REAL-TIME OPTIMIZATION OF PATIENT FLOW IN EMERGENCY DEPARTMENTS

Boaz Carmeli

REAL-TIME OPTIMIZATION OF PATIENT FLOW IN EMERGENCY DEPARTMENTS

Research Thesis

Submitted in Partial Fulfillment of
the Requirements for the Degree
of
Master of Science in Information
Management Engineering

Boaz Carmeli

Submitted to the Senate of
the Technion - Israel Institute
of Technology

Av 5772 Haifa July 2012

ACKNOWLEDGEMENTS

This Research Thesis was carried out under the supervision of Prof. Avishai Mandelbaum in the Technion Faculty of Industrial Engineering and Management.

The generous financial help of the Technion is gratefully acknowledged.

Part of this Research Thesis was published in the Medical Informatics Conference under the following:

E. Vitkin, B. Carmeli, O. Greenshpan, D. Baras, Y. Marmor,
MEDAL: Measuring of Emergency Departments Adaptive Load, MEDINFO 2010.
B. Carmeli was the corresponding author for this paper, and contributed most of the requirements, insights, domain understanding and technical architecture.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Emergency Departments (EDs) are hectic, highly stochastic environments that deal with human lives under severe resource restrictions. ED personnel must provide quality clinical service and maintain an acceptable level of patient satisfaction while using limited operational resources.

In this work we consider the required features and main characteristics of a real-time ED monitoring-and-control system. We then focus on two specific applications, namely i) monitoring the real-time ED load and ii) optimizing internal ED patient flow through real-time control.

A good real-time monitoring-and-control system provides a holistic view of the entire ED operation, emphasizing information collection, analysis and display. We analyze the ED operation from multiple dimensions and viewpoints, e.g., taking clinical, operational, and service-level aspects into account. We focus on monitoring approaches and optimization techniques that can be deployed and used within a real-time ED monitoring-and-control system.

We developed an innovative load monitoring and measurement approach based on a neural networks paradigm. We thus enable adaptation of the load function into a specific ED setting, using subjective load perception provided by a specific user or a user group.

We analyzed service policies to optimize the ED patient by addressing the following question: Which patient should a physician treat next? For that, we provide an optimal control, based on a fluid model analysis and discrete event simulation. Deploying the resulting service policies within a real-time monitoring-and-control system would enable ED management and staff to improve overall ED operations.

LIST OF ACRONYMS

ADT	Admit, Discharge, Transfer
ATD	Adaptive Threshold
ATS	Australasian Triage Scale
BAM	Business Activity Monitoring
CEP	Complex Event Processing
CLPC	Consensus Load Parameter Classification
CPUR	Care Personnel Utilization Ration
CT	Computerized Tomography
DES	Discrete Event Simulation
DFC	Deadline-Feasibility Condition
DTD	Dynamic Threshold
ED	Emergency Department
EMR	Electronic Medical Record
EMS	Emergency Medical Services
EPA	Event Processing Agent
EPN	Event Processing Network
FCFS	First Come First Serve
FT	Fast Track
GPS	Global Positioning System
GTD	Greedy Threshold
HMO	Health Maintenance Organizations
HTTP	Hyper Text Transfer Protocol
ILT	Indoor Location Tracking
IP	In Process
IT	Information Technology
IPF	IP patient First

KPI	Key Performance Indicator
LIMS	Lab Information Management System
LOS	Length of Stay
LSP	Left Sync Point
MIS	Management Information System
MRI	Magnetic Resonance Imaging
NA	Newly Arrived
NAF	Newly Arrived First
PACS	Picture Archive and Communication System
PHO	Physical Occupancy
PPE	Proactive Period Equality
PTN	Patient to Treat Next
PUR	Patient Utilization Ratio
QED	Quality and Efficiency Driven
RFID	Radio Frequency Identification
RSP	Right Sync Point
STD	Static Threshold
TCP/IP	Transmission Control Protocol/Internet Protocol
TSLD	Threshold
TTFE	Time Till First Encounter
UI	User Interface

CHAPTER 1: INTRODUCTION

An Emergency Department (ED) is a hectic, highly stochastic, environment that deals with human lives under severe resource restrictions. ED personnel must provide quality clinical service, and maintain an acceptable level of patient satisfaction while using limited operational resources.

This research stems from the need for ED management to optimize ED operations. Specifically, we focus on the optimization approaches that stem from the ability to monitor and control ED operations in real-time.

Advances in Information Technology (IT) is reflected in the extensive use of hospital IT systems such as Admit, Discharge, Transfer (ADT), Electronic Medical Record (EMR) systems, Picture Archive and Communication Systems (PACS) and alike, as well as by the implementation of new RFID-based technologies for tracking human and equipment movement. These advances suggest new monitoring and control opportunities. Such technologies may provide ED management with a holistic view of the current ED situation and enable the development of real-time optimization which aims at improving the ED clinical and operational environment.

In this work, we present the required features and main characteristics of a real-time ED monitoring-and-control system. We then focus on two specific applications, namely i) monitoring the real-time ED load and ii) optimizing internal ED patient flow via real-time control. Through these two applications, we demonstrate two fairly different research techniques. We analyze the ED load problem using artificial intelligence, i.e., neural networks, aiming at capturing the ED operational characteristics while treating it as a black-box, bypassing the need to deeply understand its internal behavior. We used a somewhat different research approach for the second application. Specifically, we address the "Which patient to treat next?" question, using a detailed queuing model that seeks to mathematically understand micro-behavior and then optimize it.

A good real-time monitoring-and-control system provides a holistic view of the entire ED operation, emphasizing information collection, analysis and display. The ED operation should be analyzed from multiple dimensions and viewpoints, taking clinical, operational, and service-level aspects into account. This research focuses on monitoring approaches and optimization techniques that can be deployed and used within a real-time ED monitoring-and-control system.

The ED is part of a broader clinical and operational ecosystem through which patients flow. Patients arrive at the ED from various places, and under various clinical conditions, with a highly varied arrival rate. From the ED, patients are either admitted to one of the hospital wards, sent back home, or transferred to another clinical facility. Thus, the input and output patient flows must also be considered while analyzing and optimizing ED operations. Influencing patient flow external to the ED by controlling

the ED arrival rate, allowing faster admissions, or reducing hospital occupancy, is beyond the scope of this research. We thus focus on the ED internal patient flow processes, taking external operations as a given.

1.1 Thesis Structure

The rest of this thesis is structured as follows. In Chapter 2, we provide a detailed description of the ED operational environment and the surroundings affecting it. We then present a classification of time-related optimization categories, namely strategic, tactical, and real-time, which characterize the focus of this research. We conclude Chapter 2 by presenting a monitoring-and-control methodology that provides the foundation for the rest of the work. In Chapter 3, we survey the most important ED key performance indicators, with special attention to indicators that affect the ED load. In Chapter 4, we introduce EdRhythm, an ED real-time monitoring-and-control system, and present its key concepts. In Chapter 5, we discuss an innovative approach for real-time ED load monitoring based on neural networks. In Chapter 6, we present two major ED patient-flow control problems, namely: "where should a patient go next?" and "which patient should a physician treat next?" (PTN); we continue with a theoretical discussion of the second question. Chapter 6 concludes by proposing a heuristic service policy that best addresses the PTN question, which is the main finding of this research. In Chapter 7, we provide a simulation-based analysis for various service policies that address the PTN question. In Chapter 8, we give a complementary mathematical analysis of a fluid-model for a stylized version of the PTN question and present an optimal control for it. We conclude this thesis with conclusions and ideas for future work.

CHAPTER 2: EMERGENCY DEPARTMENT OPERATIONS

Emergency departments serve multiple purposes in the overall hospital setting. Hospital is a central location in which a specialized staff provides the best possible treatment to patients using state-of-the-art clinical procedures and the most advanced equipment. The hospital's main goal is to provide the best possible care to patients within a controlled cost. A hospital is a complex operational environment, designed to address a wide variety of patients' clinical needs. The department, or ward, is the core hospital's clinical and operational unit. In general, each ward is specialized in treating patients under similar clinical conditions, such as oncology, cardiology, internal, and so forth. An emergency department is somewhat different. The ED is designed to provide a medical treatment facility specializing in acute care of patients who arrive without a prior appointment [7]. Due to the unplanned nature of patient attendance, the department must provide initial treatment for a broad spectrum of illnesses and injuries, some of which may be life-threatening and require immediate attention. As such, the ED also serves as the hospital's main gateway for arriving patients.

Numerous types of ED settings are found in different parts of the world, and even within the same country. ED types were developed over the years following two main models known as the Anglo-American model and the Franco-German model [15]. The Anglo-American model suggests an acute care facility, or a unit within the hospital that serves as both the gateway to the hospital and the provider of emergency medical care for arriving patients. The Franco-German model, on the other hand, emphasizes the evaluation and treatment of patients before arriving to the hospital, e.g., at the patients' home or in the ambulance by emergency medical services. In these cases, the patient is given first-aid or pre-hospital emergency medical care and, in case hospitalization is required, the ED serves as an intermediate router to the relevant hospital ward. The models differ both in their clinical settings as well as in their operational settings. Our research focuses on an ED that follows the Anglo-American model, namely, an ED that provides acute care treatment targeted at discharging patients to their homes as well as initial diagnoses for patients who will be admitted into one of the hospital wards.

2.1 The Role of the ED

The role of the ED can be analyzed along the abovementioned two complementary aspects—clinical and operational. From the clinical perspective, the ED's main goal is to provide appropriate treatment for a broad spectrum of illnesses and injuries within a broad spectrum of severity levels. From the operational perspective, the ED processes should be designed in a way that allows care personnel to efficiently reach a decision as to where a patient should go next, and then to act upon it. The ED operational process does not stand on its own but is designed to best support its clinical goals. The ED,

under its operational role, can be thought of as a patient router, routing patients safely and efficiently to suitable destinations. Processes at the ED can thus be viewed and analyzed along their clinical and operational aspects. In the next section, we provide a detailed analysis of the high-level ED operational processes, emphasizing their related clinical aspects.

2.2 ED Operational Processes

We use the conceptual input-throughput-output model suggested by Asplin et al. [17] to describe and analyze the various high-level ED operational processes.

The overall view of that model is depicted in Figure 1. The input, or arrival, processes deal with aspects related to patients' arrivals at the ED. In Section 2.2.1, we discuss in more detail the various factors affecting this process. It is important to note that the ED management, and even hospital management, has minimal control over the arrival process. The throughput, or internal, processes, are related to the actual activities happening within the ED. These processes involve the patient's clinical assessments, treatment, and routing. These processes thus make up the core of the ED and are under significant ED management control. The throughput processes are the main focus of this research, and thus are further discussed throughout this thesis. The output, or admit, discharge, and transfer (ADT) processes, deal with releasing patients from the ED, either to their home, to one of the ED wards, or to another clinical facility. Output processes are further discussed in Section 2.2.5.

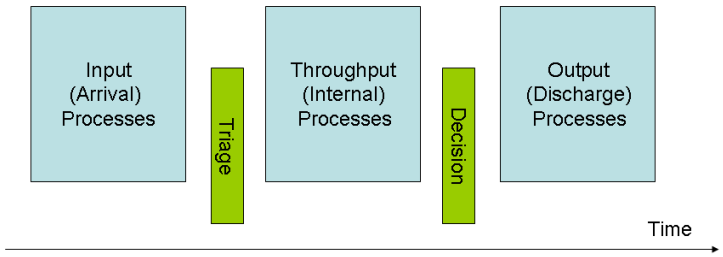


Figure 1: Schematic view of the main ED processes

In the following sections, we further discuss the input, output, and throughput processes and their relations to the various types of ED designs.

2.2.1 The ED Arrival Process

The ED arrival process deals with all aspects of patient arrivals. The main arrival process factor is the rate by which patients arrive at the ED. A typical ED arrival rate changes significantly according to the time of day. This phenomenon is important when analyzing ED operations and when trying to control them. We will further refer to this phenomenon in chapters 6, 7, and 8 in discussing the real-time ED control model.

Other relevant arrival process factors are patients' clinical severity and complexity. Patient severity can be monitored in various ways, such as through the source of arrival. Patients arriving via ambulance are usually in more serious clinical conditions than patients arriving on their own. It is important to note that a significant fraction of patients that arrive to the ED do not require immediate acute treatment. Treatment to these patients can be delayed, even up to a few hours, with no significant clinical consequences.

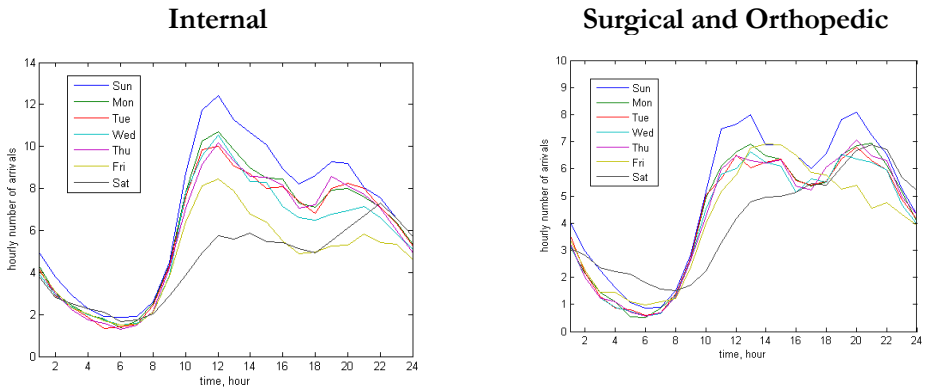


Figure 2: Hourly arrival rates per patient type (averaged over 4 years)

Figure 2 presents hourly arrival rates to the Rambam ED, collected and averaged over 4 years [44]. The left side of the figure shows arrivals to the Internal section of the ED, while the right side of the figure shows average arrivals to the surgical and orthopedic ED sections. As shown, the arrival rate has two peaks, the first just before noon and the second in the early evening. Another important observation is that arrival rates drop to nearly zero during most late night hours. A significant amount of research has been devoted to developing ED arrival rate forecasting methods. For a reference on forecasting and modeling ED arrivals and related literature, see [23].

Being able to forecast ED arrivals enables ED management to prepare for overcrowded situations, e.g., by shifting key personnel. An interesting research question relates to the freedom of choice patients have while approaching an ED. In other words, to what degree can patients' decisions affect the typical arrival rate pattern, e.g., by postponing their visit for a few hours? This question is not within the scope of this research. In

Israel, most arrivals to ED require a referral from a community physician, making up about 60% of the visits in 2009 [19]. Thus, real-time communication between the community care facilities (i.e., the four major Health Maintenance Organizations in Israel) has the potential to dramatically improve the ED arrival rate forecast.

Another interesting aspect of an ED real-time monitoring system is its ability to provide online, real-time updated status of the internal load within a specific ED, allowing patients to make autonomic decisions as to which EDs to approach. Such real-time views into the internal ED status will create a negative feedback loop that has the potential to balance the ED operational situation. Allowing patients to notify an ED about their planned arrival, e.g., through a dedicated smartphone application, offers further improvements to the arrival forecasting methods. These questions, and the implications of possible arrival-rate forecasting improvement approaches, are beyond the scope of this research. For our purposes we will assume that the arrival rate is provided to the monitoring and control system as an input that cannot be changed.

2.2.2 The ED Triage Process

One of the challenges in treating patients at the ED is to determine their right level of clinical urgency. Such levels allow the assigning of treatment priorities for patients. Assigning clinical priorities is most important for patients just arriving to the ED, who present the most uncertain clinical situation, and hence must see a physician as early as possible. The triage process, originated and first formalized in World War I by French doctors [13], is a process of prioritizing patients based on the severity of their condition. In fact, triaging used to be taught with an emphasis on the speed of the function, rather than the accuracy of the outcome.

Triage is mainly a routing process, allowing ED management to route the arriving patients into the most suitable ED section. The triage combines clinical and operational needs. The triage process ends with two complementary results: i) the patient gets a triage score, which is an indication as to the severity of their clinical condition and ii) the patient is routed to a suitable ED section, based on the triage score and other clinical conditions. Thus, setting the triage score, such as those based on the Australasian Triage Scale (ATS), allow caregivers to transform clinical urgencies into operational priorities [41]. The triage is the first interaction between the care personnel and the patient who just arrived at the ED. Thus, it provides a clear separation between the ED arrival process and the ED internal processes, as illustrated in Figure 1. Not all EDs use an explicit triage process and not all patients are assigned triage scores. Nevertheless, the triage process can be viewed as the first clinical-operational process provided for new patients just arriving to the ED. Triage is usually performed by an experienced nurse [28].

Ample work and research have been conducted for helping assign the appropriate triage categories to patients. The relationship between these categories and the ED physical structure and resource utilization is of great importance. For example, in [42], S. Mahapatra proposes a method that uses the acute categories (triage categories 1 and 2) in optimizing patient treatment scheduling, and the lighter categories (categories 3, 4, and 5) for predicting resource use.

Assigning a triage score to a newly-arrived patient is still not sufficient for allowing physicians to prioritize patients and to decide which patient to treat next. In reality, the same physician needs to both diagnose newly-arrived patients and treat patients already in process. Thus, two competing groups of patients must be prioritized accordingly. The ED Electronic Health Record (EHR) system may provide much more detailed clinical information about patients already in process and can be used for further optimization. For example, suppose that an experienced caregiver is able to accurately predict if a patient is to be discharged or to be hospitalized. This would enable the reduction of the overall average length of stay (LOS), for example, of discharged patients, by giving them precedence over patients that are about to be hospitalized. Such optimization and control is further discussed in the subsequent chapters.

2.2.3 The ED Internal Processes

The internal processes constitute the main part of the end-to-end ED patient flow. An overall view of the ED internal process activities is provided in Figure 3 [44]. As depicted in the figure, the patient flow process can be modeled as a job shop process [22], [49], [61]. It involves several processing-stations in close physical proximity, through which jobs are traveling while completing work. Applying that model to an ED setting, we correlate: i) stations with locations in the flow process where care personnel give clinical treatment to patients and ii) jobs with the patients themselves or with additional work, such as X-Ray interpretation, needed to be done by caregivers not necessarily in front of a patient.

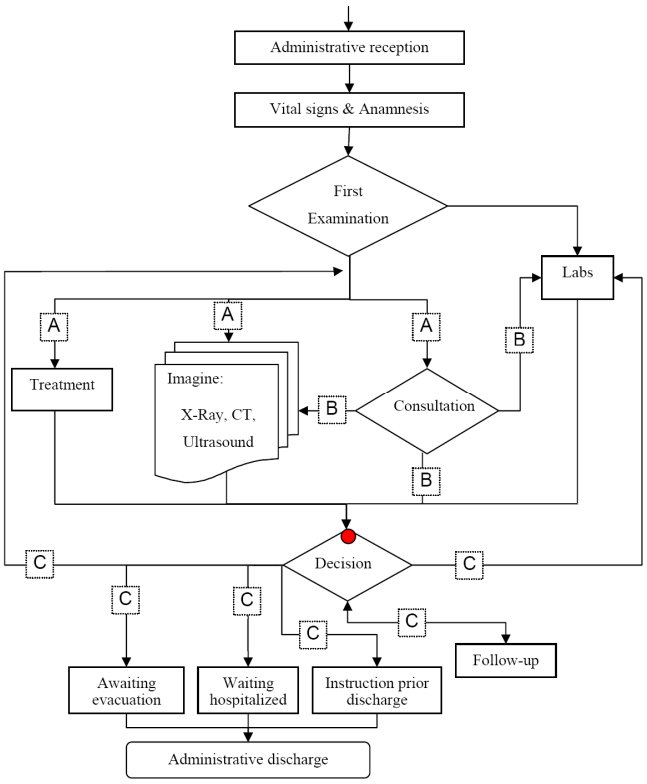


Figure 3: Internal processes activity chart. A, B, and C indicate alternative operations. The red dot indicates the merging point of all alternative operations.

There is no standard ED setting. Thus, the ED can be viewed as a composition of the following stations from several types, listed in an arbitrary order (see Figure 4) [44]:

The **arrival station** is the first station a patient encounters after arriving at the ED. Admission work is performed at this station. If a triage process exists, it is performed in an immediate subsequent station.

The **nursing station** is a station at which nurses give treatment to patients and perform related work. A single ED may contain multiple nursing stations. At that station, nurses measure vital signs, give medications, and take lab tests, etc. In some situations, patients must wait after receiving treatment for it to take affect before they can continue on to other stations.

The **physician station** is a station at which physicians treat patients. An ED generally has multiple physician stations. Physician stations may have a type, i.e., for the different specialist physicians such as internist, surgeon, or orthopedist.

The **consultant station** is a station at which a physician from outside of the ED treats patients. Unlike the physician station, the consultant station is usually unoccupied. If a patient needs to see a consultant, she will probably need to wait for her arrival to the ED. Patients are occasionally sent to the relevant ward to see a consultant rather than waiting for the consultant to come to the ED.

The **imaging station** is a station at which X-Rays, CTs, Ultrasounds, MRIs, and other similar tests are performed. Some EDs have an integral imaging station, while others use one in the main part of the hospital. Patients usually need to queue up for the imaging station. The station's service is composed of two sub-service processes: first, an image is taken, and second, a radiologist interprets the image and sends the results back to the referring physician station. The patient only needs to be physically present during the first service part. Thus, a patient may visit other stations while a radiologist interprets her images.

The **lab station** is a station at which laboratory tests are performed. Usually, a nurse conducts the lab test at the nursing station and then sends the samples to the lab station. The lab station sends back lab results to the ED after a significant period of time—usually up to half an hour or longer. Patients are able to go to other stations during that time, assuming that these other stations can perform their work without requiring the lab results.

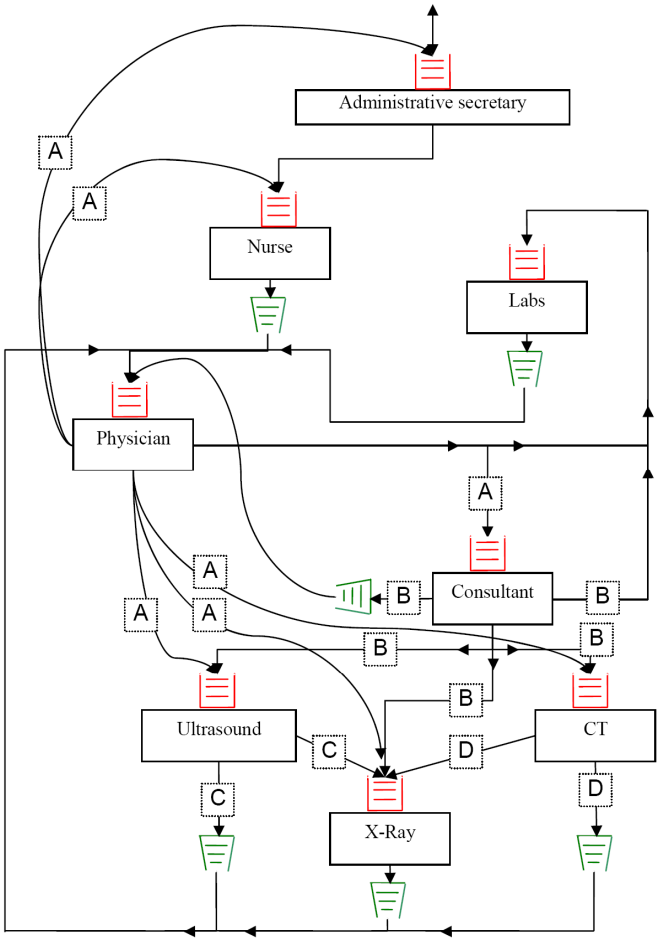


Figure 4: Internal processes station (i.e., resources) chart. A, B, C, and D – alternative operations; Resource queus – in red; synchronization queues – in green.

Figure 5 [44] illustrates the combined activity-station flow. It thus provides a high-level template for possible patient routes within an ED.

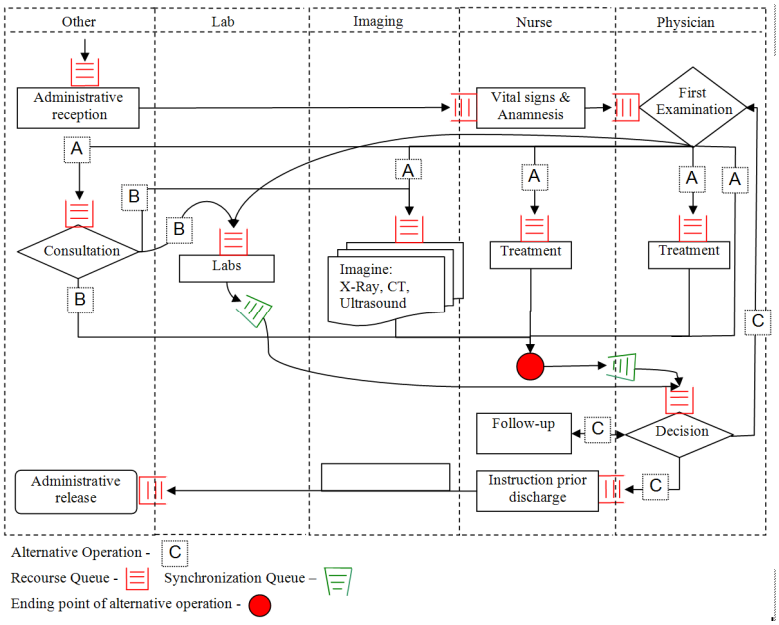


Figure 5: The internal processes combined activitiy-station chart.

As stated, patients flow through the ED stations with just a partial order. Patients may visit each station more than once, or not at all. Thus, the main goal of a patient flow monitoring and control system is to monitor the patient flow process and possibly control it for improved and perhaps optimal patient routing. Achieving this goal is the core focus of our research. Thus, we extensively discuss these issues in the subsequent chapters.

2.2.4 Reaching a Decision

The ED internal process results in the decision whether to: i) admit a patient to one of the hospital wards, ii) discharge a patient home, or iii) transfer a patient to another care delivery organization. Thus, reaching an admit, discharge, transfer (ADT) decision indicates the separation between the ED internal process and the ADT execution process. Reaching an accurate ADT decision in a timely manner is one of the most important ED goals. EDs strive to obtain an accurate ADT decision as fast as possible. The decision process itself is mainly clinical, though some operational aspects are still involved in it. The actual decision is usually made by the physician that has the main

responsibility for the specific patient's case. In some situations, it is not clear who the responsible physician is—an issue that may cause much delay in reaching a decision. Delays sometimes also occur in situations in which a decision is clinically difficult to make, thus requiring consultation from an expert physician. There are even situations in which physicians tend to delay clinically difficult decisions, hoping that things become clearer with time. Identifying these situations and alerting upon them will potentially improve ED operations, but are beyond the scope of this research.

2.2.5 The ADT Process

The ADT process starts after the ADT decision is made. The discharge process is mainly operational. During our research, we made several observations related to this process that potentially improve ED operations. Optimally, patients should leave the ED immediately after the decision is made, either to go home, to check in to one of the hospital wards, or to transfer to another care delivery organization. This does not happen for a significant fraction of the cases due to several reasons, most of them not under ED management control. The most significant cause for delays is the admitting process. The admitting process is often delayed because wards are, usually, highly occupied, tending to delay the admission of new patients. Various processes and ideas [57] are offered for accelerating and improving this process. Other delays occur in situations in which the decision is made to discharge patients back home, but only after some additional clinical treatment. As a result, patients remain in the ED for several hours, consuming the ED resources and affecting the ED measured performance. In other words, in certain situations physicians delay the discharge decision, knowing that patients need to stay in the ED for a few more hours anyway.

Being able to accurately monitor the ADT process suggests interesting control options. Knowing in advance which patients are about to be discharged back home without further delays would allow to give them some operational priority over patients that are about to be admitted or need additional treatment. This prioritizing would enable one to shorten the overall time these patients spent at the ED. We discuss such issues further in the following chapters.

2.3 ED Operational Efficiency

There are many aspects of ED operational efficiency and many operational processes affecting them. Operational processes can affect operational efficiency along three time scale categories—strategic, tactical, and real-time. The strategic category includes operational processes that affect efficiency in a years time scale. Such processes are, for example, the physical planning and setting of the ED, its capacity, and the overall resources such as rooms, beds, physicians, and nurses allocated to it. The tactical category includes processes that affect efficiency in an hours-to-days time scale, e.g.,

forecasting ED arrival rates, setting the staffing level at a specific day, and other related short-term ED policies. The real-time category includes processes that affect efficiency in a minutes-to-hours time scale, e.g., monitoring current ED load and deciding which patient to treat next or where patients should go next. Our research focuses on understanding issues related to the real-time category of operational processes. Nevertheless, in the next section, we provide a short overview of various operational efficiency aspects related to operational processes from the other two categories.

2.3.1 ED Strategic Operation Category

The term ED strategic operation refers to operational processes in the strategic category, namely those processes that relate to the way the ED is designed and set to operate. Several designs of ED operations have evolved over the years to best support operational efficiency while taking clinical aspects into account. The following list represents the most common designs and briefly describes the main pros and cons of each.

Triage is an ED design in which an experienced care giver, e.g., an experienced nurse, examines newly-arrived patients and assigns them a clinical severity triage score. The assigned score is then used for setting priorities among the patients who wait for treatment. The triage score is also used for setting deadlines for first patient-physician encounters. The score is typically not used beyond the first encounter, as the clinical status is assumed to be partially-known once the physician examines the patient. The triage process obviously improves patient routing and the ability to associate between operational aspects, such as deadlines, and clinical severity. The downside of triage is that it requires an additional resource. This resource is reduced from the overall care resources. Triage also adds a station, with a possible queue preceding it, possibly adding to the overall length of stay within the ED as well as to the time till first encounter.

Fast track (FT) is an ED design in which priority is being given to patients who require minimal ED clinical resources. Two types of patients may benefit from the FT setting—acute patients, for whom hospitalization is obviously necessary and thus the admitting decision can be made immediately, and patients with mild conditions, for whom the discharge decision can also be made immediately. The fast track seems to be the most attractive ED setting. The FT challenge is to identify those candidate patients who will potentially benefit from the fast track and then give them the appropriate precedence. Thus, FT goes hand-in-hand with triage. An efficient triage process upon patient arrival may result in a much more efficient operation in both designs.

