Dimensioning large-scale service systems with returning patients

Sloothaak, F.

Award date:
2015

Technische Universiteit Eindhoven University of Technology
Department of Mathematics and Computer Science

Dimensioning large-scale service systems with returning patients

Fiona Sloothaak

Supervised by:
ir. Teun van den Heuvel (Philips)
prof.dr. Johan S.H. van Leeuwaarden (TU/e)
ir. Britt W.J. Mathijsen (TU/e)

In cooperation with:
PHILIPS
Philips Research

Eindhoven, March 2015

Abstract

Designing cost-effective service processes in hospitals presents tremendous challenges, and due to tightening budgets this topic is very much alive. We introduce a new queueing model that can be used for dimensioning both the number of beds and the number of nurses in an Emergency Department. We scale this system in the Quality-and-Efficiency-Driven regime which can achieve both server efficiency and timely service delivery for patients. Our model accommodates patients who return to service several times during their sojourn, with an additional restriction on the total number of patients that can reside in the system simultaneously. Moreover, we translate our staffing policy into time-varying square-root staffing policies based on the modified offered-load approximation and show that this leads to stable performance.

Acknowledgement

When I started my master, I could have never imagined how these years would turn out. I believe this is an excellent opportunity to thank these people without whom I never would have been able to finish this thesis.

My deepest gratitude goes to my supervisor Johan. Your enthusiasm has let me to love this project and with your guidance, I was able to learn so much. You were and still are an absolute inspiration. Secondly, I would like to thank Britt. Not only did you teach me some simulating skills, but more importantly, you listened to me when I needed it. That made you the perfect office mate and supervisor.

Teun, you provided me the freedom to take charge of this project. I would like to thank you for the opportunity to experience working at Philips and all your advice on project planning. Also, a special thanks goes to Gerrit-Jan and Igor, for letting me peek into the world of healthcare and helping me start up. I must say that your enthusiasm was very catchy.

Jasmijn and Bart, thank you for taking the time to go through this work (I'll admit, it is a lot). I would also like to thank Sem and Edwin, for being on my graduation committee and taking the time to read my thesis. Moreover, thanks to all my friends, my old roommates an my mom for enduring me when I was babbling about my project and other stuff. Also, thanks to the PhD-group for the great atmosphere, conversations, activities and laughters.

Especially, I would like to express my gratitude to Ed en Jeanine for the overwhelming support and love during the time I most needed it. And yes, that also counts for you, Bart.

Contents

Chapter 1

Introduction

We develop a mathematical framework based on stochastic models for queueing systems that can be used in hospital settings to study the trade-off between server efficiency and timely service delivery. Inspired by the situation in a large top clinical hospital in the Netherlands, we introduce a new model that can be used for dimensioning the Emergency Department in terms of both the number of beds and the number of nurses in the Quality-and-Efficiency-Driven (QED) regime.

1.1 Topic orientation

Queueing models are used by many organizations to help determine the capacity levels needed to respond to offered demands in a timely fashion. Delays arise when demand for a service and available capacity differ. For example, patients can encounter delay when no bed is available in the Emergency Department, a surgery cannot be performed yet due to a lack of available operation rooms or follow-up examinations need to be postponed until some diagnostic test results return from the lab. Delay is a crucial aspect for hospitals in terms of having an effective and efficient process and is identified as one of the six key aims for improvement of the quality of healthcare by the Institute of Medicine (IOM) in 2001 [17].

This idea of balancing quality and efficiency complies with the so-called QED regime, a widely used scaling regime in the area of asymptotic optimization of queueing systems. A suitable staffing level, i.e. the number of servers, can be determined by choosing the minimal staffing level such that certain performance levels are met. Alternatively, one can consider some optimization problem that trades off server costs with service quality, where service quality can be quantified in terms of waiting costs, blocking rates or the number of patients in queue. These optimal staffing levels can be approximated by considering the system's asymptotic behavior, i.e. the behavior of the system as the offered traffic and the size of the service system grow large.

Many scenarios in the hospitals are often modeled by classical queueing models, such as the Erlang-C, Erlang-B and the Erlang-A model or more advanced models. The asymptotic behavior of these systems in the QED regime has already been analyzed [20, 26, 27, 14, 6], which gives rise to approximations for the staffing levels in case of a finite number of servers. These staffing levels are often of the so so-called square root staffing form

$$s = R + \beta\sqrt{R}, \tag{1.1}$$

where s is the number of servers, R the offered load and β a tunable parameter. This theory has already been successfully applied to telecommunication systems [6, 14, 28], and more recently also to more complex environments such as hospitals [27, 7, 1, 33].

1.2 Literature overview

Capacity planning in order to reduce delays in healthcare systems has been studied extensively; several papers on this topic can be found in [22]. In particular, [16] explains how basic queueing analysis

in healthcare systems can help to determine more effective policies for bed allocation and staffing and moreover, that it can serve as a key tool in estimating capacity requirements for possible future scenarios.

There exist two recent handbooks on System Scheduling and Operations Management [21, 11]. Both include chapters on capacity planning, nurse scheduling, bed assignment and management and queueing perspectives on healthcare delivery. Most natural and commonly used queueing models involve the Erlang-C, Erlang-B and the Erlang-A model. These relatively simple models can often capture complex realities [14], and serve as a basis of more complex models for healthcare operations.

An extended version of the Erlang-B model is for example considered in [7]. This model can be used to evaluate the current size of a nursing unit and incorporates time-varying arrival rates and non-exponentially distributed service times, based on the work of [4]. Bekker et al. [7] consider the advantages of merging clinical wards as it can result in economies of scales. However, merging can have undesired side effects such as higher waiting times for prioritized patients and the necessity of multi skilled medical staff. Alternatives of merging, such as earmarking, threshold policies and optimization policies are considered in [5].

In addition, Bekker et al. [7] determined the impact of a time dependent arrival pattern on the required number of beds and fraction of refused admissions. They found that the daily pattern, where most patient arrive during office hours, has a limited effect on the average number of occupied beds in the clinical wards, whereas the effect of the week/weekend pattern is much more significant. On the other hand, the daily pattern is crucial for capacity planning in the Emergency Department, since the average length of stay is typically much shorter than the length of stay in a clinical ward.

The Erlang-C model often serves as a basis for queueing analysis in healthcare settings, particularly for modeling purposes in the Emergency Department [17]. The considered delays comprise for example delays for being first seen by a physician in the Emergency Department, delays for a bed in the Emergency department, delays for a medical appointment and delays for nursing care. The main goal of these studies is to find a suitable staffing schedule with a high resource utilization while preserving acceptable waiting times for the patients. A particular method that is used for this is the Lag SIPP method [18], which is based on the Erlang-C model in the QED regime.

In this thesis we will focus on recommendations for staffing levels that are determined by queueing models in the QED regime. This regime balances patients' need for timely service against the economical preference to operate at high efficiency. The queueing models adhere to some form of the square-root staffing rule and it originates from the work of Erlang in 1923 [13].

The square-root staffing rule was first analyzed in the pioneering work of Halfin and Whitt [20] and therefore also referred to as the Halfin-Whitt regime. A first result is shown for the Erlang-C model, indicating that the probability of waiting has a nondegenerate limit as the offered load R grows to infinity under a certain constraint. This constraint gives rise to the square root staffing rule (1.1). In [6], they show that by normalizing the stationary distribution of the queue length, this process converges in distribution to a diffusion process. The Erlang-C formula is replaced by a simpler function that coincides with the Erlang-C formula as R grows large. This yields an approximation of the optimal staffing level for a linear cost function in the QED regime and they show that this procedure yields costs that are optimal to a level of $o(\sqrt{R})$. Moreover, they also provide staffing recommendations for the Efficiency-Driven (ED) and the Quality-Driven (QD) regime. In the Efficiency-Driven regime the staffing costs dominate the waiting costs, whereas the staffing costs are negligible compared to the waiting costs in the Quality-Driven Regime.

The fundamental findings of Halfin and Whitt can be extended to accommodate abandonments, described by the Erlang-A model [14]. Abandonments are an important concept to determine the quality of care in an Emergency Department, since it prescribes the rate at which patients leave without being seen (LWBS) [17]. Garnett et al. provide an extensive list of approximations of possible performance measures and following [6], they provide approximations of optimal staffing levels for the QED, the QD and the ED regimes.

