare systems. Queueing models, compared to simulation in which the queueing assumptions are relaxed, require less input data and yet, they can provide insightful results, making them easier to implement than simulation models [23].

2.2.2 Characteristics of queueing models

In most queueing models, there are six basic characteristics that should be identified to describe the system of interest:

1. Arrival pattern of customers: In order to identify the arrival pattern, the characteristics of the stochastic process that generates the arrivals should be specified, as well as the number of customers at each arrival epoch if the customers arrive in batches. Another important criterion that should be defined is the reaction of the customer upon entering the system: wait, balk or renege.

2. Service pattern of customers: A probability distribution to identify the sequence of customers' service times, in addition to specifying whether the service is done in batches or for each customer.

3. Queue discipline: the manner in which customers are selected to start service. The most common discipline observed is First Come First Serve (FCFS). Other queueing disciplines might include Last Come First Serve (LCFS), preemptive priority and nonpreemptive priority.

4. System capacity: Some queueing systems include a limitation on the queue size, in this case, a queue limit should be identified. If the queue limit is reached, customers arriving to the system are lost.

5. Number of Servers: A queueing system can either be served by one server or multiple servers. If multiple servers are used, then it should be identified whether the system

is fed by a single line or multiple lines.

6. Stages of service: A queueing system might have a single stage of service, or several stages of service.

For each queueing model we develop in the next chapters, we identify the basic characteristics mentioned above. We use the real system behavior to make decisions about those characteristics.

2.2.3 Validity of queueing models assumptions for healthcare systems

In a queueing model, a number of assumptions are usually made for the above characteristics. In this subsection, we list our main assumptions and argue the validity of those assumptions in the healthcare context.

1. The system has reached steady state: Most queueing models assume that the system has been operating with the same characteristics (number of physicians, number of nurses,...) for a sufficiently long time, such that the probability distributions derived are independent of the time. Although this assumption might not be true for healthcare systems, Green [22] gives an example on how steady state queueing models are useful to effectively allocate resources in such situations. Responding to the staffing variations without a quantitative model leads to inefficient and ineffective allocation of resources.

2. Stationary arrival process: Empirical analysis of healthcare system arrivals indicates that arrival processes are non-stationary, which means that arrival rates depend on the time of the day, day of the week, and month of the year. Using a stationary arrival process to approximate a non-stationary arrival process for admission has been justified by Lewis [24] and Kao and Tung [25] among others. To account for nonhomogeneity, practically, Cochran and Roche [26] suggest the use of a seasonality multiplier and a peaking multiplier to adjust for seasonality and time-of-day

variation in the ED arrivals as follows:

$$\lambda_0 = \frac{\text{yearly ED arrival}}{365 * 24} * \text{seasonality multiplier} * \text{peaking multiplier} \tag{2.1}$$

3. The system is stable: This implies that the system operates strictly under 100% utilization rate. This fact is actually useful to explain the long waiting lines in front of clinics, EDs or specialist lists since those systems operate usually near 100% utilization.

4. Poisson arrival process: In healthcare, the Poisson process has been verified to be a good representation for unscheduled arrivals to various parts of hospital including EDs. See Green [22] and the references therein.

In Table 2.2, we summarize some of the work that has been done using queueing models for healthcare systems. For each article we identify the application, queueing model used, and main results.

The major objective of this research is to quantify the delays ambulances experience upon arrival to EDs. Another key objective is to assess the implications of patient re-allocation to regional hospitals. In order to capture those delays, we develop three queueing models each with different assumptions. Queueing analysis can be an extremely valuable tool for utilizing resources in the most cost effective way to reduce delays [22]. For those reasons, we use this methodology to analyze the offload delay problem.

2.3 Queueing Networks with Blocking

A queueing network is a set of interconnected nodes. Each node consists of a queue, where customers wait for service, and one or more servers [27]. If one or more queues in the network have limited capacity, blocking may occur. Queueing Networks with blocking have recently become an important and active research topic in performance evaluation because of their applicability to model real life systems. They have been used, for example,

Author (year)	Application	Model structure	Main Results
Alanis et al. (2012)	Ambulance repositioning	Markov chain model	Develop a model that can be used to identify near-optimal compliance tables for ambulance repositioning
Blair and Lawrence (1981)	Burn care in New York State	$M/G/s/s$ and Markov chain model	Describe the operations of a system of burn care facilities linked together by a referral policy
Cochran and Broyles (2010)	Emergency department capacity	$M/M/1/K$ queue	Find the ED capacity requirements based on patient reneging percentage as a measure for patient safety
Cochran and Roche (2009)	Emergency department performance	queueing network model	Derive ED performance measures to reduce patient walk-aways and increases ED access
Deo and Gurvich (2011)	Ambulance diversion decisions	queueing game	Show that centralized ambulance diversion decisions outperform decentralized decisions due to resource pooling
Gorunescu et al. (2002)	Bed use in a department of Geriatric Medicine	$M/PH/c/N$ queue	Optimize the allocation of beds in order to maintain an acceptable delay
Kao and Tung (1981)	Bed allocation in public heathcare system	$M/G/\infty$ queue	Reallocate beds to services to minimize the expected overflow of patients in a healthcare system
Restrepo et al. (2009)	Ambulance allocation to bases	$M/G/c/c$ queue	Develop two models to allocate a fleet of ambulances such that the response times are minimum
Su and Zenios (2004)	Patient choice in kidney allocation	$M/M/1$ queue	Maximize social welfare by changing the queueing discipline from FCFS to LCFS
Vericourt and Jennings (2011)	Nurse staffing in medical units	$M/M/c//n$ queue	Find effective staffing policies should deviate from threshold-specified nurse-to-patient ratios

Table 2.1: Summary of literature for queuing theory for healthcare systems

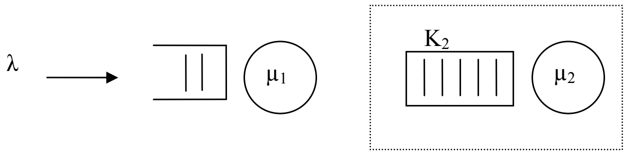

Figure 2.1: A tandem queue with finite capacity

to model production lines and telecommunication networks where capacity limitations may affect the performance of the system.

To illustrate the blocking process in more detail, we consider a queueing network that consists of two nodes as depicted in Figure 2.1. We denote the first node as the upstream node and the second as the destination or downstream node. The destination node has finite capacity (K_2) including the customer in service, while the upstream node has infinite capacity. If a customer at the upstream node finishes service and finds the destination node queue full, he will wait at the upstream server until there is space for him in the destination node. Thus, the upstream server acts as an extra waiting spot for the destination node, and at this moment we say that the upstream server is *blocked.* Blocking implies that the server is not able to serve additional customers, and the customer at the blocked server is delayed. We emphasize here that the blocking we consider in this work is not related to the blocking in the literature that assumes blocked customers are lost, where the later was the assumption made by e.g. Tahilramani et al. [28], Kouvatsos and Xenios [29], Dijk and Wal [30], and Smith [31].

Blocking may happen in different mechanisms, depending on when blocking and unblocking occur. The main blocking mechanisms used in literature are:

- Blocking After Service (BAS): The customer at the blocked node finishes service and then waits for a space in the destination node. We use this blocking mechanism in our work to model the blocking of ambulances. If an ambulance, after transferring a patient to the ED, finds the ED full, it will not be able to transfer other patients until there is space for the current patient in the ED.

- Blocking Before Service (BBS): The upstream server checks the queue of the destination node; if the queue is full, it stops and does not serve the current customer unless there is space for him in the destination node. This type of blocking is mostly incurred in telecommunication networks.

- Repetitive Service Blocking (RS): The customer receives repeated services until there is space for him in the destination node.

Analysis of queueing networks with blocking is challenging. Some small networks have exact analytical solutions, while approximations are mostly used to analyze more complicated networks. The main techniques used in literature for the analysis of these networks can be grouped into three broad categories as follows:

1. Analytical Solutions: Analytical solutions exist only for special networks, for example, the open two node queueing network, with single and multiple servers, when the service times are exponential and the arrival process is Poisson, e.g. Perros [27].

2. Numerical Solutions: Numerical solutions that are based on constructing a Markov chain model for the system have been used to solve simple queueing networks with blocking, e.g. Houdt and Alfa [32], and Latouche and Neuts [33].

3. Approximate Solutions: Most of the available literature on queueing networks with blocking is in the form of approximate solutions that utilize basic, one or two node configurations, to decompose the network into smaller blocks.

In the healthcare systems, although queueing networks with blocking can be suitable to model scarce resources, its use is limited. Koizumi et al. [34] applied queueing networks with blocking to analyze patient flow in mental health institutions in Philadelphia. Recently, Osorio and Bielaire [35] developed an approximation scheme to find the queue length distribution for a general topology queueing network with blocking and multiple servers. They apply their results to study patient flow for hospital units. However, their model does not consider multiple patients with different priorities as we assume in our work.

Bretthauer et. al [36] use the concept of blocking to model patient flow in a hospital. From a queueing perspective, they develop a heuristic for tandem queues to evaluate the effect of blocking on the entire system performance.

In Chapter 5, we model the ambulances as servers that might experience blocking. We use a computational stochastic method which is the Matrix Analytic Method, to solve a simplified model for the ambulance offload delay problem. The resulting queueing network has a general configuration with multiple priority classes and multiple servers that has not yet been analyzed. We develop an approximation algorithm to analyze the network and its performance. Since our work is related only to queueing networks with blocking and multiple servers, we review only the literature in queueing networks with blocking that consider multiple servers, or that consider multiple customer priorities.

2.3.1 Queueing Networks with Blocking and Multiple Servers

There is a considerable amount of literature that has analyzed queueing networks with blocking when one or more nodes have multiple servers. Some of the literature use the Expansion Method to analyze open, general topology networks. To account for blocking, they introduce a holding node between finite capacity nodes. Han and Smith [37] use the Expansion Method to calculate the throughput of a queueing network that has Poisson arrivals and exponential service times at each server. They approximate the service time at the finite node by a Coxian distribution. Jain and Smith [38] use the same method to derive the network throughput and investigate the optimal ordering of servers. Lately, Cruz and Smith [39] use the Expansion method to derive the blocking probability and the expected waiting time and number of customers in the system for a network that has no buffer space for customers. Andriansyah et al. [40] extend Cruz and Smith's work by optimizing the number of servers and the network throughput using genetic algorithms.

While the previous authors consider general topology networks, there are some articles that analyze tandem queues that consist of two or more nodes. Akyildiz [41] approximates the throughput for a closed tandem network. Latouche and Neuts [33] derive the exact probability distribution for the number of customers in the system for a two-node tandem queue using the Matrix Analytic Method, while van Vuuren el al. [42] decompose the network into two station subsystems to find the approximate mean sojourn time and network throughput. Table 2.2 summarizes the main research that has analyzed queueing networks with blocking and multiple servers.

2.3.2 Queueing Networks with Blocking and Multiple Classes of Customers

The theory on queueing networks with blocking is mainly related our model in Chapter 5 where ambulances are modeled as servers that have exponential service time. Queueing Networks with blocking and multiple classes of customers have received less attention because they are difficult to analyze. Wagner [43] considers a single node multi-server queue with finite capacity. The arrival process for all customer classes is Poisson and the service times are exponential. He assumes non-preemptive priority discipline. If an arrival finds all the waiting spaces occupied, it is lost. For this model, Wagner derives the steady-state probability distribution explicitly for the number in the system for a two-customer class model, and the Laplace-Stieltjes Transform for the actual waiting time of each customer class using Matrix Analytic Methods.

To our knowledge, no work has modeled multiple priorities in queueing networks with blocking where customers are delayed when the server is blocked.

2.3.3 Decomposition and Approximation for Queueing Networks with Blocking

In this section, we present one of the main methodologies used to analyze queueing networks with blocking which is decomposition and approximation. The algorithm decomposes the queueing network into isolated single nodes, each with modified arrival rate, service rate, and buffer capacity. Then each node is studied in isolation based on the new approximated parameters. The main steps to perform the analysis are:

- Decomposition of the queueing network into single nodes.

Author (year)	Network Structure	Basic Assumptions	Analysis Methodology	Main Results
Akyilidiz(1989)	Closed tandem configuration network	Fixed number of jobs in the network, BAS, deadlock free network	Transform the network to a nonblocking one with appropriate total number of jobs which has equal feasible states to the blocking network	Mean number of jobs in the system, approximate the network throughput
Andriansyah et al. (2009)	Open general topology network	Zero-buffer, BAS	Generalized Expansion Method, Genetic algorithm for optimization	Network throughput, minimum number of servers and maximum throughput
Cruz and Smith (2007)	Open general topology network	M/G/c/c state-dependent service rate, no buffer space, and BAS	Generalized Expansion Method	Blocking probability, throughput, expected performance measures
Han and Smith (1991)	Open tandem, split, and merge topologies (2-3 nodes)	Poisson arrival and exponential service time at each server	Introduce a holding node between nodes to account for blocking probability and approximate the service time by Coxian distribution	Calculate the system throughput and blocking probability approximately
Jain and Smith (1994)	Open, series and parallel network topologies	$M/M/C/K$ queue, first node never starved	Expansion method	System throughput and optimal order of servers
Latouche and Neuts (1980)	2 node network in tandem	Allow a feedback loop of departures from the second node to the first	Matrix Analytic Method	Derive the conditions for stability and the probability distribution for the number in the system
Vuuren et al. (2005)	open tandem queueing network	General service time, BAS	Decompose the network into 2 station subsystems for which they approximate the arrival and service parameters, use spectral expansion method to analyze the subsystems	Approximate the network throughput and mean sojourn time

Table 2.2: Summary of literature for queuing networks with blocking and multiple servers

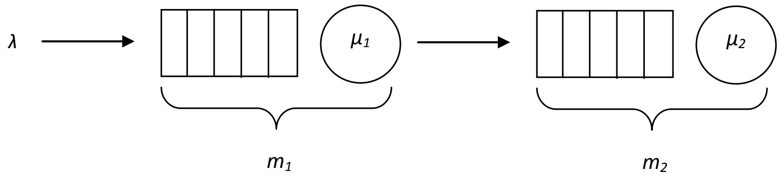

Figure 2.2: A two node tandem queueing network with blocking

- Analysis of each single node in isolation. The single nodes are related to their network surroundings by input (arrival) and output (departure) processes.
- Approximation of all nonrenewal processes by stationary renewal processes.

Consider for example the tandem queue in Figure 2.2 which consists of two nodes. The decomposition algorithms are based on determining the effective service and arrival rates for each isolated node. These approximate parameters are usually based on one of two assumptions. Either the effective service times are exponential, or they have a phase type distribution. Once the effective arrival and service rates are determined, the expected queue length and expected waiting times for customers are calculated using the $M/M/1/m_i + 1$ results. In Chapter 5, we develop an approximation scheme based on the decomposition approach described to approximate offload delays.

In the next section, we introduce some background concepts for the Matrix Analytic Methods that we use to solve our simplified version of the model.

2.4 Matrix Analytic Methods

Over the last two decades, Matrix Analytic Methods have been used to analyze a wide range of systems. These methods are popular as modeling tools because they provide the ability to construct and analyze, in a unified way and an algorithmically tractable manner, a wide class of stochastic models [44]. Matrix Analytic Methods, since their introduction in 1970s by Marcel F. Neuts, have been successfully used to model a wide variety of applications that range from queueing systems to inventory models, and most commonly, telecommunication systems. In these models, the embedded Markov chains are two-dimensional generalizations of elementary GI/M/1 and M/G/1 queues and their intersection, i.e., Quasi-Birth-and-Death (QBD) processes [45].

In this section we describe the computational procedure that was developed by Neuts in 1981 [46] to analyze QBD processes where transitions are only allowed to adjacent levels. It can be used to calculate the steady state queue length distribution for a QBD process $\{X(t) : t \geq 0\}$ that has a tridiagonal generator matrix P of the form:

$$P = \begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \tag{2.2}$$

The essential problem in determining the steady state probability distribution of a Markov process is solving a set of linear flow balance equations, where there is an equation associated with each state of the system. For systems with a large or possibly infinite number of states, exact solutions can only be obtained if one can exploit structural properties of these balance equations. Neuts developed a body of results that allows one to exploit repetitive structures. If the states of the Markov process can be grouped into vectors which possess a certain repetitive structure as in (2.2), then a recursive procedure can be used to determine the stationary state probabilities of any vector in terms of the probabilities for the previous vector [47]. If the QBD process is ergodic, Neuts shows that a nonnegative solution exists for the set of linear equations:

$$\begin{aligned} \boldsymbol{\pi} P &= 0 \qquad (2.3) \\ \boldsymbol{\pi} \boldsymbol{e} &= 1 \end{aligned}$$

where $\boldsymbol{e}$ is a column vector of ones of appropriate size. And $\boldsymbol{\pi}$ is the limiting probability vector associated with the QBD process. This solution has the geometric form:

$$\pi_n = \pi_{n-1} R, \quad \text{for } n \geq 2 \tag{2.4}$$

where the rate matrix R is the minimal nonnegative solution to the nonlinear equation:

$$A_0 + RA_1 + R^2 A_2 = 0$$

To calculate the boundary probabilities, (π_0, π_1), we solve the set of equations below:

$$(\boldsymbol{\pi_0}, \boldsymbol{\pi_1}) \begin{pmatrix} A_{0,0} & A_{0,1} \\ A_{1,0} & A_1 + RA_2 \end{pmatrix} = 0 \tag{2.5}$$

$$\boldsymbol{\pi_0} \boldsymbol{e} + \boldsymbol{\pi_1} (I - R)^{-1} \boldsymbol{e} = 1 \tag{2.6}$$

The method described only applies for infinite QBD processes with independent levels, whereas for our models, we show that the resulting Markov processes are dependent on one another (Model 1 and Model 2). In the next chapter, we model the ambulance offload delay problem using a queueing network and develop a Markov chain representation for its steady state distribution when the service discipline is preemptive.

Chapter 3

Model 1: Multiple ED Network with Preemptive Priority Discipline and Zero Transit Time

The first model we develop for the offload delay problem is constructed for multiple EDs network. This model is based on three main assumptions: First, regional EDs are dependent among each others; second, ambulance patients have higher preemptive priority over walk-in patients; third, the time for an ambulance to pick up a patient, load him into the ambulance, transfer him to the ED, and unload him from the ambulance is negligible.

In this chapter, we first describe the stochastic model in details along with the model assumptions in Section 3.1. We analyze the model with ambulance patients only in Section 3.2. Then we investigate the model with both ambulance patients and walk-in patients in Section 3.3. Numerical analysis and some case studies are shown in Section 3.4. Two of the model assumptions are validated in Section 3.5. Finally, we conclude at Section 3.6.

3.1 The Stochastic Model

We consider a queueing network for a system with one EMS provider that serves K hospitals, each with a single ED. The nodes represent hospital EDs all served by a common

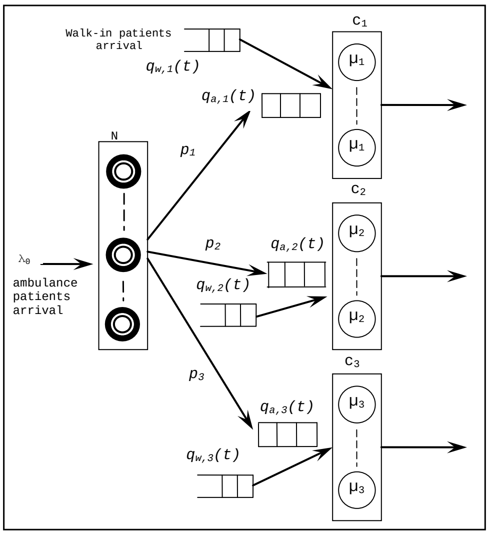

Figure 3.1: EMS-ED Queueing Network Diagram for $K = 3$

EMS provider. Figure 3.1 illustrates a network consisting of three hospitals. Ambulance patients arrive and obtain service from the EMS provider which has N ambulances. We assume that patients arrive according to a Poisson process with rate λ_0.

When an arrival occurs and an ambulance is available, the patient is brought to the k^{th} ED with probability p_k. In practice, the hospital to which the patient is taken may depend on the type of complaint they have, or on which hospital is the closest. We have constructed our model so that it does not reflect these characteristics as we wanted to understand the pattern of overall patient flows in steady state. In our model, when an arrival occurs and an ambulance is not available, the patient demand is assumed to be lost. In reality, the EMS operators monitor very carefully the number of ambulances available to respond to emergency calls. When the number becomes critically low, they will contact neighboring EMS providers to request assistance. This does not happen frequently, and the event of having no ambulances available is extremely rare. Hence we feel that our assumption of lost customers is reasonable as it will not have a large impact on the quality of our solutions.

Finally, we assume that the transit time to the hospital is small in comparison to the time a patient spends at the hospital. This simplification permits us to obtain many insights without overly complicating our model. More importantly, Offload delays, which are the focus of this work, only depend on ED capacities. In section 6 of this paper, we show how adding the EMS transit time into the model has minimal impact on offload delays and other performance measures of interest for the ambulance patients. The k^{th} ED has a service capacity of c_k. This can be viewed as the combination of resources (e.g. nurse, doctor, and bed) needed to serve an individual patient. Each unit of capacity operates independently of others. Note that we are modeling the area of the ED that deals with acute and intermediate care patients - those that have more severe ailments.

From the ED perspective, there are two arrival streams: ambulance patients, and walk-in patients. When patients arrive to an ED, they are triaged in order to assess the acuity of their illness. Generally, patients who call for an ambulance have higher acuity levels than walk-in patients. Figure 3.2, constructed with data from a local hospital, shows that patients arriving by an ambulance are most of the time are those with high acuity conditions. In the figure, CTAS 1 (Canadian Triage and Acuity Scale) represents patients with the most severe conditions who require immediate attention. For this reason we have assumed that ambulance patients have preemptive priority over walk-in patients. Preempting the service of a walk-in patient can be interpreted as preempting their care, as is the case when a severely ill patient arrives to the ED.

Walk-in patients arrive to the k^{th} ED according to a Poisson process with rate λ_k. The service time for both ambulance and walk-in patients at the k^{th} ED is assumed to be exponentially distributed with parameter μ_k. Since the service time of walk-in patients has an exponential distribution, when a walk-in patient regains service, it does not matter whether its service is resumed or repeated. Thus, both preemptive repeat and preemptive resume cases for walk-in patients are considered.

We summarize the model parameters as follows:

- K: Number of regional hospitals;
- λ_0: Patient arrival rate to the EMS system;
- p_k: Probability that an EMS arrival is sent to the k^{th} ED, for $k = 1, 2, ..., K$;
- μ_k: Service rate per server in the k^{th} ED, for $k = 1, 2, ..., K$;

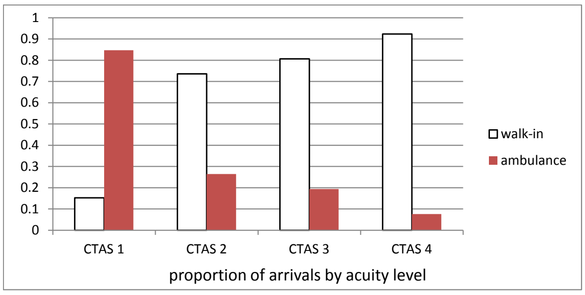

Figure 3.2: Arrivals to an ED by acuity level and mode of arrival

- λ_k: arrival rate of walk-in patients at the k^{th} ED, for $k = 1, 2, ..., K$;
- c_k: Number of servers in the k^{th} ED, which corresponds to the service capacity at the k^{th} ED, for $k = 1, 2, ..., K$.
- N: Total number of ambulances available in the system;

In order to analyze the queueing network, we introduce a Markov chain that can be used to analyze system performance. The Markov chain allows us to derive various probability distributions which we use later to derive system performance measures. We start by defining 2 sets of state variables to describe the system state:

1. $q_k(t)$: The number of ambulance patients at the k^{th} ED, including the ambulance patients in service, at time t, for $k = 1, 2, ..., K$;
2. $q_{w,k}(t)$: The number of walk-in patients in service and waiting in the k^{th} ED, at time t, for $k = 1, 2, ..., K$.

The total number of state variables we need to represent the queues in the network is $2K$. Based on the definition, if $q_k(t) \geq c_k$, then all walk-in patients are waiting in the queue; if $q_k(t) < c_k$, then there are $c_k - q_k(t)$ servers available to serve the walk-in patients at the k^{th} ED. The fact that the service discipline at each hospital ED is assumed to be preemptive, where walk-in patients have lower priority and ambulance arrivals are assigned higher priority allows us to analyze the queue of ambulance patients separately without the need to include the walk-in arrivals. We use this observation as a building block in the next layer to analyze the queue of walk-in patients.

The value of this methodology in constructing the Markov chain will become evident when we solve for real life instances of the model where the size of the problem increases. Due to splitting the high priority patients from low priority ones, we don't need to solve the entire model to find the steady state probabilities and performance measures related to ambulance patients. Solving only for the K dimensional Markov chain $\{(q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$ gives us all the results pertaining to ambulance patients and offload delays.

3.2 High priority ambulance patients

In this section, we analyze the stochastic model with only ambulance patients. The analysis consists of five steps. First, a recursive method is introduced for constructing the infinitesimal generator for $\{(q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$. Then matrix-analytic methods are used for computing the stationary distribution of that continuous time Markov chain. A number of performance measures are derived. A Markov chain is also constructed for the waiting times of ambulance patients. Finally, at the end of this section, some real cases are studied using the methods developed.