Walking acute is an ED design in which the ED is divided into two main sections: i) for treating acute patients who are not able to roam about the ED on their own, and ii) for patients who are able to walk by themselves from one station to another. In some settings, a third trauma section exists. There are significant differences in the physical

settings of these sections. The acute section is much larger and contains a bed for every patient, while the walking section is set up like a clinic. In the acute section, physicians walk from one patient to another, while in the walking section, patients wait outside to be called into the physician's office. Assigning patients to sections is usually done by assessing the way they arrive to the ED—if they arrived by themselves, i.e., walked in, they are sent to the walking section. If they arrived via ambulance or on a stretcher, they are usually sent to the acute section. Dividing the ED into acute and walking sections allows ED management to significantly increase the ED static capacity: walking patients require less space then patient in beds. Adding a formal triage process upon patient arrival may reduce errors in patient placement into sections.

Illness based is an ED design in which the ED is divided into sections according to physician specialty. There seems to be no advantage to this type of ED setting from an operational perspective, as it mainly follows clinical needs. An interesting observation [15] resulting from this design is the need to train physicians to specialize in emergency medicine. Such physician proficiency may dramatically reduce the ED resources needed for treating patients, as all physicians will have the same proficiency and will be able to share rooms and treat most patients. An emergency medicine physician is able to give initial clinical treatment to most patients arriving to the ED. For patients with clinical conditions that are beyond his reach, an emergency medicine physician may consult an expert physician through the consultancy protocol, as is already being done in the current ED setting.

2.3.2 ED Tactical Operation Category

The term ED tactical operation refers to operational processes that affect operational efficiency on a daily basis. Issues relating to the standard level of resources e.g., physicians, nurses and life-saving equipment; the definition of roles and separation of duties among physicians, nurses, and administrative staff fall within the tactical operation category. Probably the most interesting question in the tactical category is staff scheduling, specifically, how to assign the right ED staffing level for meeting ED goals. Much work [64], [29], [25] has been devoted to ED scheduling and staffing. ED scheduling and staffing and the other aspects of tactical operation are beyond the scope of this research and thus will not be further considered.

2.3.3 ED Real-time Operation Category

The term ED real-time operation refers to operational processes that affect operational efficiency on a minutes-to-hours time scale. These processes are the focus of our research. Specifically, we are interested in issues related to monitoring the real-time ED operation and in controlling it. In the subsequent sections, we further analyze and

examine these processes in detail, focusing on two complementary issues—monitoring the overall ED load and controlling the real-time patient flow within it.

2.4 Monitoring and Control Methodology

"Monitoring and control system" is a common term in the industrial environment [56]. Monitoring and control systems are used for controlling manufacturing processes to ensure adequate throughput and quality levels. With the proliferation of digital systems for process coordination and documentation and with advances in wireless communication and tracking techniques, it has become practical to provide monitoring and control systems in services-based environments such as theme parks, bank branches, telecom service centers, and hospitals [27], [50]. Advancements and progress in hospital IT have made it possible to collect an accurate view of the current state even in hectic and dynamic environments such as the ED. The methodology illustrated in Figure 6 can be followed while introducing a monitoring and control system into an ED.

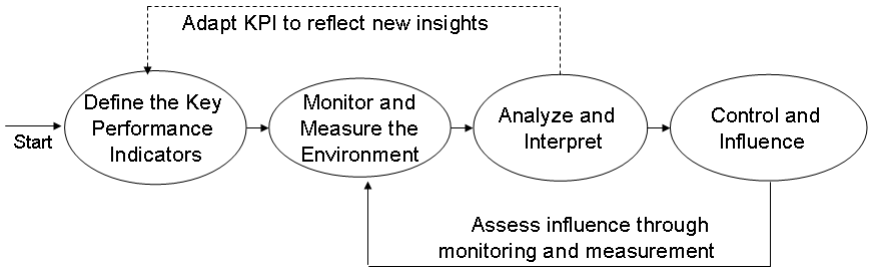


Figure 6: Monitoring and control methodology

Monitoring and control methodology consists of four main stages. In the first stage, ED management and other relevant stakeholders define the list of relevant key indicators with the expected performance of each. In the second stage, a system with monitor-and-measure capabilities is introduced into the environment. The monitoring system provides an accurate view of the tracked indicators. In the third stage, an analysis-and-understand process takes place to better understand measurement results. Last, as results from the third stage, ED management may decide to put in controls to influence the performance indicators. Alternatively, stakeholders may decide to adapt and modify the performance indicators, so that they better reflect the environmental needs, requirements, and possible achievements.

The rest of this document follows the abovementioned methodology. We discuss key performance indicator (KPI) definition in Chapter 3, in which we survey the most important ED KPI while giving special attention to indicators that affect the ED load. In Chapter 5, we propose an innovative way to monitor and measure the ED

environment. In Chapter 6, we analyze and interpret a specific ED operations question, namely "Which patient should a physician treat next?" and present a heuristic control that addresses it. In chapters 7 and 8, we further analyze the proposed control through simulation and mathematical analysis. The back arrows in Figure 6, namely those that suggest a modification to the KPI, and the measurement of the influence of a new control on the operation, require deployment of a system within a real ED and thus are not addressed further in this thesis.

CHAPTER 3: ED KEY PERFORMANCE INDICATORS

A performance indicator or key performance indicator (KPI) is a measure of performance. Such measures are commonly used to help an organization define and evaluate how successful it is, typically in terms of making progress towards its long-term organizational goals [11]. Thus, performance indicators provide the foundation of the monitoring-and-control system. Defining the KPIs of most importance enables stakeholders to measure them and then to strive to improve them in a methodological and consistent manner. The act of monitoring KPIs in real time is known as business activity monitoring (BAM) [1]. A common practice for implementing BAM systems is to use complex event processing (CEP) architecture [3], [53]. A CEP-based BAM system generates its output from collecting and analyzing streams of events. As part of the present research, we developed EdRhythm, a real-time CEP-based BAM system. The main goal of EdRhythm is to monitor the ED environment and to generate the required KPIs. We describe the key concepts and core components of the EdRhythm monitoring and control system in Chapter 4.

The ED is a complex environment in which a wide variety of indicators should be considered. Monitoring the whole set of indicators is a demanding and complex task. Indicators can be classified into various categories. Some indicators may contradict one another. Some present different levels of granularity and are measured using different units. In the following sections, we describe some of the most important KPIs used for measuring and controlling the ED. In Section 3.2, we introduce the ED load KPI, a complex KPI that presents difficulties in terms of monitoring and measurement as well as in analysis and interpretation. In Chapter 5, we offer an original approach for dealing with ED load monitoring and measurement.

3.1 Referred Key Performance Indicators

As part of this research, we developed the EdRhythm monitoring and control prototype. In Chapter 4, we describe EdRhythm in details. EdRhythm is designed to monitor and measure a wide variety of KPIs. In the following sections, we survey some of the most important KPIs generated by EdRhythm. Obviously, each KPI has a specific meaning. The EdRhythm system calculates each KPI from a set of low-level events it collects from the monitored environment. EdRhythm then generates the output KPI to be presented to ED management, staff, and other stakeholders. We describe below each KPI using three elements: i) its core meaning, ii) the set of input events from which it is being calculated, and iii) the actual output that describes it. We consider additional information and real-time alerts that the EdRhythm system is able to

generate based on the core input events. We provide a complete list of input events and output KPIs in Appendix A.

3.1.1 Time Till First Encounter (TTFE)

The Time Till First Encounter (TTFE) indicator measures the time from a patient's arrival at the ED until the time the patient is first seen by a physician. This indicator is one of the most important indicators for ED management, as it measures an operational indicator, i.e., time, with strong clinical aspects. Notably, the actual patient's clinical condition is not known until the first physician's examination; thus, delaying a physician's examination may result in clinical deterioration. Consequently, the control of this indicator is one of the main goals of this research.

Input events

The TTFE KPI is based on two input data events, namely the patient registration and the start of the first encounter event.

Output event

The system generates a TTFE output event for each patient upon the beginning of the first encounter with a physician. An optional threshold event is available for the system to report only these times that exceed the threshold.

Optional additions and alerts

Information derived from TTFE output events can be used to benefit the control and decision support part of the EdRhythm in two important ways:

- Provides an alert if patient waiting time is higher by some predefined percentage than the average treatment time, and treatment has not yet begun. Such an alert will warn ED management in advance of situations in which the TTFE deadline is about to be violated.
- Compares average TTFE for the last period with the average TTFE of some historical period (i.e., an average of the same time on the same day during the last year period). This may require additional information coming from a database. Such an alert may serve as additional input for the overall ED load KPI.

3.1.2 Total Length-of-Stay

The length-of-stay (LOS) indicator measures the overall time from a patient's arrival to the ED until admission or discharge. The LOS is calculated for each patient. Aggregative views for various categories may be provided as a derivative output event. This KPI is one of the most important indicators towards improving ED performance. The LOS KPI indicates ED clinical, operational, and service quality levels. Much work

has been devoted to monitoring and control of this KPI [26], [39]. A typical ED sets a specific LOS threshold, e.g., four hours, and aims on keeping the LOS below that threshold for the majority of the patients. The LOS can be viewed as a complex KPI calculated from several lower-level KPIs that measure the various phases of the care process. Thus, the TTFE described in the previous section measures the first clinical process stage. Similarly, a time from decision to release KPI may be generated for measuring the time from reaching an ADT decision, until the time a patient actually leaves the ED. Interestingly, an inherent tradeoff exists between the LOS and the TTFE KPIs. Controlling the LOS indicator and balancing it with the TTFE indicator are among the main goals of our research and are addressed in greater detail in the following sections.

Input events

The LOS KPI is based on two input data events, namely the patient registration event and the patient left event.

Output event

System generates LOS output events for each patient who is discharged home from the ED or admitted to one of the hospital wards. An optional threshold event is available for the system to report only those patients with an LOS that exceeded the threshold.

Optional additions and alerts

Generates an alert when the average LOS exceeds a certain predefined configuration level. This may indicate that the overall ED load is increasing above the desired threshold.

3.1.3 Patient Utilization Ratio

The patient utilization ratio (PUR) indicator measures the ratio between treatment time and the overall LOS for each patient. This KPI is calculated for every preconfigured time period. Upon patient discharge, a treatment ratio summary event is generated. Aggregative views may be provided as a derivative KPI output event

PUR Input Events

The PUR is based on multiple input events. In addition to patient registration and patient left events, the PUR monitors all start and stop treatment events generated at the various stations.

PUR Output Event

- **Treatment period** – the cumulative time period in which a patient is actually under treatment by care personnel.

- **Total length-of-stay** – the amount of time a patient spends in the ED from registration to discharge. This measurement is taken from the LOS KPI.

Optional additions and alerts

- Generates an alert if the average treatment ratio decreases below a given threshold. This may indicate that the overall ED load is increasing.
- Generates an alert if the treatment ratio for a given patient decreases below a given threshold. Various thresholds may be set for various patient severity levels. This alert may indicate that a specific patient receives low quality service, from both the clinical and operational perspectives.

3.1.4 Care Personnel Utilization Ratio

The care personnel utilization ration (CPUR) indicator measures the ratio between work time and the sum of work time with idle time for each care personnel type. This indicator is calculated for every preconfigured time period. At the end of each shift, a work-ratio summary event is generated. An individual CPUR indicator can be generated for a specific individual as required by ED management. The ED care personnel may object to such an indicator, and thus, careful attention is required as to its implementation. However, these issues are beyond the scope of this research and will not be further addressed.

Input events

The CPUR is based on multiple input events. The CPUR monitors all start and stop treatment events by a care personnel group or by a specific individual. These events cover both events associated with the actual patient treatment as well as events associated with the additional work performed by care personnel.

Output event

- **Work ratio** – the amount of time caregivers actually give treatment to patients during a given calculation period, including additional work not performed in front of a specific patient.

Optional additions and alerts

- Generates an alert if the average work ratio increases above a given threshold. This alert may indicate that the ED load is exceeding some certain desired level.

3.1.5 Physical Occupancy

The physical occupancy (PHO) indicator measures the number of patients in each ED room or section at preconfigured points in time.

Input events

To calculate PHO, the EdRhythm system tracks in/out location events and counts the number of patients currently in the monitored room or section.

Output event

- **Occupancy level** – generates an occupancy level report for each room, section, and for the overall ED, for each preconfigured time period.

Optional additions and alerts

- Provides alerts if room capacity is exceeded. These alerts may indicate that the ED load is exceeding some certain desired level.
- An accurate location tracking mechanism may provide many valuable alerts and controls beyond basic capacity monitoring. Some of these capabilities are:
- Monitors all people in a room (e.g. family members) and alert security for potential overcrowding or escort rules violation.
- Provides additional information about a patient that is assigned to a room but is not physically located within the room. For example, a patient that needs to be in the radiology section for a CT exam but is still waiting for transfer in the main ED section.

3.2 The ED Load KPI: A Complex KPI

ED load is certainly one of the most significant and interesting KPI to monitor and control [24], [54]. Reducing the ED load is a major day-to-day ED management challenge. High ED load leads to excessive waiting times and an unpleasant environment, which in turn cause: i) poor service quality from both clinical and operational perspectives; ii) unnecessary pain and anxiety for patients; iii) negative emotions in patients and escorts that sometimes could even lead to violence against staff; iv) increased risk of clinical deterioration; v) ambulance diversion; vi) patients leaving without being seen; vii) inflated staff workload; and more [58].

Measuring ED load serves multiple purposes and may prove to be useful from various aspects. The most common need found in the literature, for knowing the actual ED load, is to allow ED management to determine the situations that call for ambulance diversion [21], [35]. That is, ED management should be able to identify situations in which an ED needs to be closed due to overcrowding, and ambulances must be redirected to other hospitals. In addition, consistent measurement of ED load may enable ED management to identify trends in load and to adjust and adapt the ED operations accordingly. Displaying a real-time snapshot of the ED load allows managers

and staff to act in extreme situations by making prompt decisions, e.g., in cases of high or extreme load peaks.

Real-time ED load measurement turns out to be a challenging, multidimensional task. First, one must decide which parameters contribute to the load. Second, one must define how to calculate load on the various parameters (i.e., resources). Third, one must assign a level of contribution to each of the parameters while integrating all measurements into a single load score.

The fact that no "Standard ED" exists adds to the load measurement complexity. No one physical ED setting can be identified as standard. EDs are varied from one another in many dimensions, such as physical size, the population they serve, staffing levels, and clinical and operational protocols. In addition, EDs involve various entities, e.g., physicians, nurses, patients, managers; each of whom may define the load function differently, and may require periodically adjusting the load definition to accommodate changes that occur over time.

As a result, ED Load is defined as an integrative KPI, the calculation of which is based on an extensive set of low-level data events collected from the monitored environment. High ED load depends on a wide variety of clinical and operational parameters. Some of these parameters are already being monitored and displayed as stand-alone KPIs, as described in previous sections. Others may serve purely as input parameters for calculating and monitoring the ED Load indicator.

3.2.1 A Consensus Load Parameter Classification

We found the consensus load parameter classification (CLPC), suggested by Solberg et al. [55], most useful for serving as a baseline for the EdRhythm load monitoring and control functionality. The CLPC was defined by a panel of 74 national experts who assessed 113 measures and chose 38 through a discussion and rating process. The CLPC follows the Input-Throughput-Output operational model introduced in Section 2.2. This model permits most identified load parameters to be grouped into one of three stages:

- **Input or Arrival** stage (15 parameters) – includes factors such as the volume of ill and injured people in the community and the capability of the rest of the health care system to address the needs of individuals not requiring emergency care.
- **Throughput or Internal** stage (9 parameters) – includes factors that affect the efficiency of an ED to cope with its input, ranging from ED beds and staffing to the efficiency of ancillary services and consultant access.
- **Output or ADT** stage (14 parameters) – includes factors that affect the ability of the inpatient system to admit patients requiring hospital care, and of the ambulatory care system to provide timely post-discharge care.

To clarify their purposes, we have further grouped the parameters within each stage by the main concept they represent:

- **Patient demand** (6 items) – refers to the volume of patients arriving to the ED for receiving medical care.
- **Patient complexity** (3 items) – refers to patient's clinical factors, such as the urgency and potential seriousness of the presented complaint, the stability of the clinical condition, and the baseline medical and psychosocial burden of illness.
- **ED capacity** (5 items) – refers to the ability of the ED to provide timely care for the level of patient demand, according to the adequacy of physical space, equipment, personnel, and the organizational system.
- **ED workload** (6 items) – refers to the demand and complexity of patient care that is undertaken by the ED within a given period.
- **ED efficiency** (4 items) – refers to the ability of the ED to provide timely, high-quality emergency care, while limiting waste of equipment, supplies, and effort.
- **Hospital capacity** (6 items) – refers to the ability of the hospital to provide timely inpatient care for ED patients who require hospitalization, according to the adequacy of physical space, equipment, personnel, and the organizational system.
- **Hospital efficiency** (8 items) – refers to the ability of the hospital to provide timely, high-quality inpatient care while limiting waste of equipment, supplies, and effort.

The list of all 38 parameters is provided in Appendix B.

In the next chapter, we discuss the core concepts of the EdRhythm system. In Chapter 5, we describe how real-time business activity monitoring system, such as EdRhythm, can be used to address the ED load monitoring challenge in an innovative way.

CHAPTER 4: REAL TIME ED MONITORING

In the previous chapter, we described the KPIs' role within a real-time business activity-monitoring environment and surveyed some of the most important ED KPIs. Next, we will describe the key concepts of an ED's real-time monitoring-and-control system; how such a system is being used for monitoring and measuring the ED environment and for generating, through calculation, the required KPIs.

A typical ED monitoring-and-control system comprises three major layers: i) the collection layer, which provides the inputs to the system, ii) the logic layer, and iii) the display layer, which presents system output to its users. Figure 7 presents the high-level structure of a typical monitoring-and-control system. Note that the figure illustrates a comprehensive view. Not all components are mandatory for every solution.

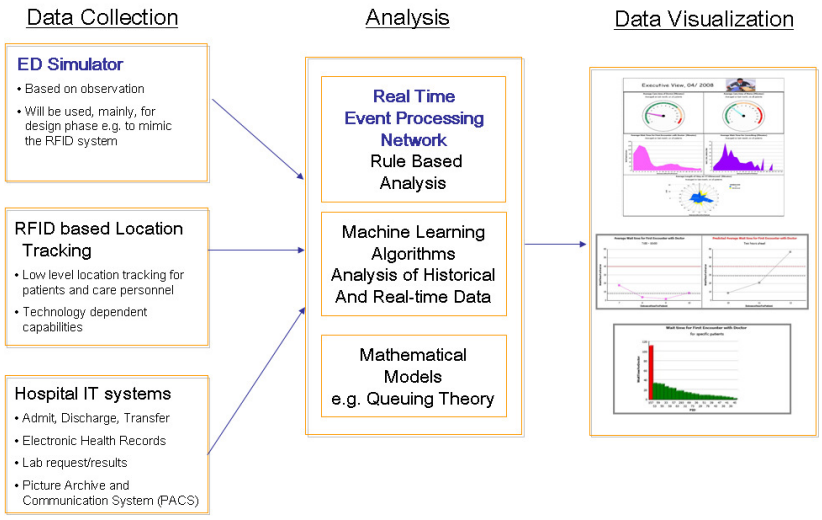


Figure 7: Typical ED monitoring-and-control component structure

4.1 EdRhythm: Real-time Monitoring-and-Control System

During our research, we developed EdRhythm, a prototype real-time monitoring-and-control system. EdRhythm has three main layers: i) the input layer, ii) the logic layer, and iii) the output layer.

The EdRhythm input and output layers are implemented as an event processing network (EPN) [3], [40]. Implementation is done in StreamBase V6.2 [12]. The EdRhythm logic layer extends the EPN with custom-made logic and analytic components.

The input layer, i.e., the data collection layer, is connected to the data sources and receives the data events through these connections. Next, we describe the various data sources typically found within an ED environment. We then follow by describing the three layers of the EdRhythm prototype system.

4.2 The Data Sources and the Data Collection Layer

The data collection layer provides the input to the system. The collection layer uses various communication means to transfer raw events from its generating sources into the system. Computer networks become pervasive and various network types can be found in every modern ED. Most commonly, a TCP/IP-based network is used by most information technology applications. In recent years, a proliferation of wireless networks has accelerated the use of mobile computers and hand-held devices. Most recently, RFID technologies have been introduced, making accurately locating people and objects possible. All such networks allow the extensive collection of events from the monitored environment and the transmission of these events into the logic and analysis layer of the monitoring-and-control system. In the subsequent sections, we provide more details on some of these applications and technologies, emphasizing the differences between clinical and operational systems. We address the ways in which both types contribute to ED real-time monitoring.

4.2.1 Clinical Applications as Operational Data Source

Hospital environments utilize IT systems at an ever-growing pace. A typical hospital manages hundreds of IT applications. Among the main applications, one can usually count the admit-discharge-transfer (ADT) system, the electronic health record (EHR) system, the picture archive and communication (PACS) system, and the lab information management system (LIMS). These systems improve the clinical process by recording and documenting it, and by providing instant access to patients' clinical information. As such, these systems provide a rich source of crucial clinical information, assisting ED personnel in performing day-to-day tasks. Coincidentally, these systems can be used as an indirect source of information regarding the operational situation in the ED. Specifically, The EHR system is the major management information system (MIS) used within hospitals for documenting clinical treatment of patients. Most hospitals are in the process of replacing their paper-based documentation system with computer-based EHR. Hospitals' EDs however, are less suited for deploying and adapting EHR systems, due to the urgency and mobile nature of the work within them. Nevertheless, recent experiences have shown that hospitals are extending the reach of EHR systems and deploying them within EDs as well [52]. One such EHR can be found in the Rambam Medical Center's ED. Rambam's ED management, together with Rambam's IT team, decided during 2009 to extend EHR functionality into the ED. Currently, Rambam's EHR is fully functioning and provides the needed information to nurses, physicians, and

management within the ED. EHR also provides useful information about clinical aspects of patients' treatment. Through EHR, a physician is able to document diagnosis and request further treatment, such as prescribing a specific drug or asking for expert consultation on an unresolved case. The EHR system is designed to support and document clinical processes. As such, it does not provide the out-of-the-box, required view on the operational status of the ED.

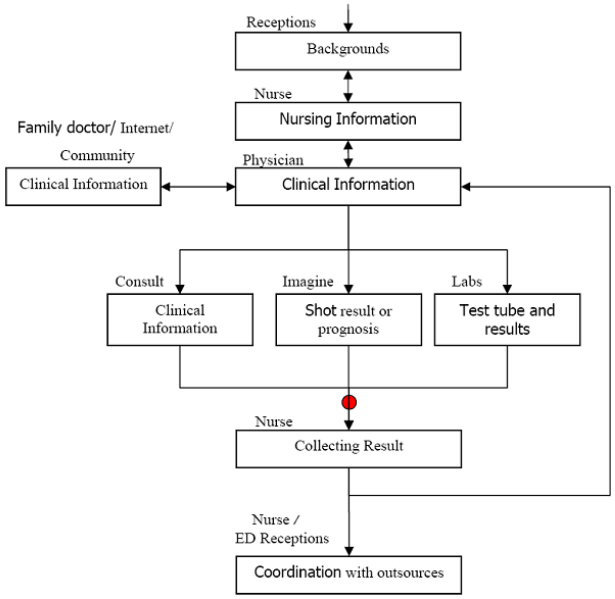


Figure 8: Information generated and collected during the clinical process

Still, as shown in Figure 8, many operational events can be generated by EHR as well as by other MIS, e.g., by monitoring the time physicians take to enter specific clinical orders for a patient. In the following section, we discuss IT systems that are designed to collect operational events from the monitored environment.

4.2.2 Operational Data Sources

Complementary to the clinical applications, operational applications are designed to assist in managing the operational environment. Indoor location tracking (ILT) [9], [32], [33], [48] is a specific class of operational applications that enables tracking the exact locations of patient and personnel within the ED and accurately identifies the start and end of patient/care personnel interactions. Recently, the global positioning system (GPS) has become the de-facto standard for outdoor tracking, and it serves as the foundation for many location-tracking applications [8]. In parallel, significant efforts

have been devoted to develop an efficient and accurate ILT system. However, as yet, no optimal, standard technology exists that is suitable for indoor location tracking.

ILT systems are also referred to as RFID systems, after the technology of radio frequency identification. RFID technology has recently become widespread, due to its many merits. Basically, RFID provides unique identifications to objects; hence, it can be used as the foundation for tracking, monitoring, and controlling object movements [32], [33].

RFID has traditionally been used for tracking objects such as consumer-packaged goods, medications, and medical equipment. Yet this same technology can be used for uniquely identifying humans, e.g., patients and care personnel in hospitals. Applying RFID for indoor-location tracking requires an additional layer to associate the RFID tag with a specific location. This association can be implemented via two conceptually different approaches:

- **Cell-based location tracking** – location identified through the location of the reader of the RFID tag.
- **Triangulation** – location calculated from radio frequencies, used in the communication between the RFID tag and scattered RFID readers.

Extensive research is being devoted to better understand the pros and cons of each approach and their various aspects; thus, we will not address these aspects further within our research.

4.2.3 Location Monitoring of ED Entities

Using RFID or a similar technology allows a monitoring-and-control system to collect real-time location information of relevant entities within the ED. Real-time entity-location monitoring enables the real time measurement and calculation of most required KPIs:

Specifically, the following entities play a significant role within the ED process flow; some entities are further grouped into entity types.

- **Patients**: Each patient is identified by a unique ID. Patients are further grouped by clinical condition.
- **Physicians**: Physicians are grouped into several types by proficiency. Each proficiency type has a unique identifier. A limited number of physicians from each proficiency type are in the ED at any given point in time.
- **Nurses**: Nurses are grouped into several types according to their specific roles. Each role type has a unique identifier. A limited number of nurses from each type are in the ED at any given point in time.

- **Rooms**: Each room in the ED is monitored. Rooms may have limited capacity, e.g., if serving acute patients, or unlimited capacity for serving ambulatory patients.
- **Beds**: Each bed has a unique ID. A room's capacity for acute patients is monitored by the number of beds it contains.

A combination of clinical and operational monitoring data sources provides sufficient infrastructure for a real-time ED monitoring-and-control system.

For the sake of this research, the raw clinical and operational data events are provided by the ED Simulator [43]. The ED Simulator generates events similar to those expected to be generated from clinical and operational IT systems in a typical ED. Rambam ED management and other subject matter experts validated the patterns and distribution of information generated by the ED Simulator and found them similar to those that are typically found in a real ED operational environment. The simulator generates about 100 different event types. Most events are related to the operational processes and patient flow within the ED. A comprehensive event list can be found in Appendix A.

4.3 The EdRhythm Input Layer

The EdRhythm input layer is implemented using StreamBase Input Connectors. The Input Connector client is designed to be embedded within the clinical and operational data sources that are available in a specific ED. Similarly, the Input Connector client is embedded within the ED Simulator code and provides the main EPN's event channel between the data sources and the EdRhythm. On startup, the ED Simulator connects to the EPN using the StreamBase (SB) internal protocol, which is implemented on top of TCP/IP. Using that protocol, the Input Client sends events to the logic layer, which is implemented as an EPN within SB.

The ED Simulator uses two parallel channels over which it sends two types of events, data events and clock events.

The current version of the ED Simulator generates just operational data events. The need to consider clinical data events and the exact data pieces are described in chapters 6 and 7. To overcome this difficulty, we added clinical data attributes to the EdRhythm logic directly, bypassing the EdRhythm input layer.

4.3.1 The Input Events

There are three main input event types in the EdRhythm: i) data events, ii) control events, and iii) clock events. Each event type is further described below.

Data Events

Each data event has an event-ID field that encodes the meaning of the event. A list of all possible input data events is provided in Appendix A.

Each data event contains the following attributes:

- **Event ID** – Unique identifier for the event.
- **Resource Type** – Can be a room (e.g., in which a CT scan is being taken), a bed, or a lab test.
- **Resource ID** – a unique resource identifier.
- **Care Giver Type** – e.g., physician, nurse, consultant, etc.
- **Care Giver ID** – unique care giver identifier.
- **Patient Type** – e.g., orthopedic, surgical, or trauma.
- **Patient ID** – a unique patient ID.
- **Time** – a time stamp in which an event was generated.

Control Events

Several control events are used to set various thresholds and calculate time periods and other configuration parameters within the EPA. Control events are configuration-dependent. Only minimal control event functionality is currently implemented in the EdRhythm. The most important implemented control event is the time till first encounter (TTFE) duration. TTFE is an ED-dependent parameter, based on the triage categories and their associated deadlines, as defined by the ED management.

Clock Events

EPN uses clock events to execute logic that is not triggered by specific events. The clock event provides the granularity level of the EdRhythm KPI calculation and display. The granularity level depends on the typical processes pace, the rate and accuracy of the input data events, and the monitoring-and-control granularity level required by ED management. In addition, different KPIs may require different granularity levels. The minutes scale is found to be an appropriate level of granularity for real-time monitoring of ED environments.

The simulation-based EdRhythm system requires synchronization between the simulator clock and the real-world clock used by the EPN. The calculation period of the EPN can be configured and is currently set to one-minute intervals. The internal ED simulator uses one-second intervals between successive ticks. Hence, the clock event is generated by the ED simulator every minute of simulation time, e.g., every 60 simulation ticks.

4.4 The EdRhythm Logic Layer

The EdRhythm logic layer provides the core KPI monitoring and measurement functionality. Each indicator receives the relevant input events and generates a suitable

output event to be displayed on the EdRhythm dashboard. Indicators can be easily added to the EPN. Each event processing agent (EPA) implements a specific indicator or a small group of related indicators. Most indicators are calculated per period, i.e., as a set at a-time. Few indicators are calculated for each event, i.e. as an event-at-a-time.

Each EPA relies on a dedicated data structure. Some elementary data structures (e.g., utilization counter) can be reused between various EPAs. Each EPA uses its own data structure for saving its state. No internal communication takes place among EPAs, and each EPA acts as a stand-alone component. This architectural decision results in sub-optimal performance, but is much easier to maintain and extend.