A different approach is taken in [1]. From a data-based queueing science perspective, Armony et al. [1] determine what regimes are relevant for modeling the patient flow in the Emergency Department, patients' length of stay (LOS) in internal wards, and the transfer of the Emergency Department to the Internal Wards. They validate their models by use of data from an Israeli Hospital. Many theoretical and practical research questions are raised, also in order to motivate the reader to perform their own

(exploratory) data analysis. They emphasize the need for carefully setting assumptions on service times and interdependences.

For example, a feature that the commonly used queueing models does not account for is the concept of returns, i.e. patients can return for service after a certain amount of time. This is accommodated in the Erlang-R model, introduced in [33]. The return-to-service phenomenon was first considered by Jennings and De Véricourt in [27]. They used a closed queueing model in order to provide recommendations on nurse-to-patient ratios. These two models will play an important role in our research and we will study them in Chapter 4.

1.3 Problem description

Although delay is a commonly encountered phenomenon in many different settings in healthcare operation, the focus has traditionally been on clinical innovation and not timely delivery of service to improve service quality [17]. More recently, with increasing cuts in healthcare, the need for suitable capacity levels that secure timely service delivery is growing.

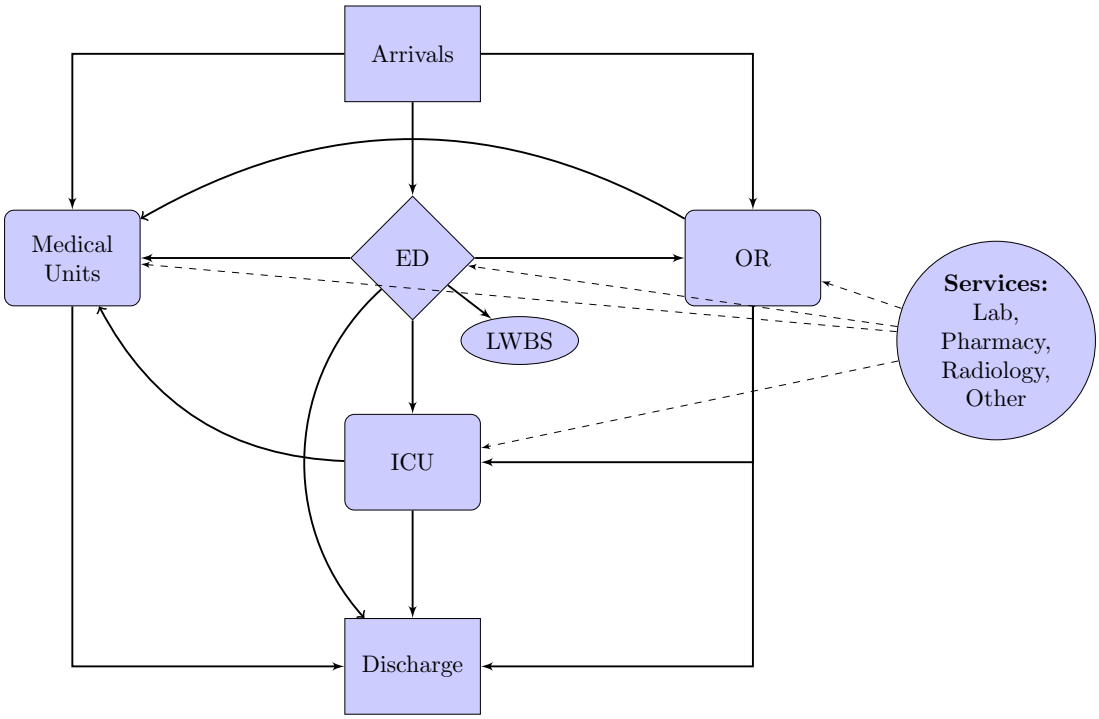


Figure 1.1: General overview of the patient flow within the hospital.

1.3.1 Patient flow in hospitals

Hospitals can be considered as a large complex stochastic network with many processes and many interdependencies. Figure 1.1 provides a general overview of the patient flow between units within a hospital [22, p.16], and particularly maps the dependencies between the Emergency Department (ED), the Operation Rooms (OR), the Intensive Care Unit (ICU) and the other medical units. Interestingly, the patient flow entering the Emergency Department only originates from an external environment. That is, patients from other medical units will never (re)enter the Emergency Department and therefore we can isolate the Emergency Department for our analysis.

The patient flow through an Emergency Department is typically modeled based on a work flow diagram similar to Figure 1.2 [11, p.53] and note that patients can encounter delay at each node. There are roughly two types of patients that enter the Emergency Department. The majority of the arriving

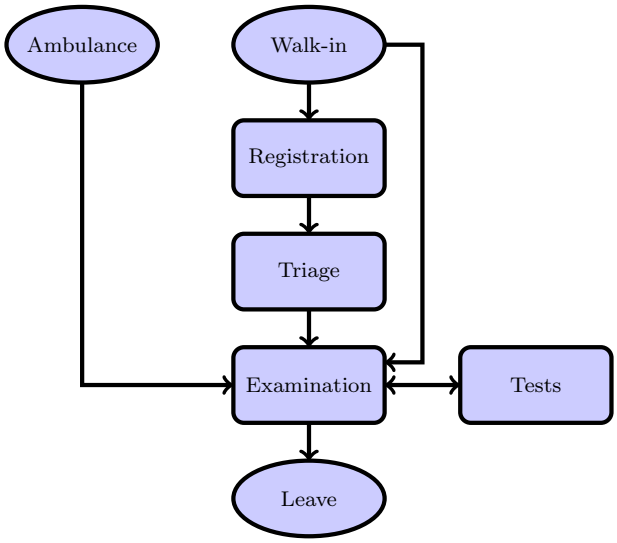


Figure 1.2: Basic work flow through the Emergency Department.

patients walk into to the Emergency Department or are referred by their General Practitioner. After registration, the severity of their condition is established during triage, indicating the urgency of treatment. The other type of patients arrive by ambulance, where the patient's condition is established by the ambulance personnel.

When the urgency of treatment is established, patients need to be assigned to a bed and receive service from the medical staff. This is seen as a major bottleneck where patients experience most delay. Patients assigned to a bed are frequently visited by a doctor or a nurse who take tests, conduct physical examinations, consider medical histories and possibly provide additional care. After a diagnosis is established, patients can be discharged or referred for further medical treatment. The main goal in an Emergency Department is to diagnose patients, which entails determining possible further treatment or help in case only minor treatment is needed.

1.3.2 Research challenges

Queueing analysis based on the Erlang-C model is often used for modeling purposes in the Emergency Department. However, the Erlang-C model does not account for the essential feature that patients return for service several times before leaving the system. The Erlang-R model [33] is the first model that incorporates this crucial phenomenon in hospital settings. This model provides an analysis to determine the number of nurses needed to achieve a predetermined service level, which typically comprises predetermines values for the waiting probabilities, waiting times or nurse utilization levels.

In collaboration with a top clinical hospital in the Netherlands, we found that the number of beds creates an additional bottleneck on the service levels. When all beds are occupied, the other patients wait in a holding room before entering the examination phase. Therefore, we introduce an extension of the Erlang-R model with the essential feature that restricts the total number of patients that can reside in the Emergency Department simultaneously and call it the Semi-Open Erlang-R Model with Waiting (SERW).

The model is a three-station open queueing network, with a restriction on the number of present patients in two of the three stations, see Figure 1.3. More explicitly, we consider a model where patients require service from a nurse. When all beds are occupied upon arrival, patients wait in a holding room till another patient leaves the system and a bed becomes available. Once a bed becomes available, the first patient holding will be assigned to this bed and moves for service. After a service completion, with probability $1-p$ they exit the system and with probability p they return for further service after a random delay time.