3.2.1 The Markov chain

Consider the process $\{(q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$. Since the arrival process of ambulance patients to the system EDs is Poisson and the service times are exponential, it is easy to see that the stochastic process $\{(q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$ is a continuous time Markov chain. The queue lengths, $q_1(t)$, $q_2(t)$, ..., and $q_K(t)$ are finite such that $q_1(t) + q_2(t) + ... + q_K(t) \leq N + c_1 + c_2 + ... + c_K$, which implies that the process $\{(q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$ is a continuous time Markov chain with a finite state space. For convenience, we use i_k for the value of $q_k(t)$. The state space Ω of the Markov chain can be organized as follows:

- $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots \cup \Omega_{N+c_K}$;

- $\Omega_{i_K} = \Omega_{i_K,0} \cup \Omega_{i_K,1} \cup \ldots \cup \Omega_{i_K,N+c_{K-1}-\max(0,i_K-c_K)}$, for $0 \leq i_K \leq N + c_K$;
- $\Omega_{i_K,i_{K-1}} = \Omega_{i_K,i_{K-1},0} \cup \Omega_{i_K,i_{K-1},1} \cup \ldots$ $\cup\, \Omega_{i_K,i_{K-1},N+c_{K-2}-\max(0,i_K-c_K)-\max(0,i_{K-1}-c_{K-1})}$, for $0 \leq i_{K-1} \leq N + c_{K-1}$, $0 \leq i_K + i_{K-1} \leq N + c_K + c_{K-1}$;
- $\Omega_{i_K,i_{K-1},\ldots,i_j} = \Omega_{i_K,i_{K-1}\ldots,i_j,0} \cup \Omega_{i_K,i_{K-1},\ldots,i_j,1} \cup \ldots$ $\cup\, \Omega_{i_K,i_{K-1},\ldots,i_j,N+c_{j-1}-\max\{0,i_K-c_K\}-\ldots-\max\{0,i_j-c_j\}}$, for $0 \leq i_j \leq N + c_j$, and $0 \leq i_K + \ldots + i_t \leq N + c_K + \ldots + c_t$, $j + 1 \leq t \leq K$;
- $\Omega_{i_K,i_{K-1},\ldots,i_2} = \{0, 1, \ldots, c_1, c_1+1, \ldots, c_1+N-\max\{0, i_K-c_K\}-\ldots-\max\{0, i_2-c_2\}\}$, for $0 \leq i_2 \leq N + c_2$, and $0 \leq i_K + \ldots + i_j \leq N + c_K + \ldots + c_j$, $2 \leq j \leq K$;

We observe here that each of the state variables $q_K(t), q_{K-1}(t), \ldots,$ and $q_1(t)$ changes its value by at most one whenever an arrival or a service completion occurs. Thus, $\{(q_K(t), q_{K-1}(t), \ldots, q_1(t)),\ t \geq 0\}$ is a level dependent quasi-birth-and-death (QBD) process with a finite number of levels. See Neuts [46], and Latouche and Ramaswami [44] for more details on QBD processes.

Next, we construct an infinitesimal generator for the Markov chain. We shall call $q_K(t)$ the level variable and $(q_{K-1}(t), \ldots, q_1(t))$ the phase variable. The states in Ω_i, $0 \leq i \leq c_K + N$, will be called level i states. The infinitesimal generator of the Markov chain $\{(q_K(t), q_{K-1}(t), \ldots, q_1(t)), t \geq 0\}$ has the following general structure:

$$Q_N^{(K)} = \begin{pmatrix} A^{(K)}_{(0,0)} & A^{(K)}_{(0,1)} & & & \\ A^{(K)}_{(1,0)} & A^{(K)}_{(1,1)} & A^{(K)}_{(1,2)} & & \\ & \ddots & \ddots & \ddots & \\ & & A^{(K)}_{(N+c_K-1,N+c_K-2)} & A^{(K)}_{(N+c_K-1,N+c_K-1)} & A^{(K)}_{(N+c_K-1,N+c_K)} \\ & & & A^{(K)}_{(N+c_K,N+c_K-1)} & A^{(K)}_{(N+c_K,N+c_K)} \end{pmatrix}. \tag{3.1}$$

Intuitively, the matrices $A^{(K)}_{(i,i+1)}$, $A^{(K)}_{(i,i-1)}$, and $A^{(K)}_{(i,i)}$ give the rate by which the number of patients at the K^{th} ED increases by one, decreases by one, or does not change, respectively. The construction of the infinitesimal generator must be done with care. The main difficulty comes from the fact that the number of states in a level depends on the level. We observe that the number of states in each level is determined by the number of ambulances available to hospitals other than K. Based on this observation, we introduce the following recursive method for constructing the matrix blocks in the infinitesimal generator $Q_N^{(K)}$. Note that, in the following construction, the variable k, $1 \leq k \leq K$, represents the number of hospitals involved (i.e., hospitals 1, 2, ..., and k), and the variable n, $0 \leq n \leq N$, represents the number of available ambulances. For $k = 1$, we have, for $n = 0$,

$$Q_0^{(1)} = \begin{array}{c|cccc} 0 & 0 & & & \\ 1 & \mu_1 & -\mu_1 & & \\ \vdots & & \ddots & \ddots & \\ c_1 & & & c_1\mu_1 & -c_1\mu_1 \end{array} ; \tag{3.2}$$

and for $n \geq 1$,

$$Q_n^{(1)} = \begin{array}{c|ccccccc} 0 & -p_1\lambda_0 & p_1\lambda_0 & & & & & \\ 1 & \mu_1 & -\mu_1 - p_1\lambda_0 & p_1\lambda_0 & & & & \\ \vdots & & \ddots & \ddots & \ddots & & & \\ c_1 & & & c_1\mu_1 & -c_1\mu_1 - p_1\lambda_0 & p_1\lambda_0 & & \\ \vdots & & & & \ddots & \ddots & \ddots & \\ c_1+n-1 & & & & & c_1\mu_1 & -c_1\mu_1 - p_1\lambda_0 & p_1\lambda_0 \\ c_1+n & & & & & & c_1\mu_1 & c_1\mu_1 \end{array} . \tag{3.3}$$

Note that, if $n = 0$, there is no ambulance available. Thus, there can be no arrival of patients in $Q_0^{(1)}$. If $n \geq 1$, the total arrival rate of patients is λ_0 and the arrival rate to the first ED is $p_1\lambda_0$. The service rate is determined by $\min\{c_1, q_1(t)\}$.

We also define the following matrices:

$$U_0^{(1)} = (0)_{(c_1+1)\times(c_1+1)}; \qquad U_n^{(1)} = \begin{pmatrix} I_{(c_1+n)\times(c_1+n)} & 0 \\ 0 & 0 \end{pmatrix}_{(c_1+n+1)\times(c_1+n+1)}, \text{ for } n \geq 1. \tag{3.4}$$

$$V_n^{(1)} = \begin{pmatrix} I_{(c_1+n)\times(c_1+n)} \\ 0 \end{pmatrix}_{(c_1+n+1)\times(c_1+n)}, \text{ for } n \geq 1. \tag{3.5}$$

To indicate the size of a matrix, we have used subscripts. For example, $(0)_{(c_1+1)\times(c_1+1)}$ is a square matrix of zeros of size $c_1 + 1$.

We define

$$U_n^{(k)} = \begin{array}{c} 0 \\ \vdots \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n \end{array} \left(\begin{array}{cccccc} U_n^{(k-1)} & & & & & \\ & \ddots & & & & \\ & & U_n^{(k-1)} & & & \\ & & & U_{n-1}^{(k-1)} & & \\ & & & & \ddots & \\ & & & & & U_0^{(k-1)} \end{array}\right), \quad \text{for } n \geq 0. \tag{3.6}$$

$$V_n^{(k)} = \begin{array}{c} 0 \\ \vdots \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n-1 \\ c_k+n \end{array} \left(\begin{array}{cccccc} V_n^{(k-1)} & & & & & \\ & \ddots & & & & \\ & & V_n^{(k-1)} & & & \\ & & & V_{n-1}^{(k-1)} & & \\ & & & & \ddots & \\ & & & & & V_1^{(k-1)} \\ & & & & & U_0^{(k-1)} \end{array}\right), \quad \text{for } n \geq 1. \tag{3.7}$$

For $2 \leq k \leq K$, we have, for $n \geq 0$ and $0 \leq i \leq n + c_k$,

$$A_{n(i,i)}^{(k)} = Q_{n-\max(0,i-c_k)}^{(k-1)} - \min(i, c_k)\mu_k I - p_k\lambda_0 U_{n-\max(0,i-c_k)}^{(k-1)}. \tag{3.8}$$

If $i_k = i$, the number of ambulances available to hospitals 1, 2, ..., and $k-1$ is $\max\{0, i - c_k\}$. Thus, the transitions for $(q_{k-1}(t), ..., q_1(t))$ are described by $Q_{n-\max\{0,i-c_k\}}^{(k-1)}$. The transitions of $q_k(t)$ are determined by $\min\{i, c_k\}\mu_k I$ for decreasing its value by one, and by $p_k\lambda_0 U_{n-\max\{0,i-c_k\}}^{(k-1)}$ for increasing its value by one.

$$A_{n(i,i+1)}^{(k)} = \begin{cases} p_k\lambda_0 U_n^{(k-1)}, & \text{for } 0 \leq i \leq c_k - 1; \\ p_k\lambda_0 V_{n-(i-c_k)}^{(k-1)}, & \text{for } c_k \leq i \leq n + c_k - 1. \end{cases} \tag{3.9}$$

Note that, for levels i and $i+1$, if $i \geq c_k$, they have different number of states. The reason is that if $i \geq c_k$, for level $i+1$, there is one less ambulance available for hospitals 1, 2, ..., and $k-1$.

$$A^{(k)}_{n(i,i-1)} = \begin{cases} \min(i, c_k)\mu_k I, & \text{for } 1 \le i \le c_k; \\ \min(i, c_k)\mu_k (V^{(k-1)}_{n+1-(i-c_k)})', & \text{for } c_k + 1 \le i \le n + c_k. \end{cases} \tag{3.10}$$

where $(V^{(k-1)}_{n+1-(i-c_k)})'$ is the transpose of $(V^{(k-1)}_{n+1-(i-c_k)})$. Then $Q^{(K)}_N$ is constructed from $A^{(K)}_{n(i,i)}$, $A^{(K)}_{n(i,i+1)}$, and $A^{(K)}_{n(i,i-1)}$ by equation (3.1) when $n = N$. We summarize the steps to construct the infinitesimal generator for the stochastic model with only the high priority ambulance patients in algorithm 1.

Algorithm 1 Computing matrix blocks in $Q^{(K)}_N$

```
1. Based on equations (3.3), (3.4), and (3.5), compute matrices {$Q^{(1)}_n$, for $0 \le n \le N$}, {$U^{(1)}_n$, for $0 \le n \le N$}, and {$V^{(1)}_n$, for $1 \le n \le N$}. Set $k = 2$.
2. If $k \le K$, go to step (3); Otherwise, Stop.
3. Based on equations (3.8), (3.9), and (3.10), compute {$A^{(k)}_{n(i,i)}$, for $0 \le n \le N$ and $0 \le i \le n + c_k$}, {$A^{(k)}_{n(i,i+1)}$, for $0 \le n \le N$ and $0 \le i \le n + c_k - 1$}, {$A^{(k)}_{n(i,i-1)}$, for $0 \le n \le N$ and $1 \le i \le n + c_k$}. Then compute {$Q^{(k)}_n$, for $0 \le n \le N$}, {$U^{(k)}_n$, for $0 \le n \le N$}, and {$V^{(k)}_n$, for $1 \le n \le N$}. Set $k =: k + 1$, Go to step (2).
```

3.2.2 Matrix-Geometric Solution

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{N+c_K})$ the stationary distribution of $Q^{(K)}_N$. Since the Markov chain is irreducible, $\boldsymbol{\pi}$ exists and is the unique non-negative solution for the linear system:

$$\boldsymbol{\pi} Q^{(K)}_N = 0; \quad \text{and} \quad \boldsymbol{\pi}\boldsymbol{e} = 1, \tag{3.11}$$

where $\boldsymbol{e}$ is a column vector of ones. Since the infinitesimal generator $Q^{(K)}_N$ has a block tridiagonal structure, a matrix-geometric solution can be obtained. First, for the levels $N + c_K$ and $N + c_K - 1$, we obtain

$$\boldsymbol{\pi}_{N+c_K} = \boldsymbol{\pi}_{N+c_K-1} R(N + c_K), \tag{3.12}$$

where

$$R(N+c_K) = -A^{(K)}_{(N+c_K-1,N+c_K)}\left(A^{(K)}_{(N+c_K,N+c_K)}\right)^{-1}. \tag{3.13}$$

Then we solve recursively starting from level $N+c_K-1$ down to level 1 to obtain:

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1}R(i), \text{ for } 1 \leq i \leq N+c_K, \tag{3.14}$$

where

$$R(i) = -A^{(K)}_{(i-1,i)}(A^{(K)}_{(i,i)} + R(i+1)A^{(K)}_{(i+1,i)})^{-1}. \tag{3.15}$$

In order to find $\boldsymbol{\pi}$, we need to find the boundary $\boldsymbol{\pi}_0$. The boundary balance equations and the normalization condition lead to the following linear system for $\boldsymbol{\pi}_0$:

$$\begin{aligned} &\boldsymbol{\pi}_0\left(A^{(K)}_{(0,0)} + R(1)A^{(K)}_{(1,0)}\right) = 0; \\ &\boldsymbol{\pi}_0(\boldsymbol{e} + R(1)\boldsymbol{e} + R(1)R(2)\boldsymbol{e} + ... + R(1)...R(N+c_K)\boldsymbol{e}) = 1. \end{aligned} \tag{3.16}$$

We summarize the solution steps as follows:

Algorithm 2 Stationary distribution of $Q^{(K)}_N$

```
1. Find $R(N+c_K)$ using equation (3.13).
2. Find $R(i)$ for $1 \leq i < N+c_K$ recursively starting from the higher level using equation (3.15).
3. Find the vector $\boldsymbol{\pi}_0$ using the boundary and normalization conditions in (3.16).
4. Find $\boldsymbol{\pi}_i$ starting from $i = 1$ up to $i = N+c_K$ using equation (3.14).
```

3.2.3 Performance Measures

A number of performance measures can be derived directly from $\boldsymbol{\pi}$. We shall focus on the performance measures for the K^{th} ED. Performance measures for other hospitals can be obtained by changing the role of another ED and the K^{th} ED in the analysis.

1. In steady state, the distribution of the number of ambulance patients q_K in the K^{th}

ED is given by

$$\pi^{(K)}(i) = \boldsymbol{\pi}_i \boldsymbol{e}, \text{ for } i = 0, 1, \ldots, N + c_K. \tag{3.17}$$

2. The mean number of ambulance patients in the K^{th} ED is given by

$$E[q_K] = \sum_{i=0}^{N+c_K} i\pi^{(K)}(i). \tag{3.18}$$

3. The probability distribution of the number of ambulances in offload delay at the K^{th} ED. We define random variable $O^{(K)}$ as the number of ambulances in offload delay at the K^{th} ED. We note that there are ambulances in offload delay at the K^{th} ED if and only if $q_K(t) > c_K$. Thus, we have $O^{(K)} = \max\{0, q_K(t) - c_K\}$. The probability distribution for the number of ambulances in offload delay can be calculated as follows:

$$P\{O^{(K)} = m\} = \begin{cases} \sum_{i=0}^{c_K} \pi^{(K)}(i), & \text{for } m = 0; \\ \pi^{(K)}(m + c_K), & \text{for } 0 < m \leq N. \end{cases} \tag{3.19}$$

The mean number of ambulances in offload delay $E[O^{(K)}]$ can be obtained accordingly.

4. For state $(i_K, ..., i_1)$, we denote by $\pi_{i_K,...,i_1}$ its steady state probability, which is an element in the vector $\boldsymbol{\pi}$. The probability distribution of the total number of ambulances in offload delay, denoted by O, is given by

$$P\{O = m\} = \sum_{(i_K,...,i_1)\in\Omega:\ \sum_{k=1}^{K} \max\{0, i_k - c_k\} = m} \pi_{i_K,...,i_1}, \quad \text{for } 0 \leq m \leq N; \tag{3.20}$$

5. The loss probability: We refer to the probability that all ambulances are in offload delay as the loss probability, denoted as P_L. Then the loss probability is given by

$$P\{O = N\} = P_L = \sum_{(i_K,...,i_1)\in\Omega:\ \sum_{k=1}^{K} \max\{0, i_k - c_k\} = N} \pi_{i_K,...,i_1}. \tag{3.21}$$

3.2.4 Waiting times of ambulance patients

The waiting time w_K of an ambulance patient arriving to the K^{th} ED depends on the number of ambulance patients waiting at the K^{th} ED. Denote by $\eta_i(K)$ the probability that i ambulance patients are in the K^{th} ED when an ambulance patient arrives in the K^{th} ED. Note that an arriving patient can reach the K^{th} ED if and only if there is an ambulance available at the time of arrival. By definition, we have, for $0 \leq i \leq c_K + N - 1$,

$$\eta_i(K) = \frac{\sum_{(i,i_{K-1},...,i_1)\in\Omega:\ \max\{0,i-c_K\}+\sum_{k=1}^{K-1}\max\{0,i_k-c_k\}<N} \pi_{i,i_{K-1},...,i_1}}{1-P_L}. \tag{3.22}$$

Define $\boldsymbol{\alpha}(K) = (\eta_{c_K}(K), ..., \eta_{c_K+N-1}(K))$. Then the i-th component of $\boldsymbol{\alpha}(K)$ gives the probability that an arriving ambulance patient to the K^{th} ED has to wait for the service completion of i patients before getting a bed. Since there are c_K beds for all patients in the K^{th} ED, each with an exponential service time with parameter μ_K, if all beds are occupied, the time to serve one patient has an exponential distribution with parameter $c_K\mu_K$. Thus is all c_k servers are busy, the total time to serve i patients has an Erlang distribution of order i. Consequently, when an ambulance patient arrives to hospital K, the waiting time has a generalized Erlang distribution with a phase-type representation $(\boldsymbol{\alpha}(K), c_K\mu_K J_N)$, where

$$J_N = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}_{N\times N}. \tag{3.23}$$

The distribution function of the waiting time w_K is given by

$$P\{w_K < t\} = 1 - \boldsymbol{\alpha}(K)\exp\{-c_K\mu_K J_N t\}\boldsymbol{e}. \tag{3.24}$$

By routine calculations, we obtain

$$E[w_K] = \sum_{i=1}^{N} \frac{i\eta_{c_K-1+i}(K)}{c_K\mu_K}. \tag{3.25}$$

The mean waiting time $E[w_K]$ and the mean queue length $E[q_K]$ satisfy Little's law as follows: $E[q_K] = \lambda_0(1 - P_L)p_K(E[w_K] + 1/\mu_K)$, which is useful for a computational accuracy check.

Denote by w the waiting time of an arbitrary ambulance patient who enters the system (i.e., is not lost). Since arriving ambulance patients are sent to individual hospitals with probabilities $\{p_1, ..., p_K\}$, the mean waiting time of an arbitrary ambulance patient who actually enters a hospital is given by $E[w] = \sum_{k=1}^{K} p_k E[w_k]$.

Since the service time in the k^{th} ED has an exponential distribution with parameter μ_k, the mean sojourn time of an ambulance patient to the k^{th} ED is given by $E[w_k]+1/\mu_k$. The mean sojourn time of an arbitrary ambulance patient who enters the system can be calculated by $\sum_{k=1}^{K} p_i(E[w_k] + 1/\mu_k)$.

3.3 Low priority walk-in patients

To account for the walk-in patients who arrive to the hospitals' EDs with lower acuity problems, we utilize the Markov chain outlined in section 4 to develop a new Markov chain model that includes both arrival streams. Due to the fact that walk-in patients across hospitals are independent, we can focus on one hospital at a time without loss of generality. We also recall that the service discipline is preemptive. Since the service time of walk-in patients is assumed to be exponential, its overall service time will be exponential regardless of whether its service is preempt-resume or preempt-repeat.

3.3.1 The Modified Markov chain

We add $q_{w,K}(t)$ to the Markov chain considered in Section 4 to obtain a continuous time Markov chain $\{(q_{w,K}(t), q_K(t), q_{K-1}(t), ..., q_1(t)), t \geq 0\}$, which has an infinite state space. Since the level variable $q_{w,K}(t)$ changes its value by at most one, decreasing by one or increasing by one, the process $\{(q_{w,K}(t), (q_K(t), q_{K-1}(t),\ ..., q_1(t))), t \geq 0\}$ is a QBD process with an infinite number of levels. Since the service discipline is preemptive, walk-in patients have no impact on the service of ambulance patients. Thus, the infinitesimal generator Q_{wK} has the following structure:

$$\begin{aligned} Q_{wK} &= I \otimes (Q_N^{(K)} - \lambda_K I) \\ &+ \begin{pmatrix} 0 & \lambda_K I & & & & & \\ M_{K,1} & -M_{K,1} & \lambda_K I & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & M_{K,c_K} & -M_{K,c_K} & \lambda_K I & & \\ & & & M_{K,c_K} & -M_{K,c_K} & \lambda_K I & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}, \end{aligned} \tag{3.26}$$

where $I \otimes (Q_N^{(K)} - \lambda_K I)$ is the Kronecker product of I and $Q_N^{(K)} - \lambda_K I$, $Q_N^{(K)}$ is defined in equation (3.1), and $M_{K,n}$ are diagonal matrices that include service rates for walk-in patients conditioning on the number of ambulance patients in the K^{th} ED: (Note: $n = q_{w,K}(t) \geq 1$)

$$M_{K,n} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c_K - 1 \\ c_K \\ \vdots \\ c_K + N \end{array} \begin{pmatrix} \min\{n, c_K\}\mu_K I & & & & & & \\ & \min\{n, c_K - 1\}\mu_K I & & & & & \\ & & \ddots & & & & \\ & & & \mu_K I & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}, \tag{3.27}$$

The diagonal elements of $M_{K,n}$ indicate the number of walk-in patients with a bed, which depends on the number of available beds and the number of walk-in patients in the K^{th} ED, and is given by $\max\{0, \min\{n, c_K - q_K(t)\}\}$. We note that the Markov chain is level dependent up to level c_K. Beyond this level, the Markov chain has a level independent structure. This allows us to find its stationary probability distribution using matrix-analytic methods.

3.3.2 Matrix-Geometric Solution

Let $\boldsymbol{\phi} = (\boldsymbol{\phi_0}, \boldsymbol{\phi_1}, \ldots)$ be the stationary probability distribution of $\{(q_{w,K}(t), (q_K(t), q_{K-1}(t), ..., q_1(t))), t \geq 0\}$. The stationary distribution exists if and only if the Markov chain is ergodic. Since the Markov chain of interest is irreducible and has a QBD structure, by Neuts [46], the Markov chain is ergodic if and only if $\lambda_K \boldsymbol{\pi e} < \boldsymbol{\pi} M_{K,c_K} \boldsymbol{e}$, which can be simplified to

$$\lambda_K + p_K \lambda_0 (1 - P_L) < c_K \mu_K. \tag{3.28}$$

Intuitively, the left hand side of equation (3.28) is the total arrival rate to the K^{th} ED and the right hand side is the potential service capacity at the K^{th} ED. Equation (3.28) ensures that there is enough capacity to serve all patients arriving to the K^{th} ED. In the rest of this paper, we assume that equation (3.28) holds. The stationary probability distribution $\boldsymbol{\phi}$ can thus be obtained by solving the linear system

$$\boldsymbol{\phi} Q_{wK} = 0; \text{ and } \boldsymbol{\phi e} = 1. \tag{3.29}$$

By Neuts (1981), the stationary distribution has a matrix geometric form:

$$\boldsymbol{\phi}_n = \boldsymbol{\phi}_{c_K} R^{n-c_K}, \quad \text{for } n \geq c_K \tag{3.30}$$

where the rate matrix R is the minimal nonnegative solution to the nonlinear equation:

$$\lambda_K I + R(Q_N^{(K)} - \lambda_K I - M_{K,c_K}) + R^2 M_{K,c_K} = 0. \tag{3.31}$$

The above equation can be solved using the logarithmic reduction algorithm of [44]. For the level dependent part of the Markov chain, the probabilities can be obtained by solving a finite level QBD process of size c_K. Details for computing $\boldsymbol{\phi}$ are given in Algorithm 3.

By routine calculations, the mean queue length of walk-in patients in the K^{th} ED can

Algorithm 3 Computation of stationary distribution for Q_{wK}

```
1. Check stability of the Markov chain using the condition (3.28). If the system is
   stable, continue with step 2; Otherwise the stationary probability distribution does
   not exist.
2. Find $R$ by solving (4.19).
3. Set $R_{c_K} = R$.
4. Find $R_n$ for $1 \leq n < c_K$ recursively starting from $n = c_K - 1$ using the equation:
   $R_n = -\lambda_K(Q_N^{(K)} - \lambda_K I - M_{K,n} + R_{n+1}M_{K,n+1})^{-1}$
5. Find the vector $\boldsymbol{\phi}_0$ using the boundary and normalizing conditions:
   $\boldsymbol{\phi}_0(Q_N^{(K)} - \lambda_K I + R_1 M_{K,1}) = 0$,
   $\boldsymbol{\phi}_0(I + R_1 + R_1R_2 + \ldots + R_1R_2\ldots R_{c_{K-1}} + R_1R_2\ldots R_{c_K}(I - R)^{-1})\boldsymbol{e} = 1$.
6. Find $\boldsymbol{\phi}_n$ starting from $n = 1$ up to $n = c_K$ using equation:
   $\boldsymbol{\phi}_n = \boldsymbol{\phi}_{n-1}R_n$ for $1 \leq n \leq c_K$.
7. Find $\boldsymbol{\phi}_n$ for $n > c_K$ using equation (3.30).
```

be obtained as

$$E[q_{w,K}] = \sum_{n=0}^{c_K-1} n\boldsymbol{\phi}_n\boldsymbol{e} + \boldsymbol{\phi}_{c_K}\left(R(I-R)^{-2} + c_K(I-R)^{-1}\right)\boldsymbol{e}. \tag{3.32}$$

3.3.3 Sojourn Times of Walk-in Patients

We now construct a continuous time Markov chain for analyzing the sojourn time of a walk-in patient. Since a walk-in patient may get a bed and then lose it a number of times prior to leaving the hospital, the waiting time is less meaningful than the sojourn time, $w_{w,K}$, the total time that a walk-in patient is in the ED.