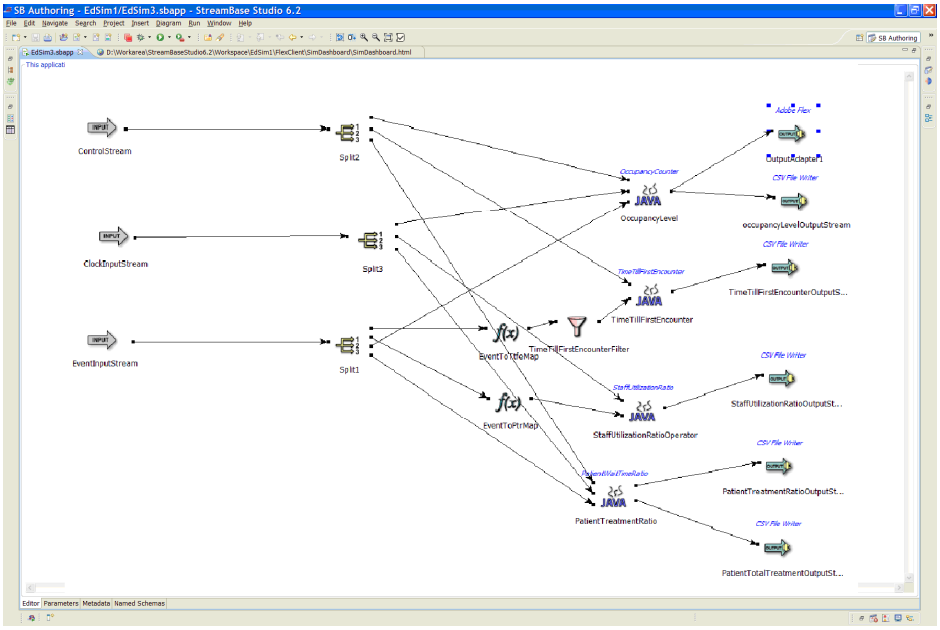


Figure 9: The EdRhythm EPN implemented by using StreamBase 6.2

Figure 9 presents the high-level view of the EPN implemented in SB. There is minimal use of the ready-made StreamBase operators, e.g., to filter out events that are not relevant to a specific agent, or to converge similar events into one. This is mainly because we developed a Java code that implements the EPA logic.
Thus, embedding these operators directly into the EPA logic instead of maintaining them as separate entities becomes trivial in most cases. Note that adding new indicators is simple and straightforward, as there are no dependencies among the EPAs.

4.4.1 The EdRhythm Output Layer

The main purpose of the EdRhythm output layer is to communicate the calculated KPI to the situation dashboard. The EdRhythm output layer is implemented by using the StreamBase output connector. The output connector client is embedded within the ED

dashboard. On startup, the ED dashboard connects to the EPN, using StreamBase (SB) protocol, which is implemented on top of HTTP. Using that protocol, the SB Flex output client receives the output event stream from the logic layer and passes it to the Flex-based ED dashboard application. The ED dashboard receives numerous output streams. Each is implemented by using a dedicated SB Flex connector. Each output stream is designed to handle a specific output event type that is associated with a specific KPI. Thus, output events have no unified structure. All required event adaptation and modifications are done within the EdRhythm logic layer. Each output event is tailor-made to be presented on the ED dashboard without further analysis and transformation.

4.5 Monitoring Dashboard

Monitoring dashboards serve as the user interface (UI) components of real-time monitoring-and-control systems [4]. Through dashboards, users realize the whole environmental situation as well as the situations of specific entities within that environment. Public dashboards are implemented using large electronic displays. Technology for creating large electronic displays has become pervasive. Such displays are currently being used in many services-based environments, such as airports, bank branches, telecom service centers, and hospitals. The most significant advantage of big dashboards lies within their ability to serve multiple users concurrently. A flight arrivals board at the airport is designed to serve all passengers. No individual-specific information is displayed on such public boards. Monitoring and controlling patient flow within EDs requires managing individual's specific information. Thus, systems must consider patient privacy while displaying individual specific information on public dashboards.

Moreover, a targeted audience of ED public dashboards can be roughly categorized into two groups, patients and family escorts and care and management personnel. Each of these groups requires different information from the monitoring-and-control system. While designing a dashboard-based UI, careful attention must be paid to the various user groups and their different needs.

UI capabilities and the use of dashboards in a public environment, such as an ED, are subjects for a separate research [16]. In our research, we only minimally investigated the best information data required by patients for efficient patient flow control. We further developed a proof of concept ED dashboard for displaying the various KPIs that the EdRhythm generates.

4.5.1 The ED Dashboard

The ED dashboard provides the ED users—patients, care personnel, and management—with a graphical interface to the set of indicators calculated and monitored by the EdRhythm. The main purpose of the ED dashboard is to demonstrate EdRhythm capabilities and to display the various KPIs the dashboard monitors. The ED dashboard is implemented using Flex technology [2]. Flex allows fast development of rich and professional dashboards. Using Flex, the dashboard can be viewed from anywhere via most commercial Internet browsers. We utilize existing Flex widgets and components and integrate them into the ED dashboard. Figure 10 and Figure 11below show two snapshots of the ED dashboard.

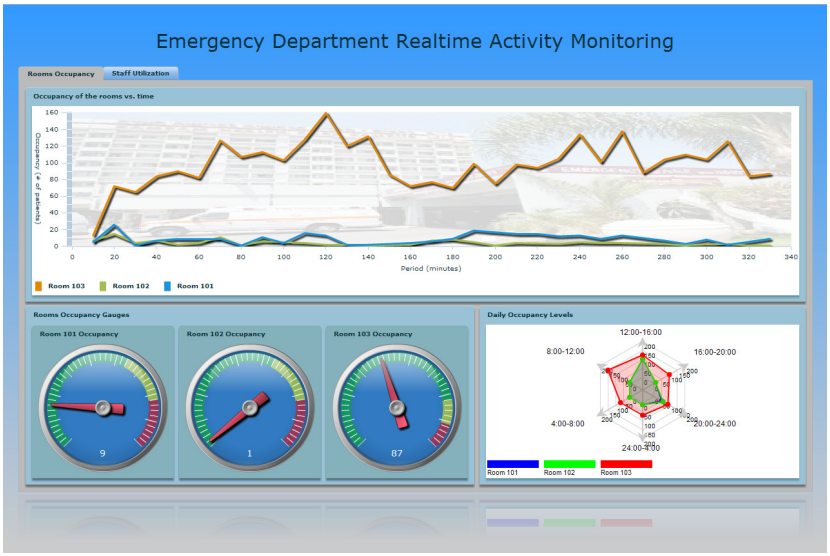


Figure 10: The Occupancy level indicator is displayed using several Flex widgets

Figure 10 demonstrates various display options for the physical occupancy KPI. The top view shows the occupancy level for the current day; the left meter view shows instant occupancy level, and the right view shows average daily occupancy level, measured annually.

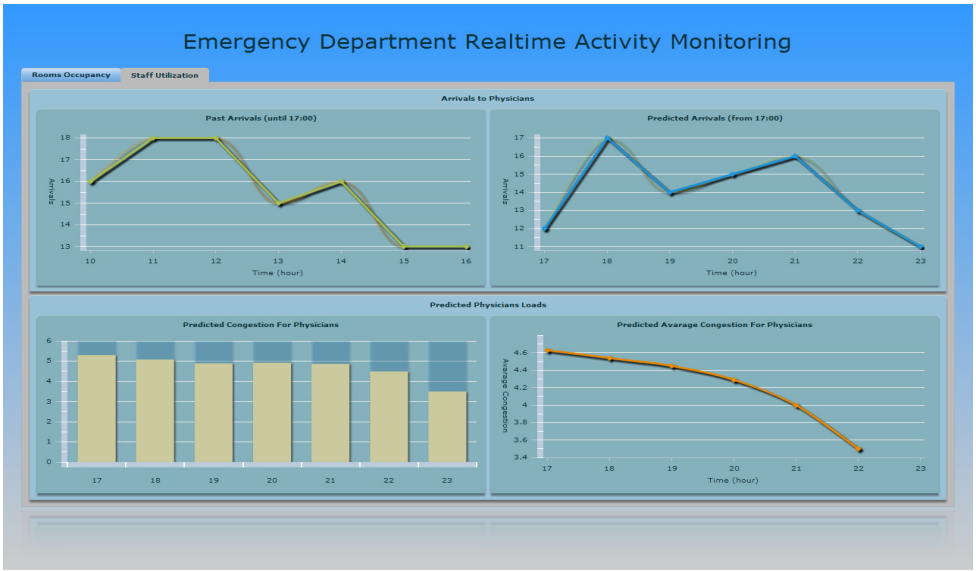


Figure 11: Arrivals and staff utilization including forecasting

Figure 11 illustrates the predictive capabilities and decision support functionality that might be incorporated into the ED dashboard. The top-left graph shows the patient arrival rate, as monitored by the EdRhythm till 17:00 for that day. The top-right graph shows the predicted arrival rate from 17:00 until 23:00 for that day. Such predictive capability, described in [30], can be incorporated into the EdRhythm. The actual physician utilization KPI, which is part of the personnel utilization KPI, can then be presented, alongside the predicted utilization, based on the expected arrival rate and the number of physicians that are expected to be present at the ED. Such a prediction assists ED management to adjust physicians' numbers for accommodating immediate ED requirements and for meeting the required KPIs. Note that predictive techniques and capabilities are beyond the scope of this research and are provided here just to complete the picture of ED dashboard capabilities.

4.5.2 Future Research into ED Situation Dashboard

Extensive research is still required and is underway to identify the various ED dashboard user roles and the specific set of requirements of each role. Such research will provide answers to crucial questions, such as:

- Which indicators should be monitored? And at what intervals?
- How much data should be presented in each screen?
- How should navigate between screens work?
- What thresholds are needed for alerting for the various KPIs?
- Is a drill down needed? And if so, to what level and by whom?
- How should the dashboard provide ED staff with real-time control capabilities?
- How should the dashboard provide "what-if" analysis and decision support?
- How can privacy be protected?
- And many more…

In the next chapter, we further describe methods and findings for monitoring the complex ED load KPI. In chapters 6, 7 and 8, we then follow with a deep analysis and suggested decision support for the question "Which patient to treat next?", seeking an appropriate clinical and operational balance between the TTFE and the LOS KPIs.

CHAPTER 5: REAL-TIME ED LOAD MONITORING AND MEASUREMENT

In previous chapters, we described the complexity of the ED environment, discussed various interesting KPIs to monitor and measure, and presented the key concepts behind the ED real-time monitoring-and-control system. One of the main promises of an ED's real-time monitoring is regarding its ability to measure the real-time ED load. In Section 3.2, we presented the ED load-monitoring and measurement challenge. During this research, we developed an innovative approach for real-time ED load monitoring and measurement. We then implemented this approach and demonstrated it using the EdRhythm system, as described in the previous chapter.

Our approach enables the measurement and calculation of user-tuned load, based on a wide spectrum of input data events and various predefined load functions. Being aware of specific user needs makes the system user-specific, i.e., resulting in a load score that reflects the relevancy of the low-level situational events to the subjective load experience of a specific user. Our approach, which is based on artificial neural networks [31], enables the following: i) a static mechanism for the definition of an explicit load function and ii) a dynamic learning mechanism that adapts the load calculation to user perception by overriding the explicit static load function definition.

The dynamic learning mechanism has two main advantages. First, it enables simple adaptation of the load function into any ED setting, bypassing the need to enforce a rigid load function definition for non-standard ED settings. Second, it allows for the calculation of different load values for the same objective situation. This is particularly useful for capturing the operational load perception differences of various user groups. Thus, by sacrificing rigid definition for high flexibility, our approach allows users to compare various situations and to reach informed decisions regarding the appropriate steps to take to reduce the ED load, by declaring an ambulance diversion situation, for example.

5.1 Dynamic ED Load Function

The first step in calculating the ED load is to define the load function. For that, we need to define the exact set of input parameters and the relative contribution of each parameter to the overall load score. Vast research has been devoted to the definition of a canonical and standard ED load function [20], [51], [55], [62]. The main goal of such research is to come up with a unified load score. A unified load score will enable ED comparison and will accelerate the development of generic methods for optimizing the ED operation. Unfortunately, until now, a consensus has not been reached on a unified

ED load score. Moreover, such a consensus may never be reached, due to the significant differences in ED types and settings. This observation leads us to develop an adaptive load function that calculates the load score based on a dynamic list of low-level input parameters. The exact set of input parameters that are relevant to a specific ED setting is chosen from an extensive set of parameters that were identified in the literature [55] as potential contributors to the overall ED load. Incorporating a dynamic and adaptive load function within the EdRhythm system allows it to provide a meaningful and valuable load score in a wide variety of ED settings. Configuring and adapting the load function into a specific ED setting can be done in two ways—statically, i.e., by explicitly defining the relative contribution of each of the input parameters to the overall load and dynamically, i.e., by using learning techniques for assigning the relative contribution level through a feedback mechanism. In the next sections we provide more details on these mechanisms and the way in which they are implemented.

5.2 Neural Network-Based Load Function

We chose to use artificial neural networks for implementing the ED dynamic-load-function-enabling mechanism. Neural networks (NN)-based functions are flexible for composition, adaptive over time, meaningful for the user, and enable the definition of complex relationships (e.g., nonlinear) between inputs and outputs.

5.2.1 Neural Networks – Theoretical Background

Artificial neural networks [31] are graphical representations of complex mathematical functions. They are composed of units called perceptrons (Figure 12(a)) and arranged as a multi-layered feed-forward network (Figure 12(b)), in which the outputs of one layer are the inputs of the next layer. This type of structure was inspired by the brain structure. Neural networks are successfully used in many applications such as pattern classification, dimensionality reduction, and function approximation [34], [36], [38]. Because of the origins of the network's design, the nodes in such networks are often called neurons. NN's greatest advantage, in comparison to other machine learning techniques, is their simplicity, both in representation and in learning. In addition, the number of required training examples relative to the network structure is not high compared to other machine learning solutions.

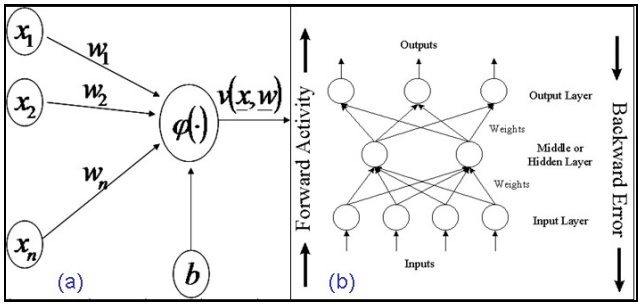


Figure 12: (a) single perceptron; (b) multi-layer network

Each perceptron is composed of n inputs $x_1, x_2, \ldots, x_n$; n weights $w_1, w_2, \ldots, w_n$; and an activation function $\varphi(\bullet)$. The output of the unit is $v(\underline{x}, \underline{w})=\varphi(\underline{x}^T, \underline{w})$, where $\underline{x}=(1, x_1, x_2, \ldots, x_n)$, $\underline{w}=(b, w_1, w_2, \ldots, w_n)$. Examples of activation functions are sign ($\varphi(u)=\text{sign}(u)$), linear function ($\varphi(u)=u$), and logistic function ($\varphi(u)=1/(1+e^{-u})$). The type of activation function affects the ability of the network to learn and is application-dependent. The units in different layers are connected in a feed-forward style to determine the network structure (see Figure 12(b)). The exact structure is also application-dependent, and in many cases domain knowledge can help to determine this structure.

Given an M element training set of the form $(\underline{x}_i, y_i)$, in which $\underline{X_i} \in \Re^n$ is the input to the network and $y_i \in \Re$ is the expected output (or target function) of the network, the back propagation algorithm [31] can be used to find a set of weights that minimizes the mean square error (MSE) between the provided output and the current calculated output. Two types of learning can occur—offline (or batch) learning and online learning. In offline learning, the entire training set is given in advance. At each iteration of the back propagation algorithm, all of the examples are taken into account when updating the weights. In online learning, the examples are given one after another, and each learning iteration depends on the current example only. Online learning is typically used when the environment changes over time, and when the network is trained to fit those changes.

5.3 Using Neural Networks to Calculate ED Load

To demonstrate the advantages of our methodology, we built a neural network that uses a wide range of load input parameters. We established the initial network structure and the set of individual parameter contributions based on the exhaustive structured-list and indicators suggested in the load parameter review paper [55]. We used the initial structure described in Figure 13 as our basic network for the load calculation.

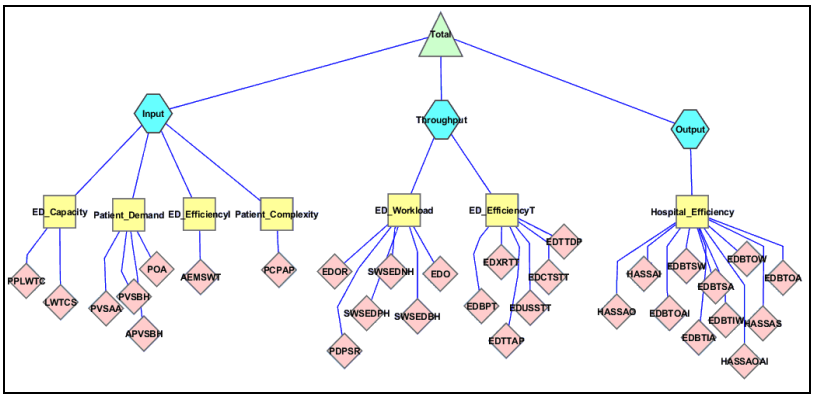


Figure 13: ED neural network view. The triangle is a total neuron, hexagons are the stage neurons, rectangles are the concept neurons, and diamonds are primitive input indicator neurons.

The hierarchy in the network consists of four main layers:

- **Indicator Layer** – This layer can take any number of low-level input parameters. Our initial implementation follows the set of indicators suggested in the review paper [55]. We modify some of the indicators so they better reflect the typical ED environment suggested by [44]. For example, indicator "ED Throughput time" was spliced into two nodes—one for admitted patients and one for discharged patients. This adaptation allows us to assign different contribution weights to each of the two more basic indicators.
- **Concept Layer** – The basic indicators from the first layer are connected to six concept nodes: patient demand, patient complexity, ED capacity, ED efficiency, ED workload, and hospital efficiency. In our basic setting, each indicator is connected to a single concept. The hospital capacity concept was omitted from the concrete implementation due to the lack of appropriate data. The ED efficiency concept was divided into two sub-concepts serving the input and the throughput separately. This modification was made to keep the tree-like structure of the network.
- **Operational Stage Layer** – The seven concepts from the second stage are connected to three operational stages, input, throughput, and output, following the ED conceptual model described in Section 2.2.
- **Load Score Layer** – This layer contains a single output node representing the total ED load score.

5.4 The ED Load Learning Mechanism

The artificial neural network learning mechanism can be implemented using either batch or online methods. Both approaches require knowledge of the "true" ED load value given a set of input parameter vectors. One way to get the set of ED load values is to present each input vector to the expert user and ask him/her to provide the ED load value in return. However, this method is not practical for two main reasons. First, input vectors are often too long for human perception and embedding. Second, the desired value of the target function cannot be explicitly calculated. In other words, there is no such thing as a "true" ED load value that the user is able to provide for a given input parameter vector. Thus, we choose a different approach, which relies on the EdRhythm situation dashboard and the user's subjective situation perception.

Instead of presenting the input vector itself, we present the current ED situation using the EdRhythm situational dashboard (Figure 14). The EdRhythm dashboard presents the current calculated ED load score (measured in percentage of the load baseline) together with additional information about the status of the ED.

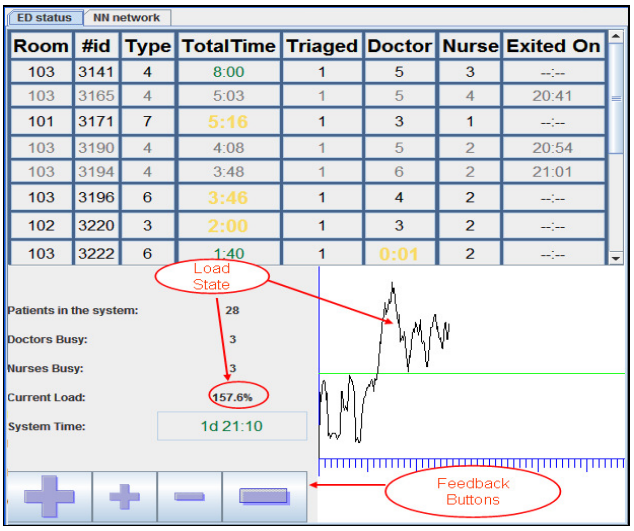


Figure 14: Dashboard snapshot;
load value (black line) is calculated as the percent of the average (green line)

By looking at the dashboard, and by physically experiencing the ED's current situation, the user gains insight into the accuracy of the calculated ED load. This insight is merely subjective and is based on a comparison of the current situation to previous situations as experienced by the user. Using the relative feedback buttons, the user then provides feedback about the discrepancy between the system's calculated load and her subjective load experience. The neural network learns from the provided feedback and adjusts its load function accordingly. For example, if the user feels that the represented load is far

below the desired value, she clicks on the larger + button, indicating that the load score should be increased by approximately 10%. A similar update process (+1%, -1%, -10%) is executed for the other three relative feedback buttons. Letting users provide feedback over long enough time periods results in having a load functions that accurately reflect the ED load. After completing this learning process, the system is able to present a real-time load score as a percentage of the load baseline, as can be seen in Figure 14.

5.5 Tracking Load on Internal Resources

One key advantage of a neural network lies in its ability to reflect a complex physical structure, i.e., by allowing every neuron to have an explicit operational meaning. For that, our neural network design keeps the tree-like neuron hierarchy instead of the usual all-to-all connections. This allows each neuron to preserve its operational meaning during the learning process. Conserving the tree-like structure allows the user to track the current load back into the environment and to gain a deeper understanding of the current load status (Figure 4). Moreover, we can get an alert from any hierarchy level in the system if a certain neuron becomes overloaded. For example, in some situations the overall ED load is only 40% of the baseline, but the CT room is overcrowded due to lack of personnel. In these cases, the appropriate neuron's status will reach the high mark and the system will thus send an alert to the situation dashboard, provided the neuron was preconfigured accordingly. As a result, the ED manager might react by sending less severe patients to the Hospital's CT room instead of to the ED's CT room, for example.

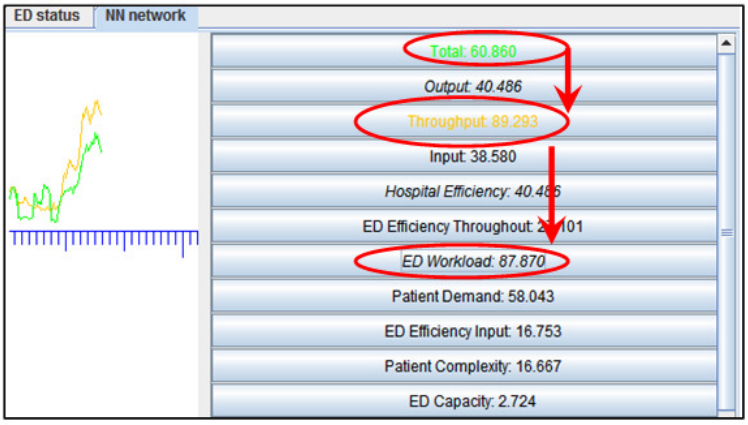


Figure 15: Tracing load. Green line indicates total load, orange line indicates throughput; the rise in the total load was clearly caused by increasing the throughput neuron; we can trace it further and deduce that the peak in throughput was caused by elevation of the ED workload concept neuron.

5.6 Multiple Views of ED Load

A key advantage of our framework is its ability to allow dynamic learning based on feedback from different user groups. Thus, we can calculate and present several different load scores for the same objective situation. This ability allows us to adapt the load values to a variety of ED settings. The only concrete requirement for providing meaningful values is to get consistent user feedback. This can only be achieved if feedback will be consistently provided by the same user. Situations might occur in which we can extend the group of users that provide feedback, assuming their feedback is somewhat consistent with the load situation. For example, measuring the current ED load as perceived by doctors, nurses, and patients, or even by a single individual such as the ED manager, could have an interesting application. Research [44] shows that each role group in the ED consistently experience varying loads due to frequent environmental changes throughout each day. Thus, the ability to establish a subjective load function that best reflects the actual load experienced by a given user group could be a useful tool for managing the day-to-day ED load.

To enable the EdRhythm system to reflect a group's subjective load, we first need to define the group's profile. Each group's profile reflects operational load as it is being experienced by a given group. Group profiles can be statically defined by fixing weights to relevant neurons, or by dynamically learning them, which is preferable. Dynamic learning involves capturing feedback from a specific user group associated with a specific profile.

To demonstrate the system's ability to reflect various load values for various user groups, we identified three possible group types—nurse, doctor, and patient. Each group has an assigned group profile and a relevant target load function. Nurse and doctor target load functions were defined as the average occupation ratio during a certain time period. Patient target load function was defined as the ratio of a patient's waiting time to the patient's total staying time in the ED. Figure 16 shows a typical day load curve for the three user groups.

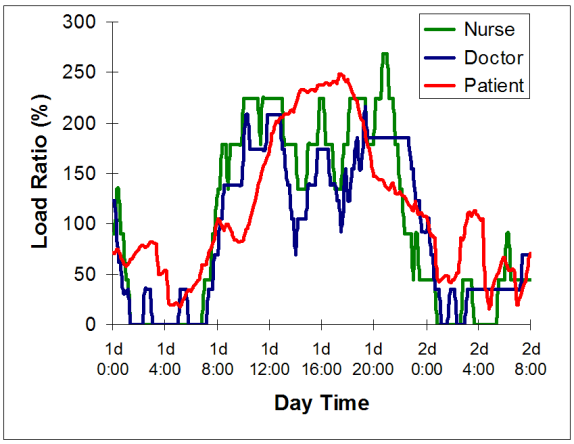


Figure 16: Simulated nurse, doctor, and patient profile behavior, when 100% is the average daily load

Figure 17 summarizes average load values for the three examined user groups: nurse, doctor, and patient. Each profile comprises the weights of major relevant neurons monitored by the system.

Raw Indicator \ Profile	Nurse	Doctor	Patient
Patient Volume standardized for Bed Hours	42%		76%
Summary Workload standardized for ED Bed Hours	18%	45%	
ED Bed Placement Time	12%	30%	8%
ED CT service Turnaround Time	28%	25%	16%

Figure 17: User profile composed of major raw indicator weights learned by the system

We can see that all three profiles show reasonable behavior when comparing time of day and when comparing the profiles to one another. When the system is overloaded, all users feel it. However, the load experience is different for each group. For example, doctors must stay later than nurses at the end of the day to close all open cases. Thus, the load on them decreases later than it does for nurses. On the other hand, triage, served by nurses, is the first station in the patient flow. Hence, the nurses' operational load starts earlier. These examples demonstrate that different weights do indeed exist on the neurons, emphasizing the need for subjective load scores for the same objective ED state.

CHAPTER 6: ED PATIENT FLOW CONTROL

In the previous chapter, we demonstrated the benefits of deploying a real-time monitoring and control system for the monitoring and measurement of ED load. The control part of the real-time monitoring-and-control system deals with the ability to influence patient flow within the ED in real time to improve the ED key performance indicators (KPIs). To do so, we first need to gain a deep understanding of its various aspects (see Figure 6). In this chapter, we examine specific ED operational aspects, analyze them, and suggest potential optimization approaches. We then follow with some simulation results in Chapter 7, then continue, in Chapter 8, with a mathematical analysis of the problem's stylized model. For the analyses, we need a set of mathematical tools. Queueing network is a natural choice for analyzing the ED environment and for gaining a deep insight into its KPIs. As described in Chapter 2, the ED internal operational processes can be analyzed using a queueing network model. The processing servers model the various ED stations and the jobs model the patients that require service at one of the ED stations. The ED usually operates in high load, thus most stations usually have waiting queues where patients wait for service.

Consider the comprehensive ED activity and resource queueing model described in Figure 5. The complexity of such models renders them intractable for mathematical analysis and unsuitable for gaining insights into service policies and optimal control. Thus we continue our analysis by breaking it into smaller queueing models and analyzing them as if they were stand-alone models. Specifically, we identify two complementary views while modeling the ED as a queueing network—the patient's view and the care personnel view. We then translate these two views into two related queueing models, each with its own decision problem. From the patient's view, we confront a routing problem, i.e., deciding to which station the patient must go next; while from the care personnel's view (i.e., the station's view), we confront a scheduling problem, i.e.,, deciding which patient should be treated next. We further simplify our models as required, by adding additional constraints, e.g., over the arrival process, to make it mathematically tractable. Furthermore, we use complementary techniques, such as simulation, and interviews with ED managers, for gaining additional insights into the operational behavior of the analyzed control policy in situations as close as possible to those expected to be found in the day-to-day ED reality.

In the following sections, we first briefly address the flow control routing problem from the patient's perspective, specifically, "Where should the patient go next?". We then continue with a detailed analysis of the flow control scheduling problem from the station's view—"Which patient should the physician treat next?".

6.1 Which Station to Visit Next?

The exact station-sequence for a new patient arriving to the ED is unknown upon arrival and is determined as treatment progresses. The treatment process is not sequential and patients often return to the same station multiple times. Notice, though, that the next station(s) in a patient's route within the ED is always known. Usually, after completing treatment at some station, a physician provides indication about the required subsequent stations. The physician may indicate the need to visit more than a single station, e.g., for ordering laboratory tests, for consulting a specialist such as gynecologist, and for issuing an ultrasound test. In some situations, physicians request partial ordering over the station sequence. For example, a gynecologist may need to see the patient only after an ultrasound test is completed. In other situations, treatment in different stations can be executed in a totally arbitrary order. For example, lab tests must be taken prior to a final decision but independent of a gynecologist's consultation.

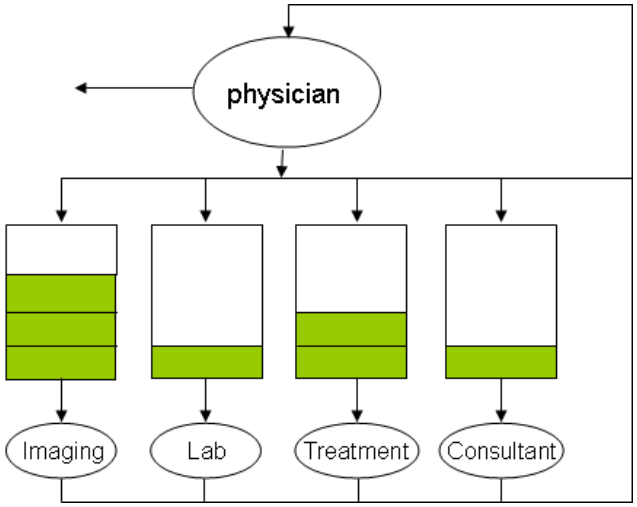


Figure 18: A stylized queuing model for the "Which station to visit next?" problem.