We are interested in the stationary distribution of the number of patients at each station. While

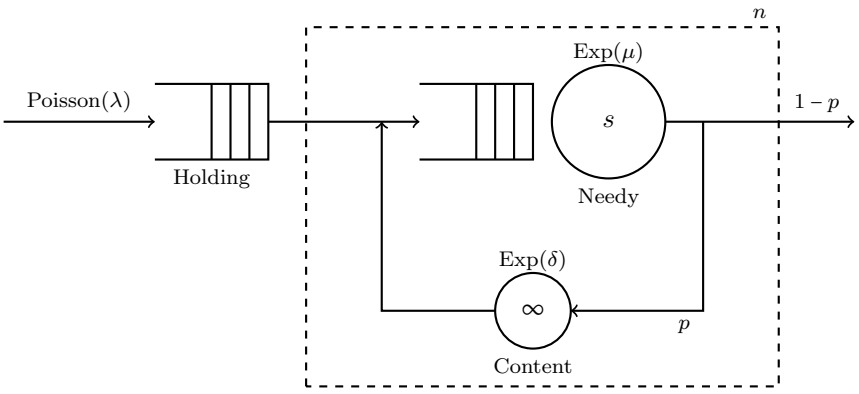


Figure 1.3: The Semi-Open Erlang-R Model with Waiting.

the Erlang-R model is a two-station Jackson network with a product-form solution for the stationary distribution, our model presents more mathematical difficulties due to the restriction on the number of simultaneously treated patients. We will study how this system can be modeled as a quasi-birth-death process and how the stationary distribution can be obtained using matrix-geometric methods.

We seek capacity management policies in terms of the number of beds n and the number of servers s in the QED regime. Interestingly, our model requires the simultaneous scaling of both n and s in order to obtain QED behavior. Since the stationary distribution cannot be written in a closed-form expression, we make stochastic comparisons with the Erlang-R model and the closed ward model by Jennings and De Véricourt [26]. Using these results, we identify a two-fold scaling policy that can be used for dimensioning the Emergency Department in the QED regime.

We also apply time-varying diffusion approximations, and translate the two-fold scaling into time-varying square-root staffing policies based on the modified offered-load (MOL) approximation. We will study whether this will lead to stable performance via simulations.

Our main objective is to provide a staffing policy for the Semi-Open Erlang-R Model with Waiting, and provide a corresponding dimensioning scheme in terms of the number of nurses and the number of beds in an Emergency Department.

1.4 Readers' guide

This thesis is divided in three parts. Part I considers the classical models and how they can be applied in hospital settings. In particular, in Chapter 2 we explain basic queueing theory and describe the classical queueing models in detail, and consider specific applications of the classical queueing models in hospitals. Chapter 3 describes the philosophy of the QED regime. Moreover, we will see how to dimension the classical models in the QED regime, including concrete examples in hospital settings.

Part II describes the more advanced queueing models motivated by healthcare systems. Chapter 4 provides an overview of existing models, such as the closed ward model [26], the Erlang-R model [33] and the Semi-Open Erlang-R Model with Blocking [32], whereas Chapter 5 introduces our new SERW model. It is illustrated how this model can be modeled as a quasi-birth-death process, and how matrix-geometric methods can be used for solving the stationary distribution. Chapter 5 is concluded by proposing a staffing policy based on an underlying intuition. In Chapter 6 we compare our new SERW model with the other models presented in Chapter 4.

Part III validates our proposed staffing algorithm by means of simulations. Chapter 7 shows that performance stabilizes and provides an overview of managerial insights in the stationary case. Moreover, we apply time-varying diffusion approximations, and translate the two-fold scaling into time-varying square-root staffing policies based on the modified offered-load (MOL) approximation. We compare it to the Pointwise Stationary Approximation (PSA) and show that only the MOL staffing procedure stabilizes the probability of waiting. We conclude this thesis in Chapter 8 by providing a dimensioning scheme that can be used for the SERW model. This practical algorithm determines the number of nurses needed

such that a predetermined value for the probability of waiting is achieved.

This thesis is written for a varied audience with different backgrounds. Here we will provide a short readers' guide for different readers, and leave it to the reader to identify the profile best fitted for her.

Students:
For students that are interested in how queueing theory can be applied in hospital settings and/or dimensioning purposes the complete thesis is of interest. If the student is unfamiliar with probability theory or queueing theory we suggest to read Appendix A first. Chapters 2 till 5 will suffice in order to obtain a general overview of this area, and proofs are included for the readers' interest. If the student is interested in the practical aspects of our model, Part III is of more interest and in Appendix B we provide a simulation program.

Researchers:
Fellow researchers familiar with the area of asymptotic optimization of queueing systems, particularly in hospital settings, can skip Chapters 2-4.

Engineers:
We think that examples of dimensioning systems in hospitals throughout this thesis can be of interest, and particularly, the example provided in Section 8.5. Our conclusions in Chapter 9 will highlight the most important findings of this thesis, and future research suggests some natural extensions of our model.

Part I

Classical patient flow models

Chapter 2

Classical queueing models

To determine appropriate capacity levels in hospitals, many studies rely on classical queueing models such as the Erlang-C, the Erlang-B and the Erlang-A model. In this chapter, we will describe these queueing models and consider how they can be used in a hospital setting. We refer to Appendix A for a short overview of some relevant basic concepts of probability theory and queueing theory that are used in this thesis.

2.1 Queueing models

In a basic queueing model, customers arrive at a service facility that provides service for a certain amount of time, after which customers leave the system. The service facility can consist of a single, multiple or even a network of servers. In healthcare delivery systems, patients are typically seen as the customers of the system. The beds in an internal ward, the Emergency Department or the Intensive Care Unit (ICU) can be viewed as the service facility for example. However, many atypical examples of patients and service facilities can be named as well, e.g. [11, p.19].

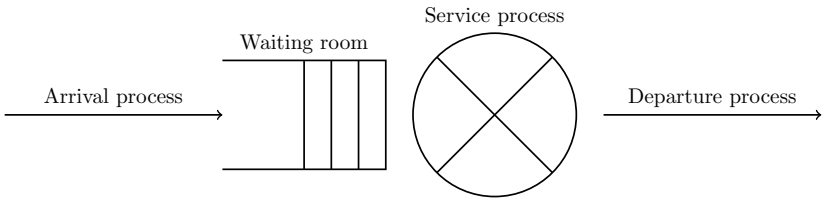


Figure 2.1: Representation of a basic queueing model.

Figure 2.1 illustrates a basic queue. In general, a queueing model is characterized by several features.

- *The arrival process.* This process describes when patients arrive over time. Typically this process is considered stochastic with a Poisson process as the traditional example.
- *The behavior of patients.* Arriving patients may be impatient and leave before service. On the other hand, they can also be very patient and wait indefinitely. The latter is for example a more realistic setting for patients that arrive for surgery.
- *The service times.* This process describes the service durations of a patient.
- *The service discipline.* This feature describes the way in which patients are served. Patients can be served one by one or in batches. Also, the order in which patients are served can be specified. That is, patients can be served in order of arrival (first come first served), randomly, last come first served, or with certain priorities (emergency arrivals first).
- *Service capacity.* The maximum number of servers available for serving the patients.

- *The waiting room.* This describes the number of patients that are allowed to wait. The capacity of the waiting room can be infinite, but sometimes there are limitations with respect to the number of patients in the queue.

A queueing model is often described using the so-called Kendall notation A/B/c. It is a shorthand notation to characterize a range of queueing models. The first letter A specifies the interarrival time distribution, the second letter B represents the service time distribution and the third letter c represents the number of servers. The three letters M, G and D are mostly used to describe the distribution of the interarrival times and the service times. The M represents the exponential distribution, G stands for a general distribution and D stands for a deterministic distribution. Sometimes the notation is extended by a fourth letter, representing a limitation on the number of patients in the waiting room. In the most basic queueing model, patients arrive one by one and are always allowed to enter the system (i.e. infinite waiting room), there are no priority rules and patients are served in order of arrival.

Queueing models are often used to analyze certain performance measures, such as waiting times and utilization levels. In this chapter, we will consider three typical queueing models and consider their properties.

2.2 Birth-death processes

The birth-death process serves as a building block of the Erlang-C, the Erlang-B and the Erlang-A model.

Definition 2.2.1. *A birth-death process is a continuous-time Markov process with state space $S = \{0, 1, ..., I\}$, where I can be infinite, and all transitions only occur between two adjacent states. That is, the only transitions from state i are to state $i-1$ (death) and to state $i+1$ (birth).*

An illustration of the birth-death process is given in Figure 2.2.