To construct the absorbing Markov chain for the sojourn time of a tagged walk-in patient, we only need to consider those walk-in patients who arrived before the tagged walk-in patient. The Markov chain is terminated when the tagged walk-in patient completes its service. If the tagged walk-in patient occupies a bed, the service is completed at the rate μ_K. The tagged walk-in patient may be pushed out of a bed a number of times by ambulance patients before the completion of service. Again, we recall that the service to ambulance patients is not affected by that of walk-in patients. We define, for

$0 \leq n \leq c_K - 1$,

$$T_{n,w} = \begin{pmatrix} Q_N^{(K)} - M_{K,1} & & & \\ M_{K,1} & Q_N^{(K)} - M_{K,2} & & \\ & \ddots & \ddots & \\ & & M_{K,n} & Q_N^{(K)} - M_{K,n+1} \end{pmatrix}, \tag{3.33}$$

and, for $n \geq c_K$,

$$T_{n,w} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c_K \\ \vdots \\ n \end{array} \begin{pmatrix} Q_N^{(K)} - M_{K,1} & & & & & \\ M_{K,1} & Q_N^{(K)} - M_{K,2} & & & & \\ & \ddots & \ddots & & & \\ & & M_{K,c_K} & Q_N^{(K)} - M_{K,c_K} & & \\ & & & \ddots & \ddots & \\ & & & & M_{K,c_K} & Q_N^{(K)} - M_{K,c_K} \end{pmatrix}. \tag{3.34}$$

Given that there are n walk-in patients already in the hospital when a tagged walk-in patient arrives, the tagged patient's sojourn time has a phase-type distribution with matrix representation $((0, ..., 0, \boldsymbol{\phi}_n/(\boldsymbol{\phi}_n\mathbf{e}), T_{n,w})$. Note that, if the phase in $T_{n,w}$ is $c_K - 1$ or less, the tagged patient is in service and may complete its service earlier than other patients in service. Then we obtain the conditional probability distribution of the sojourn time as:

$$P(w_{w,K} \leq t \mid n) = 1 - (0, ..., 0, \boldsymbol{\phi}_n/(\boldsymbol{\phi}_n\boldsymbol{e})) \exp\{T_{n,w}t\}\boldsymbol{e}. \tag{3.35}$$

The distribution of the sojourn time of an arbitrary walk-in patient can be obtained as

$$P(w_{w,K} \leq t) = 1 - \sum_{n=0}^{\infty}(0, ..., 0, \boldsymbol{\phi}_n) \exp\{T_{n,w}t\}\boldsymbol{e}. \tag{3.36}$$

By using truncation, the above formula can be used for computing the distribution of sojourn time. As for the mean sojourn time, the following explicit formula can be obtained, where the computation is finite as long as the matrix R can be obtained. Define $D_K = -(Q_N^{(K)} - M_{K,c_K})^{-1}M_{K,c_K}$, and $A_K = -(Q_N^{(K)} - M_{K,c_K})^{-1}$, and for $0 \leq n \leq c_K - 1$,

$$\begin{aligned} B_n \;=\; & -(Q_N^{(K)} - M_{K,n+1})^{-1} + (Q_N^{(K)} - M_{K,n+1})^{-1}M_{K,n}(Q_N^{(K)} - M_{K,n})^{-1} \\ & - ... + (-1)^{(n+1)}(Q_N^{(K)} - M_{K,n+1})^{-1}M_{K,n}(Q_N^{(K)} - M_{K,n})^{-1}...M_{K,1}(Q_N^{(K)} - M_{K,1})^{-1}. \end{aligned} \tag{3.37}$$

By routine calculations, the mean sojourn time can be found as, for $0 \leq n \leq c_K - 1$,

$$E[w_{w,K}|n] = -(0, ...0, \boldsymbol{\phi}_n/(\boldsymbol{\phi}_n e))T_{n,w}^{-1}\boldsymbol{e} = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n \boldsymbol{e}} B_n \boldsymbol{e}, \tag{3.38}$$

and, for $n \geq c_K$,

$$E[w_{w,K}|n] = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n \boldsymbol{e}}(A_K + D_K A_K + D_K^2 A_K + \ldots + D_K^{n-c_K} A_K + D_K^{n-c_K+1} B_{c_K-1})\boldsymbol{e}. \tag{3.39}$$

which can be reduced to,

$$E[w_{w,K}|n] = \frac{\boldsymbol{\phi}_n}{\boldsymbol{\phi}_n \boldsymbol{e}}\left((I - D_K^{n-c_K+1})(I - D_K)^{-1} A_K + D_K^{n-c_K+1} B_{c_K-1}\right)\boldsymbol{e}. \tag{3.40}$$

For an arbitrary walk-in patient at the K^{th} ED, we obtain:

$$\begin{aligned} E[w_{w,K}] \;\; = \;\; & \textstyle\sum_{n=0}^{c_K-1} \boldsymbol{\phi}_n e E[w_K|n] \\ & + \boldsymbol{\phi}_{c_K}(I-R)^{-1}(I-D_K)^{-1}A_k \boldsymbol{e} \\ & + \boldsymbol{\phi}_{c_K}(\textstyle\sum_{n=0}^{\infty} R^n D_K^n) D_K (I - D_K)^{-1} A_k \boldsymbol{e} \\ & + \boldsymbol{\phi}_{c_K}(\textstyle\sum_{n=0}^{\infty} R^n D_K^n) D_K B_{c_K-1})\boldsymbol{e}. \end{aligned} \tag{3.41}$$

The infinite summation in equation (3.41) can be transformed into the following form by using a direct sum $f(.)$:

$$f\left(\sum_{n=0}^{\infty} R^n D_K^n\right) = \sum_{n=0}^{\infty} f(I)(R' \otimes D_k)^n = f(I)(I - R' \otimes D_K)^{-1}. \tag{3.42}$$

Note: 1) The direct sum $f(X)$ of X is a row vector and is obtained by stringing out the vectors starting from the first row of X; 2) $R' \otimes D_K$ is the Kronecker product of matrices R' and D_K. Consequently, computation of $E[w_{w,K}]$ involves only finite summations and can be done efficiently if the sizes of the matrices involved are moderate. If the sizes of the matrices involved are large, the following recursive method can be used for computing

$E[w_{w,K}]$:

$$\begin{aligned}
&B_0 = -(Q_N^{(K)} - M_{K,1})^{-1};\\
&x_0 = \boldsymbol{\phi}_0 B_0 \boldsymbol{e};\\
&B_n = -(Q_N^{(K)} - M_{K,n+1})^{-1}(I + M_{K,n}B_{n-1}),\ \text{for } 1 \le n \le c_K - 1;\\
&x_n = x_{n-1} + \boldsymbol{\phi}_n B_n \boldsymbol{e},\ \text{for } 1 \le n \le c_K - 1;\\
&B_n = A_K + D_K B_{n-1},\ \text{for } n \ge c_K.\\
&x_n = x_{n-1} + \boldsymbol{\phi}_{c_K} R^{n-c_K} B_n \boldsymbol{e},\ \text{for } n \ge c_K.
\end{aligned} \tag{3.43}$$

By definition, we have $\lim_{n\to\infty} x_n = E[w_{w,K}]$. This approach requires truncation, which can be done properly since the matrix-geometric solution $\{\boldsymbol{\phi}_n,\ n \ge 0\}$ has a geometric decay. Details are omitted.

Similar to the mean queue length and mean waiting time for patients arriving by ambulance, Little's law applies to the mean queue length $E[q_{w,k}]$ and mean sojourn time $E[w_{w,k}]$, i.e., $E[q_{w,k}] = \lambda_k E[w_{w,k}]$. Thus, computing one gives the other. Little's law can be used for an accuracy check if both are computed separately. Since all computations in this section, as well as in Section 4, involve large size matrices, it is important to compute both $E[q_{w,k}]$ and $E[w_{w,k}]$ and use Little's law to check the accuracy of the computations.

Remark: We remark that the waiting time of a tagged walk-in patient (i.e., the time from the arrival of the patient until the first time that the patient gets a bed) can be studied similarly. Absorbing Markov chains can be constructed in the same way, except that only states without a bed available to the tagged patients are kept. Details are omitted.

Kao and Narayanan [48] consider a multiprocessor single node queue and two types of jobs with one having preemptive priority over the other. To find the waiting time distribution for the low priority jobs; they find two distributions: the time spent waiting in the queue until reaching a server, and the time elapsed between the epoch when the job reaches the server for the first time and the epoch it departs the system. Our approach described above is more efficient.

3.4 Numerical Analysis

Using the methods developed in Sections 3 and 4, we analyze four cases for which the models imitate the ROW EMS-ED network mentioned in the introduction and depicted in Figure 1. The four cases are mainly differ in the number of ambulances and beds available, which are selected to reflect different design scenarios in the ROW. We also test the possibility to balance workloads between EDs by adjusting the routing probabilities of ambulances. For each case study, we calculate first the stationary probability distributions of the queue lengths and then the performance measures for ambulance patients and walk-in patients, followed by a discussion on observations.

3.4.1 Parameter Selection

Parameter selection for each of the four case studies is guided by the the ROW project. The the ROW EMS-ED network consists of one EMS provider and three hospitals; Grand River General Hospital, St. Mary's General Hospital, and Cambridge Memorial Hospital. In order to mimic the real network, we utilize the available data from one of the regional hospitals. The four case studies are developed with the following features.

1. Case study 1 represents a small network that experience low offload delays.
2. Case study 2 represents a medium sized network in which significant offload delays are incurred. For this case study, we also investigate the effect of ambulance patients routing probabilities on total offload delays experienced by the EMS.
3. Case study 3 represents the case that is most close to the ROW EMS-ED system. For this case study, we investigate the effect of service rates on offload delays.
4. Case study 4 represents a similar network size as case study 3. The main difference is the higher service rates at the regional EDs.

More specifically, the system parameters for the four case studies are selected as follows:

Patient arrival rates We use different arrival rates for ambulance patients to generate different workloads to the EMS and regional EDs. On the other hand, we use available data to estimate the arrival rates of walk-in patients. To do so, we use the current EDs utilizations as provided by the regional hospitals to calibrate our model, and to find the corresponding ED walk-in patients arrival rates.

Routing probabilities To calculate the routing probability vector $\{p_1, p_2, p_3\}$ for the region's three hospitals, we use the 2006 data for the numbers of EMS visits per year for individual EDs that were (8900, 5700, 5200) visits per year, respectively. This corresponds to 45% of the arrivals being transferred to Grand River General Hospital, 29% to St. Mary's General Hospital, and 26% to Cambridge Memorial Hospital.

Number of beds at EDs The numbers of physical beds at the three EDs in ROW are 39, 34, 23. We approximate the service capacity at each ED to be about 36%, 50% and 60% of the total number of beds for case study 1, 2, and 3, respectively. The reason for the use of a smaller number of beds is that the service to patients consists of beds, nurses, doctors, and other necessary resources.

Service rates at EDs The service rate at each ED, μ_k, corresponds to the reciprocal of the Length Of Stay (LOS) of patients in the corresponding ED which is approximately 6 hours as reported by the Grand River General Hospital. We use this information for the first three cases. For case study 4, we change EDs service rates to observe their effect on EDs' performance measures.

To compare between EDs in each case study, we define two types of server utilization for the k^{th} ED, for $1 \leq k \leq K$:

- ED utilization for ambulance patients $\rho_{a,k}$: Since the service of ambulance patients is not affected by walk-in patients, we can define the server utilization for ambulance patients alone. Define $\rho_{a,k} = \min\{1, \lambda_0 p_k(1 - P_L)/(c_k \mu_k)\}$, where $\lambda_0 p_k(1 - P_L)$ is the arrival rate of ambulance patients to the k^{th} ED, and $c_k \mu_k$ is the total service capacity at the k^{th} ED.

- ED total utilization ρ_k: Considering the service of both types of patients, the server utilization can be defined as $\rho_k = \min\{1, (\lambda_0 p_k(1 - P_L) + \lambda_k)/(c_k \mu_k)\}$.

Parameter set	value
N	6
(λ_0)patients/hr	1.5
$(\lambda_1, \lambda_2, \lambda_3)$ patient/hr	$(1.7, 1.4, 0.8)$
(μ_1, μ_2, μ_3) patient/hr	$(1/6, 1/6, 1/6)$
(c_1, c_2, c_3)	$(15, 12, 8)$
(p_1, p_2, p_3)	$(0.45, 0.29, 0.26)$
$(\rho_{a,1}, \rho_{a,2}, \rho_{a,3})$	(27%, 22%, 29%)
(ρ_1, ρ_2, ρ_3)	(95% , 91.75% , 89.25%)

Table 3.1: System parameters for Case Study 1

	Matrix analytic results		
Measures	$k = 1$	$k = 2$	$k = 3$
$E[q_{a,k}]$	4.05	2.61	2.34
$E[O^{(k)}]$	$8.7 * 10^{-6}$	$5.4 * 10^{-6}$	$1.3 * 10^{-3}$
$E[w_{a,k}]$	$1.29 * 10^{-6}$	$1.25 * 10^{-6}$	$3.2 * 10^{-3}$
P_L	$1.35 * 10^{-6}$		
$E[q_{w,k}]$	24.10	16.06	10.44
$E[w_{w,k}]$	14.17	11.47	13.06

Table 3.2: Performance measures for Case Study 1

3.4.2 Case Study 1

The system parameters used in this case are recorded in Table 3.1. The results are reported in Table 3.2.

As we can see from the results, the waiting times of ambulance patients and walk-in patients are quite different. For ambulance patients, the mean waiting times (offload delays) are almost zero. For walk-in patients, the mean waiting times are more than 11 hours in all three EDs. The reason is that the utilizations for the two types of patients are quite different. As shown in Table 1, the utilizations for ambulance patients only are less than 30%, which implies that there is enough service capacity to serve all incoming ambulance patients when they arrive at an ED. On the other hand, the total utilizations for both types of patients are about 90% or higher at the EDs. The results show clearly the effect of the priority service discipline on the waiting times of all patients and the offload delays of ambulances.

The results also show the effect of routing on the waiting time of walk-in patients. Since 45% of ambulance patients are transported to the first ED resulting in the highest utilization for that ED, the corresponding ED waiting walk-in patients mean waiting time is the longest. On the other hand, the third ED has the least service capacity. Although it is small, the ambulance patients sent to the third ED has the longest waiting time.

This case study shows that different admitting policies into the EDs have a great impact on the waiting of both types of patients. Assigning a higher priority for ambulance patients decreases their waiting time and, consequently, decreases the offload delays experienced by the ambulances, but at the expense of increased waiting times for walk-in patients.

3.4.3 Case Study 2

In this case, we investigate the impact of routing probabilities $\{p_1, \ldots, p_K\}$ on system performance, which is relevant to the ROW EMS-ED network. To do so, we consider two scenarios. The first scenario represents the current unbalanced system in ROW, for which the service capacity is scaled down to 50% of the full capacity. The second scenario corresponds to a proposed balanced system for which the routing probabilities are proportional to the EDs' capacities. Specifically, we set $p_k = c_k\mu_k/(c_1\mu_1 + c_2\mu_2 + c_3\mu_3)$ for $k = 1, 2, 3$. For the two scenarios, the patient arrival rates are the same. The system parameters used in this case are recorded in Table 3.3.

The results, which are recorded in Table 3.4 for both scenarios, show how balancing the utilizations $\{\rho_{a,1}, \rho_{a,2}, \rho_{a,3}\}$, has balanced the numbers of ambulances in offload delays at EDs. More interestingly, the expected total number of ambulances in offload delay is decreased from 3.42 (i.e., $\sum_{k=1}^{3} E[O^{(k)}] = 1.68+0.16+1.58$) in the current scenario to 2.92 (=0.83+0.93+1.16) ambulances in the balanced scenario, which corresponds to a 14% decrease in the number of ambulances in offload delays. The total expected offload delay (i.e., $\sum_{k=1}^{3} p_k E[w_{a,k}]$) is decreased from 0.54 hours to 0.45 hours in the balanced scenario. This corresponds to a 9.9% decrease in the total hours of offload delays experienced in the region. The loss probability P_L is decreased from 6.93% in the current scenario to

Parameter set	value
N	9
(λ_0) patient/hr	7
$(\lambda_1, \lambda_2, \lambda_3)$ patient/hr	$(0.3, 0.6, 0.23)$
(μ_1, μ_2, μ_3) patient/hr	$(1/6, 1/6, 1/6)$
(c_1, c_2, c_3)	$(20, 17, 12)$

Table 3.3: System parameters for Case Study 2

Performance measure	Current			Balanced		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
p_k	45%	29%	26%	40.82%	34.69%	24.49%
$\rho_{a,k}$	87.95%	66.68%	84.69%	81.45%	81.44%	81.45%
ρ_k	96.95%	87.86%	96.19%	90.45%	100%	92.95%
$E[q_{a,k}]$	19.27	11.50	11.74	17.12	14.78	10.93
$E[O^{(k)}]$	1.68	0.16	1.58	0.83	0.93	1.16
$E[w_{a,k}]$	0.60	0.09	0.93	0.32	0.43	0.71
P_L	6.93%			4.98%		
$E[q_{w,k}]$	18.12	7.46	15.34	5.33	–	7.75
$E[w_{w,k}]$	60.40	12.43	66.70	17.77	–	33.70

Table 3.4: Performance measures for Case Study 2

4.98% in the balanced scenario.

In addition, from the EMS perspective, decision makers are interested in finding the routing probabilities for which the total number of ambulances in offload delay is reduced. Figure 5.4 presents the distributions of ambulances in offload delay under both the current and balanced scenarios. Under the current scenario, the probability of zero ambulances in offload delay is 29% while under the balanced scenario this probability increases to 35%, which is a significant increase in the availability of ambulances to deal with sudden events for which a number of ambulances has to be used.

While the benefit to ambulance patients is clear, the impact of balancing the utilization of ambulance patients on walk-in patients is negative for the second ED. As shown in Table 3.3, the total utilizations of the second ED is 100% for the balanced scenario. Then the queue of walk-in patients can be very long. Consequently, the routing mechanism has to be adjusted for implementation in practice. Nevertheless, the results indicate a possible direction for reducing offload delays of ambulance patients, without increasing service

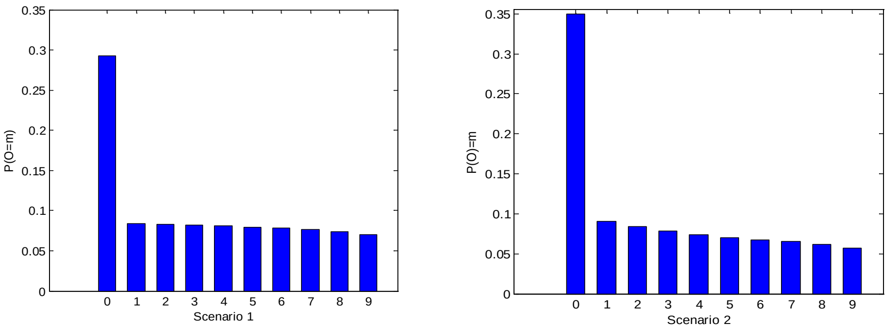

Figure 3.3: The distribution for the total number of ambulances in offload delay

Parameter set	value
N	16
(λ_0) patient/hr	7
$(\lambda_1, \lambda_2, \lambda_3)$ patient/hr	$(0.75, 0.9, 0.5)$
(c_1, c_2, c_3)	$(24, 21, 16)$
(p_1, p_2, p_3)	$(0.45, 0.29, 0.26)$

Table 3.5: System parameters for Case Study 3

capacity.

3.4.4 Case Study 3

In this case study, we increase the number of ambulances to 16, which is close to 18, the total number of ambulances available in ROW. We set the service capacity to be 60% of the numbers of beds available at ROW. We vary the mean service time from $(1/6, 1/6, 1/6)$ to $(1/5, 1/5, 1/5)$ to observe the effect of increasing the service capacity on the model output. Increasing the service rate or increasing the number of servers have similar effects on the performance measures because both variations correspond to increasing the service capacity at the destination EDs. The system parameters for this case study are reported in Table 3.5.

The results, which are recorded in Table 3.6, indicate that the EMS provides enough ambulances and three hospitals provide ample capacities to serve ambulance patients. Thus, the loss probability P_L is quite small. The waiting times of ambulance patients are short as well. This is consistent with the actual situation in ROW. On the other hand,

Performance measure	Current			Increased capacity		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
μ_k	1/6	1/6	1/6	1/5	1/5	1/5
$\rho_{a,k}$	78.75%	58.00%	68.25%	65.63%	48.33%	56.88%
ρ_k	97.50%	83.71%	87.00%	81.25%	69.76%	72.50%
$E[q_{a,k}]$	19.52	12.19	11.14	15.82	10.15	9.14
$E[O^{(k)}]$	0.64	0.02	0.23	0.07	0.00	0.04
$E[w_{a,k}]$	0.20	0.01	0.13	0.02	$9.32 * 10^{-5}$	0.04
P_L	$9.01 * 10^{-4}$			$1.6 * 10^{-5}$		
$E[q_{w,k}]$	20.85	7.10	5.98	4.74	4.69	2.90
$E[w_{w,k}]$	27.80	7.89	11.95	6.32	5.21	5.79

Table 3.6: Performance measures for Case Study 3

both the queue lengths and waiting times of walk-in patients are significant.

We also record the results when the service rate of each of the three EDs is increased from 1/6 to 1/5 in Table 3.6. As it is shown in the table, the total offload delays, walk-in patients waiting and expected queue lengths decreases as the service capacity increases. Compared to the ambulance patients, the walk-in patients waiting time decreases more drastically. More interestingly, we notice that the EDs with higher utilizations benefit more from adding more capacity to the system (e.g. the first ED performance change is the highest and the second ED change is the lowest among the three EDs).

This case study shows how our model can be used to assess the effect of adding more capacity to the system. It also shows where to add resources in order to maximize system performance.

3.4.5 Case study 4

This case study is similar to case study 3 in terms of the number of ambulances and number of beds at each ED. Unlike the previous cases, where we set the mean service time at EDs to be identical, we set (μ_1, μ_2, μ_3) to be $(1/5, 1/6, 1/5)$ respectively to observe the effect of different service rate on the system performance measures. The input parameters for this case study are reported in Table 3.7.

Table 3.8 shows the analytic results. We notice that the first ED, which has the

Parameter set	value
N	16
(λ_0) patient/hr	7
$(\lambda_1, \lambda_2, \lambda_3)$ patient/hr	$(1.4, 1.2, 1.0)$
(μ_1, μ_2, μ_3) patient/hr	$(1/5, 1/6, 1/5)$
(c_1, c_2, c_3)	$(24, 21, 16)$
(p_1, p_2, p_3)	$(0.45, 0.29, 0.26)$
$(\rho_{a,1}, \rho_{a,2}, \rho_{a,3})$	$(65.63\%, 58.00\%, 56.88\%)$
$(\rho_1, \rho_2.\rho_3)$	$(94.75\%, 91.95\%, 88.19\%)$

Table 3.7: System parameters for Case Study 4

Performance measure	Matrix analytic results		
	$k = 1$	$k = 2$	$k = 3$
$E[q_{a,k}]$	15.82	12.20	9.14
$E[O^{(k)}]$	0.07	0.02	0.04
$E[w_{a,k}]$	0.02	0.01	0.02
P_L	$1.86 * 10^{-5}\%$		
$E[q_{w,k}]$	18.73	13.98	8.86
$E[w_{w,k}]$	13.38	11.65	8.86

Table 3.8: Performance measures for Case Study 4

highest utilization, experiences the most offload delays, while the second ED experiences the least. In the previous case studies, where the service rates were set to be identical, we noticed that the expected offload delays experienced by each ED are directly related to the specific ED utilization $\rho_{a,k}$. In this case study, this observation does not hold, for example, the second ED utilization is higher than the third ED utilization ($\rho_{a,2}\rho_{a,3}$, while the expected offload delays are lower for the second ED than the third ED. This is due to the difference in the service rate among the EDs.

From walk-in patients perspective, the expected queue length and the expected waiting time are the highest for the first ED which has the highest utilization ($\rho_{w,k}$).

We note here that for cases with a small and moderate state space (e.g., Case 1 and Case 2), the matrix-analytic methods are effective and efficient. On the other hand, for large size problems, the efficiency of matrix-analytic methods is limited by the computer physical memory needed for storing matrices. For such cases, the classical Gauss-Seidel iteration can be used for computing the stationary distributions of queue lengths. However, the matrix-analytic methods is more efficient than the Gauss-Seidel iteration for small and moderate cases.

In the next section, we validate two main assumption made by this model using discrete event simulation.

3.5 Model Validation

In this section, through simulation, we validate two assumptions made in Section 3.2:

1. the transit time of an ambulance patient is zero; and
2. the service times in EDs have an exponential distribution.

We use the three case studies in Section 5 as the base models for model validation. Then we add transit time into the queueing network or change the service time distribution from exponential to more general distributions. The extended models are analyzed through simulation. Performance measures are collected for the original models (Section 2) and for the extended models. Then we compare the results in order to validate the assumptions. Of course, the assumptions are validated if the performance measures collected for the two groups of models are close to each other.

The assumption on transit time First, we consider an extended model in which the transit time of ambulance patients is nonzero. Real data on transit times from the ROW EMS database is used. By using the Stat-Fit Package, it is found that the transit time can be approximated by a beta distribution with parameters ($\alpha = 2.75$, $\beta = 22.9$) and a coefficient of variation of 0.5. See Figure (3.4) for the fitted data from the ROW EMS. We also use the exponential distribution to model that transit time since it was used in the literature, e.g. ([49], [50]). We use a parameter of $\mu = 1/0.73$. We define u_A the utilization of ambulances in the EMS, which is the long-term percentage of ambulances being used. For the zero transit time case, ambulances are busy only when they are experiencing offload delays. Mathematically, $u_A = E[O]/N$. While for the nonzero transit time case,

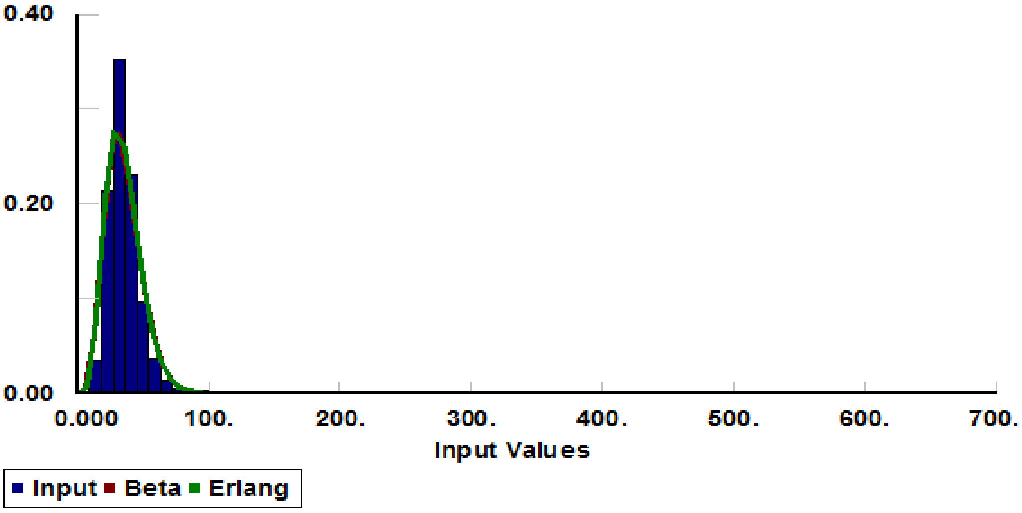

Figure 3.4: The fitted distribution for ambulance transit time

an ambulance is busy if it is either transferring an patient or waiting outside an ED. The EMS utilization in this case is collected from the simulation output.