Thus, the dynamic, flexible, and unpredicted ED environment offers the potential for significant routing optimization. Figure 18 describes a stylized model for the "Which station to visit next?" decision problem. Such models were mathematically analyzed by A. Zviran [65] in the context of Healthcare. Indeed, even an elementary real-time control of patient flow would likely yield a significant improvement of the ED operational environment. For example, an immediate ideal for flow control is to direct patients to the station with the shortest queue. For an improved control, we could look further into the anticipated patient's route. Such controls may require forecasting of the future stations on a patient's treatment route. Our analysis does not deal with such

optimal routing issues. Clearly, however, such decisions must play a significant role in the overall optimization of the ED patient flow.

6.2 Which Patient to Treat Next?

The "Which patient to treat next?" (PTN) decision problem stands at the core of our research. This question emerged as most interesting within the ED day-to-day operation. The stochastically high load under which the ED usually operates results in an operational queue of waiting patients prior to each treatment station. In other words, patients must typically wait for treatment at each and every station along their treatment route. Thus, the improvement of the ED operational performance, gained by optimally addressing the PTN question, is expected to be significant.

PTN improvements may affect the patient's final clinical results, which is the most important ED performance measure. The final clinical results may be complex to measure. Measuring the effect of PTN optimization on ED final clinical results is out of the scope of our research and is subject to future work. Nevertheless, and to partially address this difficulty, we considered multiple views while analyzing the PTN question, namely the clinical view, the service level view, and the operational view. Together, these views provide a reasonable estimate for overall ED performance with respect to the problem at hand. Next, we describe the core aspects of these views and their implications on the PTN question.

6.2.1 Clinical View

The clinical view is the most important view while addressing the PTN question. It is important to keep in mind that most patients' clinical aspects are hidden to physicians during the ED care process. The main purpose of the ED care process is to reveal the patients' clinical situations and to act upon them. The known clinical aspects of the PTN questions are encapsulated within the triage score. The triage score presents the patient's clinical status as assessed by the triage nurse prior to the first physician encounter. Best practices at most EDs do not require physicians to update the triage score after the first encounter, nor do they require providing other "running" clinical scores to patients, even though such a procedure may lead to further improvements in answering the PTN question.

6.2.2 Operational View

The operational view is our main focus while addressing the PTN question. Our sought-after policy seeks to improve various EDs' operational KPIs, including time till first encounter, overall patient length-of-stay, patient waiting-time to service-time ratio, and the overall number of patients within the ED. By limiting ourselves to operational

optimization, we must translate clinical criteria into their operational proxies. This is done mainly by using the triage score as a proxy to the patient's clinical status, and by providing cost functions that take into account parameters with clinical relevancy, such as patient age, waiting times, and the likelihood of a patient to be admitted to the ward or discharged from the hospital.

6.2.3 Service Level View

The service level refers to a specific operational aspect that focuses on the customer experience during the ED care process. Recent works in service science, such as Armony et al. [14], suggest a greater focus on patients' needs while providing service. Improving service level, in our context, means improving the service experience from the patient's point of view. We translate the service level view into reducing waiting time and the overall length-of-stay at the ED, without compromising the level of clinical treatment. One interesting option for service level improvement is to assign priority to patients who are about to be discharged from the hospital over patients who are about to be admitted. Since admitted patients will stay at the hospital anyway, they are thus more agnostic to the overall length-of-stay. Patients who are about to be discharged back to their homes are anxious to leave the ED, and the hospital, as soon as possible, which is also clinically safer for them. Fairness is a particularly interesting service performance indicator. For example, taking patients' age into account while determining which patient should be treated next may consider favorably from the service level fairness perspective. First come first serve (FCFS) is a widely-used fairness policy at EDs, which most patients would accept. Changing this policy while trying to address a wider KPI set may result in patient disappointment and objection, namely service level deterioration. The affect of applying a service policy other than FCFS on patient satisfaction and on other service level KPIs is subject to future research.

6.3 Operationally Optimal PTN Control

Discussions with ED managers led to the following clinical requirements, while addressing the PTN question—patients' length-of-stay should be minimized and meeting triage deadlines is a must. Furthermore, these discussions also suggested differentiation in the waiting costs of different patient classes. Based on these discussions, we formulated a clinically optimal control that yields the following guidelines:

A clinically optimal PTN control meets triage deadline constraints with the least minimal effort for newly arrived (NA) patients. It then serves in process (IP) patients so as to minimize clinical costs. The cost functions combine clinical, operational, and service level aspects, i.e., by using waiting time, triage score, patient age, and ADT status as input parameters. The exact values for these parameters depend on a specific ED setting

and are thus to be provided by the ED manager. The optimal PTN controls with their associated aspects are further described in the following sections.

6.4 The PTN Stylized Queueing Model

The core of the PTN question deals with the need to serve several competing work items, e.g., the patients' encounter, by the same server, e.g., physician. Thus, the question becomes which work item should the server cater to next. Obviously, the work item parameters should be considered. Mapping the PTN into a queueing network model leads to the following formalism:

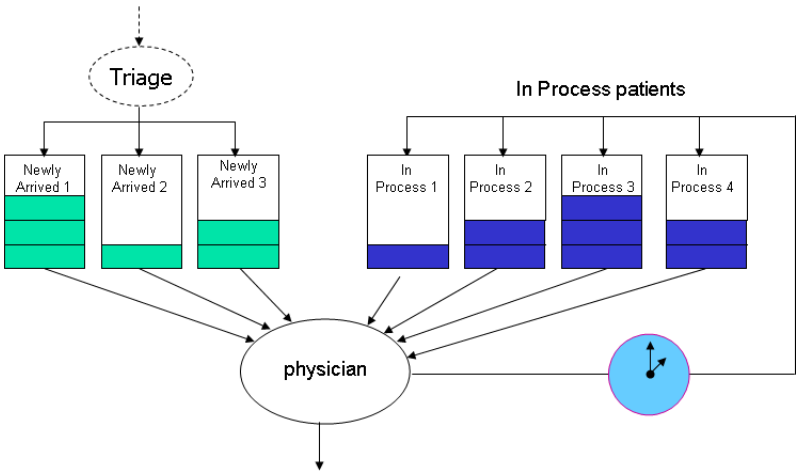


Figure 19: A stylized queueing model for the PTN question.

Consider a multiclass single station queueing system with feedback. Each station is comprised of a pool of statistically identical servers (physicians). Newly arrived work items (patients) arrive to the queue (waiting area) in a generic arrival process. Service time (patient-physician encounters) is modeled by an independent and identically distributed (i.i.d.) random variable s, with mean E(s). Service rate is thus given by $\mu=1/E(s)$. The system has feedback. That is, after finishing service, the item (patient) either leaves the system (discharged home or admitted to the hospital) or continues to a latent time period, i.e., a period at which it is being served at other station(s), and returns to one of the in process (IP) queues for additional service.

The system has multiple priority queues. Each item has a priority index based on a variety of parameters. Thus, items are not necessarily served by their arrival order but are based on their priority. The scheduling algorithm chooses the next item to process based on some service policy that may take item's priority into account.

Special attention is given to the encounter sequence number. Items requiring initial service (NA patients) are treated differently than items already within the process (IP patients).

6.5 Queueing Theory and Queueing Networks

Queueing theory is the mathematical study of waiting lines, or queues. The theory enables the mathematical analysis of several related processes, including arriving at the (back of the) queue, waiting in the queue, and being served at the front of the queue. The theory permits the derivation and calculation of several performance measures, including the average waiting time in the queue or the system, the expected number waiting or receiving service, and the probability of encountering the system in certain states, such as empty, full, having an available server, or having to wait a certain amount of time to be served.

Networks of queues are systems that contain an arbitrary but finite number of queues. Customers, sometimes of different classes, travel through the network and are served at the nodes. In open networks, customers can join and leave the system, whereas in closed networks, the total number of customers circulating within the system remains fixed. In queueing theory, a queueing model is used to approximate a real queueing situation or system, so the queueing behavior can be analyzed mathematically. However, the queueing model, e.g., the one described in Figure 19, often turns out too complex for exact analysis. We can then resort to approximations, as was done in this case. The referenced model was analyzed asymptotically in heavy traffic, by Huang, Carmeli and Mandelbaum [37]. Next, we provide further description of several queueing model concepts that are relevant to the ED patient flow.

6.5.1 System Utilization and Operation Regimes

The system utilization, ρ, is determined by the proportion between the system's capacity, μ, and the arrival rate, λ. Arrival rate is defined as the rate by which jobs arrive to the system. Time-varying arrival rates, as found in the ED environment, result in a transient network. A transient network is characterized by its alternation between low utilization and high utilization [45].

In present research of queueing systems, various operational regimes have emerged, which place a different focus on resource efficiency vs. service quality. Moreover, most of this research is for queueing systems with many servers that operate in steady state (an exception is the work done by Yom-Tov [63]); thus, this research is not directly relevant to ours since EDs, modeled as queueing systems, require time-varying analysis for small number of servers.

We thus further focus our research on a situation of rigid capacity. We assume that the ED resource level (e.g., number of servers) is fixed and search ways for a real-time improvement of ED performance under time-varying arrival rate by optimizing operational aspects, such as the PTN service policy.

6.5.2 Service Policies and Priority Queues

Service policy is defined as the policy by which the server caters to jobs that wait in queue. The most straightforward service policy is first come first serve (FCFS), in which jobs are being served by their arrival order. Note that in the FCFS, queue arrival time is the only parameter that is being considered for deciding upon the next job to serve. As a result, FCFS can easily be modeled as a single queue single server queuing system. This is not the situation in most service environments, including in the ED. A multiclass queueing system allows for classifying jobs into classes by using various job characteristics and then to assign a queue for each of these classes. The service policy then must determine which queue to serve next. We assume jobs within each queue are being served along the FCFS policy.

In our model, we thus need to decide upon the characteristics to consider while classifying jobs into classes. Specifically, within clinical environments, we consider clinical characteristics, such as patient age and triage score, alongside with operational characteristics, such as the triage deadline and the expected ADT status. Next, we describe several known multiclass service policies.

6.5.3 Cost-based Service Policy and the Generalized Cμ Rule

A cost-based policy associates a cost function with each queue. Specifically, following [59], we consider a general single-server multiclass queueing system that incurs a delay cost at rate $C_k(\tau_k)$ for each class k job that resides τ_k units of time in the system. We denote the marginal delay cost and (instantaneous) service rate functions of class k by $c_k = C'_k$ and μ_k, and we let $a_k(t)$ be the "age" or time that the oldest class k job has been waiting at time t. We call the service policy that at time t serves the oldest waiting job of that class k with the highest index $\mu_k(t)c_k(a_k(t))$, the Generalized cμ (gcμ) Rule. Van Mieghem further shows that, with non-decreasing convex-delay costs, the gcμ rule is asymptotically optimal if the system operates in heavy traffic. The gcμ rule suggests an attractive policy for serving IP-patients given that suitable cost functions are provided. The cost-based policy seems less suitable, at least directly, for meeting the deadline requirements of NA-patients.

6.5.4 Deadlines-constrained Service Policy

A time limit or deadline is a narrow interval of time, or particular point in time, by which an objective or task must be accomplished. Deadlines are types of operational constraints provided by operation managements as a control mean. A deadline-constrained service policy, within a multiclass, single-station queueing system, strives to serve any class k job arriving at time t by its deadline $t+d_k$. As shown by [60], the deadline-constrained service policy can be approximated by sequence of convex-increasing delay cost functions. This formulation reduces the intractable optimal scheduling problem into one for which the gcμ scheduling rule is known to be asymptotically optimal. Moreover, such an approach allows translating a deadline-constrained service policy into a cost service policy.

6.6 Multiple Decisions

Assessing the PTN clinical requirements reveals that the PTN decision process can be viewed as a decision tree in which multiple decisions are required at multiple levels, as shown in Figure 20.

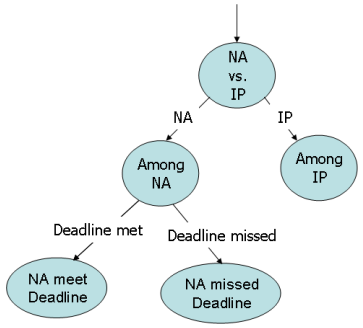


Figure 20: The PTN conceptual decision tree.

First one must decide whether to choose from the NA-patient queues or from the IP-patient queues. This first decision is derived from the different clinical requirements associated with the different patient types. That is, NA patients should be served by their deadline, while IP patients should be served based on their overall lengths-of-stay.

The second decision to make, once a patient type is chosen, is which exact queue to serve within that type. Lastly, different policies are necessary for situations in which the triage deadline can be met and for situations in which it cannot.

6.6.1 Main Service Policy Options

We can thus apply many different service policy combinations and test them all. Our analysis starts by considering the major service policy alternatives for the first level of the PTN decision tree. These options are described in Table 1.

Policy	Main Approach
First come first serve (FCFS)	Serve the patient with the longest queue time next, while ignoring the total length-of-stay as well as all other clinical and operational parameters.
Cost	Serve the patient with the highest waiting cost next. Waiting costs are based on suitable convex functions, i.e., as provided by the ED manager, following the gcμ rule.
Deadline-constraint	Set deadlines for both NA patients (triage deadline) and IP patients (total length-of-stay, e.g., four hours) and try to meet them both using the deadline-constraint policy suggested by Van Mieghem [60].
Hybrid	Strive to first meet NA patient deadline constraints, then chose among IP-patient using a cost function

Table 1: Major service policy classes for the PTN.

6.6.2 The Hybrid-Approach Service Policy

The hybrid-approach to the PTN question applies different service policies to NA and IP patients. The hybrid-approach starts by deciding whether to next treat NA or IP patients, and for that it applies some threshold mechanisms. The hybrid-approach leads to a two-step decision problem. The service policy must first determine whether to treat NA patients or IP patients. After this first decision, the policy then determines which specific patient to select out of these two groups.

6.6.3 The Constraint-based Service Policy

A specific class of hybrid-approach policies is the constraint-based approach, which handles the triage time till first encounter (TTFE) deadline as a constraint that must be met. The constraint hybrid-approach thus seeks to minimize situations in which NA patients first see a physician after their assigned triage deadline. It then seeks to reduce the overall waiting cost of IP patients. Intuitively, such a hybrid-approach may result in

very long lengths-of-stay (LOS) because it gives total precedence to NA patients. LOS can still be controlled through the admission-control policy that refers patients to other hospitals in case of ED overload. In addition, a situation could still occur in which the TTFE constraint cannot be met. Such a situation requires that the scheduling algorithm be able to choose the next patient to treat from a list of patients who already missed the deadline. A complementary cost-approach methodology for NA patients is suggested to handle this situation.

6.6.4 The Minimal-effort Due-date Service policy

A particularly interesting class of constraint-based policies is the minimal-effort due-date class. Policies in this class seek to meet NA-patient deadline constraints using minimal processing effort, thus allowing maximal processing of IP patients. The minimal-effort class of service policies best combines clinical and operational aspects and thus provides the core focus of our research.

6.6.5 List of Analyzed Hybrid-Approach Policies

Table 2 lists the various threshold algorithms we analyzed, under the various hybrid-approach policies, and provides key aspects of each.

Policy	Main Aspects
Newly arrived first (NAF)	NAF serves any existing NA patients and otherwise serves IP patients.
Static threshold (STD)	STD chooses between NA patients and IP patients using a computed static threshold. The STD threshold assumes a fixed arrival rate and heavy traffic conditions. It then computes the threshold as follows: $Treshold = \sum_{j \in J} \lambda_j d_j m_j$, where j is a specific triage queue out of J queues, λ_j the arrival rate to the queue, d_j the queue's deadline and m_j the total service time of the patient in the system. The STD algorithm follows the PTN mathematical analysis suggested by Huang and Mandelbaum [37].
Adaptive threshold (ATD):	ADT is a modification of the STD approach, which handles varied arrival rate. The ADT estimates the arrival rate in an hour time slot and adapts the static threshold accordingly.
Greedy threshold (GTD)	GTD chooses an NA patient who is about to miss the deadline from the NA queue, if one exists, and otherwise chooses an IP patient. The GTD only looks at head-of-the-line NA patients for making a decision.
Dynamic threshold (DTD)	DTD suggests a heuristic improvement over the GTD approach. The DTD starts as GTD, i.e., serves the NA patient if at least one such patient is about to cross the triage deadline. DTD then further invokes a look-ahead mechanism. It looks into the NA queue and estimates through simulation the expected "start of treatment" time for each patient already in the queue, assuming all processing capacity is allocated to serve NA patients, and that each treatment will take an average service time. It then determines which NA patient to treat if the look-ahead process suggests that at least one of the NA patients already waiting in the queue will miss his deadline.

Table 2: List of Hybrid service policies.

6.6.6 The NA Internal Competition

We considered two policies while choosing among NA patients in constraint-based policies, namely the portion policy and the difference policy. A portion service policy (w/d) gives precedence to jobs with a higher portion between the waiting time and the deadline. A difference policy gives precedence to jobs with a lower difference (w-d). A combined approach may also be considered. Such an approach uses difference if the deadline can be met and portion if it cannot, as indicated for DTD in Table 3.

6.6.7 The IP Internal Competition

We considered several service policy classes for the internal IP competition, namely FCFS, cost, and even a deadline-based class. We found the cost approach most suitable for resolving the IP internal competition and thus used it within all analyzed algorithms.

6.6.8 Hybrid-approaches with NA and IP Service Sub-Policies

Hybrid-approach service policies comprise several service sub-policies, as shown in Figure 20. Table 3 summarizes key aspects of the core subset of the hybrid-approach policies we analyzed during our research, along with their service sub-policy components.

Service policy	**Threshold methodology**	**Choosing among IP patients**	**Choosing among NA patients who meet deadline constraint**	**Choosing among NA patients who missed deadline constraint**
NAF	NA patient first	FCFS	Difference (w-d)	Difference (w-d)
STD	Calculating using the queues length	Cost	Portion (w/d)	Portion (w/d)
DTD	Calculating using waiting time of the "oldest" waiting patient	Cost	Difference (w-d)	Portion (w/d)

Table 3: Hybrid-approach service policies and their service sub-policy components.

6.7 The "Best" Service Policy

In Chapter 7, we provide detailed simulation results for the various service policies described in Table 3. Our simulation suggests that the dynamic threshold (DTD) hybrid service policy, which applies the threshold algorithm described below along with the service sub-policies described in Table 3, best addresses clinical needs under realistic ED situations.

Below, we provide the pseudo code of the DTD threshold algorithm.

```
At every service completion do{
 If the head of the line patient at any of the triage queues
 already missed the deadline then{
      • serve NA-patients
 }Else{ // perform a look ahead into the triage queues
   Save a copy of the queueing system and the simulation time;
   While there are patient at any of the triage queue
   Do{
        •Pick the patient who is the closest to her deadline
        •If that patient already missed the deadline then{
           ▪  Exit and serve NA-patient
         }
        •Serve this patient, assuming average service time
        •Increase simulation running time by the service time
           amount
   }//do
      •  Exit and serve IP-patient
  }//else
   Restore queueing system to its original status and
   simulation time to it original time
 }
```

Figure 21: DTD threshold algorithm pseuso code.

Note that the main strengths of the DTD service policy are i) its robustness with respect to the varied arrival rate and ii) its ability to serve as a real-time control in a real-life ED environment.

Analyzing encounter data from a real ED setting suggests that NA patients comprise about 30% of encounters. Thus, only 30% from the system capacity is required for serving NA patients. As a result, we can reasonably assume that in most situations, the system has enough capacity in the system to serve NA patients just before their deadlines. The look-ahead capability of the algorithm allows it to proactively compensate for short arrival rate peaks that exceed maximal system capacity.

We may also consider an interesting improvement to the DTD threshold algorithm, that is, to also simulate arrivals into the triage queue based on the expected arrival rate,

during the simulated look-ahead process. The current DTD threshold algorithm does not exploit this enhancement. We envision that such an enhancement will only have a minor effect on the results. We thus do not discuss it further.

In the next chapter, we provide detailed simulation-based analysis results for the DTD service policy, as well as for other important service policies. We then provide a comparison analysis between the DTD and the STD service policies. In Chapter 8, we provide fluid model analysis for a close stylized variant of the DTD threshold algorithm.

CHAPTER 7: WHICH PATIENT SHOULD BE TREATED NEXT? SIMULATION-BASED ANALYSIS

In this chapter, we describe the discrete event simulation (DES) [5] we developed for analyzing the PTN question and the simulation analysis results. The PTN simulation works in batch mode and is designed to evaluate potential PTN service policies for their performance. Nevertheless, the chronological nature of DES allows the developed service policy to serve as a real-time control within the EdRhythm system, described in Chapter 4, while prototyping a real-life ED setting. Moreover, the DTD service policy, described in Section 6.7, uses the simulation engine for real-time control. Collecting events from the monitored environment and using real-time simulation to control it represents a methodology known as symbiotic simulation [18].

7.1 The Simulation Environment

The PTN simulation environment comprises three main functionality modules: the ED process simulator, the scheduling engine and the result processor. The scheduling engine, which provides the core algorithms of the PTN simulator, is implemented using the Java [10] programming language on top of the Eclipse framework [6]. The ED process simulator generates a list of patients visits based on a set of preconfigured parameters. The scheduling engine then applies various scheduling algorithms to the visit list and logs its output into an Excel file using the results processor. We then analyze the Excel file and compare the results of the various scheduling algorithms.

7.1.1 The PTN Simulation Input

We use the EdRhythm platform, described in Chapter 4, to generate the time-based list of patient visit events. The EdRhythm is calibrated to generate low-level visit events according to the visits expected to find in a typical ED. To this end, we validated the patient arrival profile, generated by the EdRhythm system, with several ED managers. The EdRhythm generates all kinds of low-level events, related to various ED activities. Thus, we first filter the list of low-level events generated by the EdRhythm, allowing only relevant events to pass through. The list of relevant low-level events is summarized in Table 4.

	First Encounter	**Interim Encounter**	**Last Encounter**
Waiting starts	33	41	47
Encounter starts (waiting ends)	34	42	48
Encounter ends	35	43	49

Table 4: List of processed low-level events.

All patients have at least two encounters, namely first encounter and last encounter. The first encounter is indicated by events 33, 34; the last encounter is indicated by events 47, 48, and 49. Many patients have more than two encounters. In rare situations, patients may have up to six encounters. Interim encounters are indicated by events 41, 42, and 43. Table 5 describes a typical number-of-encounters distribution.

2 encounters	**3 encounters**	**4 encounters**	**5 encounters**	**6 encounters**
28%	30%	28%	11%	3%

Table 5: Distribution of the number of encounters in a typical input set.

We aggregated the list of low-level encounter events into a list of encounter records. Each encounter record is composed of 4 elements that are described Table 6.

Encounter Number	Arrival Time	Service Time	Latent Time
Sequential encounter number	Indicates the arrival time to the queue	The time of physician-patient encounter	The time patients spend in other ED stations before returning to the physician queue

Table 6: Patient's encounter record.

The time of the first encounter event (33) indicates the patient's arrival time at the ED. Service times are calculated by subtracting the time in the 34-35, 42-43, and 48-49 pairs, respectively. Latent time is calculated by subtracting the encounter end time (either 35 or 43) from the following encounter's start waiting time (41 or 47, respectively) for the same patient. The latent time is then used for calculating the next patient's arrival time into the physician's queue.

7.1.2 Additional Input Parameters

The PTN simulator further assigns several clinical characteristics to each simulated patient, namely her triage score with its associated deadline, age, and the expected ADT status based on a realistic distribution. It then uses these characteristics for setting deadlines and waiting costs for waiting patients. The distribution of these parameters can be configured into the simulator. Below, we provide a typical realistic distribution of the various input parameters.

Table 7 describes a typical distribution of patients along triage scores.

Triage 3	Triage 4	Triage 5
10%	40%	50%

Table 7: Typical triage score distribution.

Table 8 describes the typical triage deadlines for the three triage scores.

Triage 3	Triage 4	Triage 5
30 minutes	60 minutes	120 minutes

Table 8: Typical triage deadlines.

Table 9 describes the age groups, as defined by an ED manager, and the respective distribution of patients among these groups.

Under 45	45-65	65-75	Over 75
40%	30%	20%	10%

Table 9: Patients' age groups with their distribution.

Table 10 describes the expected ADT distribution. We could further seek the ADT distribution for each of the triage groups. Notably, expected ADT status does not currently exist in most ED settings, and the option for caregivers to update the ADT status throughout the course of the treatment seems even more far-fetched. Furthermore, the ability to provide accurate ADT status heavily depends on the proficiency level of the caregiver.

Admitted	Discharged	Unknown
30%	60%	10%

Table 10: Distribution of the ADT expected status.

More importantly, the input parameters, with their respective distribution, provide just a baseline for the PTN simulation. In a real environment, the exact distribution can directly be calculated from the historical data, collected by a system such as the EdRhythm. We tested all service policies under a wide variety of input parameters to ensure their robustness functionality under the expected day–to-day ED conditions.

7.1.3 Arrival Processes

The arrival process has a significant impact on the performance of the service policy being tested. The PTN simulator provides two types of arrival processes—a Poisson arrival process and realistic arrivals. We use the Poisson arrival process for analyzing the various service policies for robustness. The realistic arrival pattern is provided in Figure 22. The PTN simulator provides a means to scale up the realistic arrivals while maintaining the arrival pattern itself. This ability allows scaling up the system-under-test into any desired size. Thus, for our tests, we generated an arrival function with scaled rates based on these two patterns.

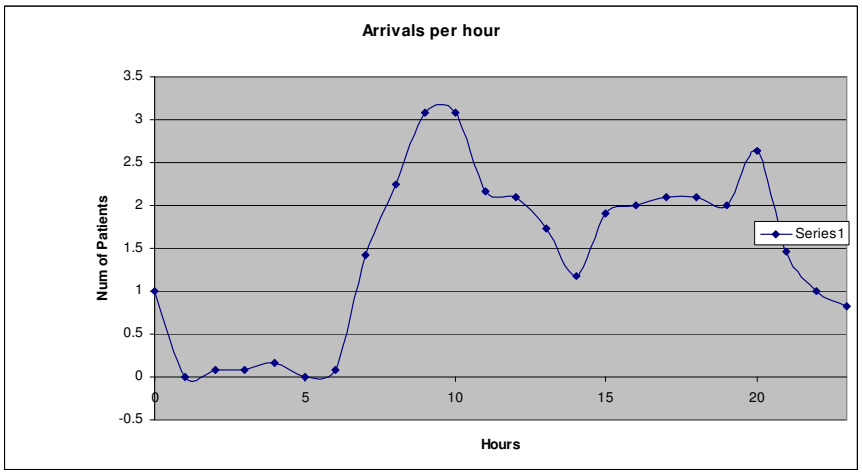


Figure 22: A typical realistic arrival rate patern for a small ED.

7.1.4 Setting the Cost parameters

Service policies use cost functions for deciding which patient to treat next. Together with an ED manager, we defined the cost functions c(t), t≥0, along with their parameters, as described in the following tables.

Table 11 describes the triage-related costs. Note that the important factor is the relative costs among the various triage groups.

Triage 3	Triage 4	Triage 5
$c_1(t)=4*t$	$c_1(t)=2*t$	$c_1(t)=t$

Table 11: Triage-related costs.

Table 12 describes the age-related costs, represented in terms of the triage costs.

Under 45	45-65	65-75	Over 75
$c_2(t)= c_1(t)$	$c_2(t)=2*c_1(t)$	$c_2(t)=3*c_1(t)$	$c_2(t)=5*c_1(t)$

Table 12: Age-related costs.

Table 13 describes the costs for the Expected ADT status. The current cost function provides the same cost for unknown and admitted statuses. This yields higher priority to patients with high discharge probability.

Discharged	Admitted or Unknown
$c_3(t)= 2*c_2(t)$	$c_3(t)=c_2(t)$

Table 13: ADT-related costs.

Table 14 describes the polynomial cost that is being given to patients approaching the length-of-stay soft deadline. We use the term soft deadline to refer to a deadline that does not represent a hard constraint, rather, a soft one, which causes significant increase in cost if not met. The function further takes into account the differences in the length-of-stay deadlines for different expected ADT status.

If a patient needs to be discharged and is waiting more than 3.5 hours (210 minutes)	If a patient needs to be admitted or ADT status is unknown and is waiting more than 5 hours (300 minutes)
$c(t)= c_3(t)+(t-210)^2$	$c(t)= c_3(t)+(t-300)^2$

Table 14: Polynomial waiting time costs, taking Expected ADT status into account.

Note that the value $c_i(t)$, for type i cost function, indicates the rate at which cost increases with time. These rates are all with positive slops, thus representing the reasonable satiation in which the longer the LOS the higher the cost of an addition time unit of LOS. In particular, the cumulative cost can be calculated by taking the integral over the cost rates.

Table 15 summarizes the cost rate functions $c_{i,j,k,}(t)$ for the various patient classes using the following notion: t_i, $3 \le i \le 5$ stands for triage score 3 to 5, respectively; a_j, $1 \le j \le 4$, stands for the 4 age groups, and d_k, $1 \le k \le 2$, stands for the 2 expected ADT groups.

Patient class	**Rate function before the soft deadline**	**Rate function after the soft deadline**
t_3, a_1, d_1	c(t)= 8t	c(t)= 8t+(t-210)^2
t_4, a_1, d_1	c(t)= 4t	c(t)= 4t+(t-210)^2
t_5, a_1, d_1	c(t)= 2t	c(t)= 2t+(t-210)^2
t_3, a_2, d_1	c(t)= 16t	c(t)= 16t+(t-210)^2
t_4, a_2, d_1	c(t)= 8t	c(t)= 8t+(t-210)^2
t_5, a_2, d_1	c(t)= 4t	c(t)= 4t+(t-210)^2
t_3, a_3, d_1	c(t)= 24t	c(t)= 24t+(t-210)^2
t_4, a_3, d_1	c(t)= 12t	c(t)= 12t+(t-210)^2
t_5, a_3, d_1	c(t)= 6t	c(t)= 6t+(t-210)^2
t_3, a_4, d_1	c(t)= 40t	c(t)= 40t+(t-210)^2
t_4, a_4, d_1	c(t)= 20t	c(t)= 20t+(t-210)^2
t_5, a_4, d_1	c(t)= 10t	c(t)= 10t+(t-210)^2
t_3, a_1, d_2	c(t)= 4t	c(t)= 4t+(t-300)^2
t_4, a_1, d_2	c(t)= 2t	c(t)= 2t+(t-300)^2
t_5, a_1, d_2	c(t)= t	c(t)= t+(t-300)^2
t_3, a_2, d_2	c(t)= 8t	c(t)= 8t+(t-300)^2
t_4, a_2, d_2	c(t)= 4t	c(t)= 4t+(t-300)^2
t_5, a_2, d_2	c(t)= 2t	c(t)= 2t+(t-300)^2
t_3, a_3, d_2	c(t)= 12t	c(t)= 12t+(t-300)^2
t_4, a_3, d_2	c(t)= 6t	c(t)= 6t+(t-300)^2
t_5, a_3, d_2	c(t)= 3t	c(t)= 3t+(t-300)^2
t_3, a_4, d_2	c(t)= 20t	c(t)= 20t+(t-300)^2
t_4, a_4, d_2	c(t)= 10t	c(t)= 10t+(t-300)^2
t_5, a_4, d_2	c(t)= 5t	c(t)= 5t+(t-300)^2

Table 15: Cost-rate functions for the various patient classes.