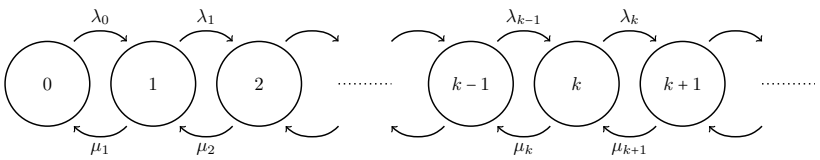


Figure 2.2: Transition diagram of a birth-death process.

We are interested in the stationary distribution of the birth-death process being in a certain state (when the stationary distribution exists). In other words, we want to find the probability π_i that the process is in state i. Let λ_i and μ_i represent the transition rate from state i to $i+1$ (birth) and i to $i-1$ (death) respectively.

Proposition 2.2.2. *If*

$$\sum_{i=0}^{I} \sum_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} < \infty,$$

then the stationary distribution exists and is given by

$$\pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \tag{2.1}$$

with

$$\pi_0 = \left[\sum_{i=0}^{I} \sum_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right]^{-1}. \tag{2.2}$$

The proof of Proposition 2.2.2 can be found for example in [10, p.74] and for completeness we included it in Appendix C. In the next sections, we will see how this process can be used to determine the stationary distributions of some classical queueing models.

2.3 Erlang-C model

This model is also referred to as the $M/M/s$ model. The first M refers to the arrival process, which is Poisson with rate λ, where λ is the expected number of arrivals in one time unit.

The second M refers to the service distribution, that is exponentially distributed service with rate μ. Moreover, there are s servers available and each server can help one patient at a time. If an arriving patient finds all servers occupied, he will wait till one is available. Patients are served in accordance with the FCFS (First Come First Served) discipline. An illustration of this model is given in Figure 2.3.

Denote the load intensity by $\rho = \lambda/(\mu s)$, which is equal to the fraction of time that a server is busy. Thus for stability we must have $\rho < 1$.

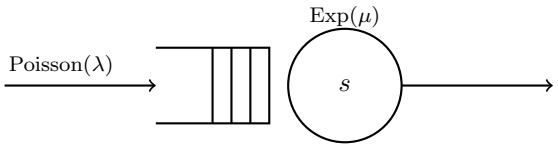


Figure 2.3: The Erlang-C model.

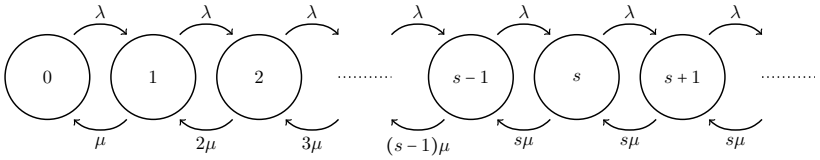


Figure 2.4: Transition diagram of the Erlang-C model.

2.3.1 Properties

We are interested in the number of patients in the system, which we refer to as the queue length. So if the process is in state i, then there are i patients in the system, either waiting or being served. Let π_i denote the stationary probability of the queue length.

Proposition 2.3.1. *The stationary distribution of the queue length in the Erlang-C model is given by*

$$\pi_i = \begin{cases} \pi_0 \frac{(\rho s)^i}{i!} & \text{for } 0 \le i \le s-1, \\ \pi_0 \frac{\rho^i s^s}{s!} & \text{for } i \ge s \end{cases} \tag{2.3}$$

with

$$\pi_0 = \left[\sum_{j=0}^{s-1} \frac{(\rho s)^j}{j!} + \frac{(\rho s)^s}{(1-\rho)s!} \right]^{-1}. \tag{2.4}$$

This and the following results can be found in [10, pp.90-92] and we rederive Proposition 2.3.1 in Appendix C. An important quantity is the Erlang-C formula that defines the probability that an arriving patient finds all servers occupied, which we dote by $C(s, \rho)$. Since the PASTA property holds for patients arriving according to a Poisson process, we have

$$C(s,\rho) = \sum_{i=s}^{\infty} \pi_i = \frac{(\rho s)^s}{(1-\rho)s!} \pi_0. \tag{2.5}$$

Equivalently, denoting the offered load by $R = \rho s = \lambda/\mu$, we obtain

$$C(s,R) = \frac{s}{s-R} \frac{R^s}{s!} \pi_0$$

with

$$\pi_0 = \left[\sum_{j=0}^{s-1} \frac{R^j}{j!} + \frac{s}{s-R} \frac{R^s}{s!} \right]^{-1}.$$

Lemma 2.3.2. *The following recursion holds for the Erlang-C formula:*

$$C(s+1,R) = \frac{(s-R)RC(s,R)}{s(s+1-R) - RC(s,R)} \tag{2.6}$$

with $C(0,R) = 1$.

Recursion (2.6) proves to be very useful for numerical computations.

2.3.2 The Emergency Department

Providing timely service is important in order for the patient's condition not to worsen. To reach this goal for the Emergency Department, typically the Erlang-C model is used.

It has been shown that patient arrivals to the Emergency Department are well approximated by a time-homogeneous Poisson process [22, 21, 18, 4]. Moreover, it is reasonable to assume patients originate from a large population, arrive one at a time and independently of one another.

In [18], the Erlang-C model is used with arrival rates dependent on the hour of the day and delay measured from the moment of admission/registration until the first moment a provider (physician, or other) is available to see the patient. Therefore, the waiting time of the queueing model does not equal the actual waiting time experienced by the patient since the triage process is included in the delay. Nevertheless, by using certain properties of this model one can provide a high-level approach to determine an appropriate staffing level with respect to the number of care providers.

In addition, delays can arise due to an insufficient number of beds. This is named as the single biggest cause of Emergency Department overcrowding according to the American Hospital Association [2, 17]. An insufficient number of beds interferes with the ability of physicians and nurses to take care of their patients and even worse, it can cause the condition of patients to worsen throughout this delay. In this context, the beds are considered as the servers of the system and delay arises when there is no bed available after the patient is triaged.

2.4 Erlang-B model

This model is also referred to as the $M/M/s/s$ model, or the Erlang loss model. Again, patients arrive according to a Poisson process with rate λ and require exponential service time with rate μ. However, in this case when patients find all servers occupied upon arrival, they will leave the system immediately without being served, see Figure 2.5. Therefore we do not have a stability condition in this setting. Again, let $\rho = \lambda/(\mu s)$ denote the load intensity and $R = \lambda/\mu$ the offered load.

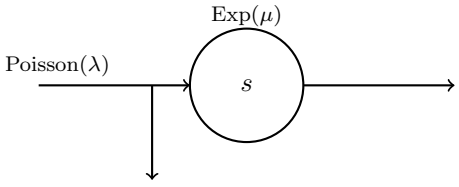


Figure 2.5: The Erlang-B model.

2.4.1 Properties

We are interested in the number of patients in the system. Let π_i denote the stationary probability of the queue length.

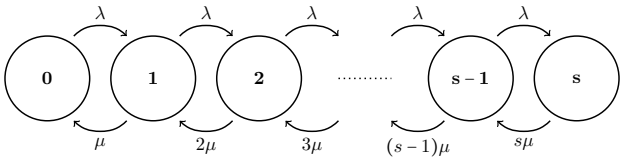


Figure 2.6: Transition diagram of the Erlang-B model.

Proposition 2.4.1. *The stationary distribution of the queue length in the Erlang-B model is given by*

$$\pi_i = \pi_0 \frac{(\rho s)^i}{i!} \quad \text{for } 0 \leq i \leq s \tag{2.7}$$

with

$$\pi_0 = \left[\sum_{j=0}^{s} \frac{(\rho s)^j}{j!} \right]^{-1}. \tag{2.8}$$

Contrary to the Erlang-C model, we do not have patients waiting. Instead, an important quantity is the probability that an arriving patient finds all servers occupied and will be blocked, denoted by $B(C, \rho)$. By the PASTA property this equals

$$B(s, \rho) = \pi_C = \pi_0 \frac{(\rho s)^s}{s!}.$$

This can also be expressed in terms of R. Then

$$B(s, R) = \pi_0 \frac{R^s}{s!} \tag{2.9}$$

with

$$\pi_0 = \left[\sum_{j=0}^{s} \frac{R^j}{j!} \right]^{-1}.$$

Lemma 2.4.2. *The following recursion holds for the Erlang-B formula:*

$$B(s+1, R) = \frac{R\, B(s, R)}{R\, B(s, R) + s + 1} \tag{2.10}$$

with $B(0, R) = 1$

Recursion (2.10) will be very useful for numerical computations. We will also provide an expression for the relation between the Erlang-B and Erlang-C formula.