Through simulation, performance measures of the system with nonzero transit time are collected (including the EMS utilization). Results are presented in Table 3.9. Also presented in Table 3.9 are the results for the zero transit time case. Results are presented for all three cases considered in Section 3.5. We have the following observations.

System performance measure	zero transit time			Beta $(2.75, 22.9)$			Exponential (0.73 hr)		
	Case study 1								
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
$E[q_{a,k}]$	4.05	2.61	2.34	4.04(0.01)	2.61(0.01)	2.33(0.01)	4.04(0.01)	2.61(0.01)	2.33(0.01)
$E[O^{(k)}]$	$8.7*10^{-6}$	$5.4*10^{-6}$	$1.3*10^{-3}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[w_{a,k}]$	$1.29*10^{-6}$	$1.25*10^{-6}$	$3.2*10^{-6}$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[q_{w,k}]$	24.10	16.06	10.44	24.99(0.56)	16.01(0.24)	10.44(0.19)	25.23(0.61)	15.96(0.19)	10.51(0.24)
$E[w_{w,k}]$	14.17	11.47	13.06	14.70(0.33)	11.42(0.18)	13.05(0.24)	14.84(0.36)	11.40(0.14)	13.14(0.29)
P_L	$1.35*10^{-6}$			$8.88*10^{-4}(0.43*10^{-4})$			$8.10*10^{-4}(0.41*10^{-4})$		
u_A	0.02%			18.61%(0.03%)			18.27%(0.03)		
Case study 2									
$E[q_{a,k}]$	19.27	11.50	11.74	17.10(0.03)	10.72(0.03)	10.20(0.03)	17.17(0.02)	10.74(0.03)	10.25(0.02)
$E[O^{(k)}]$	1.68	0.16	1.58	0.59(0.01)	0.07(0.01)	0.66(0.01)	0.62(0.01)	0.07(0.00)	0.68(0.01)
$E[w_{a,k}]$	0.60	0.09	0.93	0.23(0.01)	0.05(0.01)	0.43(0.01)	0.22(0.01)	0.04(0.01)	0.42(0.02)
$E[q_{w,k}]$	18.12	7.46	15.34	5.25(0.07)	5.56(0.04)	5.08(0.12)	5.43(0.06)	5.63(0.05)	5.26(0.11)
$E[w_{w,k}]$	60.40	12.43	66.70	17.53(0.25)	9.26(0.07)	22.11(0.51)	18.08(0.23)	9.38(0.08)	22.87(0.47)
P_L	6.93%			12.58%(0.03%)			12.40%(0.04%)		
u_A	38.00%			65.22%(0.05%)			64.83%(0.05%)		
Case study 3									
$E[q_{a,k}]$	19.52	12.19	11.14	19.33(0.05)	12.14(0.05)	11.07(0.04)	19.33(0.06)	12.14(0.02)	11.07(0.04)
$E[O^{(k)}]$	0.64	0.02	0.23	0.52(0.01)	0.02(0.00)	0.20(0.01)	0.52(0.01)	0.02(0.01)	0.20(0.01)
$E[w_{a,k}]$	0.20	0.01	0.13	0.17(0.01)	0.01(0.01)	0.11(0.02)	0.17(0.01)	0.01(0.01)	0.11(0.02)
$E[q_{w,k}]$	20.85	7.10	5.98	26.98(0.25)	6.83(0.04)	5.89(0.08)	27.29(1.39)	6.82(0.05)	5.86(0.08)
$E[w_{w,k}]$	27.80	7.89	11.95	36.50(0.40)	7.71(0.03)	12.08(0.06)	36.93(0.47)	7.66(0.02)	12.11(0.06)
P_L	$9.01*10^{-4}$			$4.8*10^{-3}(0.00)$			$4.7*10^{-3}(0.00)$		
u_A	5.56%			36.42%(0.08)			36.42%(0.02)		

Table 3.9: Effects of nonzero transit time (Note: The 95% confidence interval half widths for simulation in parentheses)

- The results in Table 3.9 supports the assumption that zero transit time has negligible effect on the offload delays experienced by ambulances for case studies 1 and 3, where the ambulance utilization u_A is small or moderate (i.e., 18% and 36%).

- When the ambulances are highly utilized as in case study 2 (i.e., 65%), the probability of losing patients increases significantly when the transit time becomes nonzero. The offload delays does not change significantly, but the waiting times of walk-in patients are changed dramatically. In fact, due to losing about 12% of the ambulance patients, walk-in patients get their service more quickly (i.e., $E[w_{w,k}]$ is smaller).

- Both the beta and exponential distributions have given similar results in terms of system performance measures.

- For case study 1 (low offload delays case) the simulation did not show any possibility of offload delays at the three EDs. While the analytic method gives, for example, that the third ED expected offload delay is $3.2 * 10^{-3}$, which corresponds to 13.82 ambulance hours per month. This demonstrates a limitation of the simulation approach, which is the difficulty in capturing rare events.

The assumption on service times The second assumption we want to validate is the exponential service time for serving patients at the EDs. The data we have from one of the regional hospitals in ROW is for the flow time of patients, so it includes patients' delays in addition to service time. To approximate the service time distribution, we fit flow time data using the Stat-Fit package. The resulting distribution is Erlang and is shown in Figure 3.5. We assume that the service time has a similar distribution to the flow time but with different parameters. Then the Erlang distribution can be a good candidate for the service time distribution.

Since the Erlang distribution does not have the memoryless property, the preemptive repeat and the preemptive resume give different results. We assume preemptive resume in this section, which is closer to the practice in the EDs.

In Table 3.10, analytical and simulation results are reported for the first three case studies in Section 5, where the service time is exponential or Erlang with the same mean.

We have the following observations.

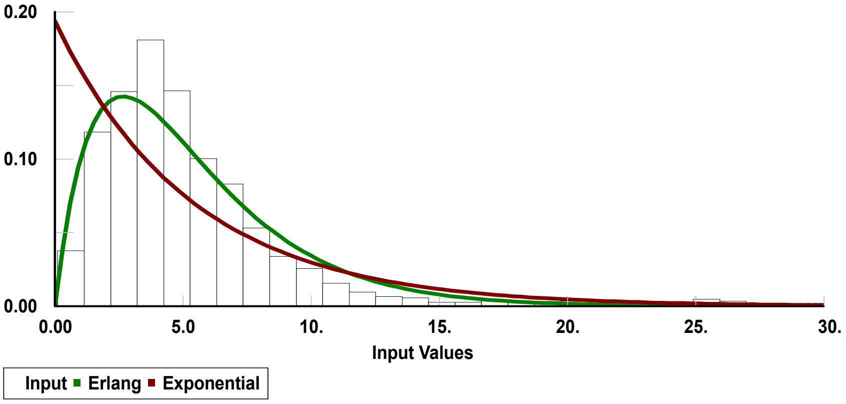

Figure 3.5: The fitted distribution for patient flow time

- Due to the lower coefficient of variation for the Erlang distribution, expected queue lengths and consequently, expected waiting times for both ambulance and walk-in patients are slightly lower under the Erlang service time distribution (for case 1 and 3 only). Thus, our assumption of exponentially distributed service time gave an upper bound on the system performance measures.

- Another observation we have with respect to case study 2 is the significant increase in walk-in patients expected sojourn time and queue lengths at all EDs. This is because under the Erlang distribution service time, which has less coefficient of variation, more high priority ambulance patients are accepted (P_L decreased). As a result, the low priority walk-in patients queue lengths and consequently, waiting times are affected significantly.

In summary, if the loss probability is small, performance measures for both types of patients are not affected significantly by adding the transit time or by changing the service time distribution. In reality, ambulances usually operate at around $u_A = 35\%$ utilization [11] including transit time which is similar to case study 3. For such a case, the loss probability is small. This indicates that the queueing network introduced in the paper is robust as long as the system of interest is working under normal operating conditions. In

Performance measure	Exponential			Erlang $M = 2$		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Case study 1						
$E[q_{a,k}]$	4.05	2.61	2.34	4.05(0.01)	2.61(0.01)	2.34(0.01)
$E[O^{(k)}]$	$8.7 * 10^{-6}$	$5.4 * 10^{-6}$	$1.3 * 10^{-3}$	0.00(0.00)	0.00(0.00)	0.00(0.00)
$E[w_{a,k}]$	$1.29 * 10^{-6}$	$1.25 * 10^{-6}$	$3.2 * 10^{-6}$	0.00(0.01)	0.00(0.01)	0.00(0.01)
$E[q_{w,k}]$	24.10	16.06	10.44	22.44(0.43)	14.78(0.24)	9.63(0.17
$E[w_{w,k}]$	14.17	11.47	13.06	13.20(0.24)	10.56(0.14)	12.03(0.21)
P_L	$1.35 * 10^{-6}$			$1.00 * 10^{-6}(1.17 * 10^{-6})$		
Case study 2						
$E[q_{a,k}]$	19.27	11.50	11.74	19.5(0.03)	11.63(0.03)	11.83(0.02)
$E[O^{(k)}]$	1.68	0.16	1.58	1.69(0.01)	0.15(0.01)	1.54(0.01)
$E[w_{a,k}]$	0.60	0.09	0.93	0.57(0.01)	0.08(0.01)	0.90(0.01)
$E[q_{w,k}]$	18.12	7.46	15.34	27.76(1.65)	7.54(0.11)	21.15(0.93)
$E[w_{w,k}]$	60.40	12.43	66.70	86.39(2.34)	12.57(0.20)	92.00(3.98)
P_L	6.93%			5.77%(0.05%)		
Case study 3						
$E[q_{a,k}]$	19.52	12.19	11.14	19.43(0.03)	12.20(0.02)	11.09(0.03)
$E[O^{(k)}]$	0.64	0.02	0.23	0.54(0.01)	0.02(0.00)	0.18(0.01)
$E[w_{a,k}]$	0.20	0.01	0.13	0.17(0.01)	0.01(0.00)	0.10(0.01)
$E[q_{w,k}]$	20.85	7.10	5.98	32.63(1.97)	6.94(0.05)	5.58(0.07)
$E[w_{w,k}]$	27.80	7.89	11.95	44.29(2.57)	7.74(0.04)	11.52(0.11)
P_L	$9.01 * 10^{-4}$			$4.5 * 10^{-4}\%(3.1 * 10^{-5}\%)$		

Table 3.10: Service time distribution effect(95% confidence interval half widths in parentheses)

other words, the analysis in this section indicates that the assumptions made in Section 2 are valid as long as the ambulance utilization is not too high, which is the actual condition under which the EMS operates.

3.6 Conclusion

In this chapter, we modeled ambulance offload delays for a multiple ED network. We assumed that the transit times of ambulance patients are negligible. We also assumed that ambulance patients have higher preemptive priority over walk-in patients. We developed a Markov chain that captured the number of ambulance and walk-in patients at each ED. Subsequently, we presented an exact solution methodology based on matrix-analytic methods to find the probability distribution of the number of patients at each hospital. We derived various queueing performance measures to evaluate the system performance under different model parameters. Moreover, we derived the waiting time distributions for both patient classes using an absorbing Markov chain methodology. Discrete event simulation approach was used to validate two model assumptions. Validation results show that our model is robust with low to medium ambulance utilization, which is the actual operating conditions for EMS.

Although the main cause of ambulance offload delays is serious congestion in the Emergency Departments in particular and the healthcare system in general. We show that even small changes in routing decisions can have great impact on the total offload delay experienced. This is the most important practical contribution of this model. A second contribution, more theoretical in nature, is that we have taken advantage of the problem structure to create an efficient algorithm that solves a complex queueing model with priorities.

The main challenge we faced with this model was computational in nature. Once the model with walk-in patients increases in size to represent real cases, the time to get walk-in patients results is long. But, for ambulance patients, who are the main concern of this work, results were collected quickly and efficiently. This is due to the simplifying assumption of preemptive priority discipline. In the next Chapter, we analyze a similar model network structure with different modeling approach. Instead of using the concept of service capacity as being the combination of (doctor, nurse, bed), we model the beds as being the servers.

Chapter 4

Model 2: Multiple ED Network with Non-Preemptive Priority Discipline and Zero Transit Time

From queueing perspective, in a preemptive priority discipline, a patient with a higher priority is allowed to enter service immediately even if another patient with lower priority is already present in service [51]. On the other hand, in a nonpreemptive discipline, the highest priority patient just goes to the head of the queue to wait for his turn. In the previous model, we assumed that patients arriving to the EDs by an ambulance have higher preemptive discipline over patients arriving to the EDs by themselves. Interrupting the service for a walk-in patient can be explained as follows: when an ambulance patient with higher acuity level arrives to the ED, and if there are any walk-in patients already in service, the service of the walk-in patient is interrupted by moving the nurse and the doctor to treat the more seriously sick ambulance patient. The preemptive priority discipline is justified when we think of the number of servers in the queueing network as the service capacity. Another modeling approach would be to model the physical beds available in an emergency department as the number of servers. As a result, we assume that ambulance patients have higher nonpreemptive priority discipline over walk-in patients. That is, if a walk-in patient is occupying a bed upon an arrival of an ambulance patient, the ambulance patient will have to wait until a bed is available for him.

In this chapter, we present a similar model to the one developed in Chapter 3 but we assume nonpreemptive priority discipline. We first describe the stochastic model in details along with the model assumptions in Section 4.1. We analyze the model with both types of patients in Section 4.2. Then we find the matrix geometric solution in Section 4.3. In Section 4.4, we derive some performance measures of interest. Numerical analysis and some case studies are shown in Section 4.5. Finally, we conclude at Section 4.6.

4.1 The Stochastic Model

We consider a queueing network for a region that consists of K hospitals and one EMS provider. Figure 4.1 is an example of a region that consists of 3 hospitals. For each ED, there are two types of arrivals into the system; a generic arrival stream of patients who arrive by an ambulance and a specific arrival stream for each ED or walk-in patients who decide to go to a specific hospital by themselves. When an ambulance patient calls for an ambulance, and if an ambulance is available, it transfers the patient to the k^{th} ED with probability p_k. The EMS provider has a finite number of ambulances N. If all ambulances are busy, the patient is lost. Losing the patient in our model mimics the cases in which adjacent regions' ambulances are called to back up the fully utilized system. We assume that the transit time for patients is zero in comparison to the time spent at the ED waiting or in service. The k^{th} ED has a capacity of c_k beds that operate independently of each other.

We assume that ambulance patients arrive to the EMS according to a Poisson process with rate λ_0. Patients who arrive by an ambulance are assigned a high priority before getting admission to the hospital ED. They possess a higher nonpreemptive priority over walk-in patients due to the fact that ambulance patients generally are assigned higher CTAS levels than walk-in patients (see Figure 3.2). And walk-in patients who usually arrive with lower acuity conditions are assigned a low priority. They arrive according to a Poisson process with rates λ_k for the k^{th} ED, for $k = 1, 2, \ldots, K$.

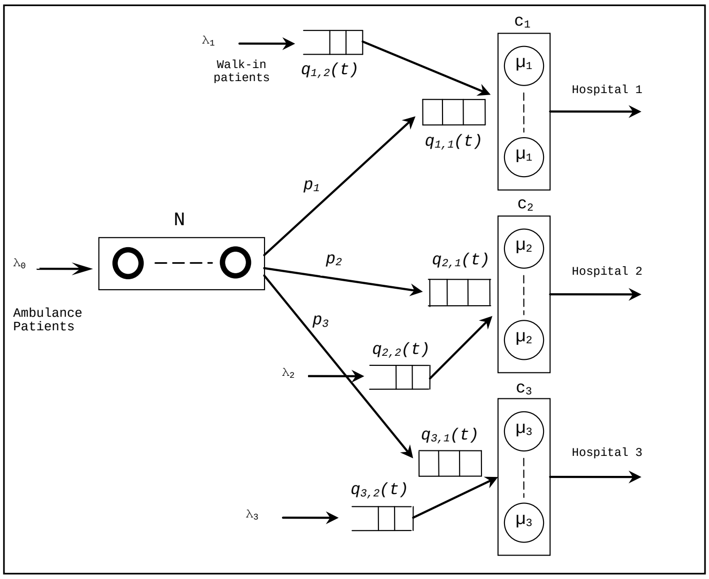

Figure 4.1: ROW EMS-ED Queueing Network Diagram

The service time at each ED has an exponential distribution with parameter μ_k for hospital $k = 1, 2, \ldots, K$. The service times for both types of patients are identically distributed.

We summarize the model parameters as follows:

- K: The number of regional hospitals;
- λ_0: Patient arrival rate to the EMS system;
- p_k: Probability that an EMS arrival is sent to the k^{th} ED, for $k = 1, 2, ..., K$;
- μ_k: Service rate per server in the k^{th} ED, for $k = 1, 2, ..., K$;
- λ_k: arrival rate of walk-in patients at the k^{th} ED, for $k = 1, 2, ..., K$;
- c_k: Number of servers in the k^{th} ED, which corresponds to the number of beds available at the k^{th} ED, for $k = 1, 2, ..., K$.
- N: Total number of ambulances available in the system;

In order to analyze the queueing network, we establish a Markov chain representation that can be useful to analyze system performance. The Markov chain allows us to derive various probability distributions which we use later to derive system performance measures. We start by defining 2 sets of state variables to describe the system state at any point in time:

1. $q_{k,1}(t)$: The number of patients that are in service (from both arrival streams) and waiting in an ambulance, at time t, for $k = 1, 2, \ldots, K$;

2. $q_{k,2}(t)$: The number of walk-in patients waiting in the queue, at time t, for $k = 1, 2, \ldots, K$.

Since the service-time distribution for both priorities is identical, we don't need to differentiate between the two patients classes when they are in service, thus $q_{k,1}(t)$ includes both walk-in and ambulance patients. Based on our definition, if $q_{k,1}(t) \geq c_k$, no walk-in patients are admitted to the ED, and there are $q_k(t) - c_k$ ambulances in offload delay in front of the k^{th} ED.

4.2 The Markov Chain

In this section, we a introduce a method for constructing the infinitesimal generator for the continuous time Markov chain that represents the stochastic model. Consider the process $\{(q_{1,1}(t), q_{1,2}(t), \ldots, q_{K,1}(t), q_{K,2}(t)), t \geq 0\}$ or in a shorter format $\{(q_{k,1}(t), q_{k,2}(t)), t \geq 0, k = \{1, \ldots, K\}\}$. The value of organizing the state variables in this manner will become evident as we illustrate the Markov chain construction process. Since the arrival processes to each ED node are Poisson and the service times are exponential, it is easy to see that the stochastic process $\{(q_{k,1}(t), q_{k,2}(t)), t \geq 0, k = \{1, \ldots, K\}\}$ is a continuous time Markov chain. The state variables associated with the ambulance patients and the patients in service, $q_{k,1}(t)$, have a finite state space. While the state variables associated with the walk-in patients have an infinite state space if we assume that waiting rooms are big enough to accommodate all the arriving patients. Table (4.1) illustrates the possible transitions in the system along with the corresponding transition rates.

In order to construct the Markov chain infinitesimal generator, Q_K, we observe the

Possible event	Rate	From	To	Condition
Ambulance arrival to ED k	$p_k\lambda_0$	$(q_{k,1}, q_{k,2})$	$(q_{k,1}+1, q_{k,2})$	if $\sum_{k=1}^{k=K}(q_{k,1}-c_k)^+ < N$
Walk-in patient arrival to ED k	λ_k	$(q_{k,1}, q_{k,2})$	$(q_{k,1}, q_{k,2}+1)$	if $q_{k,1} \geq c_k$
		$(q_{k,1}, q_{k,2})$	$(q_{k,1}+1, q_{k,2})$	if $q_{k,1} < c_k$
Service completion at ED k	$c_k\mu_k$	$(q_{k,1}, q_{k,2})$	$(q_{k,1}-1, q_{k,2})$	if $q_{k,1} > c_k$
	$q_{k,1}\mu_k$	$(q_{k,1}, q_{k,2})$	$(q_{k,1}-1, q_{k,2})$	if $q_{k,1} \leq c_k$ and $q_{k,2} = 0$
	$c_k\mu_i$	$(q_{k,1}, q_{k,2})$	$(q_{k,1}, q_{k,2}-1)$	if $q_{k,1} = c_k$ and $q_{k,2} > 0$

Table 4.1: Transition rates for the model Markov chain

state of the EDs sequentially and based on that we construct the Markov chain infinitesimal generator with $2K$ state variables in a very structured process. Namely, we pair each ED state variables together and set the state variable for the walk-in patients as the level and the state variable that corresponds to the ambulance patients as the phase for the ED layer as follows: $\{(q_{K,2}(t), q_{K,1}(t)), (q_{K-1,2}(t), q_{K-1,1}(t)), \ldots, (q_{1,2}(t), q_{1,1}(t)), t \geq 0\}$. The K pairs are only connected when all the ambulances for a region are consumed.

The state variables $q_{k,1}(t)$, $k = \{1, \ldots, K\}$, have a finite range of $\{0, \ldots, N + c_k\}$; while the state variables $q_{k,2}(t)$, $k = \{1, \ldots, K\}$, have infinite range. To facilitate the construction of the Markov chain infinitesimal generator, we truncate $q_{k,2}(t)$ for $k < K$ at M. We choose the value of M large such that the stationary probabilities that the system is in state M is negligible. Or $P(q_{k<K,2} = M) \approx 0$. We divide the states into subgroups according to the ED layer where we choose the K^{th} ED walk-in patients state variable as the level: $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots \cup \Omega_\infty$, where, for $i_K = 0, 1, ..., \infty$,

$$\begin{aligned} \Omega_{i_K} = \{(i_K, j_K, i_{K-1}, j_{K-1}, ..., i_1, j_1) :\ & i_K \geq 0; \\ & \textstyle\sum_{k=1}^{k=K}(j_k - c_k)^+ \leq N; \\ & i_k = 0 \text{ if } j_k < c_k, k = 1, \ldots, K; \\ & 0 \leq i_1, i_2, \ldots, i_{K-1} \leq M\}. \end{aligned} \tag{4.1}$$

Our methodology in constructing the model Markov chain is based on pairing each ED state variables together. We use K layers to construct the Markov chain infinitesimal generator such that each layer represents one ED. We start the first layer with the first ED, and then adding layer by layer of the other EDs. All EDs service and walk-in arrivals are independent. While the ambulance patients' arrival to the K EDs are only connected when there are no ambulances available to transfer a patient who called for an ambulance. Thus, each layer only affects the size of the inner layers.

Another benefit for organizing the state space as explained above would be the future addition of other hospitals into the Markov chain model, if needed. If the queueing network of interest, for example, consists of more hospitals instead of K, then all what we need to do is to add another layer to the constructed Markov chain for the new ED. Next, we describe the process of constructing the Markov chain layers in three main steps.

Step 1: First ED Layer

In the first step, we start by the first ED state variables $\{q_{1,2}(t), q_{1,1}(t)\}$. Given the number of ambulances available for ED1, n, where $0 \leq n \leq N$. $\{q_{1,2}(t), q_{1,1}(t), t \geq 0\}$ is a Markov chain. Its infinitesimal generator, $Q_1^{(n)}$, has the following tri-diagonal structure:

$$Q_1^{(n)} = \begin{pmatrix} Q_{1(0,0)}^{(n)} & Q_{1(0,1)}^{(n)} & & & & & \\ Q_{1(1,0)}^{(n)} & Q_{1(1)}^{(n)} & Q_{1(2)}^{(n)} & & & & \\ & Q_{1(0)}^{(n)} & Q_{1(1)}^{(n)} & Q_{1(2)}^{(n)} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & Q_{1(0)}^{(n)} & Q_{1(1)}^{(n)} & Q_{1(2)}^{(n)} \\ & & & & & Q_{1(0)}^{(n)} & Q_{1(M,M)}^{(n)} \end{pmatrix} \tag{4.2}$$

where M is the truncation limit for $q_{k,2}(t)$.

The state space for the finite random variable, $q_{1,1}(t)$, is reduced by the number of ambulances in offload delay at the other EDs, or mathematically: $q_{1,1}^{max} = N + c_1 - \sum_{k=2}^{k=K} (q_{k,1} - c_k)^+$ or simply $n + c_1$. Next, we specify the transition blocks in equation (4.2). First, we note that the level variable, $q_{1,2}(t)$, increases by one only when the phase variable, $q_{1,1}(t)$, is greater or equal to c_1. Otherwise, it does not increase. In general:

$$Q^{(n)}_{1(0,1)} = \lambda_1 \cdot \begin{array}{cc} & \begin{array}{ccc} c_1 & \dots & n+c_1 \end{array} \\ \begin{array}{c} 0 \\ \vdots \\ c_1 \\ \vdots \\ n+c_1 \end{array} & \begin{pmatrix} 0 & & \\ \vdots & & \\ 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \end{array} \tag{4.3}$$

$$Q_{1(2)} = \lambda_1 \cdot I_{n+1} \tag{4.4}$$

where I_{n+1} is an identity matrix of size $n+1$. The level variable, $q_{1,2}(t)$, decreases by one only when the phase variable, $q_{1,1}(t)$, equals c_1, otherwise it does not change. This is because of the lower priority assigned to the walk-in patients, so if there was a line of ambulance patients or if $q_{1,1}(t) > c_1$, then ambulance patients will get served before walk-in patients.

$$Q^{(n)}_{1(1,0)} = c_1\mu_1 \cdot \begin{array}{cc} & \begin{array}{cccccc} 0 & \dots & c_1 & c_1+1 & \dots & n+c_1 \end{array} \\ \begin{array}{c} c_1 \\ c_1+1 \\ \vdots \\ n+c_1 \end{array} & \begin{pmatrix} 0 & \dots & 1 & & & \\ \vdots & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \end{array} \tag{4.5}$$

$$Q^{(n)}_{1(0)} = c_1\mu_1 \cdot \begin{array}{cc} & \begin{array}{cccc} c_1 & c_1+1 & \dots & n+c_1 \end{array} \\ \begin{array}{c} c_1 \\ c_1+1 \\ \vdots \\ n+c_1 \end{array} & \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \end{array} \tag{4.6}$$

The diagonal matrices $Q_{1(0,0)}$ in equation (4.2) include transitions for the phase variable, $q_{1,1}(t)$. The upper diagonal elements of $Q_{1(0,0)}$ specify the rates at which $q_{1,1}(t)$

increases by one. If all ED1 beds are full, then the rate of arrival for the ED is just $p_1\lambda_0$, that is only the high priority patients will be admitted. While if there is at least one empty bed at ED1, then both arrival streams (walk-in and ambulance patients) can be admitted to the ED, this results in a total arrival rate of $p_1\lambda_0 + \lambda_1$.