For clarification, consider the following examples:

- The cost rate for a 78-year-old patient, with a triage score of 3, who waits for four and a half hours (270 minutes) and is expected to be admitted to one of the hospital wards is given by:
  $c_{3,4,2}(270)=20*270 = 5400$
- The cost rate for a 40-year-old patient, with a triage score 5, who waits for three and a half hours (240 minutes) and is expected to be discharged home is given by:
  $C_{5,1,1}(240)= 2*240+(240-210)^2 = 9480$

The graphs below describe the waiting cost rates for the various patient classes using the same notation as above.

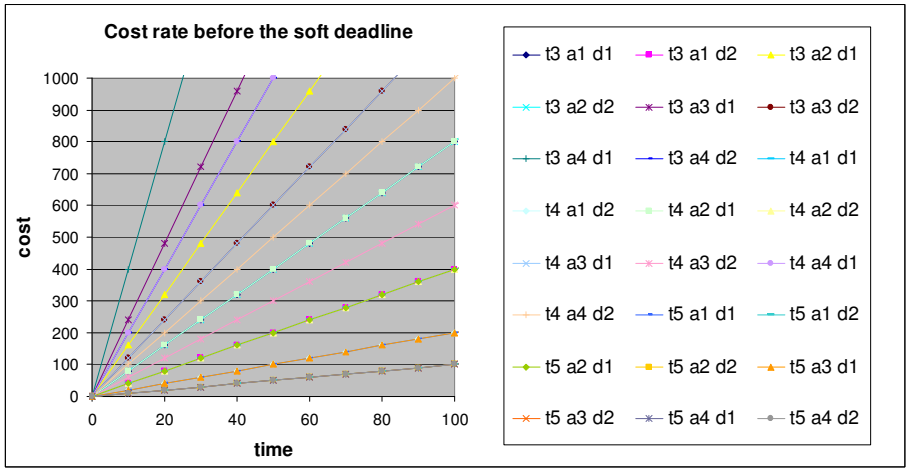


Figure 23: Cost for admitted patients is linear before reaching the four-hours-deadline.

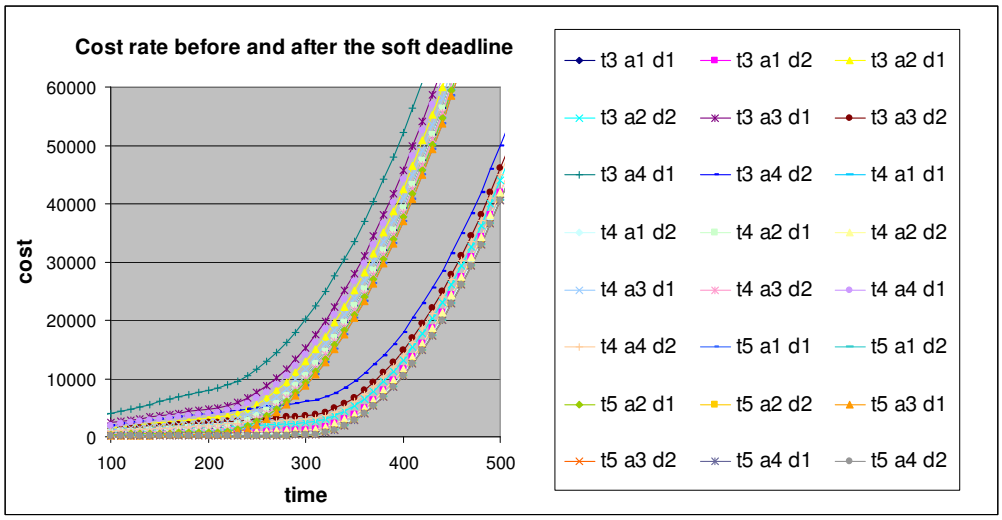


Figure 24: Cost for all patients, emphasizing admitted and discharged differences.

Note that the service policy algorithm uses configured cost functions. Thus, the ED manager is able to easily change them in order to meet desired performance indicators.

7.1.5 The PTN Simulation Model

The PTN simulation model is based on a queueing network, following the model illustrated in Figure 19. The queues are organized into a two-dimensional structure. The first dimension indicates the triage level; the second dimension indicates the patient's encounter number with a physician. The scheduling algorithm selects a patient from the appropriate queue and transfers her to the next queue after the treatment and latent time, or out of the queueing system after treatment time in the case of a patient's last encounter.

Most scheduling algorithms use FCFS within queues. Our simulation is able to simulate service policies that take individual identities into account, e.g., in situations in which cost function is complex, and to prioritize patients by calculating their waiting cost within the ED.

Service times and latent times are taken as is from the input plan. Our simulation is able to take arrival time from the input or to generate the arrival times based on known process distributions, such as the Poisson process.

The simulation simulates a physician pool with a configured number of physicians. Each physician spends time treating a patient, but physicians may also spend a configured amount of time in additional work, e.g., filling out treatment orders and discharge

summaries. This physician work profile is consistent with observational studies of physician work within the ED made by [44].

A physician becomes available to treat the next patient following the service time period and the additional work period. The "amount of additional physician work" parameter allows adjusting the overall system load without changing the arrival rates.

Note that the main focus of the PTN simulation is on the arrival patterns and the waiting queues. Thus, accurately simulating the physician work profile is not part of the PTN simulator.

7.1.6 The Simulation Output

The simulation's raw output contains the input encounter record with an additional element, specifically the encounter start time. This element is the output of the scheduling algorithm. The next arrival time to one of the waiting queues is then calculated from the encounter's service time and latent time. The rest of the KPI's are calculated after the simulation, using Excel.

7.2 Results

The PTN simulator allows simulating and analyzing various service policies under various arrival rates and staffing levels. In the following sections, we describe the main results for some of the analyzed policies. We use the FCFS service policy as the baseline for composition of our results and further compare the tested algorithms along two key performance indicators:

Time Till First Encounter (TTFE) is the average time, taken over all patients, from the time patients arrive at the ED until the time they first meet a physician. The TTFE is calculated with respect to the required deadline associated with each triage score. The simulation does not take into account additional processes that take place between patient arrival to the ED and patient arrival to the physician's waiting queue, i.e., the ED admission process, the triage process, and other processes carried out by nurses prior to the patient-physician encounter.

Total Length-of-Stay (LOS) is the time period from the arrival time of patients to the ED till the end of their last encounter with a physician. The total length-of-stay does not take into account processes that take place after the last encounter with a physician, i.e., the discharge and admitting processes and any other processes carried out by nurses after the physician's last decision has been made.

7.2.1 Comparing PTN Policies

We simulate three priority-based service policies for the PTN question, specifically i) first come first serve (FCFS), ii) NA patients first (NAF), and iii) a threshold policy (TSLD). We run the simulation many times and over various time periods, e.g., multiple days, with multiple arrival-rate patterns. We specifically emphasize a real-life scenario that assumes varied arrival rate during the day. We test this scenario under several ED scales. Our summary results for a realistic ED setting are provided in Table 16. All numbers are provided in minutes and represent average measurements.

	Time till first encounter	**Service time**	**Overall waiting time**	**Latent time**	**Total length–of-stay**
FCFS	31	14	103	59	176
NAF	8	14	116	59	189
TSLD	29	14	105	59	178

Table 16: PTN service policies comparison.

Note the following:

- FCFS is the straightforward PTN policy currently used by most EDs. FCFS in this regard refers to a specific encounter. As a result, FCFS does not give priority to returning patients. Thus, these patients need to wait in line for their turn with all other patients, including NA patients and other IP patients. Moreover, FCFS does not take triage scores, or any other patient clinical characteristics into account; thus it is obviously not good enough for addressing combined clinical and operational aspects. It is considered here just as a baseline, allowing us to qualify the results of other PTN policies.
- The NA patient first policy (NAF) gives the highest priority to newly-arrived patients. This policy can be considered as the simplest threshold policy. It applies a simple binary threshold—whether patients are in one of the triage queues or not. Results from a complementary IP patient first policy (IPF) are not provided. IPF is expected to reduce the total LOS, at the expense of a longer TTFE. Thus, is not expected to perform better then the FCFS in meeting the triage deadlines.

- Service times and latent times are not affected by the service policy, thus have the same values along the various policies. Furthermore, the service times only include patient-physician encounters, while the latent times includes the time a patient spent at stations other than the physician's station.

While the FCFS and the TSLD policy result in fairly similar performance indicators, the NAF policy yields a significantly different result. From assessing these results, we gain several interesting insights:

- The PTN decision policy has a significant effect on performance indicators.
- No single PTN service policy stands out as superior. While the FCFS provides a minimal LOS, the NAF policy provides much better TTFE.
- Giving higher priority to newly-arrived patients significantly reduces the TTFE indicator, but increases the overall LOS at the ED. This phenomenon justifies the use of triage deadlines for balancing clinical and operational needs.
- Comparing the FCFS and NAF policies indicates that the gcμ rule holds for the simulated environments. Giving precedence to newly-arrived patients increases the total LOS. The need to optimize the TTFE indicator, which is a distinctive ED performance indicator, calls for further analysis.
- Comparing the FCFS and the TSLD policies reveals no significant performance differences. The FCFS seems to be superior in the total LOS aspect, at the cost of a bit longer TTFE. A more detailed comparison of the LOS and TTFE distribution suggests that this is not the case. Such an analysis reveals that the TSLD policy provides much greater control and allows ED management to adjust the ED operation towards specific performance indicators. We further describe these insights in the following sections.

7.2.2 Meeting Predefined Deadlines

A reasonable service policy aims at reducing the averages of the TTFE and LOS indicators. Clinical needs suggest that aiming to reduce the average TTFE and LOS is not good enough, and that taking the exact distribution into account is also necessary. Specifically, clinical needs impose deadlines for both TTFE and LOS. Consequently, the optimal distribution associated with these indicators needs not to be uniform nor even symmetric. A longer TTFE is acceptable, assuming triage deadlines are not violated for any patient. Having many patients' go over the soft deadline (e.g., 4 hours) in the ED is considered not good enough, even though the average LOS is kept below the soft deadline. Table 17 shows the percentage of the patients that meet TTFE and LOS deadlines for the three above-mentioned PTN service policies.

	Percentage of patients that meet the TTFE deadline	Percentage of patients that meet the 4-hour LOS deadline
FCFS	88%	75%
NAF	100%	74%
TSLD	94%	78%

Table 17: Meeeting deadline constraints.

Table 17 reveals the following observations:

- The TSLD algorithm is superior to FCFS in both categories.
- The NAF policy allows meeting triage deadlines at all times, but comes at the expense of a longer LOS. A natural question thus emerges—what is the minimal effort required for meeting triage deadlines? Allocating the minimal effort for meeting triage deadlines allows reduction of the total LOS. An optimal TSLD policy should do exactly that. In the following sections, we further analyze the various parameters that affect the various TSLD policies, while seeking the optimal one.

In Table 18, we describe results from analyzing the two KPIs across triage categories. Such an analysis provides additional insights into the interplay and tradeoffs between the two KPIs. FCFS does not take triage categories into account; thus the LOS performance does not depend on the triage categories. Both NAF and TSLD take triage deadlines into account. As a result, triage 5 patients spend more time waiting for their first encounter, and eventually spend more time at the ED and are more likely to miss their total LOS deadlines.

	Fraction of patients that meet the TTFE deadline			Fraction of patients that meet the 4-hour LOS soft deadline		
Triage Class	3	4	5	3	4	5
FCFS	75%	82%	95%	75%	81%	72%
NAF	100%	100%	100%	92%	83%	66%
TSLD	88%	95%	95%	86%	80%	75%

Table 18: Meeting deadline constraints along triage categories.

7.2.3 Comparing the Threshold Algorithms

We further examine several variants of the threshold algorithm under several arrival processes. Specifically, we compare the adaptive threshold (ATD) algorithm with the dynamic threshold (DTD) algorithm described in Table 2. We compare these two algorithms with respect to two related parameters, the system capacity and the arrival rate. These parameters are summarized in Table 19.

	Fixed arrival rate	Varied (realistic) arrival rate
Average arrival rate that results in below heavy traffic conditions	Not interesting, thus not provided	Table 21, same performance No real issue as all patients are served before their deadline
Average arrival rate that results in heavy traffic condition	Table 20, same performance	Table 22, DTD performs better

Table 19: Summary table for the various STD and DTD tested scenarios

Fixed arrival rates are generated using a Poisson process with a fixed rate denoted by λ. Heavy traffic conditions are generated by letting λ/μ go to 1. This is done by adjusting the number of physicians and their work loads according to the provided λ.

We analyze the system under various scales. We scale up the system by increasing the number of physicians and arriving patients simultaneously, while maintaining the desired load. For a large-scale system, we use about 22 physicians, specifically, 21 physicians result in an overloaded condition, 22 in a critically loaded system, and 23 in an under-loaded one. For a more realistic system, we use five physicians. Our analysis shows that no significant differences exist between the two system scales.

For clarity reasons, we provide analysis results for just a single triage category in most of the cases.

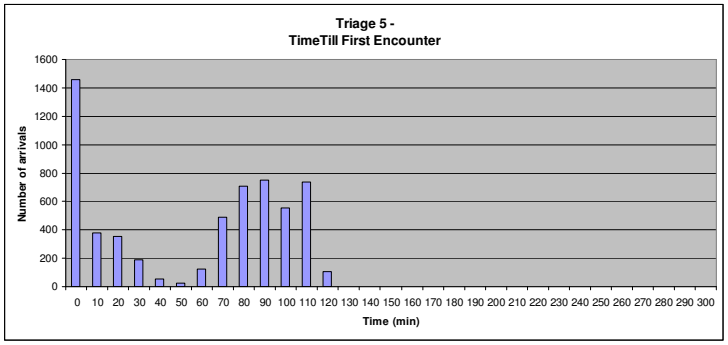
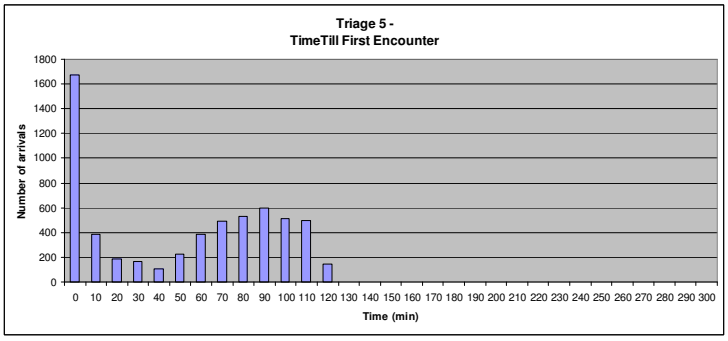
	Poisson arrival rate with fixed λ that averages in heavy traffic conditions; triage 5 patients.
Dynamic Threshold (DTD)	
Adaptive Threshold (ATD)	
Main observation	The algorithms perform similarly.

Table 20: Fixed arrival rate; heavy traffic.

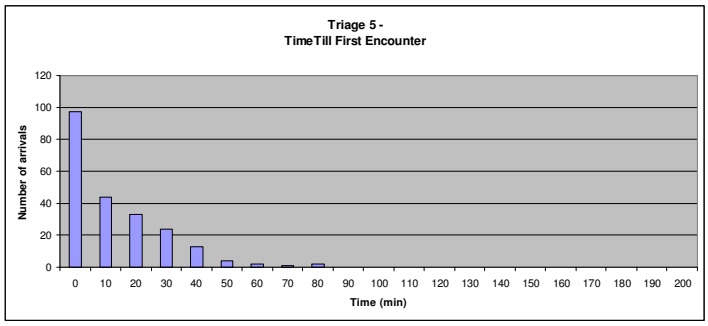
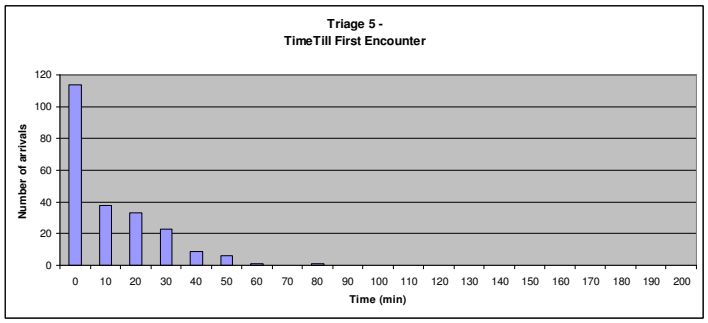
	Varied arrival rate with average below heavy-traffic conditions; medium size system.
Dynamic Threshold (DTD)	
Static Threshold (STD)	
Main observation	The algorithms perform similarly. No deadline issues occur.

Table 21: Varied arrival rate with average below heavy-traffic conditions.

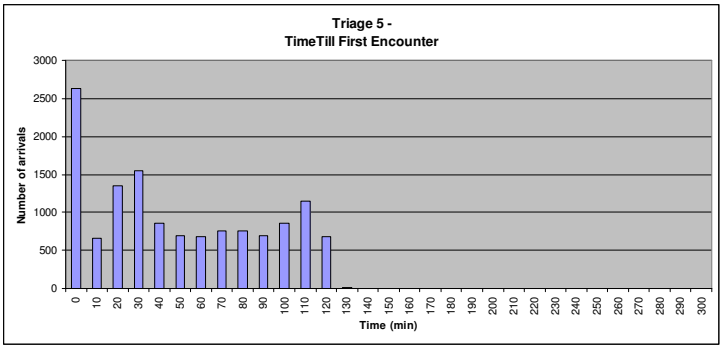
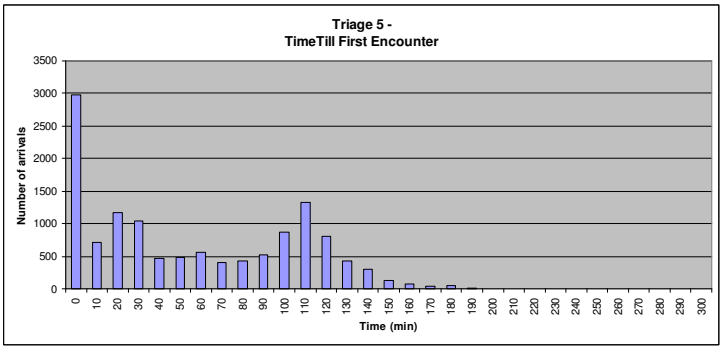
	Varied arrival rate with average at heavy-traffic conditions.
Dynamic Threshold (DTD)	
Adaptive Threshold (ATD)	
Main observation	DTD performs better due to its better resilience to the time-varying arrival rates.

Table 22: Varied arrival rate with average at heavy traffic.

7.2.4 Assessing Different Triage Categories

We further compare the ATD and DTD algorithms across the various triage categories. Recall that triage categories differ from one another in their times-to-deadline. The comparison is done for small and large system sizes and for fixed and realistic arrival rates that result in heavy-traffic conditions. We note that both algorithms perform better for a large-scale system than for a small one. The DTD performs better, or at least as well as the ATD, under all tested characteristics.

	DTD	**ATD**
Large system; fixed arrival rate; heavy-traffic conditions	Table 24	Table 25
Small system; realistic arrival rate averages in heavy-traffic conditions	Table 26	Table 27

Table 23: Summary of comparison tables.

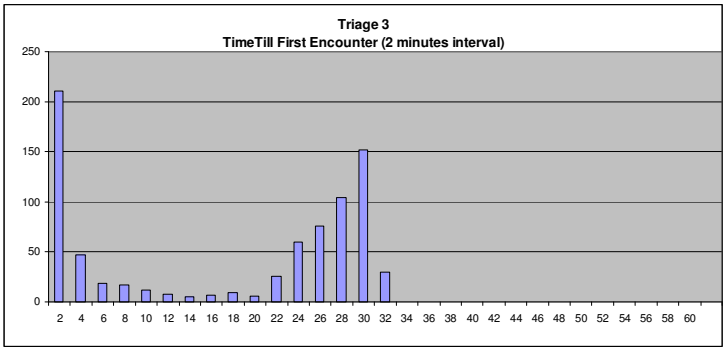
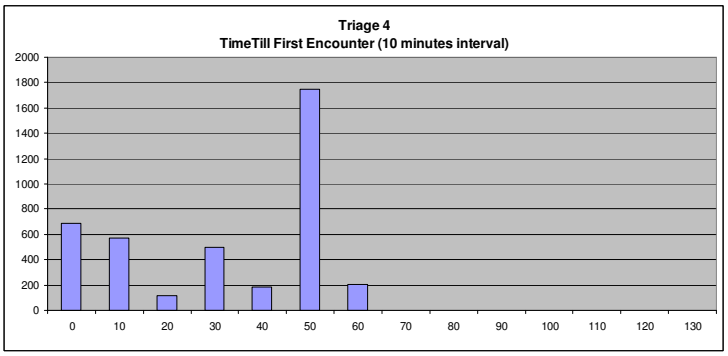
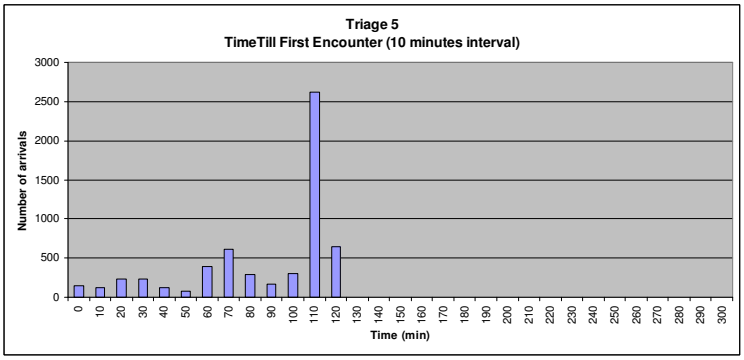
DTD	Large system; fixed arrival rates; heavy traffic conditions.
Triage 3 (30-minute deadline)	
Triage 4 (60-minute deadline)	
Triage 5 (120-minute deadline)	

Table 24: DTD, large system; fixed arrival rates; heavy traffic conditions.

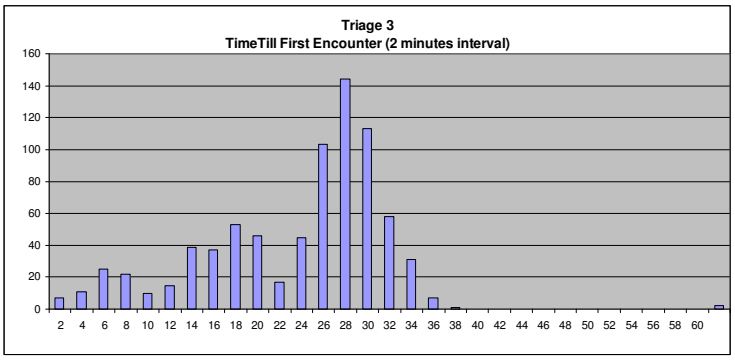
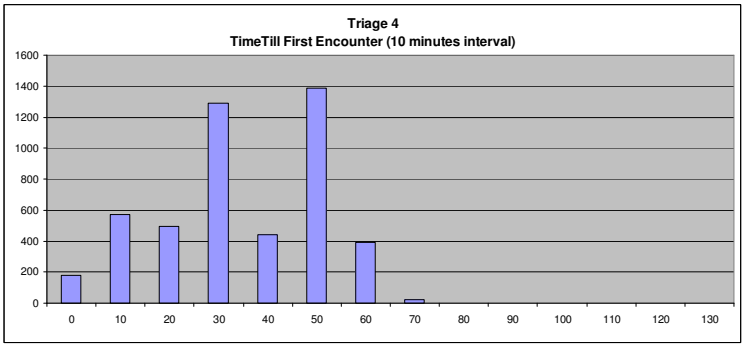
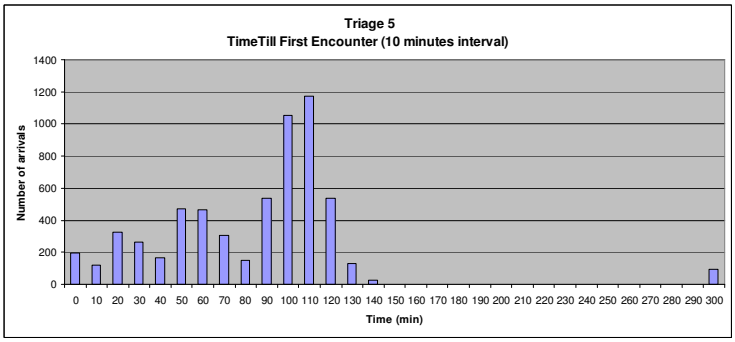
ATD	Large system; fixed arrival rate; heavy traffic conditions.
Triage 3 (30-minute deadline)	
Triage 4 (60-minute deadline)	
Triage 5 (120-minute deadline)	

Table 25: ATD; large system; fixed arrival rates; heavy traffic conditions.

Further analysis of the system under more realistic conditions, namely few physicians and a realistic arrival situation, revealed the difficulty of both the ATD and DTD algorithms to meet triage deadline at all times. However, the DTD performed significantly better under these conditions, due to its dynamic adaptation to the time-

varying arrival rate. Note that a significant part of triage 3 patients still miss their deadlines by up to 30 minutes. Similarly, about 10% of triage 4 and 5 patients also miss their deadlines by up to 30 minutes, manifesting the stochastic level of the system.

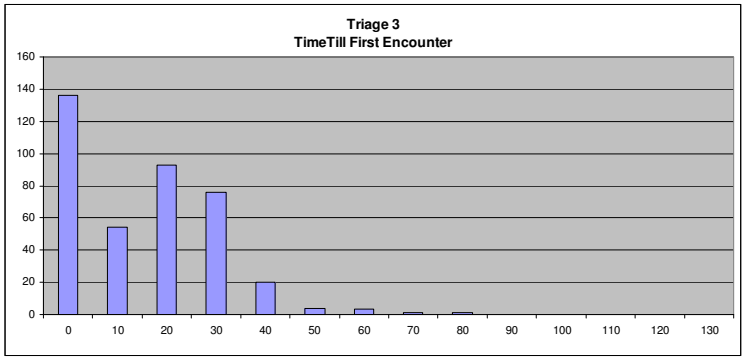
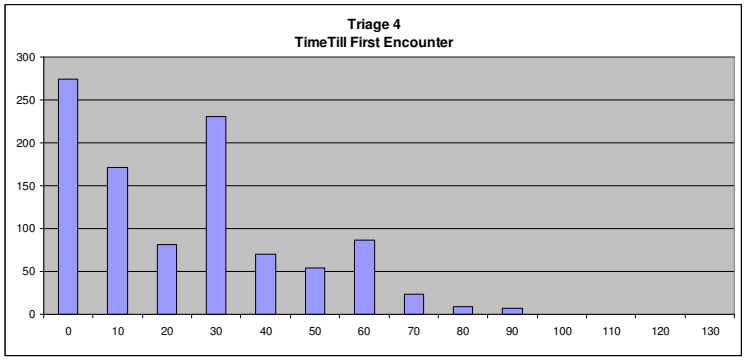
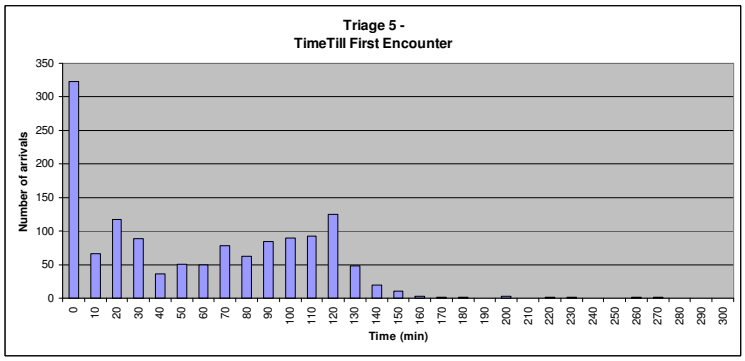
DTD	**Small system (1 physician); realistic arrival rate that averages in heavy traffic conditions.**
Triage 3 (30-minutes deadline)	
Triage 4 (60-minutes deadline)	
Triage 5 (120-minutes deadline)	

Table 26: DTD; small system; realistic arrival rate that averages in heavy-traffic conditions.

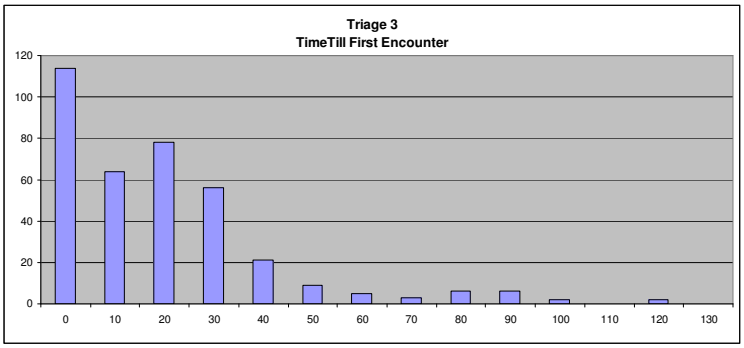
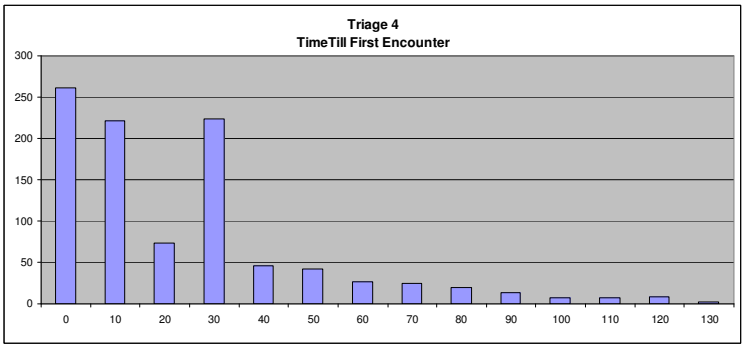
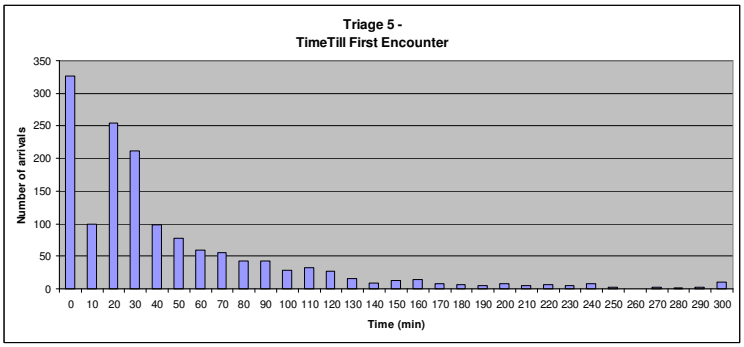
ATD	Small system (1 physician); realistic arrival rate that averages in heavy traffic conditions.
Triage 3 (30 minutes deadline)	
Triage 4 (60 minutes deadline)	
Triage 5 (120 minutes deadline)	

Table 27: ATD; small system; realistic arrival rate that averages in heavy-traffic conditions.