Lemma 2.4.3. *The following identity holds:*

$$C(s, R)^{-1} = \rho + (1 - \rho) B(s, R)^{-1}. \tag{2.11}$$

The proofs of the Proposition 2.4.1 and the two lemmas can be found in [10, pp.79-81], and we included them in Appendix C.

2.4.2 Bed capacity in clinical wards

One way to measure the capacity within a clinical ward is in terms of the number of operational beds. This is mostly done via a staffing ratio per operational bed [7], which is generally fixed and evaluated on a yearly basis.

A structural model of the patient flow through a clinical ward is shown in Figure 2.7. A patient arrives to a clinical ward and if the ward is fully occupied, the patient is refused and redirected to a less

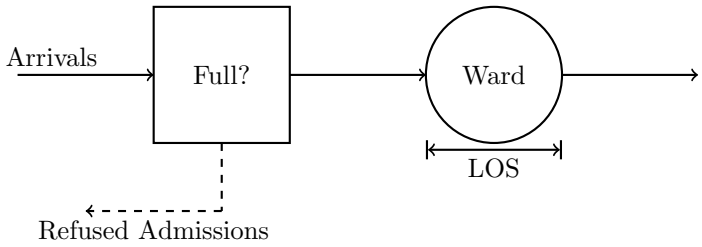


Figure 2.7: Structural model of patient flow through a clinical ward.

preferable ward or another hospital. If there is a bed available for the patient, she will be treated here for a certain amount of time, after whish the patient will leave the ward (discharged, redirected, etc.). The amount of time the patient is in the ward is referred to as the patient's Length of Stay (LOS).

When we assume a Poisson arrival process, an exponentially distributed LOS with a fixed rate and a FCFS service discipline, this corresponds to an Erlang-B model. A patient refused due to unavailable beds can be considered 'lost' or 'blocked'. This undesired result can be taken as a performance measure of the system. In practice, refused patients are diverted to another hospital or diverted to a non-preferable ward. This number is hard to approximate in practice [7].

However, for some clinical wards modeling the arrival process by a Poisson process is not appropriate. For instance, wards with a high percentage of scheduled admissions, such as Hematology and Ophthalmology cannot be modeled with a Poisson arrival process. Moreover, the Erlang-B model also makes the assumption of an exponentially distributed service time. This is an invalid choice in practice, since LOS distributions are often characterized by heavy tails and high variability [8]. For example, the coefficient of variation is often larger than one, while for the exponential distribution the coefficient equals one. This effect is even more significant when there are many patients ready to be transferred but no available beds, and thus remain at the current ward waiting for a free bed. An Erlang loss model can still be used for this scenario, but requires a more elegant treatment of the service distribution. This is for example done in [7].

2.5 Erlang-A model

Suppose an arriving patient finds all servers occupied. In the Erlang-C model the patient will wait until a server is available, while in the Erlang-B model the patient leaves immediately. In practice, the patient might wait for a certain amount of time and if no server becomes available during this time, the patient abandons the system. This model is referred to as the $M/M/s+M$ model, or the Erlang-A model. Observe that the number of servers is given by $s+M$ in Kendall's notation. That is, the capacity of the system is given by s servers plus an additional exponentially distributed time that patients are willing to wait for service.

A structural representation is given in Figure 2.8. Again, patients arrive according to a Poisson process with rate λ and require exponential service times with parameter μ. In this case, arriving patients finding all servers occupied will wait for an exponentially distributed time with rate θ. If no server becomes available during this period, the patient abandons the system. Note that there is no stability condition in this setting. Again, let $\rho = \lambda/(\mu s)$ and $R = \lambda/\mu$. An illustration of the transition rates is given in Figure 2.9.

2.5.1 Properties

We are interested in the number of patients in the system. Let π_i denote the stationary probability of the queue length.

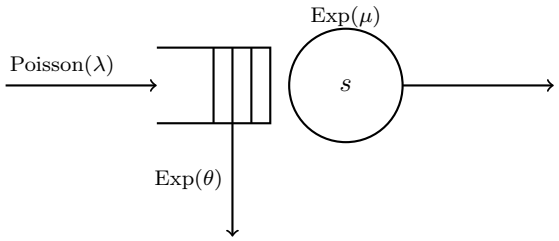


Figure 2.8: The Erlang-A model.

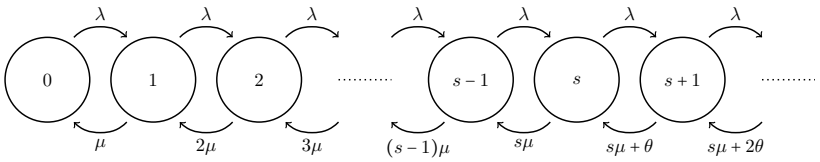


Figure 2.9: Transition diagram of the Erlang-A model.

Proposition 2.5.1. *The stationary distribution of the queue length in the Erlang-A model is given by*

$$\pi_i = \begin{cases} \pi_0 \frac{(\rho s)^i}{i!} & \text{for } 0 \le i \le s, \\ \pi_0 \frac{(\rho s)^s}{s!} \prod_{j=s+1}^{i} \frac{\lambda}{s\mu + (j-s)\theta} & \text{for } i \ge s \end{cases} \tag{2.12}$$

with

$$\pi_0 = \left[\sum_{j=0}^{s} \frac{(\rho s)^j}{j!} + \sum_{j=s+1}^{\infty} \frac{(\rho s)^s}{s!} \prod_{j=s+1}^{i} \frac{\lambda}{s\mu + (j-s)\theta} \right]^{-1}. \tag{2.13}$$

Observe that this distribution is not of a closed form, since it includes an infinite sum that can cause numerical problems. In [28] an alternative way to express the stationary distribution is deduced, which can be used for numerical computations. Recall that the Erlang-B formula $B(s, R)$ denotes the blocking probability in the Erlang-B model with s servers and load intensity ρ.

Lemma 2.5.2. *Let*

$$A(x, y) = 1 + \textstyle\sum_{j=1}^{\infty} \frac{y^j}{\prod_{k=1}^{j}(x+k)}, \quad x > 0, y \ge 0.$$

The following holds for the Erlang-A formula:

$$\pi_i = \begin{cases} \pi_s \cdot \frac{s!}{i!(\rho s)^{s-i}} & \text{for } 0 \le i \le s, \\ \pi_s \cdot \frac{(\lambda/\theta)^{i-s}}{\prod_{j=1}^{i-s}\left(\frac{s\mu}{\theta} + j\right)} & \text{for } i \ge s \end{cases} \tag{2.14}$$

with

$$\pi_s = \frac{B(s, \rho)}{1 + \left[A\left(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}\right)\right] \cdot B(s, \rho)}. \tag{2.15}$$

2.5.2 LWBS patients

Studies show that there exists a strong link between long waiting times and the fraction of patients that leave without being seen (LWBS), particularly in case of the Emergency Department [17]. Those patients can still be in need of care and it is not uncommon they return to be hospitalized within a short period, resulting in an overall decreasing quality of care.

Under the assumption that patients do not leave once they are in service, the Erlang-A model can be used to account for the fraction of LWBS patients. Although both the Erlang-C model and the Erlang-A model cannot capture all characteristics of an actual operational setting, many studies demonstrated the usefulness of these models in providing decision support that can greatly improve performance, e.g. [18].