The lower diagonal elements of the diagonal matrices $Q_{1(0,0)}$ specify the rate at which $q_{1,1}(t)$ decreases by one. The rate of service completion at ED1 depends on the number of occupied beds or simply it is the $\min\{q_{1,1}, c_1\}$. $Q_{1(0,0)}$ has the following structure:

$$Q^{(n)}_{1(0,0)} = \begin{array}{c|cccccccc|}
 & 0 & 1 & \dots & c_1-1 & c_1 & c_1+1 & \dots & n+c_1 \\
\hline
0 & * & p_1\lambda_0+\lambda_1 & & & & & & \\
1 & \mu_1 & * & p_1\lambda_0+\lambda_1 & & & & & \\
\vdots & & \ddots & \ddots & \ddots & & & & \\
c_1-1 & & & (c_1-1)\mu_1 & * & p_1\lambda_0+\lambda_1 & & & \\
c_1 & & & & c_1\mu_1 & * & p_1\lambda_0 & & \\
c_3+1 & & & & & \ddots & \ddots & \ddots & \\
\vdots & & & & & & \ddots & * & p_1\lambda_0 \\
n+c_1 & & & & & & & c_1\mu_1 & *
\end{array} \tag{4.7}$$

where $*$ is calculated such that the row sums of $Q^{(n)}_1$ are zeros.

When the level variable, $q_{1,2}(t) > 0$, the diagonal matrices $Q_{1(1)}$ have a different structure than the previous equation. Having $q_{1,2}(t) > 0$ means that there is a queue of walk-in patients waiting for admission. The queue exists because all the beds are occupied, or $q_{1,1}(t) \geq c_1$. This means that the states $q_{1,1} = 0, 1, \dots, c_1 - 1$ do not exist when $q_{1,2} > 0$. The general structure for $Q_{1(1)}$ is as follows:

$$Q^{(n)}_{1(1)} = \begin{array}{c|ccccc|}
 & c_1 & c_1+1 & \dots & n+c_1-1 & n+c_1 \\
\hline
c_1 & * & p_1\lambda_0 & & & \\
c_1+1 & c_1\mu_1 & * & p_1\lambda_0 & & \\
\vdots & & \ddots & \ddots & \ddots & \\
n-1 & & & c_1\mu_1 & * & p_1\lambda_0 \\
n+c_1 & & & & c_1\mu_1 & *
\end{array} \tag{4.8}$$

Step 2: the k^{th} ED Layer

To construct the k^{th} ED layer, we define the Markov chain $\{q_{k,2}(t), q_{k,1}(t), \ldots, q_{1,2}(t), q_{1,1}(t), t \geq 0\}$. The infinitesimal generator for the this Markov chain has a similar structure as $Q_1^{(n)}$ as follows:

$$Q_k^{(n)} = \begin{pmatrix} Q_{k(0,0)}^{(n)} & Q_{k(0,1)}^{(n)} & & & & & \\ Q_{k(1,0)}^{(n)} & Q_{k(1)}^{(n)} & Q_{k(2)}^{(n)} & & & & \\ & Q_{k(0)}^{(n)} & Q_{k(1)}^{(n)} & Q_{k(2)}^{(n)} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & Q_{k(0)}^{(n)} & Q_{k(1)}^{(n)} & Q_{k(2)}^{(n)} \\ & & & & & Q_{k(0)}^{(n)} & Q_{k(M,M)}^{(n)} \end{pmatrix} \tag{4.9}$$

where n is the number of ambulances available for the k^{th} ED, or mathematically, $n = N - \sum_{\hat{k}=k+1}^{\hat{k}=K} (q_{\hat{k},1} - c_{\hat{k}})^+$. The state space for the finite random variable $q_{k,1}$ is reduced by the number of ambulances in offload delay at the upper layers EDs only, or mathematically: $q_{k,1}^{max} = N + c_k - \sum_{\hat{k}=k+1}^{\hat{k}=K} (q_{\hat{k},1} - c_{\hat{k}})^+ = n + c_k$. Next, we specify the transition blocks in equation (4.9). First, we note that the level variable, $q_{k,2}$, increases by one only when the phase variable, $q_{k,1}$, is greater or equal to c_k. Otherwise, it does not increase. In general:

$$Q_{k(0,1)}^{(n)} = \lambda_k \cdot \begin{array}{c} \\ 0 \\ \vdots \\ c_k \\ \vdots \\ c_k + n \end{array} \begin{array}{c} \begin{array}{ccc} c_k & \ldots & c_k + n \end{array} \\ \begin{pmatrix} 0 & \ldots & 0 \\ & & \\ I_n & & \\ & \ddots & \\ & & I_0 \end{pmatrix} \end{array} \tag{4.10}$$

$$Q_{k(1)}^{(n)} = \lambda_k \cdot I \tag{4.11}$$

where I_n is an identity matrix of size $Q_{k-1}^{(n)}$. The level variable decreases by one only when the phase variable equals c_k, otherwise it does not change. This is because of the lower priority assigned to the walk-in patients, so if there were at least one ambulance in offload delay, or if $q_{k,1} > c_k$, then ambulance patients will get served before walk-in patients.

$$Q_{k(1,0)}^{(n)} = c_k\mu_k \cdot \begin{array}{c} \\ c_k \\ c_k+1 \\ \vdots \\ n+c_k \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & \dots & c_k & c_k+1 & \dots & n+c_k \end{array} \\ \begin{pmatrix} 0 & & I_n & & & \\ 0 & & & 0 & & \\ \vdots & & & & \ddots & \\ 0 & & & & & 0 \end{pmatrix} \end{array} \tag{4.12}$$

$$Q_{k(0)}^{(n)} = c_k\mu_k \cdot \begin{array}{c} \\ c_k \\ c_k+1 \\ \vdots \\ n+c_k \end{array} \begin{array}{c} \begin{array}{cccc} c_k & c_k+1 & \dots & n+c_k \end{array} \\ \begin{pmatrix} I_n & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \end{array} \tag{4.13}$$

The diagonal matrices $Q_{k(0,0)}$ in equation (4.9) include transitions for both the phase variable, $q_{k,1}$, and the transitions associated with the inner layers which are included in the matrix $Q_{k-1}^{(n)}$. The upper diagonal elements of $Q_{k(0,0)}$ specify the rates at which $q_{k,1}$ increases by one. If all the k^{th} ED beds are full, then the rate of arrival is just $p_k\lambda_0$, that is only the high priority patients will be admitted. While if there is at least one empty bed at the k^{th} ED, then both arrival streams (walk-in and ambulance patients) can be admitted, this results in a total arrival rate of $p_k\lambda_0 + \lambda_k$ when $q_{k,1} < c_k$.

The lower diagonal elements of the diagonal matrices $Q_{k(0,0)}$ specify the rate at which $q_{k,1}$ decreases by one upon a service completion. The rate of service completion at the k^{th} ED depends on the number of occupied beds, or simply it is the $\min\{q_{k,1}, c_k\}$. The diagonal elements in $Q_{k(0,0)}$ include the inner layers details. $Q_{k(0,0)}$ has the following general structure:

$$Q_{k(0,0)} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ c_k-1 \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n \end{array} \begin{array}{cccccccc} 0 & 1 & \dots & c_k-1 & c_2 & c_k+1 & \dots & c_k+n \\ \ast + Q_{k-1}^{(n)} & p_k\lambda_0 I^{(k-1)} + \lambda_k I & & & & & & \\ \mu_k I & \ast + Q_{k-1}^{(n)} & p_k\lambda_0 I^{(k-1)} + \lambda_k I & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & (c_k-1)\mu_k I & \ast + Q_{k-1}^{(n)} & p_k\lambda_0 I^{(1)} + \lambda_k I & & & \\ & & & c_k\mu_k I & \ast + Q_{k-1}^{(n)} & p_k\lambda_0 I^{(k-1)} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & c_k\mu_k I & \ast + Q_{k-1}^{(1)} & p_k\lambda_0 I^{(k-1)} \\ & & & & & & c_k\mu_k I & \ast + Q_{k-1}^{(0)} \end{array} \tag{4.14}$$

where $\ast$ is calculated such that the rows of $Q_k^{(n)}$ sum to zeros. $\ast = -p_k\lambda_0 I^{(k-1)} - I_{c_k > q_{k,1}}\lambda_k I - \mu_k \min\{q_{k,1}, c_k\} I$. And $I^{(k-1)}$ has the following structure:

$$I^{(k-1)} = \begin{pmatrix} I^{(k-2)} & & & \\ & \ddots & & \\ & & I^{(k-2)} & \\ & & & \mathbf{0} \end{pmatrix}$$

This matrix is used to show that there is no ambulance arrival to the k^{th} ED when all ambulances are being used. $I^{(1)}$ is an identity matrix that have the last lower left $M+1$ diagonal elements are zeros. note: The identity matrix associated with transitions down one level has the form $(0\ I)$ when $q_{k,1} > c_k$ since the size of Q_1^n decreases as n decreases. The same argument holds for the identity matrix associated with going up one level but with the structure $(I; 0)$.

When the level variable is strictly positive, $q_{k,2} > 0$, the diagonal matrices $Q_{k(1)}$ have a different structure than the previous equation. Having $q_{k,2} > 0$ means that there is a queue of walk-in patients waiting for admission in front of the k^{th} ED. The queue exists because all the beds are occupied, or when $q_{k,1} \geq c_k$. This means that the states $q_{k,1} = 0, 1, \dots, c_k - 1$ do not exist when $q_{k,2} > 0$. The general structure for $Q_{k(1)}$ is as follows:

$$Q_{k(1)} = \begin{array}{c} \\ c_k \\ c_k+1 \\ \vdots \\ c_k+n-1 \\ c_k+n \end{array} \begin{array}{c} \begin{array}{ccccc} c_k & c_k+1 & \dots & c_k+n-1 & c_k+n \end{array} \\ \begin{pmatrix} *+Q_{k-1}^{(n)} & p_k\lambda_0 I^{(k-1)} & & & \\ c_k\mu_k I & *+Q_{k-1}^{(n-1)} & p_k\lambda_0 I^{(k-1)} & & \\ & \ddots & \ddots & \ddots & \\ & & c_k\mu_k I & *+Q_{k-1}^{(1)} & p_k\lambda_0 I^{(k-1)} \\ & & & c_k\mu_k I & *+Q_{k-1}^{(0)} \end{pmatrix} \end{array} \tag{4.15}$$

Step 3: The last Layer

The last layer is achieved by adding the pair $(q_{K,2}, q_{K,1})$ into the previous layer Markov chain, the resulting Markov chain is $\{q_{K,2}(t), q_{K,1}(t), \dots, q_{1,2}(t), q_{1,1}(t), t \geq 0\}$. The Markov chain infinitesimal generator, Q_K, has a similar structure to the previous layer but with one main distinction. It has an infinite tri-diagonal structure as follows:

$$Q_K = \begin{pmatrix} Q_{K(0,0)} & Q_{K(0,1)} & & & & \\ Q_{K(1,0)} & Q_{K(1)} & Q_{K(0)} & & & \\ & \ddots & \ddots & \ddots & & \\ & & Q_{K(2)} & Q_{K(1)} & Q_{K(0)} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \tag{4.16}$$

The blocks in equation (4.16) can be generated similar to the previous inner layers. We summarize the main steps to find the transition blocks for a network that consists of K hospitals in the following algorithm:

Algorithm 4 Computing matrix blocks in Q_K

```
1. Construct $Q_1^{(n)}$ and its blocks using equations (4.3), (4.4), (4.5), (4.6), (4.7), (4.8).
   Set $k = 2$.
2. if $k < K$ go to step (3); otherwise, stop.
3. Construct $Q_k^{(n)}$ matrices using equations (4.10), (4.11), (4.12), (4.13), (4.14), (4.15).
4. if $k < K$ go to step (3); otherwise, stop.
```

4.3 Matrix-Geometric Solution

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \ldots)$ the stationary probability distribution of Q_K. The stationary distribution exists if and only if the Markov chain is ergodic. Since the Markov chain of interest is irreducible and has a QBD structure, by Neuts [46], the Markov chain is ergodic if and only if $Q_{K(0)}\boldsymbol{\pi e} < Q_{K(2)}\boldsymbol{\pi e}$, where $\boldsymbol{e}$ is a column vector of ones. If the stability condition is satisfied, then $\boldsymbol{\pi}$ exists and it is the unique non-negative solution for the linear system:

$$\boldsymbol{\pi} Q_K = 0; \quad \text{and} \quad \boldsymbol{\pi e} = 1, \tag{4.17}$$

Since the infinitesimal generator Q_K has a block tri-diagonal structure, a matrix-geometric solution can be obtained and it has the geometric structure:

$$\boldsymbol{\pi_i} = \boldsymbol{\pi_{i-1}} R, \quad \text{for} \ \ i > 1 \tag{4.18}$$

where the rate matrix R is the minimal nonnegative solution to the nonlinear equation:

$$Q_{K(0)} + RQ_{K(1)} + R^2 Q_{K(2)} = 0. \tag{4.19}$$

The boundary probabilities can be calculated as outlined in §2.4. In the next section we derive a number of performance measures of interest.

4.4 Performance Evaluation

In this section, we derive a number of performance measures that can be useful to assess current system parameter decisions. As we have done in the previous chapter, we focus on the last layer (K^{th} ED) performance measures. Other EDs performance measures are found by replacing the (K^{th} ED) with another ED. The performance measures of interest are:

1. The steady state probability distribution of the number of ambulances in offload delay at the K^{th} ED is given by: Let $\{\eta_K(j), 0 \le j \le c_K + N\}$ be the distribution of $q_{K,1}(t)$. Recall that $q_{K,1}(t)$ is defined as the number of both ambulance patients and walk-in patients in service and in an ambulance at the K^{th} ED. If $q_{K,1}(t)$ is greater than c_K, then there are $q_{K,1}(t) - c_K$ ambulances experiencing offload delays outside of the K^{th} ED. We use this

fact to find the probability distribution of the number of ambulances in offload delay as follows:

$$\xi_2(m) = \begin{cases} \sum_{j=0}^{j=c_K} \eta_2(j), & \text{for } \ m = 0 \\ \eta_2(m + c_K), & \text{for } \ m = 1, \ldots, N \end{cases} \tag{4.20}$$

The mean queue number of ambulances in offload delay $E[q_{K,1}]$ can be found accordingly.

2. The probability distribution of the total number of ambulances in offload delay, denoted by O, is given by

$$P\{O = m\} = \sum_{(j_K,\ldots,j_1)\in\Omega:\ \sum_{k=1}^{K} \max\{0,j_k-c_k\}=m} \pi_{j_K,\ldots,j_1}, \quad \text{for } 0 \le m \le N; \tag{4.21}$$

The mean total number of ambulances in offload delay, $E[O]$, can be obtained accordingly.

3. The distribution of the number of walk-in patients in the queue at the K^{th} ED can be found using the matrix geometric solution as $\{\boldsymbol{\pi}_i \boldsymbol{e}, \ \ i = 0, 1, \ldots\}$. The mean number of walk-in patients in the queue can be calculated by:

$$E[q_{K,2}] = \sum_{i=0}^{\infty} i\boldsymbol{\pi}_i \boldsymbol{e} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_1 R^{i-1}\boldsymbol{e} = \boldsymbol{\pi}_1 \left(\sum_{i=1}^{\infty} iR^{i-1} \right) \boldsymbol{e} = \boldsymbol{\pi}_1 (I - R)^{-2} \boldsymbol{e}. \tag{4.22}$$

4. The loss probability is given by:

$$P_L = P\{O = N\} = \sum_{(j_K,\ldots,j_1)\in\Omega:\ \sum_{k=1}^{K} \max\{0,j_k-c_k\}=N} \pi_{j_K,\ldots,j_1} \tag{4.23}$$

5. Waiting Time Distribution for the K^{th} ED ambulance patients. The waiting time $w_{K,1}$ of an ambulance patient arriving to the K^{th} ED depends on both the number of ambulance patients at the K^{th} ED and the number of walk-in patients already in service, which are captured in the state variable $(q_{K,1})$. Denote by $\alpha_i(K)$ the probability that i patients are in the K^{th} ED when an ambulance patient arrives. Note that an arriving patient can reach the K^{th} ED if and only if there is an ambulance available at the time of arrival. For $0 \le j_K \le c_K + N - 1$, we define:

$$\alpha_{j,1}(K) = \frac{\sum_{(j,j_{K-1},\ldots,j_1)\in\Omega:\ \max\{0,j-c_K\}+\sum_{k=1}^{K-1}\max\{0,j_k-c_K\}<N} \pi_{j_{K,1},j_{K-1,1},\ldots,j_{1,1}}}{1 - P_L}. \tag{4.24}$$

Define $\boldsymbol{\xi}(K) = (\alpha_{c_K,1}(K), ..., \alpha_{c_K+N-1,1}(K))$. Then the i-th component of $\boldsymbol{\xi}(K)$ denotes the probability that an arriving ambulance patient to the K^{th} ED has to wait for the service completion of i patients before getting a bed. Since there are c_K beds for all patients in the K^{th} ED, each with an exponential service time with parameter μ_K, if all beds are occupied, the time until the next service completion is exponential with parameter $c_K\mu_K$. Thus, the total time to serve i patients has an Erlang distribution of order i. Consequently, when an ambulance patient arrives to hospital K, the waiting time has a generalized Erlang distribution with a phase-type representation $(\boldsymbol{\xi}(K,1), c_K\mu_K J_N)$, where

$$J_N = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}_{N\times N} . \tag{4.25}$$

The distribution function of the waiting time $w_{K,1}$ is given by $P\{w_{K,1} < t\} = 1 - \boldsymbol{\xi}(K,1)\exp\{-c_K\mu_K J_N t\}\boldsymbol{e}$. The expected waiting time is,

$$E[w_{K,1}] = \sum_{i=1}^{N} \frac{i\alpha_{c_K-1+i}(K)}{c_K\mu_K}. \tag{4.26}$$

6. The waiting time distribution for walk-in patients, $w_{K,2}$, arriving to the K^{th} ED has a similar phase type structure to that of the ambulance patients. Recall that the state variable $q_{K,2}$ is defined such that it includes only walk-in patients in the waiting rooms. If a tagged walk-in patient can get admission to the corresponding ED, then that is captured in $q_{K,1}$. While if the tagged walk-in patient arriving to the K^{th} ED finds i walk-in patients ahead of him in the queue and j ambulance patients waiting outside the ED; then he has to wait for $i + j$ service completions plus the service completion for ambulance patients who arrived during his waiting time. The events that affect the tagged walk-in patient waiting time are service completions of both patients types, and arrivals of high priority ambulance patients during waiting time. While subsequent arrivals of walk-in patients do not affect the tagged walk-in patient waiting time because same priority patients are admitted based on a FCFS rule.

   If the state of the K^{th} ED at time t upon arrival of the the tagged walk-in patient was

(i_K, j_K) as described by the duplet $(q_{(K,2)}, q_{(K,1)})$ where $i_K \geq 0$, and $c_K \leq j_K \leq c_K + N$, then the waiting time constitutes the time till absorption for the continuous time Markov chain that has the infinitesimal generator as follows:

$$Q_{w,K} = \begin{pmatrix} Q_{K(0,0)} & & & & & \\ Q_{K(1,0)} & Q_{K(1,1)} & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & Q_{K(j-1,j-2)} & Q_{K(j-1,j-1)} & \\ & & & & Q_{K(j,j-1)} & Q_{K(j,j)} \end{pmatrix} \tag{4.27}$$

Q_{wk} has a similar structure as $Q_1^{(N)}$ but without the arrival of walk-in patients, and $Q_{K(0,0)}$ does not have the states $q_{K,1} = 0, 1, \ldots, c_K - 1$. Then we obtain the conditional probability distribution of the waiting time as:

$$P(w_{K,2} \leq t \mid j) = 1 - (0, ..., 0, \boldsymbol{\pi}_{i_K}/(\boldsymbol{\pi}_{i_K}\boldsymbol{e})) \exp\{Q_{w,K}t\}\boldsymbol{e}. \tag{4.28}$$

The distribution of the waiting time of an arbitrary walk-in patient can be obtained as,

$$P(w_{K,2} \leq t) = 1 - \sum_{i_K=0}^{\infty} (0, ..., 0, \boldsymbol{\pi}_{i_K}) \exp\{Q_{w,K}t\}\boldsymbol{e}. \tag{4.29}$$

By using truncation, the above formula can be used to compute the distribution of the waiting time of walk-in patients.

4.5 Numerical Analysis

In this section, we use the derived performance measures in the previous section to analyze a network that consists of one EMS provider and two EDs. Compared to the preemptive model of Chapter 3, the nonpreemptive model of this chapter requires more computational resources. Recall that for a network that consists of three EDs in the preemptive model, we need three finite state variables to derive ambulance patients performance measures, and four state variables to derive performance measures pertaining to walk-in patients. While, in the nonpreemptive

Parameter set	value
N	5
(λ_0) patients/hr	1.3
(λ_1, λ_2) patient/hr	$(0.2, 0.2)$
(μ_1, μ_2) patient/hr	$(1/5, 1/6)$
(c_2)	(6)
(p_1, p_2)	$(0.55, 0.45)$

Table 4.2: System parameters for Case Study

model, six state variables are needed to model a three-ED network, three of which are infinite. For the above reason, we consider a network that consists of two EDs in the numerical analysis. However, the methodology and analysis developed in the previous sections can be used to analyze a network of an arbitrary size.

To compare between EDs, we define two types of parameters for the k^{th} ED, for $1 \leq k \leq K$:

- The proportion of time the servers are busy with ambulance patients $(\rho_{k,1})$: Since the service of ambulance patients is not affected by walk-in patients, we can define the server utilization for ambulance patients alone. Define $\rho_{k,1} = \min\{1, \lambda_0 p_k(1-P_L)/(c_k\mu_k)\}$, where $\lambda_0 p_k(1-P_L)$ is the arrival rate of ambulance patients to the k^{th} ED, and $c_k\mu_k$ is the total service capacity at the k^{th} ED.

- ED utilization $\rho_{k,2}$: Considering the service of both types of patients, the server utilization can be defined as $\rho_k = \min\{1, (\lambda_0 p_k(1-P_L) + \lambda_k)/(c_k\mu_k)\}$.

The system parameters for the case study are reported in Table 4.2. For this case study, we truncate the walk-in patients state variable at 50. We note here that the truncation limit should be increased to increase the solution accuracy. Also, as the system utilization increases, the truncation limit should be increased. The results for this case study are reported in Table 4.3.

As we can see from the results, the waiting times for ambulance patients and walk-in patients are different but are closer than those when the service priority is preemptive. Although ambulance patients spend less time in the first ED getting service ($\mu_1 = 1/5$), the offload delays experienced by ambulances at the first ED are similar to those of the second ED. This is because the second ED has more beds than the first ED (6 compared to 5). Consequently, speeding up the service at the first ED have balanced the offload delays at the second ED. From the walk-in

	Scenario 1 ($c_1 = 5$)		Scenario 2 ($c_1 = 6$)	
Measures	$k = 1$	$k = 2$	$k = 1$	$k = 2$
$\rho_{k,1}$	66.57%	54.47%	56.97%	55.94%
$\rho_{k,2}$	86.57%	74.47%	73.64%	75.94%
$E[q_{k,1}]$	0.02	0.01	0.01	0.01
$E[w_{k,1}]$	0.02	0.02	0.02	0.02
$E[q_{k,2}]$	0.05	0.04	0.03	0.05
$E[w_{k,2}]$	0.06	0.05	0.04	0.06
P_L	0.074		0.044	

Table 4.3: Performance measures for Case Study

patients perspective, the faster service at the first ED have less impact because of the lower priority of walk-in patients.

As shown in Table 4.3, the first ED have higher utilization compared to the second ED. This have affected both ambulance patients and walk-in patients queue lengths. We notice that the probability of an emergency call being served by other regions' EMS is high (P_L=7.4%). This is because of the ambulances being used by ED patients.

To assess the effect of increased capacity on the system performance, we compare the results when the first ED number of beds are 5 (Scenario 1) versus 6 (Scenario 2). As expected, when the first ED capacity is increased, the expected queue lengths and expected waiting times are decreased. More interestingly, the second ED queue length and waiting times for walk-in patients are increased. This is because when the first ED has more capacity to serve ambulance patients, the probability of losing patients is decreased and as a result more patient are being served by the EMS. Although one might think that increasing the capacity on the first ED should not affect the system performance at the second ED, the results reveal that increasing the capacity at the fist ED have increased the utilization of servers at the second ED ($\rho_{k,1}$ and $\rho_{k,2}$). This observation supports the idea of dependence among regional EDs if and only if the EMS system is highly utilized, otherwise the EDs behave independently from each other.

In the next chapter, we used the idea of independence among regional EDs to develop a decomposed model for the ambulance offload delay problem.

4.6 Conclusion

In this chapter, we developed a stochastic model for a network that consists of one EMS provider and K hospitals. We modeled the beds at the regional EDs as the servers and hence, we assumed that patients arriving by an ambulance have a higher nonpreemptive priority over patients arriving by themselves. We developed an efficient algorithm to construct the model Markov chain for a network of arbitrary size. Then used the matrix analytic methods to derive its limiting probabilities. Similar to the previous model in Chapter 3, we developed a number of performance measures to evaluate offload delays and walk-in patients waiting.

One of the limitations of the above model is computational in nature; when we look into realistic problem size, the computation time is very high. Another limitation of this model is the truncation that is performed for the infinite state variables related to walk-in patients. As the number of EDs increases in a network, the number of variables that should be truncated increases as well. This truncation, if not performed carefully, might affect the accuracy of the model results. From an application perspective, the truncation limit can be viewed as the waiting room capacity inside the ED.

In the next chapter, we consider a model with nonpreemptive priority and Markovian transit time in which analysis is performed first at hospital level and then results are combined to get regional level results. This decomposition is practical because it does not have computational limitations.

Chapter 5

Model 3: Decomposed EMS-ED Network with Priorities and Transit Time

The third model we develop to analyze offload delays that ambulances experience in front of EDs is based on one main assumption; queues of ambulances in front of EDs are independent. Based on this assumption, we decompose the K EDs network into separate networks. This simplifying assumption makes the size of the problem manageable for realistic size networks. Another assumption we make is with respect to the transit time of ambulances when they transfer patients to the EDs. In the previous two models we assumed that transit time is negligible, in this model we consider the transit time into the stochastic model. As a result of this assumption, blocking of ambulances is used to explain the offload delays compared to the previous two models where high priority ambulance patients waiting time was used to model offload delays.