7.2.5 In Process Patient Priorities

The distribution of the Length of Stay (LOS) is highly dependent on the system load. No significant differences exist between the ADT and DTD policies, since both use the same cost function. Note that if a system is overloaded, then most resources are allocated to NA patients and LOS will grow infinitely. We provide results for finite horizon, in which the system eventually remains without patients, in the following tables.

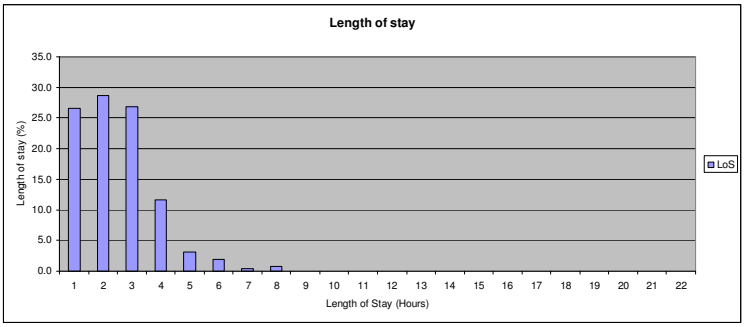
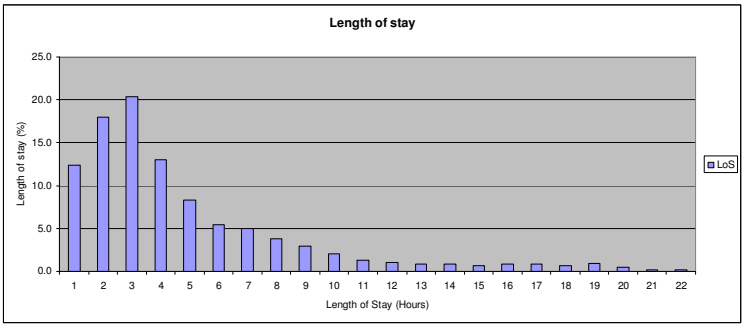
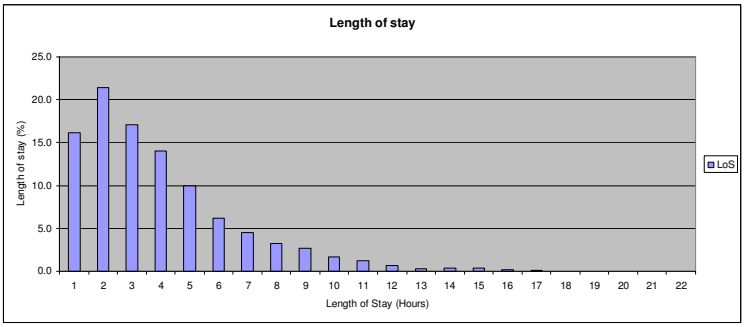
Length of Stay	Length of stay distribution for a system under various loads
Critically loaded system with fixed arrival rate	
Overloaded system with fixed arrival rate (finite horizon)	
Critically loaded system with realistic arrival rate	

Table 28: LOS distribution for a system under various loads.

Recall that our PTN model gives priority to IP patients based on a cost function. The PTN cost function takes four parameters into account: i) length-of-stay, ii) triage score, iii) patient age group, and iv) admit/discharge forecast. These parameters can be easily changed. All PTN algorithms use the same cost functions.

Figure 25 summarizes the performance indicators along the age category for the DTD algorithm, running over a realistic arrival rate that averages in heavy-traffic conditions. The cost function indeed gives precedence to old patients over younger ones, as shown in the figure.

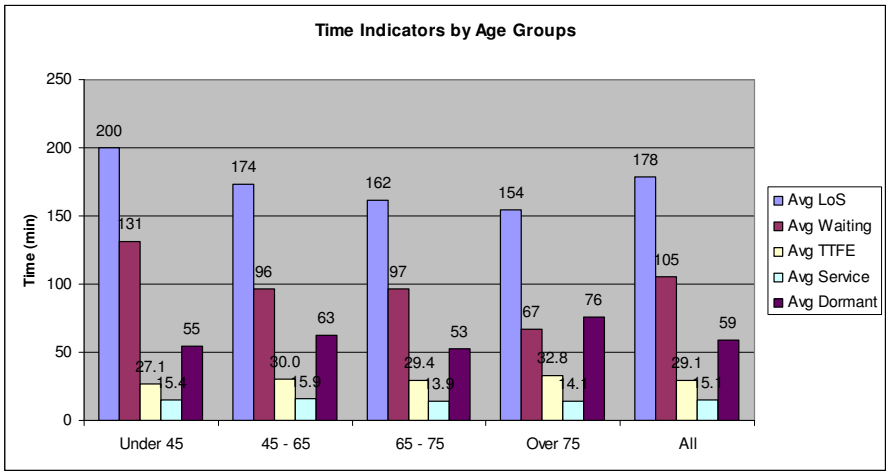


Figure 25: LOS distribution across age groups.

Figure 26 summarizes the performance indicators along the ADT category for the DTD algorithm. The results indicate the priority that the cost function gives to expected discharged patients.

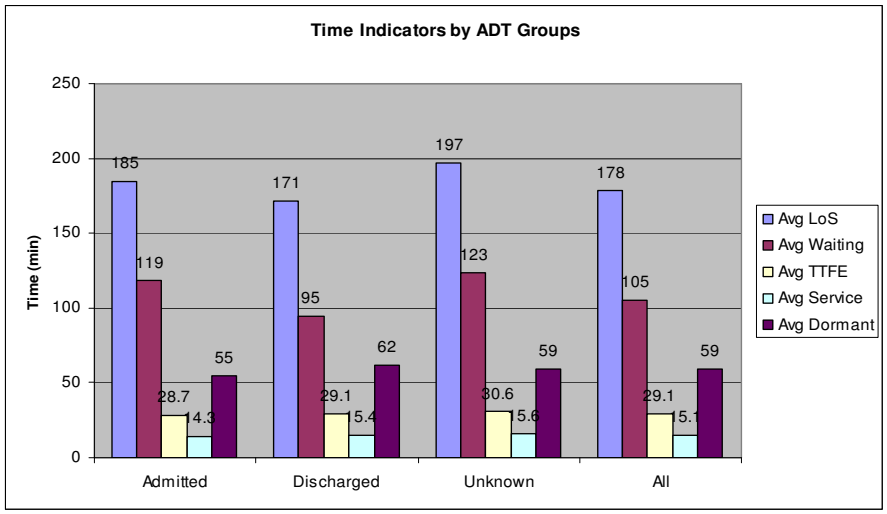


Figure 26: LOS distribution across the expected ADT statuses.

7.3 Summary of Results

Through simulation, we compared several service policies under various system conditions with respect to time-till-first-encounter and the length-of-stay KPIs. We found that the dynamic threshold (DTD) policy performs better under most system conditions. Specifically, this policy is robust against time-varying arrival rates. The ADT performs similarly to the DTD, except under time-varying arrival rate conditions. The FCFS, though resulting in the minimum lengths-of-stay, does not provide any control over triage deadlines and other service-level indicators such as age. The NAF policy best meets triage deadline constraints, but results in a below-optimal average length of stay. In the following chapter, we use a fluid model to analyze a variant of the DTD policy. For a stylized model, we show that an optimal policy exists in situations in which the arrival rate is assumed known.

CHAPTER 8: WHICH PATIENT SHOULD BE TREATED NEXT? FLUID MODEL ANALYSIS

In this chapter, we present a fluid model analysis that addresses the PTN question. In our analysis, we focus on the first decision within the PTN question, namely whether to next serve an NA patient or an IP patient. The underlying assumption is that to achieve an optimal solution to the PTN question, we must use a threshold-approach in which NA-patients are served as close as possible to their triage deadline, but not cross it. A key requirement for this analysis is to allow an arbitrary arrival rate. Thus, the goal of our fluid model analysis is to find an optimal control that, given the general arrival rate, dynamically determines whether to next treat the NA patient versus the IP patient.

8.1 Problem Definition

Our fluid model analysis is performed for a simple stylized model as depicted in Figure 27. As shown in the figure, we focus on a single NA queue and a single IP queue. The main assumption, which follows from our simulation-based analysis, is that an optimal control belongs to the family of minimal-effort due-date policies.

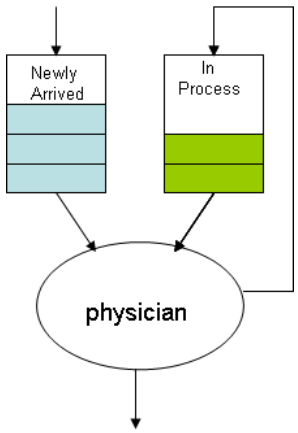


Figure 27: The stylized fluid model for the PTN question.

Thus, our goal is to find an optimal departure control. A control is optimal if it minimizes the departure process uniformly over all times subject to deadline constraints. Thus, we seek the minimal-effort control that meets deadline constraints over all times and for all possible arrival rates. For that we define physically-feasible and deadline-feasible controls. Obviously, arrival rates exist for which deadline constraints cannot be

met. Thus, we provide a necessary and sufficient deadline-feasibility condition (DFC) over an arrival rate that guarantees that deadline constraints can be met. We then provide a deadline-feasible control, namely a physically-feasible control that meets deadline constraints under the deadline-feasibility condition. We conclude by showing that our deadline-feasible control is indeed optimal.

At the end of the chapter, we provide a conjecture for the optimal control in situations in which deadline constraints cannot be met due to high arrival rates. The generalization of a single NA, single IP-queuing model into the PTN model described in Figure 19 is left for future research.

8.1.1 Mathematical Framework

We start by defining the mathematical framework for our analysis, including the following symbols and parameters:

We use $\alpha(\bullet)$ to denote the time-varying arrival rate to the system; thus $\alpha(\bullet) \geq 0$. It is technically convenient to let $\alpha(\bullet)$ be defined over the whole real time $(-\infty, \infty)$ and to assign to it the value of 0 over $(-\infty, 0)$.

For the technique of our proof, we shall impose the following additional constraints over α:

1. α should be piecewise continuous;
2. α should have finite number of local extrema.

We use A to denote the cumulative number of arrivals, or work arriving to the system. Thus A(t) denotes the amount of work arriving to the system till time t:

$$A(t) = \int_{-\infty}^{t} \alpha(u)du$$

We use $\delta(\bullet)$ to denote a time-varying departure rate from the system. Thus $\delta(\bullet) \geq 0$ and $\delta(\bullet)$ is piecewise continuous. A control policy amounts to specifying δ; we shall thus use the two interchangeably. It is technically convenient to let $\delta(\bullet)$ be defined over the whole real time $(-\infty, \infty)$ and to assign to it the value of 0 over $(-\infty, 0)$.

We use D to denote the cumulative work departing from the system; thus D(t) is the integral over $\delta(t)$.

$$D_{\delta}(t) = \int_{-\infty}^{t} \delta(u)du$$

We use μ to denote departure capacity, which is the maximal departure rate from the system; thus μ serves as upper bound for δ: $0 \leq \delta(\bullet) \leq \mu$.

We use d to denote the deadline; d is an input constant to the model, which is measured in time units.

We use the following superscript throughout: t^z, t^s, t^d, t^e, which stands for **Z**ero, **S**tart, **D**eadline, and **E**nd respectively. These subscripts are used to denote specific points over given time intervals.

We use $\delta^*(\bullet)$ to denote the optimal control policy, which is the sought-after solution for our problem.

Using the above framework, we can now provide a solution to the PTN optimization problem.

8.1.2 Physically-feasible Control

A physically-feasible control $\delta(\bullet)$ is a control that meets the following constraints:

1. $\delta(t) = 0, \ t<0,$
2. $0 \leq \delta(t) \leq \mu, 0 \leq t$
3. $D_\delta(t) \leq A(t), -\infty<t< \infty.$

The above constraints ensure that the control is physically viable. The first two constraints ensure that the departure rate is assigned a positive number that cannot exceed the physical maximum departure rate of the system. The third constraint ensures that work cannot depart from the system if not yet arrived into it.

8.1.3 Deadline-feasible Control

A deadline-feasible control $\delta(\bullet)$ is a physically-feasible control that further meets the following constraint:

4. $A(t-d) \leq D_\delta(t), -\infty<t<\infty.$

The above constraint ensures that work that arrived to the system at time t will depart no later than at time t+d.

We will refer to these four constraints (1-4) by their numbers, during proofs and discussions in subsequent sections.

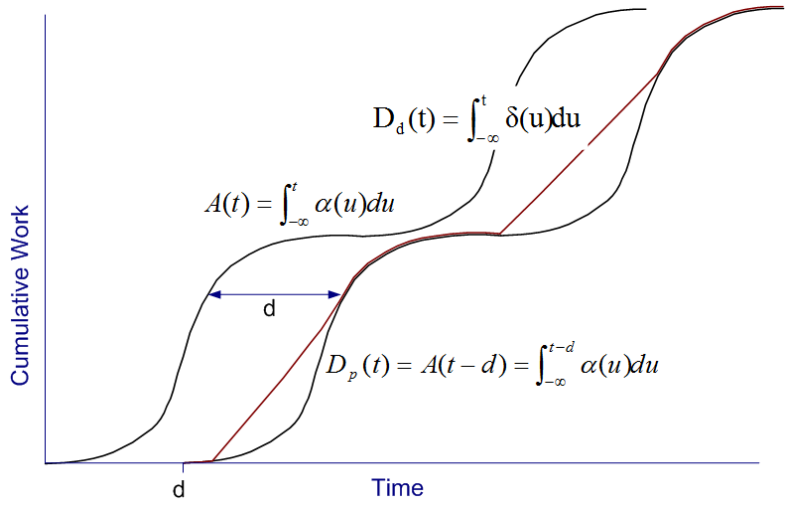


Figure 28: Cumulative arrival and departure work.

Figure 28 describes a cumulative arrival work function, A(t), and a cumulative deadline-feasible departure work function, $D_d(t)$; $D_p(t)=A(t-d)$ is the ultimate departure work function in which all work departs from the system exactly at its deadline. This $D_p(t)$ is usually not achievable due to the $\delta(\bullet) \leq \mu$ constraint.

8.1.4 An Optimal Control

Let Δ be the set of deadline-feasible controls. Our optimal control problem seeks to identify, among all $\delta \in \Delta$, the one δ^* that minimizes $D_\delta(t)$-A(t-d), simultaneously over all times t. Note that it is a priory unclear that there is indeed such $\delta^* \in \Delta$. Nevertheless, we do show that, if Δ is not empty, δ^* exist. Denote by D^* the departing work corresponding to δ^*. We shall also show that $D^* \leq D_\delta$ for all $\delta \in \Delta$, thus establishing the optimality of δ^* in the sense of minimizing efforts subject to triage constraints. Note that a physically-feasible control, for which $D_\delta(t)=A(t-d)$ at all times, is obviously optimal.

8.1.5 A Necessary and Sufficient Deadline-Feasibility Condition

A necessary and sufficient deadline feasibility condition (DFC) which ensures the existence of deadline-feasible departure policy, is as follows:

For any two time-points t_i and t_j such the $t_i<t_j$, the following must hold:

Deadline feasible conditions (DFC):

(Eq. 1) $$A(t_j) - A(t_i) = \int_{t_i}^{t_j} \alpha(t)dt \le (t_j - t_i)\mu + \mu d$$

Although checking for the DFC for a given α might be cumbersome, some observations can still be made:

1. If $\alpha(t) \le \mu$, for $-\infty<t<\infty$ then α is deadline feasible. Otherwise:
2. Identify the set of time intervals (t^s_i, t^e_i), $t^s_i \le t < t^e_i$ for which $\alpha(t) \ge \mu$. If there exists an interval (t^s_i, t^e_i) such that $\int_{t^s}^{t^e} (\alpha(t) - \mu)dt > \mu d$, then α is not deadline-feasible.

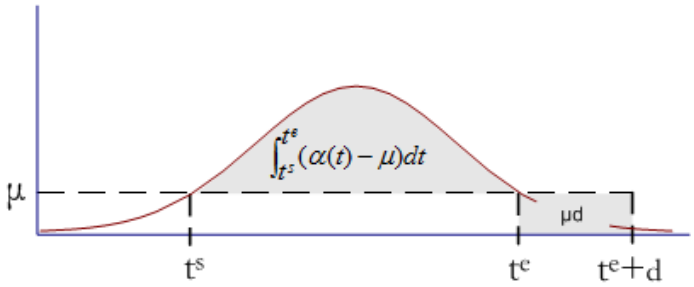


Figure 29: A not deadline-feasible α(t)

8.1.6 Stability Constraint

A stability constraint over α is required for constructing the departure policy:

There exist $T<\infty$ such that $\alpha(t) \le \mu$ for all $t>T$.

Thus, for constructing α, we need to identify the maximal time T for which $\alpha(t)>\mu$. Analyzing the system under finite horizon $(-\infty, T]$ poses no problem in that regard, as $\alpha(\bullet)$ is only defined over the half-closed interval $(-\infty, T]$ and $\alpha(t)=0$ for $t<0$. Thus the maximal time t, for which $\alpha(t)>\mu$, obviously exists within the time interval $[0, T]$.

Analysis of the system under an infinite horizon requires identifying a time T for which $\alpha(t) \le \mu$ for all $t>T$.

8.2 Proving the Deadline-Feasibility Condition

We start by proving the necessity and sufficiency of the DFC provided at (Eq. 1):

8.2.1 Proof-Necessity

For necessity we must show that if the condition does not hold, the deadline cannot be met. Thus, assume there are two time-points t_i and t_j for which the necessary condition does not hold. Namely:

$$A(t_j) - A(t_i) = \int_{t_i}^{t_j} \alpha(t)dt > (t_j - t_i)\mu + \mu d = (t_j + d - t_i)\mu.$$

The amount of work arrived at time t_j is given by:

$$A(t_j) = \int_{-\infty}^{t_j} \alpha(t)dt = \int_{-\infty}^{t_i} \alpha(t)dt + \int_{t_i}^{t_j} \alpha(t)dt.$$

The amount of work that departed from the system at time t_j+d is, at most, all the work that arrived at time t_i (i.e., fully depleting the system, also noted as $D(t_i)=A(t_i)$, which indicates the physically-feasible constraint), plus the maximal departure rate, μ, over the time period $[t_i, t_j+d]$:

$$D(t_j + d) = \int_{-\infty}^{t_j+d} \delta(t)dt = \int_{-\infty}^{t_i} \alpha(t)dt + \int_{t_i}^{t_j+d} \mu dt = \int_{-\infty}^{t_i} \alpha(t)dt + (t_j + d - t_i)\mu <$$

$$\int_{-\infty}^{t_i} \alpha(t)dt + \int_{t_i}^{t_j} \alpha(t)dt = A(t_j).$$

From the above equation, we conclude that not all the work that arrived at time t_j departed at time t_j+d. In other words, a part of the work arrived to the system at time t_j and missed its deadline.

8.2.2 Proof-Sufficient

For sufficiency, we need to prove that if the condition holds, the deadline can be met over all times. We prove it by constructing a deadline-feasible departure policy. The construction's steps are provided in the subsequent sections.

8.3 The Trivial Arrival Rate Case

A trivial arrival rate case is: $\alpha(t) \leq \mu$, $-\infty<t<\infty$.

In this case, meeting the deadline constraints is possible by simply delaying all work for exactly d units of time. Such an approach minimizes the departure process uniformly over all times. Formally:

$\delta^*(t) = \alpha(t\text{-}d)$, $-\infty \leq t \leq \infty$.

is an optimal solution for the problem.

8.3.1 Meeting Physically-Feasible Constraints

The δ* meets all physically-feasible constraints:

1. $\delta(t)=0$, $t\leq 0$

$\alpha(t) = 0$, $t\leq 0$. Thus $\delta^*(t)=\alpha(t-d)=0$, $t\leq d$.

2. $0\leq\delta(t)\leq\mu$, $0\leq t$.

$0\leq\alpha(t) \leq \mu$ at all time, so that $\alpha(t-d)\leq\mu$ at all times. Thus $0\leq\delta^*(t)= \alpha(t-d)\leq\mu$

3. $D_\delta(t) \leq A(t)$, $-\infty\leq t\leq \infty$.

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = \int_{-\infty}^{t} \alpha(u-d)du = A(t-d) \leq A(t),\ -\infty<t<\infty.$$

Thus $D(t)\leq A(t)$, $-\infty<t<\infty$.

8.3.2 Meeting Deadline-Feasible Constraint

The δ*(•) meets the deadline-feasible constraint:

4. $A(t-d)\leq D_\delta(t)$, $-\infty\leq t\leq \infty$.

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = \int_{-\infty}^{t} \alpha(t-d)du = A(t-d),\ -\infty<t<\infty.$$

Thus $A(t-d)\leq D(t)$, $-\infty<t<\infty$.

8.3.3 Proving Optimality

The δ* and the corresponding D* are obviously optimal since

$D^*(t) = A(t-d)$, $-\infty<t<\infty$

and thus

$D^*(t)-A(t-d) = 0$, $-\infty<t<\infty$,

while $D(t)-A(t-d) \geq 0$, $-\infty<t<\infty$, for all other $\delta \in \Delta$.

8.4 Variable Arrival Rates under the Deadline Feasibility Condition

As discussed earlier, it is convenient to distinguish between two arrival-rate situations, specifically those for which a deadline-feasible control exists and those for which is none exists. In the following sections, we discuss the first situation, that is, we assume that the DFC prevails at all times. We then construct a deadline feasible control δ, and prove that it is optimal.

8.4.1 Constructing an Optimal Control

Assume that the DFC and the Stability constraints hold at all times.

Introduce a finite set of m+1 time intervals $[t^s_i, t^e_i]$, $0 \le i \le m$, such that:

$\alpha(t) \ge \mu$, $t^s_i < t < t^e_i$, $0 \le i \le m$,

$\alpha(t) < \mu$ otherwise.

Note that $m < \infty$ since α has a finite set of local extrema.

Denote $[a_i]$, $0 \le i \le n$, a vector of index alignment constants, such that $a_n = m$ and $a_i < a_{i+1}$.

We now construct a sequence of n+1 time intervals $[\tau^s_i, \tau^e_i]$, $0 \le i \le n \le m$, such that:

$T = t_m = t_{a_n} = \tau^e_n > \tau^s_n > t^e_{a_{(n-1)}} = \tau^e_{n-1} > \tau^s_{n-1} \ldots > t^e_{a_1} = \tau^e_1 > \tau^s_1 > t^e_{a_0} = \tau^e_0 > \tau^s_0 > -\infty$,

for which:

1. $A(\tau^e_i) - A(t) = \int_t^{\tau^e_i} \alpha(u)du \ge \mu(\tau^e_i - \tau^s_i)$, $\tau_{ns} < t < \tau_{ne}$, $0 \le i \le n$
2. $A(\tau^e_i) - A(\tau^s_i) = \int_{\tau^s_i}^{\tau^e_i} \alpha(t)dt = \mu(\tau^e_i - \tau^s_i)$
3. There exist small $\varepsilon > 0$ such that

$$A(\tau^e_i) - A(\tau^s_i - \varepsilon) = \int_{\tau^s_i - \varepsilon}^{\tau^e_i} \alpha(t)dt < \mu(\tau^e_i - \tau^s_i - \varepsilon)$$

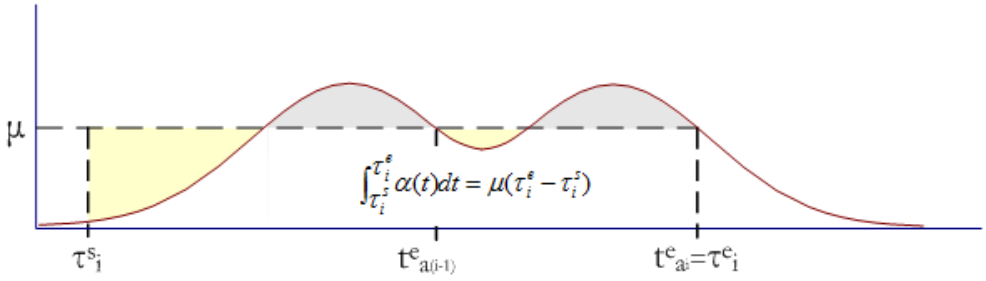


Figure 30: Finding the set of τ^s_i points

Note that:

1. The construction of the $[\tau^s_i, \tau^e_i]$ is carried out in an iterative manner starting from t=T backwards.
2. The process must eventually stop as $a_i < a_{i+1}$, $0 \le i \le n$, and the number of time intervals $[t^s_i, t^e_i]$ is finite.

3. τ^s_0 exists, and $-\infty < \tau_{0s}$ since $\alpha(t) = 0$ for $t<0$, thus $\int_{\tau^s_0}^{\tau^e_0} \alpha(t)dt$ is bounded while $\mu(\tau^e_0 - \tau^e_0)$ tends to ∞ as τ_{0s} goes to $-\infty$.

4. $t^e_{a_{(i-1)}} < \tau^s_i < t^s_{a_i}$. That is, τ^s_i falls within an interval in which $\alpha(t)<\mu$. Assume, in contradiction, there is $\varepsilon'<\varepsilon$ such that $\alpha(t)\geq\mu$, for $\varepsilon'<\varepsilon<\tau^s_i$. We then have:

$$A(\tau^e_i) - A(\tau^s_i - \varepsilon') = \int_{\tau^s_i - \varepsilon'}^{\tau^e_i} \alpha(t)dt = \int_{\tau^s_i - \varepsilon'}^{\tau^s_i} \alpha(t)dt + \mu(\tau^e_i - \tau^s_i)$$

$$\geq \mu(\tau^e_i - (\tau^s_i - \varepsilon)'$$

while contradicting the above third condition.
Thus, we conclude that $\alpha(t)<\mu$ within the time interval $t^e_{ai-1}=\tau^e_{i-1}<t<\tau^s_i$.

Denote a derived set of time intervals $\Gamma = \{\gamma^s_i, \gamma^e_i\}$ defined by:

$\gamma^s_i = \tau^s_i + d$

$\gamma^e_i = \tau^e_i + d$

Let δ^* be defined over Γ as follows:

- $\delta^*(t) = 0$, $t<0$,
- $\delta^*(t) = \mu$, $t \in \Gamma$,
- $\delta^*(t) = \alpha(t-d)$, otherwise.

See Figure 31 for a graphical demonstration of δ^* and the way we construct it.

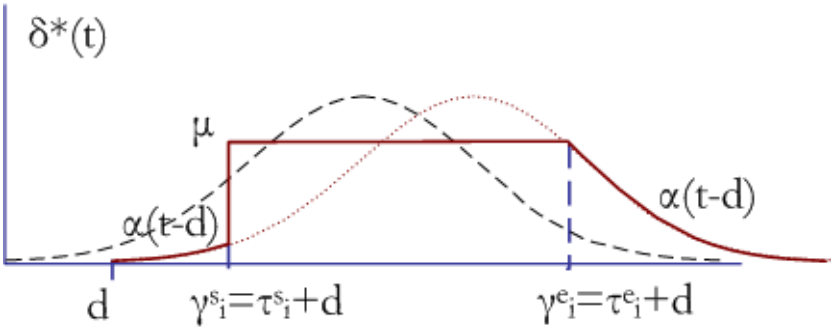


Figure 31: Constructing δ*(t)

We argue that the above δ^* is physically-feasible, deadline-feasible, and constitutes optimal control. We prove these claims in the following sections.

8.4.2 The Proactive Period

Next, we define the proactive period equality (PPE). The PPE is defined for every two time points τ^s_i, τ^e_i, $\tau^s_i<\tau^e_i$ for which the below equality holds:

PPE:

(Eq. 2) $A(\tau_i^e) - A(\tau_i^s) = \int_{\tau_i^s}^{\tau_i^e} \alpha(t)dt = \mu(\tau_i^e - \tau_i^s)$

We will use the PPE throughout our proofs.

8.4.3 Sync Point Correlations Lemma

Sync point is a pair $\{t_1, t_2\}$ in which work arrives to the system at time t_1, then departs from the system exactly d time units afterwards, at time t_2.

We define the pair $\{t^l_1, t^l_2\}$ as left sync point (LSP), and the pair $\{t^r_1, t^r_2\}$ as right sync point (RSP), respectively, if there exist small ε1, ε2 such that work arriving to the system at the interval $\varepsilon_1 < t < t^l_1$, $t^r_1 < t < \varepsilon_2$, respectively for RSP, departs from the system exactly at its deadline (i.e., d time units afterwards), while work arriving to the system at $t^l_1 < t < \varepsilon_2$, $\varepsilon_1 < t < t^r_1$, respectively for RSP, departs from the system either before or after its deadline. See Figure 32 for a graphical demonstration of LSP and RSP.

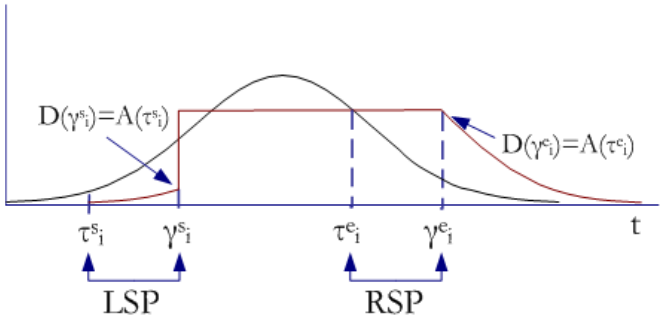


Figure 32: Left sync point and right sync point.