Chapter 3

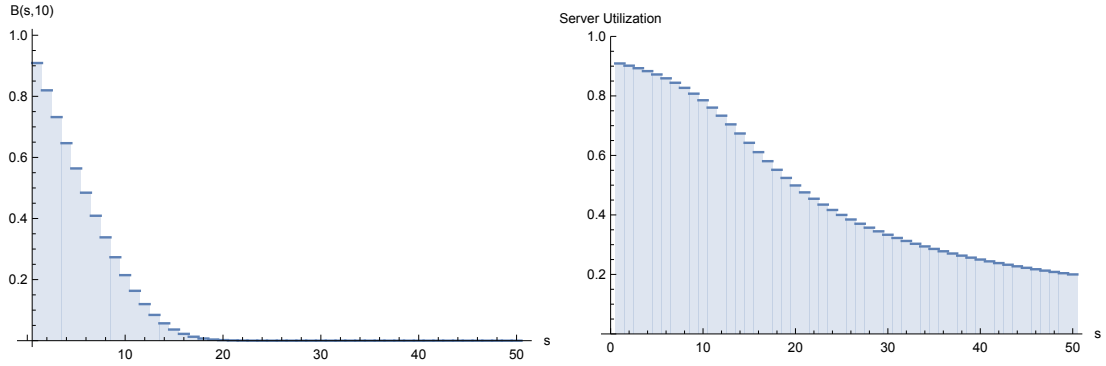
The QED regime

Hospitals would like to be both efficient in view of budgets and offer good quality of service to their patients. In this chapter we discuss the mathematical theory of the Quality-and-Efficiency-Driven (QED) regime, which can achieve this dual goal.

A main question in this area involves matching capacity and demand of the queueing model according to a certain rule of thumb. More concretely, we want to find a suitable number of servers when the system operates in a certain regime. In this chapter we will provide some results for the classical models introduced in Chapter 2 and consider them in hospital settings.

3.1 Introduction

Recall that in a basic queueing model patients arrive to a service facility that provides service for a certain amount of time, after which the patient leaves the system. The number of servers, which we refer to as the staffing level, directly influences the level of efficiency and the level of quality of the system. To provide an intuitive feeling of this concept, we consider the Erlang-B model with $R = 10$. More servers will result in a smaller fraction of patients that are blocked for service, i.e. a higher quality of service. However, this is at the expense of the efficiency since the utilization level of the servers decreases as the number of servers grows, see Figure 3.1.



(a) Blocking probability as a function of s.

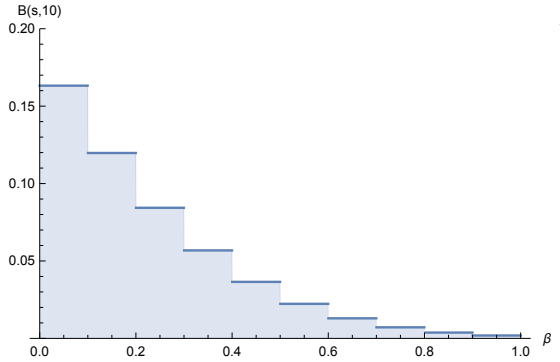
(b) Utilization level as a function of s.

Figure 3.1: Effects of s for the Erlang-B model.

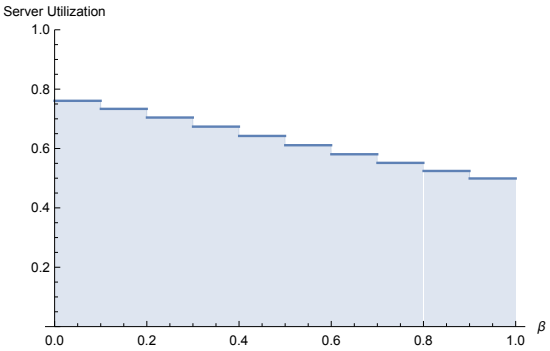
This raises the question of how many servers are needed such that a predetermined performance is achieved. It seems reasonable that the capacity should at least exceed the offered load in order to manage the demand, but in what way do we choose this slack?

Within asymptotic optimization of queueing systems, usually three different regimes are distinguished. The first regime is called the Quality-Driven (QD) regime, where the capacity exceeds the mean demand

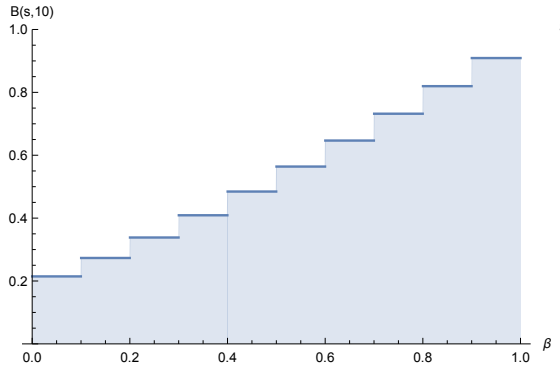
considerably. For many queueing systems the staffing rule can be written in the form $s = R+\beta R = (1+\beta)R$, where s is the number of servers and R the offered load. In other words, the capacity exceeds the offered load by a fixed percentage β. Then, the number of servers is an increasing function of β and both the probability of blocking and the utilization level are decreasing functions of β. In general, an important characteristic of this regime is that the probability of blocking or waiting converges to zero as the offered load R (and thus the staffing level) grows to infinity. On the other hand, the utilization level does not converge to one.



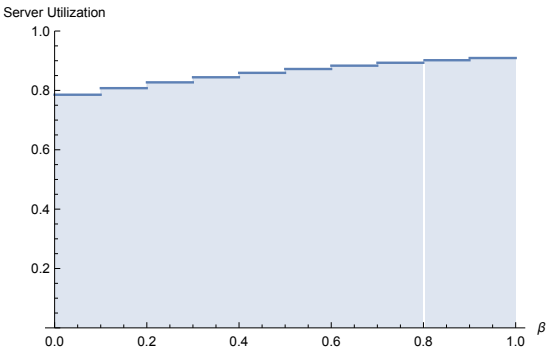
(a) Blocking probability as a function of β in the QD regime.



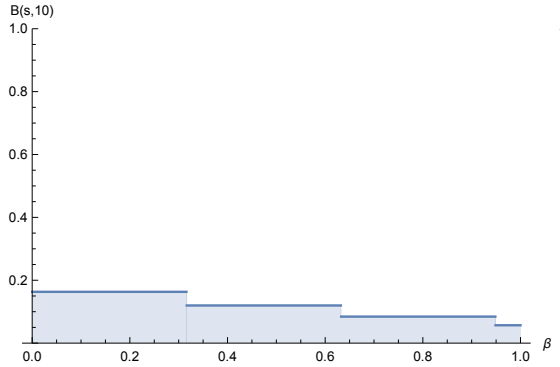
(b) Utilization level as a function of β in the QD regime.



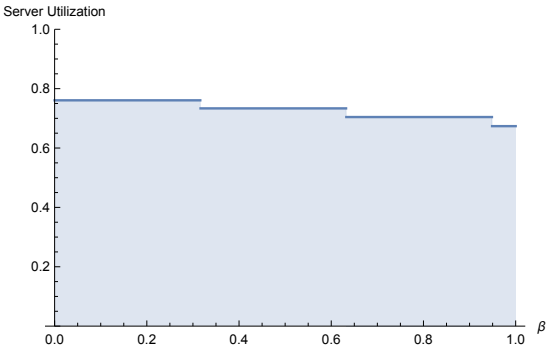
(c) Blocking probability as a function of β in the ED regime.



(d) Utilization level as a function of β in the ED regime.



(e) Blocking probability as a function of β in the QED regime.



(f) Utilization level as a function of β in the QED regime.

Figure 3.2: Effects of β for the Erlang-B model.

The second regime is called the ED-regime (Efficiency-Driven), where a high utilization of the servers is of greater importance than the probability of waiting or blocking. In some queueing models the capacity even falls short of the offered load by some percentage. For the Erlang-B model, the staffing rule can be written as $s = R-\beta R = (1-\beta)R$. Then s is a decreasing function of β and both the probability

of blocking and the utilization level are increasing functions of β. In general, the typical characteristic of this regime is that the utilization level converges to one as well as the probability of blocking or waiting as R grows to infinity.

In our situation we would like to balance the efficiency and the quality of the system. In other words, we would like to have a high occupation rate, while the probability of waiting or blocking remains reasonable. This can be achieved with the QED-regime (Quality-and-Efficiency-Driven). In this regime, for many queueing models the staffing rule is some form of the so-called square root formula, given by

$$s = R + \beta\sqrt{R}.$$

We find that s equals the offered load R plus some hedge $\beta\sqrt{R}$. The constant β can be seen as a quality parameter that determines how high the quality, and at the same time, how costly the system is. Characteristic of the QED regime is that the probability of blocking converges to zero as $R \to \infty$ at a rate such that the utilization level converges to one.