Our objective in this chapter is the optimal allocation of ambulance patients into the regional hospitals, not performance evaluation as in Chapters 3 and 4. The problem of workload allocation was studied by a number of researchers particularly in flexible manufacturing systems. This problem is also related to job shop manufacturing [52]. Calabrese [52] investigates the workload allocation problem in an open Jackson network with multiple M/M/c queues. The author proves that nodes with a higher number of servers should be loaded more heavily with respect to nodes with less number of servers (the utilization per server is higher) due to server pooling. But if the nodes have identical number of servers, they should be loaded equally. Mehrotra et al. [53] study routing in call centers to decide which calls should be handled by which agents based on the state of the system. The focus of the paper is on customer waiting time and overall resolution rate for different routing strategies.

In this chapter, we first describe the stochastic model in details along with the model assumptions and Markov chain details in Section 5.1. We present an iterative algorithm to obtain the steady state probabilities of the Markov chain, and the network performance measures at both the hospital level, and at the regional level in Section 5.2. Then we develop an approximation scheme for computing the system performance measures in Section 5.3. An optimization problem for ambulance routing decisions is developed in Section 5.4. Numerical analysis is performed in Section 5.5. Finally, we conclude in Section 5.6.

5.1 The Stochastic Model

As illustrated by Figure 5.1, when an emergency call that requires an ambulance arrives to the EMS, an ambulance is dispatched to the call scene. Upon arrival, the paramedic team apply the basic life saving procedure and upload the patient into the ambulance. Then they transfer the patient into one of the region K hospitals. Usually the time it takes to reach the patient, upload him into the ambulance and then transfer him into the ED is about 1 hour which we refer to as the transit time. In steady state, p_k of the EMS arrivals are transferred to the k^{th} ED. ED beds are also allocated to other emergency patients who arrive to the corresponding EDs independently by themselves, we shall refer to those patients as walk-in patients later.

Although the network diagram in Figure 5.1 depicts for the actual flow of patients from the community until they leave an ED bed, the network of a region is quite complicated from a stochastic modeling perspective. This is due to two reasons: first, the ambulances may experience blocking; second, there are two priority levels for patients through the network. One of the main operating characteristics for the EMS services is normally to operate at low utilization levels (35% or less [11]). Or to set the probability of having all the ambulances busy in any period of the day equal to zero. Thus, the availability of ambulances is high. Consequently, it is reasonable to assume that an ambulance is always available. Under this assumption, the

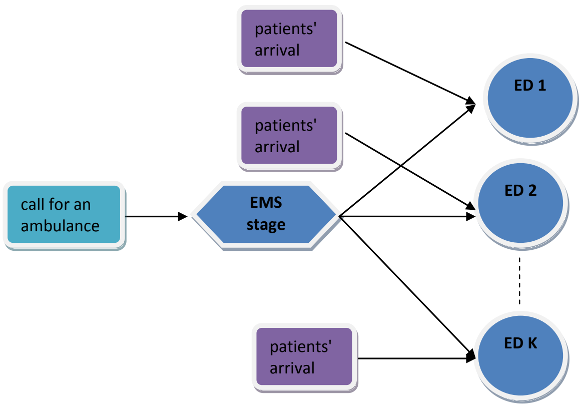

Figure 5.1: An EMS-ED flow chart for a region of K hospitals

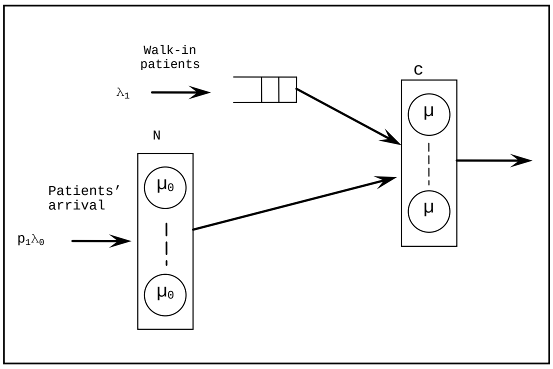

Figure 5.2: One ED Network Diagram

EMS-ED network of figure 5.1 can be decomposed into K similar networks such that each decomposed network consists of two stages; the EMS stage, and the ED stage as depicted in Figure 5.2. Once we analyze one of the decomposed networks, we can use the results to analyze the regional network utilizing the fact that all the EDs operate independently across the region.

We assume that emergency calls to the EMS arrive according to a Poisson process with rate λ_0. From the ED perspective, there are two arrival streams: ambulance patients, and walk-in patients. Once they arrive to an ED, emergency patients are assigned different acuity level score that ranges from 1-5 where 1 corresponds to the severely ill patients who require immediate care and 5 corresponds to the least ill patients. Usually the highest ill patients (level 1 and 2) arrive to the ED by an ambulance while the lower acuity patients (level 3 and 4) walk to the ED. For this reason we assume that ambulance patients have higher priority over walk-in patients. A patient that begins service, completes its service before another patient is admitted, regardless of the priorities of the patients in the queue. Walk-in patients arrive to the k^{th} ED according to a Poisson process with rate λ_k.

The Length of Stay (LOS) of patients inside the ED corresponds to the time the patient spends with the doctor, time he waits for lab results, until he is discharged by the ED doctor. We assume that this time is exponentially distributed with parameter $1/\mu_k$. The capacity of the ED is determined by the number of beds available (c_k).

In terms of available capacity at the EMS, there are N ambulances available to serve a region's emergency calls. We assume that transit times (travel times, time at scene, travel times to the hospital) are independent and exponentially distributed with mean $\frac{1}{\mu_0}$. Restrepo et al. [49] and the references therein assume that the total service time (which includes offload delays) is exponential. They also argue that the time spent by an ambulance at the scene typically dominates the travel time, so the dependence between calls is mild.

The model parameters are:

- K: Number of hospitals in a region;
- λ_0: Patient arrival rate to the EMS system;
- p_k: Probability that a patient arrival to the EMS triggers an arrival to the k^{th} ED, for $k = 1, \ldots, K$;
- μ_k: Service rate per server in the k^{th} ED, for $k = 1, \ldots, K$;
- μ_0: Service rate per server in the EMS;
- N: Total number of ambulances available;
- λ_k: Walk-in patients arrival rate to the k^{th} ED, for $k = 1, \ldots, K$;
- c_k: Number of servers (beds) at the k^{th} ED, for $k = 1, \ldots, K$.

For notational convenience, in Section 5.2, we remove subscript k from μ_k, c_k, p_k, and λ_k, which are parameters for the k^{th} ED. We still use parameters μ_0 and λ_0 for the EMS. Because there is no buffer between the EMS stage and the ED stage, upon service completion at the EMS node, a patient may get blocked if at that moment he finds all the downstream node servers (beds) are occupied. The server of that patient (ambulance) will get blocked too. When a patient departure occurs at the downstream node (ED), one of the blocked patients at the

Possible event	Rate	From	To	Condition
an ambulance patient arrival	$p\lambda_0$	(q_1, q_2, q_3)	$(q_1, q_2, q_3 + 1)$	–
a walk-in patient arrival to ED	λ_1	(q_1, q_2, q_3)	$(q_1, q_2 + 1, q_3)$	if $q_2 < c$
		(q_1, q_2, q_3)	$(q_1 + 1, q_2, q_3)$	if $q_2 \geq c$
Service completion at ED	$c\mu$	(q_1, q_2, q_3)	$(q_1, q_2 - 1, q_3)$	if $q_2 > c$
	$q_2\mu$	(q_1, q_2, q_3)	$(q_1, q_2 - 1, q_3)$	if $q_2 \leq c$ and $q_1 = 0$
	$c\mu$	(q_1, q_2, q_3)	$(q_1 - 1, q_2, q_3)$	if $q_2 = c$ and $q_1 > 0$
Service completion at EMS	$q_3\mu_0$	(q_1, q_2, q_3)	$(q_1, q_2 + 1, q_3 - 1)$	if $q_3 > 0$

Table 5.1: Transition rates for the model Markov chain

upstream node will start service and its corresponding server will be unblocked. This is referred to as Blocking After Service (BAS).

To describe the system state at any point of time t, we introduce the following state variables:

1. $q_1(t)$: The number of walk-in patients waiting in the queue at time t;

2. $q_2(t)$: The number of both ambulance and walk-in patients in service and ambulance patients blocked at time t;

3. $q_3(t)$: The number of ambulance patients in transit at time t.

It is easy to see that $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$ is a Markov chain. This Markov chain allows us to derive various probability distributions which we use later to derive system performance measures. The state space Ω of the Markov chain $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$ can be organized such that $q_1(t)$ is the level variable, $q_2(t)$ is the sublevel variable, and $q_3(t)$ is the phase variable. Thus, the state space is as follows:

- $\Omega = \Omega_0 \cup \Omega_1 \cup \ldots$;

- $\Omega_i = \Omega_{i,0} \cup \Omega_{i,1} \cup \ldots \cup \Omega_{i,N+c}$, for $i = 0$; $\Omega_i = \Omega_{i,c} \cup \Omega_{i,c+1} \cup \ldots \cup \Omega_{i,N+c}$, for $i > 0$;

- $i = 0 : \Omega_{0,j} = \Omega_{0,j,0} \cup \Omega_{0,j,1} \cup \ldots \cup \Omega_{0,j,N-(j-c)^+}$;
  $i \geq 1 : \Omega_{i,j} = \Omega_{i,j,0} \cup \Omega_{i,j,1} \cup \ldots \cup \Omega_{i,j,N-(j-c)}$, for $c \leq j \leq c + N$.

Table 5.1 illustrates the possible transitions in the system along with the corresponding transition rates. $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$ is a continuous time Markov chain with the following infinitesimal generator:

$$Q = \begin{pmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{5.1}$$

We call $q_1(t)$ a level variable and $(q_2(t), q_3(t))$ the phase vector. Based on this organization $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$ is a level independent QBD process. In the next paragraphs we provide all the details of the Markov chain. Once we construct all the Markov chain infinitesimal generator details, we can find its exact steady state probability distribution using the Matrix Analytic Method which we illustrate in the next section. The boundary matrices A_{00}, A_{01}, A_{10} details are as follows:

$$A_{0,0} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ N+c-1 \\ N+c \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & N+c-1 & N+c \end{array} \\ \begin{pmatrix} a_0 & b_0 & & & & \\ d_1 & a_1 & b_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & d_{N+c-1} & a_{N+c-1} & b_{N+c-1} \\ & & & & d_{N+c} & a_{N+c} \end{pmatrix} \end{array} \tag{5.2}$$

$$A_{1,0} = c\mu \cdot \begin{array}{c} \\ c \\ c+1 \\ \vdots \\ N+c \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & \dots & c & c+1 & \dots & N+c \end{array} \\ \begin{pmatrix} 0 & \dots & I_{N+1} & & & \\ \vdots & & & 0_N & & \\ \vdots & & & & \ddots & \\ 0 & & & & & 0 \end{pmatrix} \end{array} \tag{5.3}$$

$$A_{0,1} = \lambda_1 \cdot \begin{array}{c} \\ 0 \\ \vdots \\ c \\ c+1 \\ \\ N+c \end{array} \begin{array}{c} \begin{array}{cccc} c & c+1 & \dots & N+c \end{array} \\ \begin{pmatrix} 0 & & & \\ \vdots & & & \\ I_{N+1} & & & \\ & I_N & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \end{array} \tag{5.4}$$

The rate at which the number of walk-in patients increases is defined in the matrix A_0. We note that the queue size increases by one only when the beds at the destination ED are full, otherwise it does not change. The details of A_0 are as follows:

$$A_0 = \lambda_1 \cdot \begin{array}{c} \\ c \\ c+1 \\ \\ N+c \end{array} \begin{array}{c} \begin{array}{cccc} c & c+1 & \dots & N+c \end{array} \\ \begin{pmatrix} I_{N+1} & & & \\ & I_N & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \end{array} \tag{5.5}$$

The matrix A_1 includes transitions that do not affect the walk-in patients queue length; it includes service completions of ambulances, service completions of ambulance patients, and ambulance patients arrival to the EMS service. The details of A_1 are as follows:

$$A_1 = \begin{array}{c|ccccc} & c & c+1 & \dots & N+c-1 & N+c \\ \hline c & a_c & b_c & & & \\ c+1 & d_{c+1} & a_{c+1} & b_{c+1} & & \\ & & \ddots & \ddots & \ddots & \\ N+c-1 & & & d_{c+N-1} & a_{c+N-1} & b_{c+N-1} \\ N+c & & & & d_{c+N} & a_{c+N} \end{array} \tag{5.6}$$

The matrix A_2 represents the rate at which the walk-in patients queue decreases by one. Because those patients poses lower priority than patients arriving by an ambulance, a walk-in patient cannot get admitted unless there are no patients of the higher priority waiting for a bed, or simply when $q_2(t) > c$. The details of A_2 are as follows:

$$A_2 = c\mu \cdot \begin{array}{c} \\ c \\ c+1 \\ \\ N+c \end{array} \begin{array}{c} \begin{array}{cccc} c & c+1 & \dots & N+c \end{array} \\ \begin{pmatrix} I_{N+1} & & & \\ & 0_N & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \end{array} \tag{5.7}$$

At the boundary, the transition rates are different because when $q_1(t) = 0$ the states $\{(0, 0, \cdot), \dots, (0, c-$

$1, \cdot)\}$ exist.

$$a_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ \\ N-(i-c)^+ \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & & N-(i-c)^+ \end{array} \\ \begin{pmatrix} * & p\lambda_0 & & & & \\ & * & p\lambda_0 & & & \\ & & * & p\lambda_0 & & \\ & & & \ddots & \ddots & \\ & & & & * & p\lambda_0 \\ & & & & & * \end{pmatrix} \end{array} \tag{5.8}$$

where $*$ is calculated such that the rows of the matrix Q sum to zero. $a_i(j,j) = -(\lambda_1 + \min{(i,c)}\mu + p_1\lambda_0 + j\mu_0)$

$$b_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \dots & & N \end{array} \\ \begin{pmatrix} \lambda_1 & & & & & \\ \mu_0 & \lambda_1 & & & & \\ & 2\mu_0 & \lambda_1 & & & \\ & & \ddots & \ddots & & \\ & & & & N\mu_0 & \lambda_1 \end{pmatrix} \end{array}, \quad \text{for } i < c \tag{5.9}$$

$$b_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ N-(i-c)-1 \\ N-(i-c) \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \dots & N-(i-c)-1 \end{array} \\ \begin{pmatrix} 0 & & & & \\ \mu_0 & 0 & & & \\ & 2\mu_0 & 0 & & \\ & & \ddots & \ddots & \\ & & & (N-(i-c)-1)\mu_0 & 0 \\ & & & & (N-(i-c))\mu_0 \end{pmatrix} \end{array}, \quad \text{for } i \geq c \tag{5.10}$$

$$d_i = \min(i,c) \cdot \mu \cdot I_{N+1-(i-c)^+} \quad \text{for } i \leq c \tag{5.11}$$

$$d_i = \min{(i,c)} \cdot \mu \cdot \begin{pmatrix} I_{N+1-(i-c)} & 0 \end{pmatrix} \quad \text{for } i > c \tag{5.12}$$

5.2 The Exact Analysis

In this section, we analyze the Markov chain $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$ and find performance measures in two steps. Then we use the results for the single ED network to find performance measures for the entire EMS-ED network.

5.2.1 A Matrix Geometric Solution

We denote by $\boldsymbol{\pi} = (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \ldots)$ the stationary probability distribution of $\{q_1(t), q_2(t), q_3(t), t \geq 0\}$, where $\boldsymbol{\pi}_i$ includes all the limiting probabilities of the states in level Ω_i. The stationary distribution exists if and only if the Markov chain is ergodic.

Let $A = A_0 + A_1 + A_2$. Let $\boldsymbol{\theta}$ satisfy $\boldsymbol{\theta} A = 0$ and $\boldsymbol{\theta e} = 1$. Since the Markov chain of interest is irreducible and has a QBD structure, by Neuts [46], the Markov chain is ergodic if and only if $\boldsymbol{\theta} A_0 \boldsymbol{e} < \boldsymbol{\theta} A_2 \boldsymbol{e}$, which can be reduced to

$$\lambda + p\lambda_0 < c\mu \tag{5.13}$$

If the ergodicity condition is satisfied, then $\boldsymbol{\pi}$ exists and it can be found as described in Section 2.4.

In the next two subsections, we utilize the matrix-geometric solution to find performance measures at the hospital level and the regional level.

5.2.2 Hospital level performance measures

The performance measures of interest at the hospital level are:

1. ED utilization: The maximum arrival rate to the ED node is $\lambda + p\lambda_0$. While the available capacity for those arrivals is $c\mu$. As a result, the ED utilization, ρ, can be calculated as follows:
$$\rho = \frac{\lambda + p\lambda_0}{c\mu} \tag{5.14}$$

2. The distribution of the number of walk-in patients in the queue can be found using the matrix geometric solution as $\{\boldsymbol{\pi}_i \boldsymbol{e}, \quad i = 0, 1, \ldots\}$. The mean number of walk-in patients

in the queue can be calculated by

$$E[q_w] = \sum_{i=0}^{\infty} i\boldsymbol{\pi}_i \boldsymbol{e} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_1 R^{i-1}\boldsymbol{e} = \boldsymbol{\pi}_1 \left(\sum_{i=1}^{\infty} iR^{i-1}\right)\boldsymbol{e} = \boldsymbol{\pi}_1 (I-R)^{-2}\boldsymbol{e}. \tag{5.15}$$

3. The distribution for the number of ambulances in offload delay at an ED. Let $\{\eta_2(j), 0 \leq j \leq c+N\}$ be the distribution of $q_2(t)$. Recall that $q_2(t)$ is defined as the number of both ambulance patients and walk-in patients in service and blocked. If $q_2(t)$ is greater than c, then there are $q_2(t) - c$ ambulances blocked outside the ED. That distribution can be obtained from the two vectors $\{\boldsymbol{\pi}_0, \boldsymbol{\pi}_1(I-R)^{-1}\}$. We use this fact to find the probability distribution of the number of ambulances in offload delay as follows:

$$\xi_2(m) = \begin{cases} \sum_{j=0}^{j=c} \eta_2(j), & \text{for} \quad m = 0 \\ \eta_2(m+c), & \text{for} \quad m = 1, \ldots, N \end{cases} \tag{5.16}$$

The mean queue number of ambulances in offload delay $E[q_a]$ can be found accordingly.

4. Offload delay distribution: Offload delays occur when ambulance patients are forced to wait for a bed. Thus, the offload delay distribution is the waiting time distribution for ambulance patients who possess the high nonpreemptive priority. We use a phase-type distribution to model ambulance patients waiting times. This type of distributions allows us to capture heterogeneity in patients waiting where there may be a large variation in the amount of time patients spend in the hospital. The waiting time w_a of an ambulance patient arriving to an ED depends on the number of ambulance patients at the ED and the number of walk-in patients already in service which are captured in the state variable $q_2(t)$. Denote by $\alpha(j)$ the probability that j patients are in the ED when an ambulance patient arrives. Since there are c beds for all patients in the ED, each with an exponential service time with parameter μ, if all beds are occupied, the time to serve one patient has an exponential distribution with parameter $c\mu$. Thus, the total time to serve j patients has an Erlang distribution of order j. Consequently, when an ambulance patient arrives to the hospital ED, the waiting time has a generalized Erlang distribution with a phase-type

representation $(\boldsymbol{\alpha}, c\mu J_N)$, where

$$J_N = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}_{N \times N} . \tag{5.17}$$

$$\alpha(j) = \eta_2(c+j) \quad \text{for} \quad j = 0, 1, \ldots, N \tag{5.18}$$

The distribution function of the waiting time w_a is given by $P\{W_a < t\} = 1 - \boldsymbol{\alpha} \exp\{-c\mu J_N t\}\boldsymbol{e}$. The expected offload delays can be found using the formula:

$$E[W_a] = \frac{1}{c\mu} \sum_{j=1}^{N} (j+1)\alpha(j) \tag{5.19}$$

Little's Law holds for the queueing model, i.e., $E[q_a] = p\lambda_0 E[W_a]$, which can be used to check the accuracy of the results obtained.

5. Waiting time distribution for the walk-in patients: The waiting time w_w of a walk-in patient arriving to an ED depends on the number of ambulance patients in service and waiting outside the ED. It also depends on the number of walk-in patients already in service which is captured in the state variable $q_2(t)$. First, we find the waiting time for a tagged walk-in patient arriving to the ED. Then, we generalize the result for an arbitrary patient using conditional probability concepts. The waiting time for a tagged walk-in patient has a phase type structure. Since walk-in patients possess a lower priority with respect to patients arriving by an ambulance, a walk-in patient cannot get admission unless there is a bed available for him, or mathematically only when $q_2(t) < c$. Denote by $\alpha_w(n-1)$ the probability that the tagged walk-in patient arriving to the ED finds $n-1$ walk-in patients ahead of him in the queue, and $j-c$ patients blocking ambulances (recall that when $j < c$ the patient does not have to wait); then he has to wait for $i+j-c$ service completions plus the service completions for ambulance patients who arrived during his waiting time. The events that affect the tagged walk-in patient's waiting time are service completions of both patients types, and arrivals of high priority ambulance patients during his waiting time. While subsequent arrivals of walk-in patients do not affect his waiting

time because same priority patients are admitted based on a FCFS rule.

If the state of the ED at time t upon arrival of the the tagged walk-in patient was (i, j, l) as described by the triplet $(q_1(t), q_2(t), q_3(t))$, then the waiting time constitutes the time until absorption for the continuous time Markov chain that has the infinitesimal generator as follows:

$$Q_{w,n} = \begin{pmatrix} A_1 + \lambda_1 I & & & \\ A_2 & A_1 + \lambda_1 I & & \\ & \ddots & \ddots & \\ & & A_2 & A_1 + \lambda_1 I \end{pmatrix} \tag{5.20}$$

Note that $Q_{w,n}$ has a similar structure as Q but without the transitions associated with walk-in patients' arrival. The tagged walk-in patient's waiting time has a phase type distribution with matrix representation $((0, ..., 0, \boldsymbol{\pi}_{n-1}/(\boldsymbol{\pi}_{n-1}\boldsymbol{e})), Q_{w,n})$. Then we obtain the conditional probability distribution of the waiting time as:

$$P(W_w \leq t \mid n) = 1 - (0, ..., 0, \boldsymbol{\pi}_{n-1}/(\boldsymbol{\pi}_{n-1}\boldsymbol{e})) \exp\{Q_{w,n}t\}\boldsymbol{e}. \tag{5.21}$$

The distribution of the waiting time of an arbitrary walk-in patient can be obtained as

$$P(w_w \leq t) = 1 - \sum_{n=0}^{\infty}(0, ..., 0, \boldsymbol{\pi}_n) \exp\{Q_{w,n}t\}\boldsymbol{e}. \tag{5.22}$$

Theorem 1. *The mean waiting time for an arbitrary walk-in patient who arrives to the ED is:*

$$\begin{aligned} E[W_w] = \; & \widetilde{\pi}_0 L + \pi_1(I - R)^{-1}(I - B + \boldsymbol{e}\theta)^{-1}L \\ & -\pi_1\phi_{inv}(\phi(I)(I - R' \otimes B)^{-1})B^2(I - B + \boldsymbol{e}\theta)^{-1}L \\ & +\pi_1\left[(I - R)^{-2} + (I - R)^{-1}\right]\boldsymbol{e}\theta L \end{aligned} \tag{5.23}$$

Proof. The mean waiting time for the tagged walk-in patient who arrive to the system and find $n - 1$ walk-in patients ahead of him can be found as follows:

$$E[W_w \mid n] = -(0, \ldots, \frac{\pi_{n-1}}{\pi_{n-1}\boldsymbol{e}})Q_{w,n}^{-1}\boldsymbol{e} \tag{5.24}$$

Since $Q_{w,n}$ has a lower diagonal structure, its inverse can be found as follows:

$$\begin{aligned} Q_{w,n}^{-1} = \quad & -(A_1+\lambda_1 I)^{-1} + (A_1+\lambda_1 I)^{-1}A_2(A_1+\lambda_1 I)^{-1} - \ldots \\ & +(-1)^{n+1}(A_1+\lambda_1 I)^{-1}A_2(A_1+\lambda_1 I)^{-1}\ldots(A_1+\lambda_1 I)^{-1} \end{aligned} \tag{5.25}$$

Let $B = -(A_1+\lambda_1 I)^{-1}A_2$, and $L = -(A_1+\lambda_1 I)^{-1}\boldsymbol{e}$, then the conditional expected waiting time for the tagged walk-in patient can be found as follows when he arrives to an empty queue:

$$E[W_w \mid n=1] = [\pi_{0,c}, ..., \pi_{0,c+N}]L \tag{5.26}$$

Or from the following formula if the tagged walk-in patient arrives to a non-empty queue:

$$E[W_w \mid n>1] = \frac{\pi_{n-1}}{\pi_{n-1}\boldsymbol{e}}\left[I+B++B^2+\ldots+B^{n-1}\right]L \tag{5.27}$$

By combining the above equations we find the expected waiting time for an arbitrary walk-in patient as follows:

$$E[W_w] = \sum_{n=1}^{\infty}\pi_n\boldsymbol{e}E[W_w \mid n] = \widetilde{\pi}_0 L + \sum_{n=2}^{\infty}\pi_1 R^{n-2}\left[I+B++B^2+\ldots+B^{n-1}\right]L \tag{5.28}$$

where $\widetilde{\pi}_0 = [\pi_{0,c}, ..., \pi_{0,c+N}]$. Since B is a stochastic matrix then there exits a vector θ such that $\theta B = \theta$ and $\theta\boldsymbol{e} = 1$, to find θ, and since B has a special structure as follows: $B = \begin{pmatrix} B_{11} & 0 \\ B_{21} & 0 \end{pmatrix}$, where B_{11} is of size $N+1$

Then, we find first θ_1 such that $\theta_1(B_{11}-I)=0$ and $\theta_1 e = 1$. Then $\theta = [\theta_1, \mathbf{0}]$.