From the way we constructed δ*, we observe that all couples $\{\tau^s_i, \gamma^s_i\}$ are LSP; and that all couples $\{\tau^e_i, \gamma^e_i\}$ are RSP. Specifically, work that arrives to the system at time τ^s_i departs at time γ^s_i, and work that arrives to the system at time τ^e_i departs at time γ^e_i. Moreover, all work that arrives to the system after τ^s_i but before τ^e_i departs before its deadline, and all work that arrives to the system just before τ^s_i or just after τ^e_i departs from the system exactly at its deadline.

Based on this observation, we define the following correlations:

Left Sync Point (LSP) correlation for $\{\tau^s_i, \gamma^s_i\}$

(Eq. 3) $D(\gamma^s_i) = A(\gamma^s_i - d) = A(\tau^s_i)$

Right Sync Point (RSP) correlation for $\{\tau^e_i, \gamma^e_i\}$

(Eq. 4) $D(\gamma^e_i) = A(\gamma^e_i - d) = A(\tau^e_i)$

These relations provide additional insights into the way we constructed δ*, and will assist us in proving the optimality of δ*.

We prove these relations by induction over i.

Induction base:

Consider the two adjacent time periods for i=0:

1. $t \in (-\infty, \gamma^s_0)$
2. $t \in [\gamma^s_0, \gamma^e_0]$

From the definition of δ* we conclude that:

$$D^*(\gamma^s_0) = \int_{-\infty}^{\gamma^s_0} \delta(t)dt = \int_{-\infty}^{\gamma^s_0} \alpha(t-d)dt = \int_{-\infty}^{\tau^s_0} \alpha(t)dt = A(\tau^s_0)$$

and by using the PPE (Eq. 2) for the second interval we get:

$$D^*(\gamma^e_0) = \int_{-\infty}^{\gamma^e_0} \delta(t)dt = \int_{-\infty}^{\gamma^s_0} \alpha(t-d)dt + \int_{\gamma^s_0}^{\gamma^e_0} \mu dt =$$

$$= \int_{-\infty}^{\tau^s_0} \alpha(t)dt + \int_{\tau^s_0}^{\tau^e_0} \alpha(t)dt = A(\tau^e_0)$$

Now assume by induction that the LSP and RSP relations hold as in Eq. 3 and Eq. 4 for i, and prove them for i+1. We thus need to prove that:

$D(\gamma^s_{(i+1)}) = A(\gamma^s_{(i+1)} - d) = A(\tau^s_{(i+1)})$

and

$D(\gamma^s_{(i+1)}) = A(\gamma^s_{(i+1)} - d) = A(\tau^s_{(i+1)})$

$$D^*(\gamma^s_{i+1}) = \int_{-\infty}^{\gamma^s_{i+1}} \delta(t)dt = \int_{-\infty}^{\gamma^e_i} \delta(t)dt + \int_{\gamma^e_i}^{\gamma^s_{i+1}} \alpha(t-d)dt =$$

$$A(\tau^e_i) + \int_{\tau^e_i}^{\tau^s_{i+1}} \alpha(t)dt = A(\tau^e_i) + A(\tau^s_{i+1}) - A(\tau^e_i) = A(\tau^s_{i+1})$$

By using the PPE again:

$$D^*(\gamma_{i+1}^e) = \int_{-\infty}^{\gamma_{i+1}^e} \delta(t)dt = \int_{-\infty}^{\gamma_{i+1}^s} \delta(t)dt + \int_{\gamma_{i+1}^s}^{\gamma_{i+1}^e} \delta(t)dt =$$

$$A(\tau_{i+1}^s) + \int_{\gamma_{i+1}^s}^{\gamma_{i+1}^e} \mu dt = A(\tau_{i+1}^s) + A(\tau_{i+1}^e) - A(\tau_{i+1}^s) = A(\tau_{i+1}^e)$$

8.4.4 Meeting the Physically-Feasible Constraints

For δ* to be physically-feasible it must meet the three constraints defined in Section 8.1.2:

δ(t)=0, t≤0

Note that while constructing δ*, we enforced δ*(t) = 0, t<0. For this enforcement to hold, we need to show that $\gamma^s_0 \geq 0$. Otherwise, our definition of δ* is not consistent.

We thus need to show that if the deadline-feasible condition holds, then $\gamma^s_0 \geq 0$.

Assume, on the contrary, that $\gamma^s_0 = \tau^s_0 + d < 0$, and $\gamma^s_0 - d = \tau^s_0$. Thus, using the PPE for $[\tau^s_0, \tau^e_0]$ we get:

$$A(\tau_0^e) - A(\tau_0^s) = \int_{\tau_0^s}^{\tau_0^e} \alpha(t)dt = \mu(\tau_0^e - \tau_0^s) = \mu(\tau_0^e - 0 + 0 - (\gamma_0^s - d)) =$$

$$\mu(\tau_0^e - 0) + \mu d - \mu\gamma_0^s$$

Recall further that α(t)=0 for t<0 thus:

$$A(\tau_0^e) - A(\tau_0^s) = \int_0^{\tau_0^e} \alpha(t)dt$$

Combining the two equations we get:

$$A(\tau_0^e) - A(\tau_0^s) = \int_0^{\tau_0^e} \alpha(t)dt = \mu(\tau_0^e - 0) + \mu d - \mu\gamma_0^s > \mu(\tau_0^e - 0) + \mu d$$

This is in contradiction to the deadline-feasibility condition for time $\{0, \tau^e_0\}$, since if $\gamma^s_0 < 0$ then $\mu\gamma^s_0$ is a negative number.

1. $0 \leq \delta(t) \leq \mu$, $0 \leq t$.

Obviously the constraint holds for $t \in \Gamma$. We, thus need to show that $\delta^*(t) = \alpha(t-d) \leq \mu$ for $t \notin \Gamma$. While constructing δ*, we showed (Section 8.4.1 Remark 4) that τ^s_i falls within an interval in which α(t)<μ, and concluded that α(t)<μ within the time interval $\tau^e_{i-1} < t < \tau^s_i$. Thus, α(t-d) <μ for $\tau^e_{i-1} + d = \gamma^e_{i-1} < t < \gamma^s_i = \tau^s_i + d$.

2. $D\delta(t) \leq A(t)$, $-\infty \leq t \leq \infty$.

We will show that the constraint holds by considering four different time period categories, which together cover the whole real time:

Two boundary conditions periods

a. $t<\gamma^s_0$

b. $t>\gamma^e_n$

and two internal period categories:

c. $\gamma^e_i < t < \gamma^s_{(i+1)}$ $0 \le i < n$

d. $\gamma^s_i \le t \le \gamma^e_i$ $0 \le i \le n$

3.a) Let $-\infty < t < \gamma^s_0$.

Recall that:

$\delta^*(t) = 0$ for $t<0$

$\delta^*(t) = \alpha(t-d)$ for $0 \le t < \gamma^s_0$

Thus:

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = 0 + \int_0^t \alpha(u-d)du = A(t-d) - A(0-d) = A(t-d) \le A(t),$$

$-\infty < t < \gamma^s_0$.

3.b) Let $t>\gamma^e_n$.

Recall that $\delta^*(t) = \alpha(t-d)$ for $t>\gamma^e_n$

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = D(\gamma_n^e) + \int_{\gamma_n^e}^{t} \delta^*(u)du = A(\gamma_n^e - d) + \int_{\gamma_n^e}^{t} \alpha(u-d)du =$$

$$= A(\tau_n^e) + A(t-d) - A(\gamma_n^e - d) = A(\tau_n^e) + A(t-d) - A(\tau_n^e) = A(t-d) \le A(t)$$

for $t>\gamma^e_n$.

3.c) Let $\gamma^e_i < t < \gamma^s_{(i+1)}$, $0 \le i < n$.

Recall that $\delta^*(t) = \alpha(t-d)$, for $\gamma^e_i < t < \gamma^s_{(i+1)}$.

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = \int_{-\infty}^{\gamma_i^e} \delta^*(t)dt + \int_{\gamma_i^e}^{t} \delta^*(u)du = D(\gamma_i^e) + \int_{\gamma_i^e}^{t} \delta^*(u)du$$

$$= A(\gamma_i^e - d) + \int_{\gamma_i^e}^{t} \alpha(u-d)du = A(\tau_i^e) + A(t-d) - A(\gamma_i^e - d) = A(\tau_i^e) + A(t-d) - A(\tau_i^e)$$

$$= A(t-d) \le A(t), \quad \gamma_i^e < t < \gamma_{i+1}^e, \quad 0 \le i < n.$$

3.d) Let $\gamma^s_i \le t \le \gamma^e_i$ $0 \le i \le n$.

Recall that $\delta^*(t) = \mu$ for $\gamma^s_i \le t \le \gamma^e_i$

Thus, using the RSP relations as in Eq. 4:

$$D^*(t) = \int_{-\infty}^{\gamma_i^s} \delta(t)dt + \int_{\gamma_i^s}^{t} \delta^*(u)du = D(\gamma_i^s) + \int_{\gamma_i^s}^{t} \delta(u)du = A(\tau_i^s) + \int_{\gamma_i^s}^{t} \mu du =$$

$$A(\tau_i^s) + \mu(t - \gamma_i^s) = A(\tau_i^s) + \mu(t - (\tau_i^s + d)) = A(\tau_i^s) + \mu(t - \tau_i^s) - \mu d$$

and:

$$A(t) = \int_{-\infty}^{\tau_i^s} \alpha(t)dt + \int_{\tau_i^s}^{t} \alpha(u)du = A(\tau_i^s) + \int_{\tau_i^s}^{t} \alpha(u)du.$$

Now assume, on the contrary, that there is a time t for which D(t)>A(t).

Thus:

$$D^*(t) = A(\tau_i^s) + \mu(t - \tau_i^s) - \mu d > A(\tau_i^s) + \int_{\tau_i^s}^{t} \alpha(u)du = A(t)$$

and thus:

$$\mu(t - \tau_i^s) - \mu d > \int_{\tau_i^s}^{t} \alpha(u)du = \int_{\tau_i^s}^{\tau_i^e} \alpha(t)dt - \int_{t}^{\tau_i^e} \alpha(u)du.$$

From the way we constructed δ^* and specifically from the PPE (Eq. 2) we know that:

$$\mu(\tau_i^e - \tau_i^s) = \int_{\tau_i^s}^{\tau_i^e} \alpha(t)dt$$

Thus:

$$\mu(t - \tau_i^s) - \mu d > \mu(\tau_i^e - \tau_i^s) - \int_{t}^{\tau_{ie}} \alpha(u)du = \mu(\tau_i^e - t) + \mu(t - \tau_i^s) - \int_{t}^{\tau_i^e} \alpha(u)du$$

This results in:

$$-\mu d > \mu(\tau_i^e - t) - \int_{t}^{\tau_i^e} \alpha(u)du$$

and thus:

$$\int_{t}^{\tau_i^e} \alpha(u)du > \mu(\tau_i^e - t) + \mu d$$

which is in contradiction to the deadline-feasibility condition for points τ^e_i and t.

We conclude that $D_\delta(t) \le A(t)$, $\gamma^s_i \le t \le \gamma^e_i$ $0 \le i \le n$.

This last conclusion completes the proof that δ^* meets the $D_\delta(t) \le A(t)$ constraint over the whole real time, and that δ^* is indeed a physically-feasible control.

8.4.5 Meeting the Deadline-Feasible Constraint

For δ* to be a deadline-feasible control, it must meet the constraint defined in Section 8.1.3. Namely, we need to show that:

3. $A(t-d) \leq D^*(t)$, $-\infty \leq t \leq \infty$.

For that, consider again the four different time period categories:

Two boundary condition periods

a. $t < \gamma^s_0$

b. $t > \gamma^e_n$

and two internal period categories:

c. $\gamma^e_i < t < \gamma^s_{(i+1)}$ $0 \leq i < n$

d. $\gamma^s_i \leq t \leq \gamma^e_i$ $0 \leq i \leq n$

4.a) Let $-\infty < t < \gamma^s_0$.

Recall that

$\delta^*(t) = 0$ for $t < 0$

$\delta^*(t) = \alpha(t-d)$ for $0 \leq t < \gamma^s_0$

Thus:

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = 0 + \int_0^t \alpha(u-d)du = A(t-d) - A(0-d) = A(t-d),$$

$-\infty < t < \gamma^s_0$.

4.b) Let $t > \gamma^e_n$.

Recall that $\delta^*(t) = \alpha(t-d)$ for $t > \gamma_{ne}$. Thus, using RSP correlation (Eq. 4):

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = D(\gamma^e_n) + \int_{\gamma^e_n}^{t} \delta^*(u)du = A(\gamma_{ne} - d) + \int_{\gamma^e_n}^{t} \alpha(u-d)du =$$

$$= A(\tau^e_n) + A(t-d) - A(\gamma^e_n - d) = A(\tau^e_n) + A(t-d) - A(\tau^e_n) = A(t-d)$$

Thus $D^*(t) = A(t-d)$ for $t > \gamma^e_n$.

4.c) Let $\gamma^e_i < t < \gamma^s_{(i+1)}$, $0 \leq i < n$.

Recall that $\delta^*(t) = \alpha(t-d)$ for $\gamma^e_i < t < \gamma^s_{(i+1)}$. Thus, using RSP correlation (Eq. 4):

$$D^*(t) = \int_{-\infty}^{t} \delta^*(u)du = \int_{-\infty}^{\gamma_i^e} \delta^*(t)dt + \int_{\gamma_i^e}^{t} \delta^*(u)du = D(\gamma_i^e) + \int_{\gamma_i^e}^{t} \delta^*(u)du$$

$$= A(\gamma_i^e - d) + \int_{\gamma_i^e}^{t} \alpha(u-d)du = A(\tau_i^e) + A(t-d) - A(\gamma_i^e - d) = A(\tau_i^e) + A(t-d) - A(\tau_i^e)$$

$$= A(t-d)$$

Thus $D^*(t)=A(t-d)$ for $\gamma^e_i<t<\gamma^s_{(i+1)}$, $0\le i<n$

4.d) Let $\gamma^s_i \le t \le \gamma^e_i$ $0\le i\le n$

Recall that $\delta^*(t) = \mu$ for $\gamma^s_i \le t \le \gamma^e_i$

Thus, using RSP correlation (Eq. 4):

4.d.1
$$D^*(t) = \int_{-\infty}^{\gamma_i^s} \delta(t)dt + \int_{\gamma_i^s}^{t} \delta(u)du = D(\gamma_i^s) + \int_{\gamma_i^s}^{t} \delta(u)du =$$
$$A(\gamma_i^s - d) + \int_{\gamma_i^s}^{t} \mu du = A(\gamma_i^s - d) + \mu(t - \gamma_i^s)$$

4.d.2 $$A(t-d) = A(\gamma_i^s - d) + \int_{\gamma_i^s}^{t} \alpha(u-d)du = A(\gamma_i^s - d) + \int_{\tau_i^s}^{t} \alpha(u)du$$

From the PPE (Eq. 2) and from the definition of τ^s_i, τ^e_i, γ^s_i and γ^e_i we know that

4.d.3 $$A(\tau_i^e) - A(\tau_i^s) = \int_{\tau_i^s}^{\tau_i^e} \alpha(t)dt = \int_{\gamma_i^s}^{\gamma_i^e} \alpha(t-d)dt = \mu((\gamma_i^e - d) - (\gamma_i^s - d)) = \mu(\gamma_i^e - \gamma_i^s)$$

Recall further that from the way we constructed Γ it holds that for each $t \in \Gamma$:

4.d.4 $$A(t) = \int_{t}^{\tau_i^e} \alpha(u)du = \int_{t}^{\gamma_i^e} \alpha(u-d)du > \mu((\gamma_i^e - d) - (t-d)) = \mu(\gamma_i^e - t)$$, $\gamma^s_i \le t \le \gamma^e_i$

By combining 4.d.3 with 4.d.4 we get:

4.d.5
$$A(t) = \int_{\tau_i^s}^{t} \alpha(u)du = \int_{\tau_i^s}^{\tau_i^e} \alpha(t)dt - \int_{t}^{\tau_i^e} \alpha(u)du < \mu(\gamma_i^e - \gamma_i^s) - \mu(\gamma_i^e - t) = \mu(t - \gamma_i^s)$$
, $\gamma^s_i \le t \le \gamma^e_i$

and by combining 4.d.1, 4.d.2 and 4.d.5 we get:

4.d.6 $$A(t-d) = A(\gamma_i^s - d) + \int_{\tau_i^s}^{t} \alpha(u)du < A(\gamma_i^s - d) + \mu(t - \gamma_i^s) = D^*(t)$$, $\gamma_{is} \le t \le \gamma_{ie}$

Thus, we conclude that $A(t-d) \le D_\delta(t)$, $\gamma^s_i \le t \le \gamma^e_i$, $0\le i\le n$.

This last conclusion completes the proof that δ^* meets the $A(t-d) \le D_\delta(t)$ constraint over the whole real time, and that δ^* is indeed a deadline-feasible control.

8.4.6 Proving Optimality

We will now prove by induction over i that D*(t) defined as $D^*(t) = \int_{-\infty} \delta^*(t)dt$ is optimal.

Induction base:

Consider the two adjacent time periods for i=0:

1. $t \in (-\infty, \gamma_0^s)$

2. $t \in [\gamma_0^s, \gamma_0^e]$

We have already shown that D*(t) = A(t-d) for all $t \in (-\infty, \gamma_0^s)$; thus D*(t) is indeed optimal over that time interval.

Recall that δ*(t)=μ for all $t \in [\gamma_0^s, \gamma_0^e]$. Recall further that δ(t) is constrained by μ. That is δ(t) ≤μ.

We will thus show that if there exists a point in time t' such that δ(t')<μ, then $D'(t) = \int_{-\infty}^{t} \delta'(u)du$ does not meet the A(t-d) ≤ D(t) constraint.

Assume, on the contrary, that there is at least one point in time t' such that $t' \in [\gamma_0^s, \gamma_0^e]$ and δ'(t')<μ in some small neighborhood of t' (recall δ(•) is piecewise continuous) and that δ'(t)= δ*(t) for $t \in (-\infty, \gamma_0^s)$.

For such δ' there exists:

$$\int_{\gamma_0^s}^{\gamma_0^e} \delta'(t)dt < \int_{\gamma_0^s}^{\gamma_0^e} \delta^*(t)dt = \mu(\gamma_0^e - \gamma_0^s).$$

$$D'(\gamma_0^e) = \int_{-\infty}^{\gamma_0^e} \delta'(t)dt = \int_{-\infty}^{\gamma_0^s} \delta'(t)dt + \int_{\gamma_0^s}^{\gamma_0^s} \delta'(t)dt = \int_{-\infty}^{\gamma_0^s} \delta(t)dt + \int_{\gamma_0^s}^{\gamma_0^e} \delta'(t)dt =$$

$$A(\gamma_{0s} - d) + \int_{\gamma_0^s}^{\gamma_0^e} \delta'(t)dt < A(\gamma_{0s} - d) + \int_{\gamma_0^s}^{\gamma_0^e} \delta^*(t)dt = A(\gamma_0^s - d) + \mu(\gamma_0^e - \gamma_0^s) =$$

$$A(\gamma_0^s - d) + \int_{\gamma_0^s}^{\gamma_0^e} \alpha(t-d)dt = A(\gamma_0^s - d) + A(\gamma_0^e - d) - A(\gamma_0^s - d) = A(\gamma_0^e - d)$$

This is in contradiction to the constraint that $D'(\gamma_0^e) \geq A(\gamma_0^e - d)$.

Now assume by the induction that D*(t) is optimal for i.

We thus need to prove that D*(t) is also optimal for i+1.

Consider again the two adjacent time periods for i+1:

1. $t \in (\gamma_i^e, \gamma_{i+1}^s)$

2. $t \in [\gamma^s_{i+1}, \gamma^e_{i+1}]$

From the induction assumption we know that D*(t) is optimal for $t<\gamma^e_i$.

We have already shown that D*(t)=A(t-d) for all $t \in (\gamma^e_i, \gamma^s_{i+1})$; thus D*(t) is optimal for $t<\gamma^s_{(i+1)}$.

Recall that $\delta^*(t)=\mu$ for all $t \in [\gamma^s_{i+1}, \gamma^e_{i+1}]$. Recall further that $\delta(t)$ is constrained by μ. That is $\delta(t) \leq \mu$ for all t.

We will thus show that if there exists a point in time t' such that $\delta(t')<\mu$, then $D'(t) = \int_{-\infty} \delta'(t)dt$ does not meet the A(t-d) ≤ D(t) constrain.

Assume, on the contrary, that there is at least one point in time $t' \in [\gamma^s_{i+1}, \gamma^e_{i+1}]$ for which $\delta'(t')<\mu$, in some small neighborhood of t', and that $\delta'(t)= \delta^*(t)$ for $t \in (-\infty, \gamma^s_{(i+1)})$.

For such δ' there exist:

$$\int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta'(t)dt < \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta^*(t)dt = \mu(\gamma^e_{i+1} - \gamma^s_{i+1})$$

$$D'(\gamma^e_{i+1}) = \int_{-\infty}^{\gamma^e_{i+1}} \delta'(t)dt = \int_{-\infty}^{\gamma^s_{i+1}} \delta'(t)dt + \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta'(t)dt =$$

$$\int_{-\infty}^{\gamma_{(i+1)s}} \delta^*(t)dt + \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta'(t)dt = A(\gamma^s_{i+1} - d) + \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta'(t)dt <$$

$$A(\gamma^s_{i+1} - d) + \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \delta^*(t)dt = A(\gamma^s_{i+1} - d) + \mu(\gamma^e_{i+1} - \gamma^s_{i+1}) =$$

$$A(\gamma^s_{i+1} - d) + \int_{\gamma^s_{i+1}}^{\gamma^e_{i+1}} \alpha(t-d)dt =$$

$$A(\gamma^s_{i+1} - d) + A(\gamma^e_{i+1} - d) - A(\gamma^s_{i+1} - d) = A(\gamma^e_{i+1} - d)$$

This is in contradiction to the constrain that $D'(\gamma^e_{i+1}) \geq A(\gamma^e_{i+1} - d)$ and by that we conclude the optimality proof for δ*.

8.5 General Time-Varying Arrival Rate

In the following section, we formulate a conjecture for an optimal δ* in the situation for which deadlines can not be met at all times. We start by defining optimality for such situations and then describe a method to construct the optimal δ*. Proving that δ* is indeed a physically-feasible optimal control is left for future work.

8.5.1 An Optimal Control

Let Δ be the set of physically-feasible controls.

Identify, among all $\delta \in \Delta$ the one $\hat{\delta}$ which minimizes the cumulative time period t for which $D_\delta(t)<A(t-d)$.

Formally:

denote t^s_i, a point such that $D_\delta(t^s_i)=A(t^s_i-d)$,

denote t^e_i a second point such that $t^e_i>t^s_i$, and $D_\delta(t^e_i)=A(t^e_i-d)$,

and:

$D_\delta(t)<A(t-d)$ for all $t \in [t^s_i, t^e_i]$.

Then a physically-feasible optimal control δ^ minimizes $\sum_i (t^e_i - t^s_i)$ simultaneously over all times.

Further identify, among all $\hat{\delta} \in \Delta$ the one δ* that minimizies $D_\delta(t)$-A(t-d) simultaneously over all time intervals in which the deadline-feasible condition prevails.

Further show that for any $\delta \in \Delta$ $D^*(t) \le D_\delta(t)$, for all t.

Note that a physically-feasible control, for which $D_\delta(t)=A(t-d)$ at all times is obviously optimal.

8.5.2 Constructing Optimal Control

Denote the finite set of m+1 time intervals $[t^s_i, t^e_i]$, $0 \le i \le m$, such that:

$\alpha(t) \ge \mu$, $t^s_i<t<t^e_i$, $0 \le i \le m$

$\alpha(t) < \mu$ otherwise.

Note that $m<\infty$ since α has finite set of local extrema.

Denote $[a_i]$, $0 \le i \le n$ a vector of index alignment constants, such that $a_0=0$ and $a_i<a_{i+1}$.

We now construct a sequence of n+1 time intervals $[\varepsilon^s_i, \varepsilon^e_i]$ $0 \le i \le n \le m$ such that:

$-\infty<t^s_0=\varepsilon^s_0<\varepsilon^e_0<t^s_{a_1}=\varepsilon^s_1<\varepsilon^e_1<\ldots<t^s_{a_i}=\varepsilon^s_i<\varepsilon^e_i<\ldots<t^s_{a_n}=\varepsilon^s_n<\varepsilon^e_n<\infty$

for which:

1. $A(t)-A(\varepsilon^s_i)=\int_{\varepsilon^s_i}^{t}\alpha(u)du \ge \mu(t-\varepsilon^s_i)$, $\varepsilon^s_i<t<\varepsilon^e_i$, $0 \le i \le n$

2. $A(\varepsilon^e_i)-A(\varepsilon^s_i)=\int_{\varepsilon^s_i}^{\varepsilon^e_i}\alpha(t)dt=\mu(\varepsilon^e_i-\varepsilon^s_i)$, $0 \le i \le n$.

3. There exist small ε>0 such that

$$A(\varepsilon_i^e + \varepsilon) - A(\varepsilon_i^e) = \int_{\varepsilon_i^e - \varepsilon}^{\varepsilon_i^e} \alpha(t)dt < \mu(\varepsilon_i^e + \varepsilon - \varepsilon_i^e),\ 0 \le i \le n$$

Note that:

1. The construction of the $[\varepsilon^s_i, \varepsilon^e_i]$ is done in an iterative manner starting from $t = t^s_0$ forwards.
2. The process must eventually stop as $a_i < a_{i+1}$, $0 \le i \le n$, and the number of time intervals [tsi, tei] is finite.
3. ε^e_i exists, and $\varepsilon^e_i < \infty$ since $\alpha(t) < \mu$ for $T < t$, thus $\int_{\varepsilon_n^s}^{\varepsilon_n^e} \alpha(t)dt$ is bounded while $\mu(\varepsilon^e_n - \varepsilon^s_n)$ goes to ∞ as ε^e_n goes to ∞.
4. $t^e_{ai} < \varepsilon^e_i < t^s_{ai+1}$, that is, ε^e_i falls within an interval in which $\alpha(t) < \mu$. Otherwise, i.e., assuming $\alpha(t) > \mu$ for $\varepsilon^e_{ai} < t < \varepsilon^e_{ai} + \varepsilon$, we have

$$A(\varepsilon_i^e + \varepsilon) - A(\varepsilon_i^s) = \int_{\varepsilon_i^s}^{\varepsilon_i^e + \varepsilon} \alpha(t)dt = \mu(\varepsilon_i^e - \varepsilon_i^s) + \int_{\varepsilon_i^e}^{\varepsilon_i^e + \varepsilon} \alpha(t)dt,$$

$$> \mu(\varepsilon_i^e + \varepsilon - \varepsilon_i^s)$$

in contradiction to the above third condition. Hence, $\alpha(t) < \mu$ for $\varepsilon^e_i < t < \varepsilon^s_{i+1} = t^s_{a_{i+1}}$.

Further, denote $[b_i]$, $0 \le i \le k$ a vector of index alignment constants, such that $b_i < b_{i+1}$.

We now construct a sequence of k+1 time intervals $[\beta^s_i, \beta^e_i]$ $0 \le i \le k \le m$ such that:

$-\infty < \varepsilon^s_{b_0} = \beta^s_0 < \beta^d_0 < \beta^e_0 = \varepsilon^e_{b_0} < \ldots < \varepsilon^s_{b_i} = \beta^s_i < \beta^d_i < \beta^e_i = \varepsilon^e_{b_i} < \ldots < \varepsilon^s_{b_k} = \beta^s_k < \beta^d_k < \beta^e_k = \varepsilon^e_{b_k} < \infty$

for which there exist at least one point $t' \in [\beta_i^s, \beta_i^d]$:

1. $A(t') - A(\beta_i^s) = \int_{\beta_i^s}^{t'} \alpha(t)dt > \mu(t' - \beta_i^s) + \mu d$, $\beta^s_i < t < \beta^d_i$, $0 \le i \le k$
2. $A(\beta_i^d) - A(\beta_i^s) = \int_{\beta_i^s}^{\beta_i^d} \alpha(t)dt = \mu(\beta_i^d - \beta_i^s) + \mu d$
3. For $t \in [\beta_i^d, \beta_i^e]$ $A(t) - A(\beta_i^s) = \int_{\beta_i^s}^{t} \alpha(u)du < \mu(t - \beta_i^s) + \mu d$
4. Denote $\beta_i^z \in [\beta_{i-1}^d, \beta_i^s]$ the point for which

$$A(\beta_i^s) - A(\beta_i^z - d) = \int_{\beta_i^z - d}^{\beta_i^s} \alpha(t)dt = \mu(\beta_i^s - \beta_i^z)$$

Define Z as the set of time intervals $[\beta^z_i, \beta^s_i]$.

Define *B* as the set of time intervals $[\beta^s_i, \beta^d_i]$.

Define Θ as the set of time intervals $[\beta^d_i, \beta^z_{i+1}]$.

Define δ^*_i to be the optimal control over the Θ. Construct δ^*_i as described in Section 8.4, over each of the Θ intervals, such that α(t) is defined over the close intervals $[\beta^d_i - d, \beta^z_{i+1} - d]$.

Define δ^* to be the optimal control over the whole real numbers T as follow:

- $\delta^*(t) = 0$, $t<0$,
- $\delta^*(t) = \mu$, $t \in B$
- $\delta^*(t) = \mu$, $t \in Z$
- $\delta^*(t) = \delta^*_i$ $t \in \Theta$

δ^* is physically-feasible but not deadline feasible over B.

δ^* is deadline-feasible over $\Theta \cup Z$.

δ^* is optimal over T

8.5.3 Intuition and Claims

Note:

$\beta^z_i < \beta^s_i < \beta^d_i < \beta^z_{i+1}$

Intuition:

The interval $[\beta^s_i, \beta^d_i]$ is physically-feasible.

The system has enough capacity to deplete the queue at time β^e_i. We know that $\beta^e_i < \beta^s_{i+1}$, we also know that at time β^d_i, deadlines have just been met, namely $D(\beta^d_i) = A(\beta^d_i - d)$, and $\{\beta^d_i - d , \beta^d_i\}$ are RSP.

We argue that $\beta^d_i + d < \beta^s_i$, since we are able to deplete the queue during that time period; this inequality stems from the way we constructed $[\varepsilon^s_i, \varepsilon^e_i]$.

Similarly, the condition $\beta^d_i \leq \beta^z_{i+1}$ stems from the same fact.