For the Erlang-B model, Figure 3.2 illustrates the effect of β on the probability of blocking and the utilization level for the three regimes. Moreover, Table 3.1 summarizes the performance characteristics for the three different regimes, where ρ is the utilization level. This concept is very powerful when

Regime	Staffing Level	Quality	Efficiency
QED	$s = R + \beta\sqrt{R}$	$P(\text{Block}) \to 0$	$\rho \to 1$
QD	$s = R + \beta R$	$P(\text{Block}) \to 0$	$\rho \to 1/(1+\beta)$
ED	$s = R - \beta R$	$P(\text{Block}) \to \beta$	$\rho \to 1$

Table 3.1: System Performance for three regimes for the Erlang-B model.

used for determining suitable staffing levels in a queueing system. We will demonstrate this in the next sections for the classical queueing models described in Chapter 2.

3.2 Erlang-C model

The Erlang-C model is often used for modeling the Emergency Department of a hospital [17]. The goal is to provide a staffing level in the QED-regime.

3.2.1 QED behavior

First we construct a framework in which we can consider the asymptotic behavior of this queueing system. We consider a sequence of $M/M/s$ queues indexed by some integer n. We will adopt the notation used in Section 2.3, subscribed with index n. In our framework we have that the arrival rate $\lambda_n \to \infty$ as $n \to \infty$. Moreover, let $\mu_n = \mu$, $s_n = n$ and $\rho_n = \lambda_n/(n\mu)$. Note that for stability we must have $\rho_n < 1$ for all $n \in \mathbb{N}$.

Within this framework, it turns out that the probability of delay has a nondegenerate limit as $\rho_n \to 1$ with a certain rate. Next we will formalize this notion, from which the square root staffing rule can be obtained.

Let Φ and ϕ denote the standard normal cumulative distribution and density respectively. Let Q_n denote the number of patients in the n'th system (in its stationary state). The following theorem will formally state under what condition the probability of delay has a nondegenerate limit.

Theorem 3.2.1 (Halfin & Whitt [20])**.** *The probability of delay has a nondegenerate limit, i.e. a limit strictly between zero and one, if and only if*

$$\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta, \quad \beta > 0.$$

In this case we have

$$f(\beta) := \lim_{n\to\infty} P(Q_n \geq n) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} = \frac{\phi(\beta)}{\phi(\beta) + \beta\Phi(\beta)}. \tag{3.1}$$

The proof provided in [20] makes use of the CLT (Central Limit Theorem) and Stirling's approximation and can be found in Appendix D.

The limiting result in (3.1) gives rise to a simple approximation for the probability of delay of a finite $M/M/s$ queue. Note that from $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta$ it follows that $\rho_n \to 1$ as $n \to \infty$. Moreover $n = s_n$, thus fix β by setting

$$\beta = \frac{1-\rho}{\sqrt{\rho}}\sqrt{s}.$$

Equivalently, we derive the staffing rule

$$s = R + \beta\sqrt{R}. \tag{3.2}$$

In conclusion, Theorem 3.2.1 states that the probability of delay will converge to a probability strictly between zero and one when the square root staffing rule (3.2) is chosen.

3.2.2 Dimensioning scheme

To illustrate the power of Theorem 3.2.1 we consider two possibilities how one can dimension this system, i.e. determine the number of servers, in the QED regime. First, this can be done via a constraint based problem. For example, suppose that the fraction of patients that must wait for service cannot exceed $\alpha \in (0,1)$. Then what is the minimal number of servers to meet this performance level?

We can determine the parameter β by inverting the equation

$$\alpha = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}.$$

This equation is given in Figure 3.3. Then from the square root staffing rule (3.2) follows how many servers are needed to meet the desired performance level.

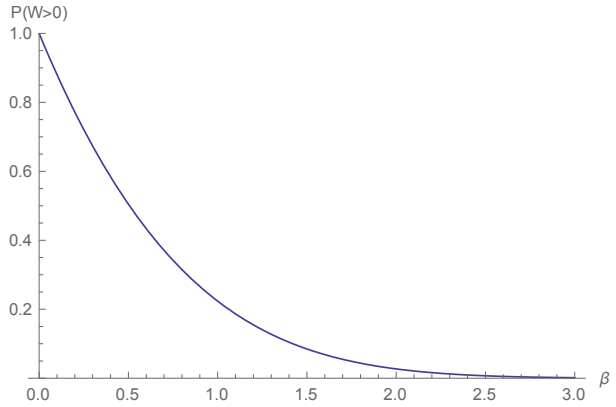


Figure 3.3: The function $f(\beta)$ defined in (3.1).

Secondly, we can dimension via a cost optimization problem, where a trade-off is made between server costs and waiting costs. The goal is to minimize the long term average costs. For example, suppose a server costs c_1 per time unit and a patient waiting costs c_2 per time unit. Let W be the stationary waiting time of an arriving patient and let $E(Q^W)$ be the expected number of patients that wait for service. Denote

$$f(\beta) := \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} \approx P(W > 0).$$

Then the total cost is given by the function $c_1 s + c_2 E(Q^W)$. By Little's law $E(Q^W) = \lambda E(W)$ and moreover, it can be shown (see e.g. [10, pp. 96-97]) that

$$P(W > t | W > 0) = e^{(1-\rho)s\mu t}.$$

Then,

$$\begin{aligned} E(W) &= E(W|W>0)P(W>0) = P(W>0)\int_0^\infty P(W>t|W>0)\,dt \\ &= P(W>0)\int_0^\infty e^{(1-\rho)s\mu t}\,dt \approx f(\beta)\cdot\frac{1}{(1-\rho)s\mu} \\ &= \frac{f(\beta)}{(s-R)\mu} = \frac{f(\beta)}{\beta\sqrt{R}\mu}, \end{aligned}$$

and therefore

$$\begin{aligned} c_1 s + c_2 E(Q^W) &= c_1(R+\beta\sqrt{R}) + \frac{c_2\lambda f(\beta)}{\beta\sqrt{R}\mu} \\ &= c_1(R+\beta\sqrt{R}) + \frac{c_2\sqrt{R}f(\beta)}{\beta} \\ &= c_1 R + \sqrt{R}\left(c_1\beta + \frac{c_2}{\beta}f(\beta)\right). \end{aligned}$$

Thus, to minimize the total cost we need to find

$$\arg\min_{\beta>0}\left\{c_1\beta + \frac{c_2}{\beta}f(\beta)\right\}.$$

This provides the asymptotically optimal β and from the square root staffing rule (3.2) we obtain the asymptotic optimal number of servers that minimizes the costs.

We note that the asymptotic probability of waiting is an approximation of the true value in a finite system. Using the analytic results of Chapter 2, we can also these address staffing problems via exact optimization. However, this lacks insight in the dependencies between the parameters and particularly how the number of servers relates to the offered load.

Summarizing, for the Erlang-C model to operate in the QED regime, Theorem 3.2.1 states that we need to apply the square root staffing rule. From this, we can determine the staffing level to meet a certain performance measure. In addition, it can be used to solve asymptotic optimization problems in order to find the corresponding optimal staffing level.

3.2.3 Dimensioning examination rooms

To provide a concrete example, we will consider how one can determine the number of examination rooms in an Emergency Department such that a predetermined value for the waiting time is achieved. We consider an average arrival stream on the number of patients in the Emergency Department of a large top clinical hospital in the Netherlands during the first half year of 2014, see Table 3.2. This particular hospital assigns examination rooms to two different types of patients, namely patients with low acuity conditions (Fast-Track) and patients whose diagnosis is often less evident upon arrival (High-Care). The service time corresponds to the time that the patient is in the examination room.

Parameters	Total	Fast-Track	High-Care
#Patients	4381	226	4155
#LWBS	172	20	152
λ	3.458	0.178	3.279
μ	0.403	0.632	0.395
λ_{LWBS}	3.322	0.163	3.159
θ	0.214	0.483	0.200

Table 3.2: Parameters from data and observations.

Suppose we want to know how many examination rooms are needed such that the expected waiting time is below a certain threshold. Then by the techniques described in the previous section we know that

$E(W) \approx f(\beta)/(\beta\sqrt{R}\mu)$, where f is the asymptotic probability of waiting. This gives rise to parameter β and hence the recommended staffing levels, which are displayed in Table 3.3.