To find the geometric sum $\left[I+B++B^2+\ldots+B^{n-1}\right]$ we multiply it with $(I-B+\boldsymbol{e}\theta)$, the detailed steps are as follows:

$$\begin{aligned} \left(I+B++B^2+\ldots+B^{n-1}\right)(I-B+\boldsymbol{e}\theta) \quad &= (I-B^n+n\boldsymbol{e}\theta) \\ \left(I+B++B^2+\ldots+B^{n-1}\right) \quad &= (I-B^n+n\boldsymbol{e}\theta)(I-B+\boldsymbol{e}\theta)^{-1} \\ \left(I+B++B^2+\ldots+B^{n-1}\right) \quad &= (I-B^n)(I-B+\boldsymbol{e}\theta)^{-1} + n\boldsymbol{e}\theta(I-B+\boldsymbol{e}\theta)^{-1} \\ \left(I+B++B^2+\ldots+B^{n-1}\right) \quad &= (I-B^n)(I-B+\boldsymbol{e}\theta)^{-1} + n\boldsymbol{e}\theta \end{aligned} \tag{5.29}$$

Next, we substitute the above result in equation (5.28). The detailed steps are as follows:

$$\begin{aligned} E[W_w] &= \widetilde{\pi}_0 L + \sum_{n=2}^{\infty} \pi_1 R^{n-2} \left[(I - B^n)(I - B + \boldsymbol{e}\theta)^{-1} + n\boldsymbol{e}\theta \right] L \\ &= \widetilde{\pi}_0 L + \sum_{n=2}^{\infty} \pi_1 R^{n-2} (I - B + \boldsymbol{e}\theta)^{-1} L - \sum_{n=2}^{\infty} \pi_1 R^{n-2} B^n (I - B + \boldsymbol{e}\theta)^{-1} L + \sum_{n=2}^{\infty} \pi_1 R^{n-2} n\boldsymbol{e}\theta L \\ &= \widetilde{\pi}_0 L + \pi_1 (I - R)^{-1} (I - B + \boldsymbol{e}\theta)^{-1} L - \pi_1 \sum_{n=2}^{\infty} R^{n-2} B^{n-2} B^2 (I - B + \boldsymbol{e}\theta)^{-1} L + \pi_1 \left[(I - R)^{-2} + (I - R)^{-1} \right] \boldsymbol{e}\theta L \end{aligned} \tag{5.30}$$

To find the infinite sum $\sum_{n=2}^{\infty} R^{n-2} B^{n-2}$, we use the direct sum as follows:

$$\phi \left(\sum_{n=2}^{\infty} R^{n-2} B^{n-2} \right) = \phi(I)(I - R' \otimes B)^{-1} \tag{5.31}$$

where $\phi(I)$ is a row vector and is obtained by stringing out the vectors starting form the first row of I. Equations (5.30) and (5.31) can be used to find the expected queue length for walk-in patients without the need for truncation. We use Little's Law to verify our results for the expected waiting time. □

5.2.3 Regional level performance measures

In this section, we consider performance measures for the entire EMS-ED network. We focus on the availability of ambulances. In Section 5.1, we assumed that individual single ED networks operate independently. Based on the assumptions, we construct performance measures from that of individual single ED networks.

1. Let $\boldsymbol{\xi}_2^{(k)} = (\xi_2^{(k)}(0), \xi_2^{(k)}(1), ..., \xi_2^{(k)}(N + c_k))$, for $k = 1, 2, ..., K$, be the distribution of the number of ambulances in offload delay in the k^{th} single ED network. Since we assume that all K single ED networks operate independently, the distribution of the total number of ambulances in offload delay can be found as the convolution of $\{\boldsymbol{\xi}_2^{(1)}, ..., \boldsymbol{\xi}_2^{(K)}\}$. Note that the total number of ambulances in offload delay found by using this method may exceed N, and it is an approximation to the actual number of ambulances in offload delay.

2. Let $\boldsymbol{\xi}_3^{(k)}$ be the distribution for the number of ambulances in transit, which is the distribution of $q_3(t)$ at the k^{th} ED. Then $\boldsymbol{\xi}_3^{(k)}$ can be obtained from $\{\boldsymbol{\pi}_0, \boldsymbol{\pi}_1(I - R)^{-1}\}$. The distribution of the total number of ambulances in transit can be obtained by the convolution of $\{\boldsymbol{\xi}_3^{(1)}, ..., \boldsymbol{\xi}_3^{(K)}\}$. Again, due to the approximation assumption, the total number of ambulances in transit obtained in this way may exceed N.

3. EMS utilization: The load faced by the EMS consists of two parts, transferring the high priority patients to the corresponding EDs, and serving patients who are blocking ambulances. It is clear that blocked ambulances consume the EMS utilization as well as new arrivals to the EMS. The total arrival for the EMS is λ_0 while the total capacity available is $N\mu_0$. In the case of no blocking, the EMS utilization, ρ_{EMS}, would be $\frac{\lambda_0}{N\mu_0}$. But since blocking exists, we modify the available capacity to take into account the effect of blocking or offload delays.

$$\rho_{EMS} = \min\left\{1, \frac{\lambda_0}{(N - \sum_{k=1}^{K} E[q_a]^{(k)})\mu_0}\right\}. \tag{5.32}$$

5.3 Approximation and Optimization

The network under consideration consists of two nodes, as illustrated in Figure 5.2; node 0 is the EMS node, and node 1 is the ED node. One of our main observations related to the project with the ROW is that one hospital ED is contributing to most of the total offload delays. This leads to the problem of load balancing or work allocation for the region EDs. Our aim here is to propose an approximation for the network to obtain explicit and simple expressions for performance measures. The approximation can serve as an effective way to determine the optimal proportion of patients that should be routed to individual EDs.

In a related work, Deo and Gurvich [4] develop a queueing game model for two EDs in which each ED tries to minimize its waiting times. They show that decentralized diversion decisions result in a depooling of the network resources. They provide a near optimal solution for the ambulance diversion problem when a centralized dispatcher coordinates diversion. In that sense. our model extends Deo and Gurvich's work by considering any number of EDs and by explicitly including offload delays into the analysis.

To approximate the performance measures of this network, we develop an approximation scheme by further decomposing the single ED network into single node queues each with modified arrival rate, service rate, and number of servers. Perros [27] used this methodology to analyze queueing networks with blocking of various topologies. Our model is different since it deals with a queueing network with blocking and multiple servers. Moreover, arrivals to the network possess different priorities. Later, we utilize the matrix geometric solution of Section 5.2 to assess the effectiveness of the approximation results.

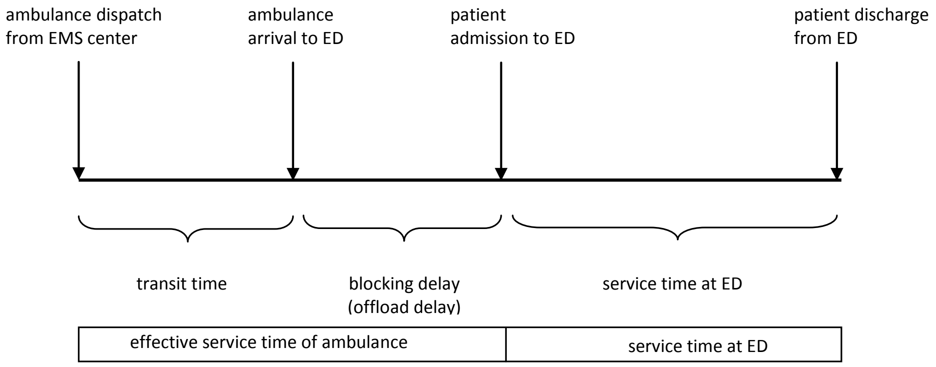
ambulance dispatch from EMS center
ambulance arrival to ED
patient admission to ED
patient discharge from ED
transit time
blocking delay (offload delay)
service time at ED
effective service time of ambulance
service time at ED

Figure 5.3: Illustration of blocking delay

5.3.1 Approximation using the $M/M/c$ queue

Our methodology is based on substituting a single ED network with two disconnected nodes each with a modified arrival rate, service rate, and number of servers that represent the effective parameter. Then, we use the existing results on single node queues with multiple servers to derive the expected queue length and the expected waiting time for each node of the queueing network. We start the analysis with the queue within an ED, called Node 1, since this node cannot get blocked. Then we use Node 1 results to approximate the parameters of Node 0.

1. Node 1. The queue within an ED can be treated as an isolated $M/M/c$ queue with c servers each with service rate μ, and two types of arrivals with nonpreemptive priority service discipline. The high priority arrival rate for node 1 is still $p\lambda_0$. While the arrival rate of the low priority patients is λ. The approximate mean waiting in the queue for the high priority arrivals, $E[\hat{W}_a]$, and the low priority arrivals, $E[\hat{W}_a]$, can be calculated as (cf. Gross et al. [51]):

$$E[\hat{W}_a] = \frac{1}{1-\sigma_1}\left(c!(1-\rho)c\mu\sum_{n=0}^{c-1}\frac{(c\rho)^{n-c}}{n!} + c\mu\right)^{-1}, \tag{5.33}$$

where $\sigma = p\lambda_0/(c\mu)$ and $\rho = (p\lambda_0 + \lambda)/(c\mu) = \sigma + \lambda/(c\mu)$.

In order to find the dispatching probability vector p to minimize system costs (expected waiting times), we prove an interesting property on $E[\hat{W}_a]$.

Theorem 2. *$E[\hat{W}_a]$ is convex in σ*

Proof. First note that equation (1) can be rewritten, in term of the Erlang delay formula $B(c, \rho)$, as follows:

$$E[\hat{W}_a] = \frac{1}{c\mu(1-\sigma)} B(c, \rho) \tag{5.34}$$

By [54] and [55], $B(c, \rho)$ is convex in ρ. Then $E[\hat{W}_a]$ is convex in σ. □

Since we are optimizing a convex function with equality constraints, this leads to a convex problem. Based on this Theorem, any optimal solution found is in deed a global optimal solution for the ambulance routing problem.

The mean waiting time of walk-in patients (lower priority patients) is given by (e.g., Gross et al. [51]):

$$E[\hat{W_w}] = \frac{1}{(1-\sigma)(1-\rho)} \left(c!(1-\rho)c\mu \sum_{n=0}^{c-1} \frac{(c\rho)^{n-c}}{n!} + c\mu \right)^{-1}. \tag{5.35}$$

The expected queue length for both arrival streams, $E[\hat{L_a}]$ and $E[\hat{L_w}]$, can be obtained by using Little's law:

$$E[\hat{L_a}] = p_1\lambda_0 E[\hat{W}_a], \qquad E[\hat{L_w}] = \lambda_1 E[\hat{W}_w] \tag{5.36}$$

2. Node 0: Node 0 can be treated as an isolated $M/M/c/c$ queue with one stream of arrivals, which is the high priority patients only. First, we need to find the effective number of servers at node 0. Due to the blocking that occurs for some of the N servers, arrivals to this node will see less available servers than the node has because some servers will be blocked when node 1 is full. If the expected queue length of the high priority arrivals exceeds c, then on average there are $E[\hat{L_a}] - c$ servers at node 0 lost due to blocking, which makes the effective number of servers at node 0, $\hat{N} = N - (E[\hat{L_a}] - c)^+$. To find the effective arrival rate at node 0, we need to account for the probability the node is full, which results in a rejection of arrival. Thus, $\hat{\lambda_0} = p\lambda_0(1 - \pi_N)$, where π_N is the probability that node 0 is full. This probability equals zero in our model, which is an operating characteristic of the EMS-ED network of interest. If, in other cases, this node

is highly utilized then this probability can be calculated from the $M/M/c$ queue results. Lastly, we need to derive the effective service rate at node 0. Although, the service rate per server at node 0 is μ_0, arrivals are expected to spend more time in the server waiting for an empty server at the destination node 1. We estimate the effective service rate per server at node 0 as follows:

$$\hat{\mu_0} = (\frac{1}{\mu_0} + E[\hat{W_a}])^{-1} \tag{5.37}$$

The above approximation for the service rate is similar to Koizumi et al. [34]. After we derive the effective arrival rate, service rate, and number of servers at node 0, we use the $M/M/c/c$ queue results to find performance measures.

5.3.2 The Ambulance Routing Problem

Offload delays occur due to the high utilization experienced at the destination hospital EDs. In order to eliminate or decrease the offload delays ambulances experience, it is trivial to suggest to add more beds at the destination hospital EDs in order to accommodate the high traffic. However, it is less trivial to try to decrease offload delays given the same network settings. i.e. by keeping the same number of beds and ambulances. One of the main observations we had with respect to the Region of Waterloo network was the different utilization levels experienced at the regions' three hospitals which resulted in higher offload delays at one hospital more than the other.

The problem of workload allocation relates to the general stream of research in call centers and flow shop manufacturing systems. Examples include [52] and [56] among others. In the manufacturing context, optimal assignment of tasks to machine centers is found either by minimizing total costs or by maximizing profit. In this section, we develop an optimization problem to find the routing probability vector that minimizes the total offload delays experienced by an EMS provider in a region. We utilize the approximation scheme developed in Section 5.2 for the expected offload delays which we aim to minimize given that all patients should be sent to a hospital. The optimization problem is:

Parameter set	value
N	17 ambulances
λ_0	3 patients per hr
μ_0	1 patient per hr
$\lambda_1, \lambda_2, \lambda_3$	(4.2,3.5,3.3) patient per hr
(μ_1, μ_2, μ_3)	$(1/6, 1/6, 1/6)$ patient per hr
(c_1, c_2, c_3)	$(35, 30, 29)$ beds
(p_1, p_2, p_3)	$(0.45, 0.29, 0.26)$

Table 5.2: Case study input parameters

$$\begin{aligned} \min_{p_1,\ldots,p_k} \quad \sum_{k=1}^{K} E[\hat{W}_a]_k = \sum_{k=1}^{K} \frac{1}{1-\sigma_k} &\left(c_k!(1-\rho_k)c_k\mu_k \sum_{n=0}^{c_k-1} \frac{(c_k\rho_k)^{n-c_k}}{n!} + c_k\mu_k \right)^{-1} \\ s.t. \qquad & \sum_{k=1}^{K} p_k = 1 \\ & p_k \geq 0 \end{aligned} \tag{5.38}$$

where $\sigma_k = \dfrac{p_k\lambda_0}{c_k\mu_k}$ and $\rho_k = \dfrac{p_k\lambda_0 + \lambda_k}{c_k\mu_k}$, $k = 1, 2, \ldots, K$. By Theorem 2, the objective function is convex in $\{\sigma_1, ..., \sigma_K\}$. Thus, the optimization problem is a convex programming problem, which can be solved effectively by existing methods. We solve the above optimization problem using the *fmincon* solver in Matlab (note: we use the *interior-point* algorithm).

5.4 Numerical Analysis

In this section, we analyze a number of examples numerically. The examples emerged from a running project with the Region of Waterloo (ROW) EMS. The Region of Waterloo, which is located in Southern Ontario Canada, is served by three hospitals and one EMS provider. The input parameters used are recorded in Table 5.2, our choice of the parameters was guided by the real data acquired from the ROW EMS and Grand River Hospital, one of the main hospitals in the region.

The steps for the analysis are:

1. Find the exact performance measures for the EMS-ED network using results from §5.2.

2. Solve the optimization problem of §5.3.2 to find optimal routing probabilities.

		ambulance patients		walk-in patients	
hospital	ρ	$E[W_a]$	$E[q_a]$	$E[W_w]$	$E[q_w]$
ED 1	93.7%	0.15	0.21	3.17	13.33
ED 2	84.8%	0.09	0.08	0.72	2.51
ED 3	81.4%	0.07	0.06	0.45	1.50

Table 5.3: Case study exact results

3. Find the approximate performance measures using results from §5.3.1 and compare the exact and approximate results to ensure that the optimal routing obtained from the approximation model is indeed optimal or near-optimal.

4. Perform extra computational analysis on the ambulance routing problem with variable arrival rates of high priority ambulance patients.

5.4.1 The exact results

As the exact results in Table 5.3 suggest, the first ED, which is the highly utilized one, experiences the highest offload delays compared to the other EDs as quantified by $E[q_a] = 0.21$ patients. At the regional level, we calculate the probability distribution for the number of ambulances in offload delay, or blocked and the probability distribution for the number of ambulances in transit. The results are shown in Figure 5.4.

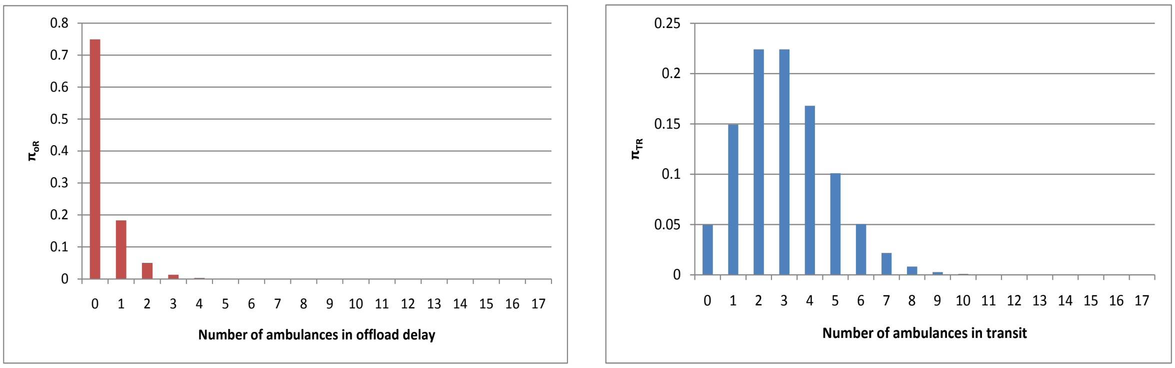

Figure 5.4: Probability distribution for the number of ambulances in offload delay and in transit at ROW

Another performance measure of interest for the EMS management is the probability distribution for the number of busy ambulances. This includes either being busy transferring patients

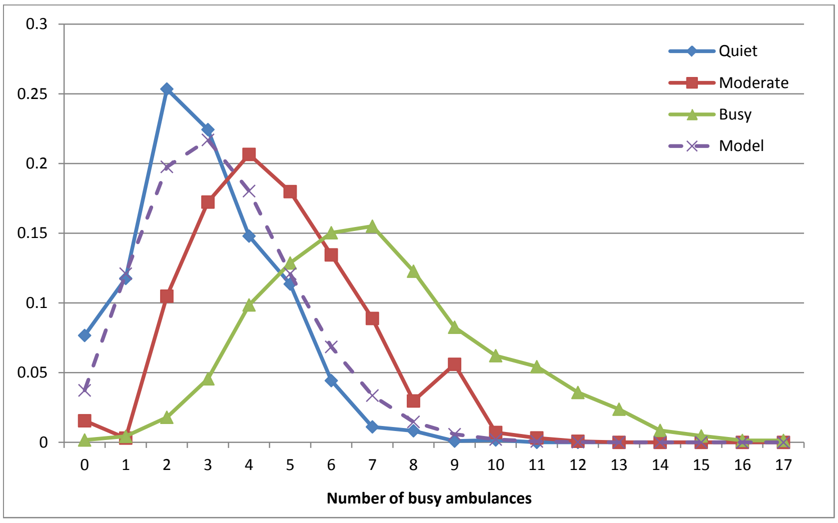

Figure 5.5: Probability distribution for the number of busy ambulances at ROW

or being in offload delay. Figure 5.5 shows the actual distribution for the number of busy ambulances gathered from the ROW EMS data for quiet, moderate, and busy days. It also includes the distribution based on our model output for the case study. As we can see, the queueing model output is close to the actual system performance at the quiet time. This result also validates our modeling approach and assumptions made to get into the steady state solution. In order to compare the system performance at the moderate and busy times, the ambulance patients arrival rate should be increased.

We also utilize the exact matrix geometric solution to calculate the EMS utilization using equation (5.32) and it is found to be 18.01%.

5.4.2 Approximation and comparison of approximation with exact results

In this subsection, we evaluate the efficiency of our approximation scheme by comparing some of the performance measures we have derived with the corresponding exact results acquired from the steady state matrix geometric solution. Namely, we compare the expected queue lengths for different instances when $N = 10$, $\mu_0 = 1$, we report the results in Table 5.4.

As the results suggest, our approximation scheme is effective and the maximum error in estimating the expected queue length is 0.12%. Another result that follows from our approximation

Input parameters	Exact measures		Approximate measures	
$(p_1\lambda_0, \lambda_1, \mu, c)$	$E[q_a]$	$E[q_w]$	$E[\widehat{L_a}]$	$E[\widehat{L_w}]$
$(1, 0.5, 1/2, 5)$	0.0149	0.0249	0.0149	0.0249
$(1.5, 0.5, 1/2, 5)$	0.0559	0.0745	0.0559	0.0745
$(1.5, 1, 1/2, 5)$	0.1619	0.7196	0.1619	0.7197
$(1.5, 1, 2/3, 5)$	0.0445	0.1250	0.0445	0.1251
$(1.5, 1, 2/3, 3)$	0.4634	4.9431	0.4640	4.9491

Table 5.4: Comparison for the expected queue length for the approximation scheme

	Exact results					Approximation results				
Hospital	ρ	$E[q_a]$	$E[q_w]$	$E(W_a)$	$E(W_w)$	$\widehat{\rho}$	$E[\widehat{q}_a]$	$E[\widehat{q}_w]$	$E(\widehat{W}_a)$	$E(\widehat{W}_w)$
ED1	93.7%	0.21	13.33	0.15	3.17	95.1%	0.21	13.33	0.15	3.17
ED2	84.4%	0.08	2.51	0.09	0.72	87.4%	0.08	2.51	0.09	0.72
ED3	81.4%	0.06	1.50	0.07	0.45	84.4%	0.06	1.50	0.07	0.45

Table 5.5: comparison for performance measures for the case study

is the applicability of Little's formula for the expected queue length results derived using Matrix Analytic Methods. More over, we compare the approximation results for the case study with the exact results in Table 5.5. The results show high accuracy of our approximation scheme. In Figure 5.6, we compare the number of busy ambulances probability distribution from the approximation scheme with the Matrix Analytic solution. Although the EMS node can be modeled as an $M/M/c$ queue, the results suggest that including offload delays by adjusting the effective service rate of ambulances have resulted in close results for both methods which is also close to the observed probability distribution in the region ROW. Such an adjustment have increased the width of the tail of the probability distribution. This example indicates that the approximation quality for low utilized first node networks is high.

5.4.3 The optimization problem

We solve the optimization problem outlined in §5.3.2 for the ROW-ED network that consists of three hospitals. Recall that the routing probability vector $(0.45, 0.29, 0.26)$ resulted in a total expected offload delays of 0.3153 hour for the region. When we supply the same network data to the optimization problem we get the optimal routing probability of $(0.3946, 0.3031, 0.3023)$. If the optimal routing probabilities are used, then the total expected offload delays decreases to 0.3086 hour, or it decreases by 2.12%. The detailed results for the network optimal performance

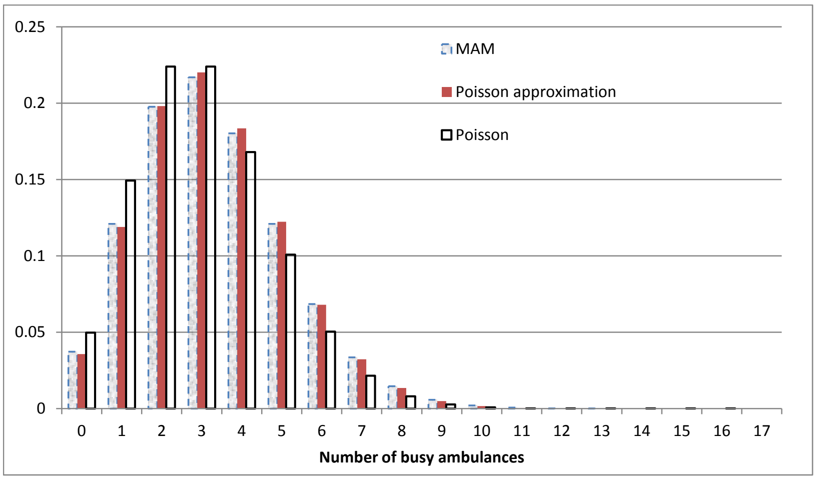

Figure 5.6: Probability distribution for the number of ambulances busy at the ROW

				ambulance patients		walk-in patients	
hospital	p^*	ρ	ρ_a	$E(W_a)$	$E[q_a]$	$E(W_w)$	$E[q_w]$
ED 1	0.3946	90.33%	20.29%	0.1167	0.1381	1.5142	6.3597
ED 2	0.3031	85.56%	18.19%	0.0981	0.0892	0.8308	2.9077
ED 3	0.3023	84.05%	18.76%	0.0939	0.0851	0.7241	2.3896

Table 5.6: Performance measures under optimal routing policy

metrics are summarized in Table 5.6.

Although the main purpose of the developed optimization problem is to decrease the overall offload delays ambulances experience in a region, it resulted also in decreasing walk-in patients waiting. Specifically, we notice that the total expected number of walk-in patients waiting in the waiting room have decreased significantly (32.77%) when the optimal routing decision is used. Decreasing walk-in patients waiting time will increase their morale and prevent their condition from deteriorating while waiting to be seen.