β^z_i is the latest point in time from which we can deplete the waiting queue such that it will become empty at time β^s_i. This will allow us to minimize the deviation from the deadline.

We define β^z_i as the point from which $\delta(t)=\mu$ to deplete the queue, thus, $\{\beta^z_i - d , \beta^z_i\}$ are LSP.

At β^z_i the queue contains all work that arrived during $[\beta^z_i - d, \beta^z_i]$, thus, to deplete the system we need to complete this work and the rest of the work that arrived at $[\beta^z_i, \beta^s_i]$, while serving at maximum capacity from β^z_i onwards. That is exactly the condition that allows us to find β^z_i.

Claim 1: $\beta^d_i < \beta^e_i$

Claim 2: $\beta^d_i \leq \beta^z_{i+1}$ (at the extreme, we need to have $\delta(t)=\mu$ from $\beta^z_{i+1} = \beta^d_i$ in order to drain the queue at β^s_{i+1}

Claim 3: the deadline-feasibility condition holds over the intervals $[\beta^d_i, \beta^s_{i+1}]$

For all $t \in [\beta_i^s, \beta_i^d]$ holds:

$$A(t) - A(\beta_i^s) = \int_{\beta_i^s}^{t} \alpha(u)du > \mu(t - \beta_i^s)$$

$$A(\beta_i^d) - A(\beta_i^s) \int_{\beta_i^s}^{\beta_i^d} \alpha(t)dt = \mu(\beta_i^d - \beta_i^s) + \mu d$$

For all $t_i, t_j \in [\beta_i^d, \beta_{i+1}^s]$ holds:

$$A(t_j) - A(t_i) = \int_{t_i}^{t_j} \alpha(t)dt < (t_j - t_i)\mu + d\mu$$

For all $t'_i \in [\beta_i^d, \beta_{i+1}^s]$ holds

$$A(t'_i) - A(\beta_i^s) = \int_{\beta_i^s}^{t'_i} \alpha(t)dt < (t'_i - \beta_i^s)\mu + d\mu$$

8.6 Summary of Results

In this chapter, we have shown that an optimal δ* can be constructed when the deterministic arrival rate function, α, is known. We have shown that if $\alpha(\bullet)<\mu$, then a trivial optimal solution exists, namely $\delta(t) = \alpha(t-d)$. We have further shown that for situations in which the DFC prevails, an optimal δ* can be found. In such situations, a proactive behavior, namely increasing the service rate for NA patients, allows meeting the deadline even in situations in which $\alpha(\bullet)>\mu$ for some time periods. Recall the NAF policy presented in Section 6.6.5. Our simulation-based analysis showed that this policy meets the triage deadline at all times under realistic arrival rates. This indicates that the DFC holds for such situations. Recall further the DTD policy presented at Section 6.7. This policy serves the triage queues at a rate equal to $\delta(t) = \alpha(t-d)$, at most times, and implements the proactive behavior in situations where $\alpha(\bullet)>\mu$. Thus, this policy suggests optimal control with realistic arrival rates, as proved by our fluid-model analysis.

CHAPTER 9: CONCLUSIONS AND FUTURE WORK

This research focuses on the benefits that a real-time monitoring-and-control system may provide for optimizing ED operations. During this work, we designated several points along the ED patient flow process that offer opportunities for promising improvements. We then analyzed and explored some of them using various techniques, such as simulation, mathematical analysis, and prototype implementations. Specifically, we identified two applications and addressed them through innovative approaches—adaptive load monitoring and the "Which patient to treat next?" service policy. While our work involved close interactions with ED management, and close assessment of ED environments in various hospitals, we did not have the opportunity to apply any of these techniques within a real ED environment. Our work was based on simulated data, which, as accurate as it can be, is still not a real ED environment. Thus, deployment of the proposed solution, along with its monitoring and control capabilities, is still required for proving its benefit to ED management in real life.

In addition, we list below some of the main challenges and aspects that we came across during our research but did not have the capacity to address. These require further future research.

9.1 Forecasting and Controlling ED Arrival Rates

The arrival rate has a significant effect on all ED operational KPIs. Being able to accurately forecast ED arrivals may allow ED management to improve the ED service level, e.g., through optimal staffing. Ample work has been devoted for arrival forecasting. Real-time monitoring and control suggest a complementary approach, in which the online load of the ED is provided to arriving patients in advance. The way by which such an online status will affect the ED arrival rate is a subject for future research.

9.2 Neural Network-Based Load Monitoring

The neural network-based approach for load monitoring, presented in Chapter 5, requires more research. Specifically, the learning mechanism that we introduced in Section 5.4, requires extensive validation, after deployment within a real ED environment. Moreover, the ability to calculate different load measurements for different ED roles, presented in Section 5.6, requires further research both for its accuracy and for its relevance.

9.3 Extending the Fluid Model Analysis

The fluid model analysis presented in Chapter 8 requires future work in various directions. First, a proof for the conjecture we presented in Section 8.5 should be worked out. Second, an extension into a multi-triage-class environment is called for. Third, a generalization for time-varying capacity, i.e., a time-dependent physician pool, is possible. Lastly, a method for combining the fluid-model analysis for a given timeframe in which arrivals are already known (i.e., from arrival to deadline) with the forecasted, time-varied, stochastic arrivals may prove to be superior over the heuristic approach that we presented in Section 6.7, and thus offers interesting potential for future research.

9.4 ED Priority Queues

Serving patients by a policy other than FCFS may be conceived as unfair and result in service-level deterioration. A better understanding of patient behavior, under a clinical-dependent priority policy, and the optimal ways to communicate such a policy to patients, requires further work

9.5 Situational Displays

An important part of a real-time monitoring-and-control system is a situational display, also known as a dashboard. The dashboard communicates various aspects of the current environment status to patients and care personnel. A complementary requirement is to provide ED management and care personnel with a means of intervention in situations in which human intervention is required in real time. Furthermore, the ED situational display may confront challenges not usually found in other service environments, such as airports, train stations, and customer service stores, namely, the need to present highly private information on public displays. Such questions, or those related to the actual information needed to be displayed, are left for further research.

BIBLIOGRAPHY

1. "Business activity monitoring." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 14 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Business_activity_monitoring>. [Chapter 3]
2. "Apache Flex." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 30 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Apache_Flex>. [4.5.1]
3. "Complex event processing." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 24 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Complex_event_processing>. [Chapter 3], [Chapter 4]
4. " Dashboard (business)." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 12 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Dashboard (business)>. [4.5]
5. "Discrete event simulation." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 2 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Discrete_event_simulation>. [Chapter 7]
6. "Eclipse (software)." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 2 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Eclipse (software)>. [7.1]
7. "Emergency department." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 24 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Emergency_department>. [Chapter 2]
8. "Global positioning system." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 15 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Global_positioning_system>. [4.2.2]
9. "Indoor positioning system." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 21 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Indoor_positioning_system>. [4.2.2]
10. "Java (programming language)." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 18 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Java_programming_language >. [7.1]
11. "Performance indicator." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 29 May 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/Performance_indicator>. [Chapter 3]
12. "StreamBase Systems." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 14 February 2012. Web. 1 June 2012. <http://en.wikipedia.org/wiki/StreamBase_Systems>. [4.1]
13. "Triage." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc. 30 May 2012. Web. 15 March 2012. <http://en.wikipedia.org/wiki/Triage>. [2.2.2]
14. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., Yom-Tov, G., 2001. Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. Stochastic Systems, November 2011. [6.2.3]
15. Arnold, J. L., 1999. International emergency medicine and the recent development of emergency medicine worldwide. Annals of Emergency Medicine, 33(1), 97–103. [Chapter 2], [2.3.1]

16. Aronsky, D., Jones, I., Lanaghan, K., Slovis, C. M., 2008. Supporting patient care in the emergency department with a computerized whiteboard system. Journal of the American Medical Informatics Association, 15(2), 184–194. [4.5]

17. Asplin, B. R., Magid D. J., Rhodes K. V., Solberg L. I., Lurie N., Camargo C. A. Jr., 2003. Conceptual model of emergency department crowding. Annals of Emergency Medicine, 42(2), 173–180. [2.2]

18. Aydt, H., Turner, S. J., Cai, W., Hean, M. Y., 2008. Low symbiotic simulation systems: An extended definition motivated by symbiosis in biology. The 22nd ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation, June 03-06, 2008, Rome, Italy. [Chapter 7]

19. Bar-Yaacov, M., Cohen, A., 2007. New Statistical projects and Publications in Israel, The Central Bureau of Statistics, Israel. Web. <http://www.cbs.gov.il/q154_eng.htm>. [2.2.1]

20. Bernstein, S. L., Verghese, V., Leung, W., Lunney, A. T., Perez, I., 2003. Development and validation of a new index to measure emergency department crowding. Acad. Emerg. Med., 10(9), 938–942. [5.1]

21. Burt, C. W., McCaig, L. F., Valverde, R. H., 2006. Analysis of ambulance transports and diversions among US emergency departments. Annals of Emergency Medicine, 47(4), 316–326. [3.2]

22. Carliera, J., 1982. The one-machine sequencing problem. European Journal of Operational Research, 11(1), 42–47. [2.2.3]

23. Channouf, N., L'Ecuyer, P., Ingolfsson, A., Avramidis, A. N., 2007. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. Health Care Management Science, 10, 25–45. [2.2.1]

24. Derlet, R. W., Richards, J. R., 2000. Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. Annals of Emergency Medicine, 35, 63–68. [3.2]

25. Feldman, Z., Mandelbaum, A., Massey, W. A., Whitt, W., 2008. Staffing of time-varying queues to achieve time-stable performance. Management Science, 54(2), 324–338. [2.3.2]

26. Forster, A. J., Stiell, I., Wells, G., Lee, A. J., Van Walraven, C., 2003. The effect of hospital occupancy on emergency department length of stay and patient disposition. Academic Emergency Medicine, 10(2), 127–133. [3.1.2]

27. Fry, E. A., Lenert, L. A., 2005. MASCAL: RFID tracking of patients, staff and equipment to enhance hospital response to mass casualty events. AMIA 2005 Symposium Proceedings, 261–265. [2.4]

28. George, S., Read, S., Westlake, L., Williams, B., Pritty, P., Fraser-Moodie, A., 1993. Nurse triage in theory and in practice. Archives of Emergency Medicine, 10(3), 220–228. [2.2.2]

29. Green, L., 2004. Capacity planning and management in hospitals. Operations Research and Health Care, 70(2), 15-41. [2.3.2]

30. Greenshpan, O., Wasserkrug, S., Marmor, Y., Carmeli, B., Vortman, P., Basis, F., Schwartz, D., Mandelbaum, A., 2009. InEDvance: Advanced IT in support of emergency department management. NGITS 2009, 5831, 86–95. [4.5.1]

31. Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, 2nd edition, Upper Saddle River, NJ Prentice Hall, Inc. [Chapter 5], [5.2.1]

32. Hightower, J., Borriello, G., 2001. A survey and taxonomy of location systems for ubiquitous computing. Computer, 34(8), 57–66. [4.2.2]

33. Hightower, J., Want, R., Borriello, G., 2000. SpotON: An indoor 3D location sensing technology based on RF signal strength. UW CSE Technical Report 2000-02-02. [4.2.2]

34. Hinton, G. E., Salakhutdinov, R. R., 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507. [0]

35. Hoot, N. R., Zhou, C., Jones, I., Aronsky, D., 2007. Measuring and forecasting emergency department crowding in real time. Annals of Emergency Medicine, 49(6), 747–755. [3.2]

36. Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feed forward networks are universal approximators. Neural Networks, 2(5), 359–366. [5.2.1]

37. Huang, J., Carmeli, B., Mandelbaum, A., 2012. Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback. In Preparation. [6.5], [6.6.5]

38. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., Meltzer, P. S. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7(6), 673–679. [0]

39. Khare, R. K., Powell, E. S., Reinhardt, G., Lucenti, M., 2009. Adding more beds to the emergency department or reducing admitted patient boarding times: Which has a more significant influence on emergency department congestion. Annals of Emergency Medicine, 53(5), 575–585. [3.1.2]

40. Luckham, D. C., 2001. The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. [4.1]

41. King, D. L., Ben-Tovim, D. I., Bassham, J., 2006. Redesigning emergency department patient flows: Application of lean thinking to health care emergency department. Emergency Medicine Australasia, 18, 391–397. [2.2.2]

42. Mahapatra, S., Koelling, C. P., Patvivatsiri, L., Eitel, D., Grove, L., Fraticelli, B., 2003. Pairing emergency severity index5-level triage Data with computer aided system design to improve emergency department access and throughput. Proceedings of the 2003 Winter Simulation Conference, Dec. 7–10, New Orleans, LA, U.S.A. [2.2.2]

43. Marmor, Y. N., Sinreich, D., 2004. Simple and intuitive simulation tool for analyzing emergency department operations. Proceedings of the 2004 Winter Simulation Conference, Dec. 5–8, Washington, D.C., U.S.A. [4.2.3]

44. Marmor, Y. N., 2010. Emergency-Departments Simulation in Support of Service-Engineering: Staffing, Design, and Real-Time Tracking. Ph.D. Thesis, Technion, February 2010. [2.2.1], [2.2.3] [5.3], [5.6], [7.1.5]

45. Massey, W. A., 2002. The analysis of queues with time-varying rates for telecommunication models. Telecommunication Systems, 21(2–4), 173–204. [6.5.1]

46. Miró, O., Antonio, M. T., Jiménez, S., De Dios, A., Sánchez, M., Borrás, A., Millá, J., 1999. Decreased health care quality associated with emergency department overcrowding. European Journal of Emergency Medicine: Official Journal of the European Society for Emergency Medicine, 6(2), 105–107. [3.2]

47. Moskop, J. C., Sklar, D. P., Geiderman, J. M., Schears, R. M., Bookman, K. J., 2009. Emergency department crowding, part 1—concept, causes, and moral consequences. Annals of Emergency Medicine, 53(5), 605–611. [3.2]

48. Ni, L. M., Liu, Y., Lau, Y. C., Pathil, A. P., 2004. LANDMARC: Indoor location sensing using active RFID. Wireless Networks, 10, 701–710. [4.2.2]

49. Pinedo, M. L., 2005. Planning and Scheduling in Manufacturing and Services. Chapter 7, Springer. [2.2.3]

50. Rao, B., Minakakis, L., 2003. Evolution of mobile location-based services. Communications of the ACM Archive, SPECIAL ISSUE: Mobile commerce opportunities and challenges, 46(12), 61–65. [2.4]

51. Reeder, T. J., Garrison, H. G., 2001. When the safety net is unsafe: Real-time assessment of the overcrowded emergency department. Acad Emerg Med., 8(11), 1070–1074. [5.1]

52. Rosenbloom, S. T., Grande, J., Geissbuhler, A., Miller, R. A., 2004. Experience in implementing inpatient clinical note capture via a provider order entry system. JAMIA, Jul-Aug, 11(4), 310–315. [4.2.1]

53. Sharon, G., Etzion, O. 2008. Event-processing network model and implementation. IBM Systems Journal, 47(2), 321–334. [Chapter 3]

54. Sinreich, D., Marmor, Y. J., 2005. Ways to reduce patient turnaround time and improve service quality in emergency departments. Health Org & Manage, 19(2), 88–105. [3.2]

55. Solberg, L. I., Asplin, B. R., Weinick, R. M., Magid, D. J., 2003. Emergency department crowding: Consensus development of potential measures. Annals of Emergency Medicine, 42(6), 824–834. [3.2.1], [5.1], [5.3], [B]

56. Tonshoff, H. K., Wulfsberg, J. P., Kals, H. J. J., Konig, W., Van Luttervelt, C. A., 1988. Developments and trends in monitoring and control of machining processes. CIRP Annals - Manufacturing Technology, 37(2), 611–622. [2.4]

57. Tseytlin, Y., 2009. Queueing Systems with Heterogeneous Servers: On Fair Routing of Patients in Emergency Departments. M.Sc. Thesis, Technion, April. [2.2.5]

58. Vitkin, E., Carmeli, B., Greenshpan, O., Baras, D., Marmor, Y., 2010. MEDAL: Measuring of emergency departments' adaptive load. Studies in Health Technology and Informatics, 160(Pt 1), 218–222. [3.2]

59. Van Mieghem, J. A., 1995. Dynamic scheduling with convex delay costs: The generalized cμ rule. The Annals of Applied Probability, 5(3), 809–833. [6.5.3]

60. Van Mieghem, J. A., 2003. Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. Operations Research, 51(1), 113–122. [6.5.4]

61. Wein, L. M., Chevalier, P. B., 1992. A broader view of the job-shop scheduling problem. Management Science, 38(7), 1018–1033. [2.2.3]

62. Weiss, S. J., Derlet, R., Arndahl, J., Ernst, A. A., Richards, J., Fernández-Frackelton, M., Schwab, R., Stair, T. O., Vicellio, P., Levy, D., Brautigan, M., Johnson, A., Nick, T. G., 2004. Estimating the degree of emergency department overcrowding in academic medical centers: Results from the National ED Overcrowding Study (NEDOCS). Acad Emergency Medicine, 11(1), 38–50. [5.1]

63. Yom-Tov, G., 2010. Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime. Ph.D. Thesis, Technion, June. [6.5.1]

64. Zeltyn, S., Marmor, Y., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Schwartz, D., Moskovitch, K., Tzafrir, S., Basis, F., Shtub, A., Lauterman, T., 2011. Simulation-based models of emergency departments: Operational, tactical and strategic staffing. ACM Transactions on Modeling and Computer Simulation (TOMACS), 21(4), 24:1–24:5. [2.3.2]

65. Zviran, A., 2011. Fork-Join Networks in Heavy Traffic: Diffusion Approximations and Control. M.Sc. Thesis, Technion, March. [6.1]

APPENDIX A: INPUT OUTPUT AND CONTROL EVENTS OF EDRHYTHM

The EdRhythm is based on event processing technology. Below we provide a description of the various events that serve as input and output to the system. There are two categories of input events—data events and control events. All data events have the same structure. Each control event has its own structure.

Output events are associated with a specific KPI. Each output event has its own structure. The tables below describe the structure of the various events and event categories and list the exhaustive input event types.

1. Data Events

All data events have the same structure. Data events are identified by the unique event type associated with each one. The list of possible event types is provided in Table 29.

Attribute	Description
Event type	See Table 36 for possible types
Room ID	The physician room number or ED section
Care giver type	Group type, e.g., triage nurse
Care giver ID	ID of the specific nurse or physician
Patient type	Patient group classification, i.e., by triage score
Patient ID	Unique ID for a patient
Time	Time of operation start

Table 29: Data event structure.

2. Control Events

Each control event has a type and a set of up to three values. The table below shows the various control events and the meaning of the values in each of them:

Type	Value	Description	Value 1	Value 2
Set Threshold	1001	Set a threshold for the TTFE KPI	The threshold value	
Set Room Threshold	1002	Set a threshold for the room occupancy KPI	The room ID	The threshold value
Set Patient Treatment Ratio threshold	1003	Set the threshold for the patient treatment ratio KPI	The threshold for all patients (in %)	
Set clock period	1100	Set clock period	The period in seconds	

Table 30: Control event structure.

3. Output Events

Output event contains the KPI value after calculation. Each output event implements a specific indicator and has its own structure and set of values. The tables below describe a small subset of the monitored KPI implemented by the EdRhythm

Patient's Time Till First Encounter:

Attribute	Description
Event Type	The event type (110)
Patient Type	Patient's group
Patient Id	Unique patient id
Waiting Time	The time from registration till first treatment
Registration time	Time of registration
First treatment time	Time of first encounter

Table 31: TTFE KPI structure.

Staff Utilization Ratio:

Attribute	Description
Event Type	The event type (111)
Staff Type	The type of care personnel
Staff Id	Specific care personnel ID
Treatment	Total treatment time for a given period
Additional Work	Additional (not in front of patient) treatment time in a given period
Period	Time (in seconds) for that period

Table 32: Stuff utilization ration KPI structure.

Occupancy Level:

Attribute	Description
Event Type	The event type (112)
Room ID	The ID for the room
Occupancy Level	Number of patients within the room
Period	Time (in seconds) for that period

Table 33: Occupancy level KPI structure.

Patient Treatment Ratio - per period:

Attribute	Description
Event type	The event type (113)
Patient type	Patient's group
Patient ID	Unique patient ID
Treatment time	Total treatment time for a given period
Period	Time (in seconds) for that period

Table 34: Patient treatment ratio KPI structure.

Patient total treatment time (114)

Attribute	Description
Event type	The type of the control event
Patient type	The type of patient
Patient ID	Patient ID
Treatment time	Total treatment time for a given period
Registration time	Time of patient arrival
Discharge time	Time of patient discharge

Table 35: Patient total treatment time KPI structure.

4. Event Types

The table below summarizes the various event types within the system. Each event is identified by a unique event ID. The table contains all event types, i.e., data, control and output events.

Type	Val	Event Type
Patient registered	10	Data
Patient starts waiting for triage	11	Data
Patient starts triage	12	Data
Patient finishes triage	13	Data
Patient waits for nurse	15	Data
Nurse starts treatment	16	Data
Nears finish of treatment	17	Data
Patient waits for additional treatment by nurse	18	Data
Nurse starts additional work	19	Data
Nurse finishes additional work	20	Data
Patient waits for nurse	21	Data
Nurse starts treatment	22	Data
Nears finish of treatment	23	Data
Patient waits for additional treatment by nurse	24	Data
Nurse starts additional work	25	Data
Nurse finishes additional work	26	Data
Patient waits for nurse before final decision	27	Data
Nurse starts treatment	28	Data
Nears finish of treatment	29	Data
Patient waits for additional treatment by nurse	30	Data
Nurse starts additional work	31	Data
Nurse finishes additional work	32	Data
Patient starts to wait for physician treatment	33	Data

Type	Val	Event Type
Physician starts treatment	34	Data
Physician finishes treatment	35	Data
Patient starts to wait for additional work by physician	36	Data
Physician starts additional work	37	Data
Physician finishes additional work	38	Data
Patient starts to wait for physician examination	41	Data
Physician starts treatment	42	Data
Physician finishes treatment	43	Data
Patient starts to wait for additional work by physician	44	Data
Physician starts additional work	45	Data
Physician finishes additional work	46	Data
Patient starts to wait for physician examination for final decision	47	Data
Physician starts treatment	48	Data
Physician finishes treatment	49	Data
Patient starts to wait for additional work by physician	50	Data
Physician starts additional work	51	Data
Physician finishes additional work	52	Data
Start lab tests	55	Data
Finish lab tests	56	Data
Start wait for a consultant	57	Data
Finish wait for a consultant and start the treatment	58	Data
Finish consultant treatment	59	Data
Start walking to CT	61	Data
Finish walking to CT	62	Data
Start waiting for CT	63	Data
Finish waiting for CT	64	Data

Type	Val	Event Type
Start CT	65	Data
Finish CT	66	Data
Start wait for CT answer	67	Data
Finish wait for CT answer	68	Data
Start return from CT	69	Data
Finish return from CT	70	Data
Start walk to US	71	Data
Finish walk to US	72	Data
Start wait for US	73	Data
Finish wait for US	74	Data
Start US	75	Data
Finish US	76	Data
Start wait for US answer	77	Data
Finish wait for US answer	78	Data
Start return from US	79	Data
Finish return from US	80	Data
Start walk to X-ray	81	Data
Finish walk to X-ray	82	Data
Start wait for X-ray	83	Data
Finish wait for X-rayand start treatment	84	Data
Start wait for X-rayanswer	87	Data
Finish wait for XRay answer	88	Data
Start return from X-ray	89	Data
Finish return from X-ray	90	Data
Bed release	91	Data
Start wait before hospitalized	92	Data
Finish wait before hospitalized	93	Data

Type	Val	Event Type
Start delay before discharge	94	Data
Finish delay before discharge	95	Data
Wait for nurse discharge	96	Data
Start nurse discharge	97	Data
Finish nurse discharge	98	Data
Patient left the ED	99	Data
TTFE	110	Output
Staff utilization ratio	111	Output
Occupancy level	112	Output
Patient wait time ratio	113	Output
Patient total treatment time	114	Output
Set TTFE threshold	1001	Control
Set room occupancy threshold	1002	Control
Set patient treatment ratio threshold	1003	Control
Set clock period	1100	Control
Time tick	2001	Clock

Table 36: Event types.

APPENDIX B: CONSENSUS ON LOAD PARAMETER CLASSIFICATION

The tables below summarizes the list of 38 load parameters as suggested by [55]

1. Input Parameters

Input Parameter	Concept Operational	Definition
1. ED patient volume, standardized for bed hours	Patient demand	Number of new patients registered within a defined period (hour, shift, day) ÷ number of ED bed hours within this period
2. ED patient volume, standardized for annual average	Patient demand	Number of new patients registered within a defined period ÷ annual mean number new patients registered within this period
3. ED ambulance patient volume, standardized for bed hours	Patient demand	Number of new ambulance patients registered within a defined period ÷ number of ED bed hours within this period
4. ED ambulance patient volume, standardized for annual average	Patient demand	Number of new ambulance patients within a defined period ÷ annual average of new ambulance patients registered within this period
5. Patient source	Patient demand	Time, arrival mode, reason, referral source, and usual care for each patient registering at an ED in a defined period (hour/shift/day)
6. Percentage of open appointments	Patient demand	Percentage of open appointments at the beginning of a day in ambulatory care clinics that serve an ED's patient population
7. Percentage of patients who leave without treatment completed*	ED capacity	Number of registered patients who leave the ED without treatment completed ÷ total number of patients who register during this period
8. Leave without	ED capacity	Average severity of patients who leave the ED

Input Parameter	Concept Operational	Definition
treatment complete severity*		without treatment completed within a defined period (shift/day/week)
9. Ambulance diversion episodes	ED capacity	Number and duration of all diversion episodes at EDs within the EMS system within a defined period (week/month/year)
10. Ambulance diversion requests denied and forced openings	ED capacity	Number of diversion requests denied or forced openings within a defined period (week/month/year)
11. Diverted ambulance patient description	ED capacity	Chief complaints and final destination of diverted EMS patients within a defined period (week/month/year)
12. Average EMS waiting time	ED efficiency	Total time at hospital for ambulances delivering patients to ED during a defined period (shift/day/week/month) ÷ number of ambulance deliveries within that period
13. Patient complexity as assessed at triage	Patient complexity	Mean complexity level as assessed at triage (using local criteria) for all
14. Patient complexity as the percentage of ambulance patients	Patient complexity	Percentage of patients registering at an ED in a defined period (shift/day/week/month) who arrived by ambulance
15. Patient complexity as assessed by coding	Patient complexity	Mean complexity level as coded at the end of the visit for all patients completed in a defined period (shift/day/week/month)

* Leave without treatment completed includes those patients who leave without being seen, leave before being finished, and leave against medical advice.

2. Throughput Parameters

Throughput Parameter	Concept Operational	Definition
1. ED throughput time	ED efficiency	Average time between patient sign-in and departure (separately for admitted vs. discharged patients) within a defined period (day/week/month)
2. ED bed placement time	ED efficiency	Mean interval between patient sign-in and placement in a treatment area within a defined period (shift/day/week/month)
3. ED ancillary service turnaround time	ED efficiency	Average time between physician's order and result report (separately for each service area) within a defined period (shift/day/week/month)
4. Summary workload, standardized for ED bed hours	ED workload	Summary of (patients treated × acuity) in a defined period (shift/day/week) ÷ number of ED bed hours within this period
5. Summary workload, standardized for registered nurse staff hours	ED workload	Summary of (patients treated × acuity) in a defined period(shift/day/week) ÷ total ED staff registered nurse hours within this period
6. Summary workload, standardized for physician staff hours	ED workload	Summary of (patients treated × acuity) in a defined period (shift/day/week) ÷ total ED staff physician hours within this period
7. ED occupancy rate	ED workload	Total number of ED patients registered at a defined time ÷ number of staffed treatment areas at that time
8. ED occupancy	ED workload	Total number of patients present in the ED at a defined time ÷ number of staffed treatment areas at that time
9. Patient disposition to physician staffing ratio	ED workload	Number of patients admitted or discharged per staff physician during a defined period (shift/day/week)

3. Output Parameters

Output Parameter	Concept Operational	Definition
1. ED boarding time	Hospital efficiency	Mean time from inpatient bed request to physical departure of patients from the ED overall and by bed type within a defined period (shift/day/week)*
2. ED boarding time components	Hospital efficiency	Mean time from inpatient bed request to physical departure of patients from the ED by bed type by component (bed assignment, bed cleaning, transfer arrival) within a defined period*
3. Boarding burden	Hospital efficiency	Mean number of ED patients waiting for an inpatient bed within a defined period ÷ number of staffed ED treatment areas
4. Hospital admission source, standardized	Hospital efficiency	Number of requests for admission within a defined period (shift/day) overall and by admission source ÷ annual mean requests for admission during that period and adjusted for day of week and season of year†
5. ED admission transfer rate	Hospital efficiency	Number of patients transferred from ED to another facility who would normally have been admitted within a defined period ÷ number of ED admissions within this period
6. Hospital discharge potential	Hospital efficiency	Number of inpatients ready for discharge at or within a defined period ÷ number of hospital inpatients at that time
7. Hospital discharge process interval	Hospital efficiency	Mean interval from discharge order to patient departure from a unit in a defined period (shift/day/week/month)
8. Inpatient cycling time	Hospital efficiency	Mean amount of time required to discharge an inpatient and admit a new patient to the same bed within this period
9. Hospital census	Hospital capacity	Mean number of inpatient beds available by bed type at a defined time ÷ number of staffed inpatient beds by bed type*

Output Parameter	Concept Operational	Definition
10. Hospital occupancy rate	Hospital capacity	Number of occupied inpatient beds overall and by bed type ÷ number of staffed inpatient beds overall and by bed type*
Output measure	Concept operational	Definition
11. Hospital supply/demand status forecast	Hospital capacity	Forecast of expected hospital admissions and discharges as reported daily at 6 AM and compared with hospital census
12. Observation unit census	Hospital capacity	Mean number of available ED observation beds at a defined time ÷ number of staffed ED observation beds
13. ED volume/hospital capacity ratio	Hospital capacity	Number of new ED patients within a defined period (shift/day) ÷ number of available hospital beds at the beginning of analysis period overall and by bed type*
14. Agency nursing expenditures	Hospital capacity	Registered nurse agency nursing expenditures (ED/overall) within a defined period ÷ total nursing expenditures (ED/overall) within this period

*Bed type=ICU/telemetry/psychiatry/ward.

†Admission source=ED/operating room/catheterization laboratory/outpatient/other.