5.4.4 More computational results for the optimization problem

In this subsection, we perform extensive numerical analysis for the optimal routing decisions in the long run for a network that consists of 3 hospitals. The test instances are generated by increasing the ambulance patients arrival rate to the region and fixing all the other model parameters. We picked those numbers based on the Region of Waterloo case. We consider the capacity at each ED to be $(35, 30, 29)$ beds respectively. The service rate at each ED is

λ_0	$p_1{}^\star$	ρ_1	$p_2{}^\star$	ρ_2	$p_3{}^\star$	ρ_3	$E[\hat{W}_a]^R$	$E[W_a]^1$	$E[W_a]^2$	$E[W_a]^3$	$E[W_a]^R$	% dev.
1.0	0.1241	74.11	0.8227	72.05	0.0532	69.36	0.0315	0.0109	0.0117	0.009	0.0316	0.32
1.1	0.1443	74.71	0.7816	72.66	0.0741	69.95	0.0347	0.012	0.0129	0.0099	0.0348	0.29
1.2	0.1612	75.30	0.7473	73.28	0.0915	70.53	0.0383	0.0132	0.0142	0.0109	0.0383	0.00
1.3	0.1754	75.89	0.7184	73.90	0.1062	71.12	0.0421	0.0145	0.0156	0.012	0.0421	0.00
1.4	0.1877	76.49	0.6935	74.52	0.1189	71.71	0.0462	0.016	0.0171	0.0131	0.0462	0.00
1.5	0.1983	77.08	0.6719	75.13	0.1298	72.29	0.0507	0.0175	0.0188	0.0144	0.0507	0.00
1.6	0.2076	77.68	0.6531	75.75	0.1393	72.87	0.0555	0.0192	0.0206	0.0158	0.0556	0.18
1.7	0.2158	78.27	0.6364	76.36	0.1478	73.46	0.0606	0.021	0.0225	0.0172	0.0607	0.17
1.8	0.2231	78.87	0.6216	76.98	0.1552	74.04	0.0662	0.0229	0.0245	0.0188	0.0662	0.00
1.9	0.2297	79.47	0.6084	77.60	0.1619	74.63	0.0721	0.025	0.0267	0.0204	0.0721	0.00
2.0	0.2356	80.06	0.5964	78.21	0.1679	75.21	0.0785	0.0272	0.029	0.0222	0.0784	-0.13
2.1	0.2410	80.66	0.5856	78.83	0.1734	75.79	0.0852	0.0295	0.0315	0.0242	0.0852	0.00
2.2	0.2459	81.26	0.5758	79.45	0.1783	76.38	0.0925	0.0321	0.0342	0.0262	0.0925	0.00
2.3	0.2504	81.86	0.5668	80.06	0.1828	76.96	0.1002	0.0348	0.0371	0.0284	0.1003	0.10
2.4	0.2545	82.45	0.5585	80.67	0.1870	77.55	0.1084	0.0376	0.0401	0.0307	0.1084	0.00
2.5	0.2583	83.05	0.5509	81.29	0.1907	78.12	0.1171	0.0407	0.0433	0.0331	0.1171	0.00
2.6	0.2618	83.65	0.5439	81.90	0.1942	78.71	0.1264	0.0439	0.0467	0.0358	0.1264	0.00
2.7	0.2651	84.25	0.5374	82.52	0.1975	79.29	0.1363	0.0474	0.0504	0.0385	0.1363	0.00
2.8	0.2682	84.86	0.5313	83.13	0.2005	79.88	0.1467	0.0511	0.0542	0.0415	0.1468	0.07
2.9	0.2710	85.46	0.5257	83.74	0.2033	80.46	0.1578	0.0549	0.0583	0.0446	0.1578	0.00
3.0	0.2737	86.06	0.5204	84.35	0.2059	81.04	0.1695	0.059	0.0626	0.0479	0.1695	0.00
3.1	0.2762	86.66	0.5155	84.97	0.2083	81.62	0.1818	0.0634	0.0671	0.0514	0.1819	0.06
3.2	0.2786	87.27	0.5108	85.58	0.2106	82.20	0.1949	0.068	0.0719	0.055	0.1949	0.00
3.3	0.2808	87.87	0.5064	86.19	0.2128	82.79	0.2087	0.0728	0.0769	0.0589	0.2086	-0.05
3.4	0.2829	88.47	0.5023	86.80	0.2148	83.37	0.2232	0.0779	0.0822	0.063	0.2231	-0.04
3.5	0.2850	89.08	0.4983	87.40	0.2167	83.95	0.2384	0.0834	0.0878	0.0673	0.2385	0.04
3.6	0.2869	89.69	0.4946	88.01	0.2185	84.53	0.2545	0.089	0.0936	0.0718	0.2544	-0.04
3.7	0.2887	90.29	0.4911	88.62	0.2202	85.12	0.2714	0.095	0.0998	0.0766	0.2714	0.00
3.8	0.2904	90.90	0.4878	89.23	0.2218	85.70	0.2891	0.1013	0.1063	0.0816	0.2892	0.03
3.9	0.2921	91.51	0.4846	89.83	0.2234	86.28	0.3077	0.1079	0.1130	0.0868	0.3077	0.00
4.0	0.2936	92.11	0.4815	90.43	0.2248	86.86	0.3273	0.1148	0.1201	0.0923	0.3272	-0.03
4.1	0.2952	92.73	0.4786	91.04	0.2262	87.45	0.3477	0.1221	0.1275	0.0981	0.3477	0.00
4.2	0.2966	93.34	0.4758	91.64	0.2276	88.04	0.3692	0.1297	0.1352	0.1042	0.3691	-0.03
4.3	0.2980	93.95	0.4731	92.24	0.2289	88.62	0.3916	0.1377	0.1433	0.1105	0.3915	-0.03
4.4	0.2993	94.56	0.4706	92.84	0.2301	89.21	0.4150	0.1461	0.1518	0.1171	0.4150	0.00
4.5	0.3006	95.17	0.4681	93.44	0.2313	89.79	0.4395	0.1548	0.1606	0.1241	0.4395	0.00

Table 5.7: Optimal routing results

$(1/6, 1/5, 1/6)$. We set μ_0 to equal 1 patient per hour. Lastly, the walk-in patients arrival rates are $(4.2, 3.5, 3.3)$ for ED1, ED2, and ED3 respectively.

Table 5.7 displays the arrival rate (λ_0), the optimal routing decision $(p_1{}^\star, p_2{}^\star, p_3{}^\star)$, the expected offload delays based on the optimization problem ($E[\hat{W}_a]^R$), the separate expected offload delays based on the analytical solution ($E[W_a]^k$), and the difference between the total expected offload delays from the approximation algorithm and the exact analytical algorithm (% dev.).

The computational results reveal the efficiency and accuracy of the approximation scheme developed in the previous section at different system utilization rates. The maximum deviation between the two solution algorithms is reported to be as low as 0.32%. See Table 5.7.

Observation 1: Hospitals with higher number of beds should be loaded more heavily than hospitals with lower number of beds when the arrival rates for low

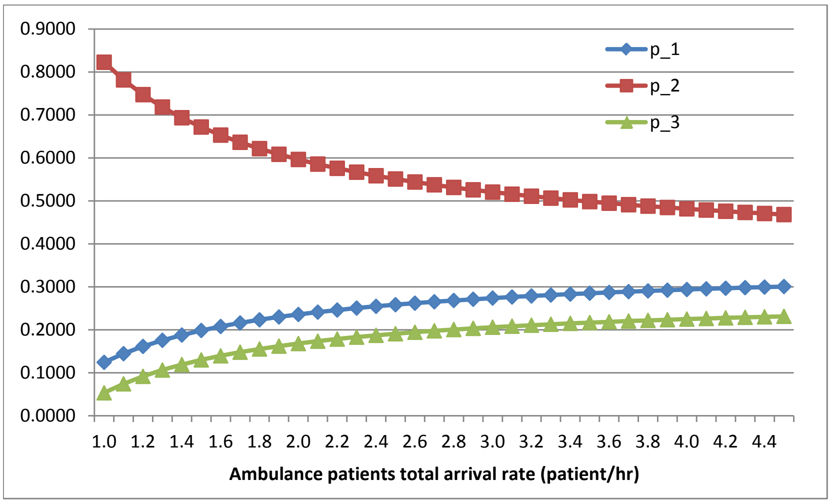

Figure 5.7: Optimal routing probabilities results

priority patients are aligned with the corresponding EDs capacity.

Careful examination of the results in Table 5.7 shows that hospitals with higher number of beds should be loaded more heavily than hospitals with lower number of beds. The intuitive explanation for this results is mainly due to make use of server pooling which reduces congestion. This finding is inline with [52] who look at workload balancing in open Jackson networks of multiserver queues with one class of customers. For the current model, and by observing the results, we find that this observation is only true when the arrival rates for low priority patients are aligned with the corresponding EDs capacity. Otherwise it does not hold. For hospitals that have equal number of beds and equal walk-in arrival rate, then they should be loaded equally with respect to ambulance patients.

Observation 2: Highly utilized networks are closer to balance than low utilized networks.

Another observation from the optimal results is with respect to how much change in routing proportion is needed to achieve optimality. We notice that when the network is highly utilized, it is closer to balance than low utilized networks. As suggested by Figure 5.7. For example, if we compare the change in p_1^* when λ_0 increases from 1.2 to 1.3 patient per hour then the change in p_1^* is from 16.12% to 17.54% corresponding to a 8.8% increase. While when λ_0 increases from 3.7 to 3.8 patient per hour, the change in p_1^* is from 28.87% to 29.04% corresponding to only 0.5% increase in the fraction of ambulance patients.

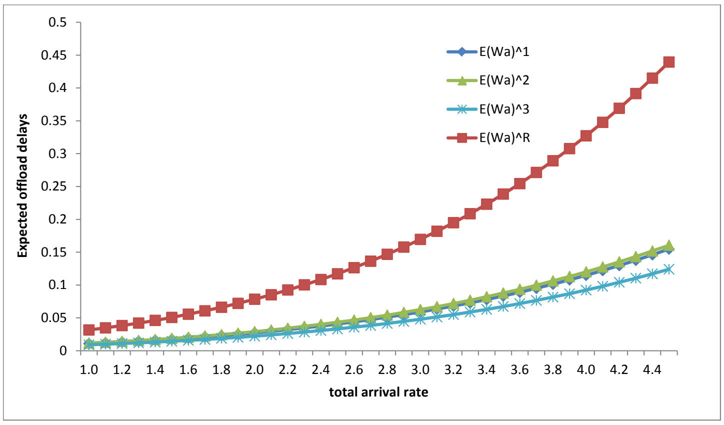

Figure 5.8: Expected waiting time results

Figure 5.8 shows a graph of total arrival rate for ambulance service (λ_0) versus the expected offload delays based on the allocation problem. Note that each point in the figure represents a unique solution for the allocation problem. We notice also that the separate and the total offload delay functions are convex with respect to the arrival rate of the high priority patients.

In Table 5.8 we change the first ED walk-in arrival rate from 2.8 to 4.5 patient per hour and keep all the other parameters constant. An empirical observation from the results is the following: when $\frac{\lambda_1}{\lambda_2} < \frac{c_1\mu_1}{c_2\mu_2}$ then the optimal allocation probabilities are set such that $p_1^* > p_2^*$.

Our results for the ambulance routing problem assume that Emergency Departments are identical. Practically, in some communities, there are specialized EDs where patients, for example, with certain conditions should be sent only to one of the region hospitals. This issue should not affect the results of the model because what we are solving is for the general flow of patients in the long run. To account for the specialized EDs case, we can update our optimization problem by adding a limiting constraint on the minimum percent of patients routed to a specific hospital ED. The minimum allowable limit would be such that it accounts for the fraction of patients that are usually sick with that condition in the corresponding region.

5.5 Conclusion

In this Chapter, we developed a decomposed model for a network that consists of K hospitals. We modeled the beds at the EDs as servers and as a result used nonpreemptive priority to model the difference between acuity levels for patients arriving by an ambulance and patients

λ_1	${p_1}^\star$	ρ_1	${p_2}^\star$	ρ_2	${p_3}^\star$	ρ_3	$E[\hat{W}_a]^R$
2.8	0.5454	76.03	0.3664	76.65	0.0882	73.74	0.0612
2.9	0.5253	76.71	0.3778	77.22	0.0968	74.27	0.0664
3.0	0.5054	77.41	0.3891	77.79	0.1055	74.81	0.0719
3.1	0.4855	78.10	0.4004	78.35	0.1141	75.34	0.0778
3.2	0.4657	78.79	0.4116	78.91	0.1227	75.88	0.0841
3.3	0.4460	79.49	0.4228	79.47	0.1312	76.40	0.0907
3.4	0.4264	80.20	0.4339	80.03	0.1397	76.93	0.0977
3.5	0.4069	80.91	0.4449	80.58	0.1481	77.45	0.1052
3.6	0.3875	81.63	0.4559	81.13	0.1565	77.97	0.1130
3.7	0.3683	82.35	0.4669	81.68	0.1649	78.50	0.1213
3.8	0.3491	83.08	0.4777	82.22	0.1732	79.01	0.1300
3.9	0.3301	83.82	0.4885	82.76	0.1814	79.52	0.1391
4.0	0.3112	84.56	0.4992	83.29	0.1896	80.03	0.1488
4.1	0.2924	85.31	0.5098	83.82	0.1978	80.54	0.1589
4.2	0.2737	86.06	0.5204	84.35	0.2059	81.04	0.1695
4.3	0.2552	86.82	0.5309	84.88	0.2139	81.54	0.1806
4.4	0.2368	87.59	0.5413	85.40	0.2219	82.03	0.1922
4.5	0.2185	88.36	0.5516	85.91	0.2299	82.53	0.2043

Table 5.8: Optimal routing results

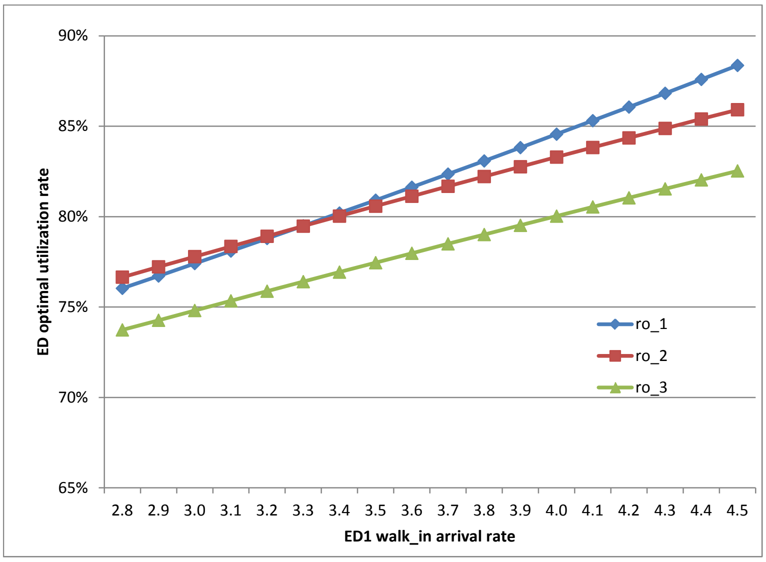

Figure 5.9: Optimal utilization rates

arriving by themselves. To optimally allocate patients to regional hospitals we developed an approximation scheme to compute the expected offload delays in terms of ED parameters. We showed that our approximation scheme is efficient and close to the exact results when the EMS operates under normal operating conditions. i.e. low to medium utilization.

Theoretically, our model looked into analyzing a queueing network with blocking and multiple servers. This kind of networks are challenging to analyze. By taking advantage of the operating conditions of the queueing network, we were able to develop an approximation scheme to analyze the queueing network.

In the next chapter, we conclude this research and discuss some future research directions.

Chapter 6

Conclusion

In this thesis we model the ambulance offload delay problem using queueing theory. We show how patient re-allocation to hospitals has great impact on the total offload delays experienced by a regional EMS provider. To achieve that purpose, we develop three distinct queueing models each with a different modeling approach.

In Chapter 3, we develop a regional network model with preemptive priority discipline where we model the service capacity at EDs as the combination of doctor, nurse and bed. We demonstrate, through case studies, how routing decisions can affect the total offload delays and walk-in patients queues. We also perform some scenario analysis to show the effect of speeding up the service at the EDs on offload delays and other performance measures of interest. From a theoretical perspective, we develop efficient algorithms to construct the stochastic model Markov chains, find the limiting probabilities, and calculate the performance measures.

Chapter 4 takes a different approach to model service capacity at the EDs. We model the beds as servers in the queueing network model which leads to the nonpreemptive assumption for admitting patients arriving by an ambulance to the ED. The nonpreemption assumption adds complexity and computational limitations into the model. Theoretically, we develop a novel approach to construct the network model Markov chain using the idea of ED layers. Through a case study, we show the effect of adding more beds on system performance measures in general, and offload delays in particular.

Contrary to the previous chapters where we analyze the EMS-ED network at the regional level, in Chapter 5 we decompose the network into multiple networks. We utilize the idea of independence across hospital EDs to decrease the size of the network and hence improve the computational capability for the model. In addition, we consider the transit time of ambulances to be Markovian. To formally address the ambulance allocation problem, we first develop an approximation scheme to the offload delays at each ED. Then, we use it to optimally re-allocate patients to regional hospitals such that the total offload delays are minimal. Through extensive numerical analysis, we show that larger EDs should be loaded more heavily to make use of server pooling.

While this thesis provides a number of insights on the offload delay problem and its relationship with capacity and routing decisions, there are several aspects that can be considered for further research. For example, variation in patients' arrival rates based on the day of the week can be modeled using the Markovian Arrival Process (MAP). This can be viewed as a generalization of the Poisson arrival process we have assumed in this research. Our models can be extended easily to include this variation but at the expense of high dimensionality. Furthermore, careful observation for the ED bed usage reveals that admitted patients block ED beds when there are no beds available for them at the destination wards. In fact, blocking is observed almost at each stage of the hospital network. Deriving more insights from a more complex queueing model that includes patients from the point of arrival to the ED to being discharged from the hospital will be valuable for decision makers.

Bibliography

[1] M. Carter, "Diagnosis: Mismanagement of resources," *OR/MS Today*, vol. 29/2, pp. 26–32, 2002.

[2] J. Goto, "Patients aren't widgets," *OR/MS Today*, vol. 35/2, pp. 24–31, 2008.

[3] M. Eckstein and L. S. Chan, "The effect of emergency department crowding on paramedic ambulance availability," *Annals Of Emergency Medicine*, vol. 43/1, pp. 100–105, 2004.

[4] S. Deo and I. Gurvich, "Centralized vs. decentralized ambulance diversion: A network perspective," *Management Science*, vol. 57/7, pp. 1300–1319, 2011.

[5] J. Prno, "Offload delays more than just an EMS issue." `http://www.cchse.org/assets/hamiltonandarea/ER_Wait_Times_Panel_Dr_John_Prno.pdf`, April 2010.

[6] J. Y. Ting, "The potential adverse patient effects of ambulance ramping, a relatively new problem at the interface between pre hospital and ED care," *Journal of Emergency, Trauma, and Shock*, vol. 1/2, p. 129, 2008.

[7] A. Drummond, "No room at the inn: overcrowding in ontario's emergency departments," *Canadian Journal of Emergency Medicine*, vol. 4/2, pp. 91–97, 2002.

[8] V. Lam, "In the ER: Fatal overcrowding," *National Post*, vol. June 1, 2005.

[9] H. E. Department and A. E. W. Group, "Improving access to emergency services : A system commitment," tech. rep., Ministry of Health And Long Term Care, 2005.

[10] P. Macintyre, "Hospital offload delay status update," tech. rep., Toronto EMS, January 2009.

[11] R. of Waterloo Public Health, “Emergency medical services master plan,” tech. rep., December 2007.

[12] M. Majedi, “A queueing model to study ambulance offload delays,” Master’s thesis, University of Waterloo, 2008.

[13] C. Taylor, D. Williamson, and A. Sanghvi, “When is a door not a door? the difference between documented and actual arrival times in the emergency department,” *British Medical Journal*, vol. 23/6, pp. 442–443, 2006.

[14] S. Silvestri, G. Ralls, L. Papa, and M. Barnes, “Impact of emergency department bed capacity on emergency medical services unit off-load time,” *Academic Emergency Medicine*, vol. 13/5, pp. 70–71, 2006.

[15] S. Silvestri, G. Ralls, K. Shah, and G. Parrish, “Evaluation of patients in delayed emergency medical services unit off-load status,” *Academic Emergency Medicine*, vol. 13/5, p. 70, 2006.

[16] M. Schull, K. Lazier, M. Vermeulen, S. Mawhinney, and L. Morrison, “Emergency department contributors to ambulance diversion: A quantitative analysis,” *Annals of Emergency Medicine*, vol. 41/4, pp. 467–476, 2003.

[17] S. Campbell and W. Patrick, “Ambulance diversion and ED overcrowding,” *Canadian Journal of Emergency Medicine*, vol. 4/4, p. 244, 2002.

[18] R. Forero, K. M. Hillman, S. McCarthy, D. M. Fatovich, A. P. Joseph, and D. B. Richardson, “Access block and ED overcrowding,” *Emergency Medicine Australasia*, vol. 22, pp. 119–135, 2010.

[19] T. V. Woensel and N. Vandaele, “Modeling traffic flows with queueing models: A review,” *Asia Pacific Journal of Operational Research*, vol. 24(4), pp. 435–461, 2007.

[20] M. Utley and D. Worthington in *Handbook of Healthcare System Scheduling* (R. W. Hall, ed.), ch. 2 (Capacity Planning), Springer, New York, 2012.

[21] S. Fomundam and J. Herrmann, “A survey of queueing theory applications in healthcare,” Tech. Rep. 2007-24, The Institute for Systems Research, 2007.

[22] L. Green in *Queueing Analysis in Healthcare, Patient Flow: Reducing delay in healthcare delivery* (R. W. Hall, ed.), ch. 10, Springer, New York, 2006.

[23] J. Wiler, R. Griffey, and T. Olsen, "Review of modeling approaches for emergency department patient flow and crowding research," *Academic Emergency Medicine*, vol. 18, pp. 1371–1379, 2011.

[24] P. Lewis, *Recent results in the statistical analysis of Univariate point processes in Stochastic point processes.* Wiley, New York, 1972.

[25] E. Kao and G. Tung, "Bed allocation in a public health care delivery system," *Management Science*, vol. 27/5, pp. 507–520, 1981.

[26] J. Cochran and K. Roche, "A multi-class queueing network analysis methodology for improving hospital emergency department performance," *Computers and Operations Research*, vol. 36, pp. 1497–1512, 2009.

[27] H. G. Perros, *Queueing Networks with Blocking.* Oxford University Press, USA, 1994.

[28] H. Tahilramani, D. Manjunath, and S. K. Bose, "$GE/GE/m/N$ queues with transfer blocking," *Seventh IEEE International Symposium on Modeling; Analysis, and Simulation of Computer and Telecommunications Systems*, 1999.

[29] D. Kouvatsos and N. Xenios, "MEM for arbitrary queueing networks with multiple general servers and repetitive-service blocking," *Performance Evaluation*, vol. 10, pp. 169–195, 1989.

[30] N. Dijk and J. Wal, "Simple bounds and monotonicity results for finite multi-server exponential tandem queues," *Queueing Systems*, vol. 4, pp. 1–16, 1989.

[31] J. M. Smith, "$M/G/c/K$ blocking probability models and system performance," *Performance Evaluation*, vol. 52, pp. 237–267, 2003.

[32] V. Houdt and A. Alfa, "Response time in a tandem queue with blocking, markovian arrivals and phase type-services," *Operations Research Letters*, vol. 33, pp. 373–381, 2005.

[33] G. Latouche and M. F. Neuts, "Efficient algorithmic solutions to exponential tandem queues with blocking," *SIAM. Journal on Algebraic and Discrete Methods*, vol. 1/1, pp. 93–106, 1980.

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

[35] C. Osorio and M. Bielaire, "An analytic finite capacity queueing network model capturing the propagation of congestion and blocking," *European Journal of Operational Research*, vol. 196, pp. 996–1007, 2009.

[36] K. M. Bretthauer, H. S. Heese, H. Pun, and E. Coe, "Blocking in healthcare operations: A new heuristic and an application," *Production And Operations Management*, vol. 20/3, pp. 375–391, 2011.

[37] Y. Han and J. M. Smith, "Approximate analysis of $M/M/c/K$ queueing networks," *R.O. Onvural and I.F. Akylldiz, Editors, Queueing Networks with Finite Capacity, Elsevier Science*, pp. 113–126, 1991.

[38] S. Jain and M. Smith, "Open finite queueing networks with $M/M/c/K$ parallel servers," *Computers and Operations Research*, vol. 12/3, pp. 297–317, 1994.

[39] F. Cruz and M. Smith, "Approximate analysis of $M/G/c/c$ state-dependent queueing networks," *Computers and Operations Research*, vol. 34, pp. 2332–2344, 2007.

[40] R. Andriansyah, T. V. Woensel, F. Cruz, and L. Duczmal, "Performance optimization of open zero-buffer multi-server queueing networks." 2009.

[41] I. F. Akyildiz, "Product form approximations for queueing networks with multiple servers and blocking," *IEEE Transactions on Computers*, vol. 38, pp. 99–114, 1989.

[42] M. van Vuuren, I. Adan, and S. Resing-Sassen, "Performance analysis of multi-server tandem queues with fininte buffers and blocking," *OR Spectrum*, vol. 27, pp. 315–338, 2005.

[43] D. Wagner, "Waiting time of a finite-capacity multi-server model with non-preemptive priorities," *European Journal of Operational Research*, vol. 102, pp. 227–241, 1997.

[44] G. Latouche and V. Ramaswami, *An Introduction to Matrix Analytic Methods in Stochastic Modeling.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[45] A. Riska and E. Smirni, “Mamsolver: A matrix analytic methods tool,” *Computer Performance Evaluation: Modelling Techniques and Tools*, vol. 2324, pp. 41–72, 2002.

[46] F. M. Neuts, *Matrix Geometric Solutions in Stochastic Methods: An Algorithmic Approach.* Dover Publications, Mineola, NY, USA, 1981.

[47] R. Nelson, “Matrix geometric solutions in markov models; a mathematical tutorial,” *IBM Research Division, T.J. Watson Research Center*, 1991.

[48] E. Kao and K. Narayanan, “Modeling a multiprocessor system with preemptive priorities,” *Management Science*, vol. 37/2, pp. 185–197, 1991.

[49] S. H. Mateo Restrepo and H. Topaloglu, “Erlang loss models for the static deployment of ambulances,” *Health Care Management Science*, vol. 12, p. 6779, 2009.

[50] S. I. Harewood, “Ambulance deployment in barbados: A multi-objective approach,” *The Journal of the Operational Research Society*, vol. 53/2, pp. 185–192, 2002.

[51] D. Gross, J. Shortle, J. Thompson, and C. Harris, *Fundamentals of Queueing Theory.* John Wiley and Sons, 2008.

[52] J. M. Calabrese, “Optimal work load allocation in open neworks of multiserver queues,” *Management Science*, vol. 38/12, pp. 1792–1802, 1992.

[53] V. Mehrotra, K. Ross, G. Ryder, and Y.-P. Zhou, “Routing to manage resolution and waiting time in call centers with heterogeneous servers,” *Manufacturing and Service Operations Management*, vol. 14/1, pp. 66–81, 2012.

[54] H. L. Lee and M. A. Cohen, “A note on the convexity of perforemance measures of $M/M/c$ queueing systems,” *Journal of Applied Probability*, vol. 20, pp. 920–923, 1983.

[55] W. Grassmann, “The convexity of the mean queue size of the $M/M/c$ queue with respect to the traffic intensity,” *Journal of Applied Probability*, vol. 20/4, pp. 916–919, 1983.

[56] B. Chen and S. Henderson, "Two issues in setting call centre staffing levels," *Annals of Operations Research*, vol. 108, pp. 175–192, 2001.