$E(W)$ (in min)	β_{Total}	n_{Total}	β_{FT}	n_{FT}	β_{HC}	n_{HC}
5	1.2987	13	1.7493	2	1.3120	13
10	1.0460	12	1.5033	2	1.0592	12
15	0.9006	12	1.3565	2	0.9135	11
20	0.8002	11	1.2515	1	0.8127	11
25	0.7245	11	1.1700	1	0.7367	11
30	0.6644	11	1.1036	1	0.6763	11
35	0.6152	11	1.0476	1	0.6267	11
40	0.5738	11	0.9994	1	0.5849	10
45	0.5383	11	0.9571	1	0.5491	10
50	0.5074	11	0.9195	1	0.5180	10
55	0.4803	10	0.8858	1	0.4906	10
60	0.4562	10	0.8552	1	0.4662	10

Table 3.3: Recommended number of examination rooms based on observational data.

Observe that the number of examination rooms for the Fast-Track patients is very low. For such small systems the asymptotic analysis provides poor approximations for the true expected waiting times. However, we observe that the recommended number of examination rooms for the High-Care patients is close to reality.

3.3 Erlang-B model

In this section we will consider the QED staffing rule for the Erlang-B model. Consider a sequence of $M/M/s/s$ queues indexed by integer n with arrival rate λ_n, service rate $\mu_n = \mu$ and traffic intensity ρ_n. Let Q_n denote the number of patients in the n'th system (in its stationary state).

Theorem 3.3.1. *The following holds for the Erlang-B model:*

$$\lim_{n\to\infty} \sqrt{n}P(Q_n \geq n) = \frac{\phi(\beta)}{\Phi(\beta)} \in (0,1),$$

if and only if

$$\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta, \quad \beta > 0.$$

The proof of this classical result (see e.g [24]) is included in Appendix D. Again, this gives rise to the square root staffing rule $s = R + \beta\sqrt{R}$, where the parameter β is found by solving

$$\sqrt{s}P\,(\text{Block}) = \frac{\phi(\beta)}{\Phi(\beta)}. \tag{3.3}$$

This relation is visualized in Figure 3.4. Consequently, in the QED-regime the probability of blocking will converge to zero as the offered load grows.

We will provide an example in a cost optimization setting for this model. Suppose that a server costs c_1 per time unit, and blocking a patient costs c_2 per time unit. We want to minimize the average total cost per time unit, which is given by

$$c_1 s + \lambda P(\text{Blocked}) \approx c_1 R + \left(c_1\beta\sqrt{R} + \frac{c_2\lambda\phi(\beta)}{\sqrt{R+\beta\sqrt{R}}\Phi(\beta)} \right).$$

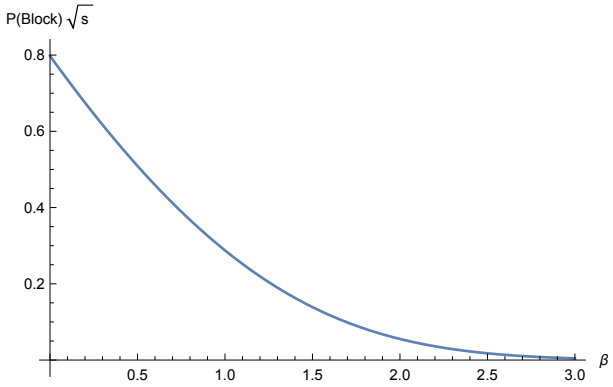


Figure 3.4: The relation given in (3.3).

Since R is large and we want to minimize the total cost, we are interested in

$$\arg\min_{\beta} \left\{ c_1\beta + \frac{c_2\mu\phi(\beta)}{\Phi(\beta)} \right\}.$$

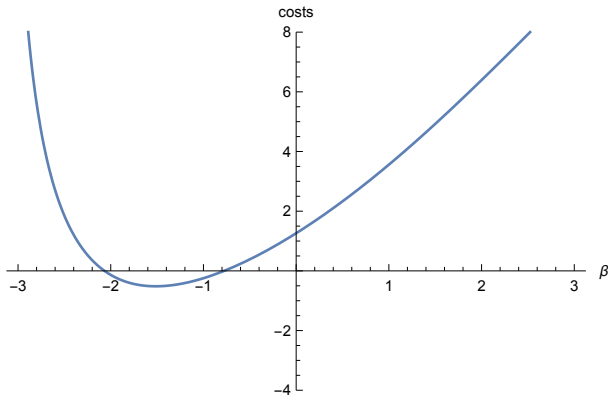


Figure 3.5: Total cost approximation example for the Erlang-B model.

This equation is illustrated in Figure 3.5 for $c_1 = 1$, $c_2 = 5$, $\lambda = 10$ and $\mu = 1$. We find that a minimum is obtained for $\beta = -1.51454$. Note that a negative value of β is possible, since we do not have the stability condition for the Erlang-B model. Thus the optimal staffing level in this example is $s = 6$.

3.4 Erlang-A model

The Erlang-C model does not take into account the fact that patients can leave the system when they do not want to wait any longer. For the Emergency Department, the percentage of patients that leaves without being seen is considered as an important performance measure for the quality of care. To incorporate this abandonment phenomenon, we will consider the Erlang-A model. This section will provide a result on the relation between waiting, abandonment and the number of servers in the QED regime.

In the framework of asymptotic optimization of this queueing system, we consider a sequence of $M/M/s + M$ queues, indexed by integer n. We have that $\lambda_n \to \infty$ as $n \to \infty$. Moreover, in this setting $\mu_n = \mu$, $s_n = n$, $\rho_n = \lambda_n/(n\mu)$ and the rate of abandonment $\theta_n = \theta \in (0, \infty)$. Note that we do not have a stability condition in this system.

The results are derived in [14]. In order to present the results, we first introduce the hazard function

h and the function w, given by

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)},$$

and

$$w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}.$$

Let W_n denote the stationary waiting time distribution of the n'th system and Ab_n the event of a patient abandoning the system before service in the n'th system.

Proposition 3.4.1. *The probability of delay has a nondegenerate limit*

$$\lim_{n\to\infty} P(Q_n > 0) = w(-\beta, \sqrt{\mu/\theta}) \in (0, 1),$$

if and only if

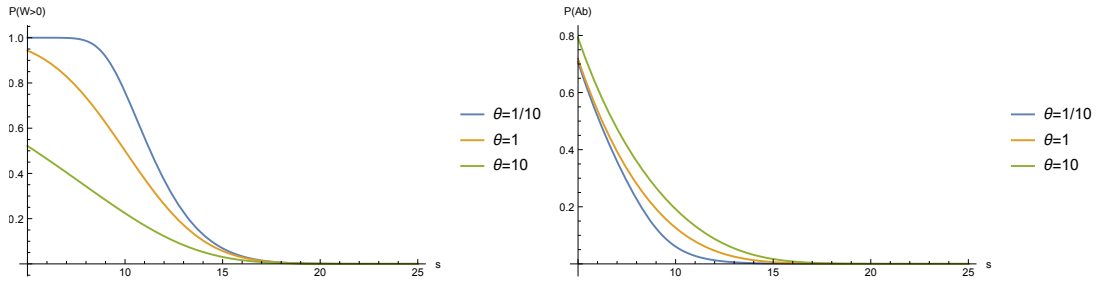
$$\lim_{n\to\infty} \sqrt{n} P(Ab_n) = \left[\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta\right] \cdot w(-\beta, \sqrt{\mu/\theta})$$

if and only if

$$\lim_{n\to\infty} (1 - \rho_n)\sqrt{n} = \beta, \quad -\infty < \beta < \infty.$$

The proof of Proposition 3.4.1 can be found in [14].

From Proposition 3.4.1 follows again the square root staffing rule $s = R + \beta\sqrt{R}$, where β can be considered as the service grade. Figure 3.6 illustrates the approximations for the probability of waiting and abandonment as a function of the number of servers s for $\lambda = 10$, $\mu = 1$ and $\theta \in \{1/10, 1, 10\}$. The different abandonment rates can be interpreted as patients that on average are willing to wait for 10 minutes, one minute and 6 seconds on average respectively.



(a) Approximation for the probability of waiting. (b) Approximation for probability of abandonment.

Figure 3.6: Performance approximations for the Erlang-A model.

The Erlang-A model can be used to account for the LWBS patients, i.e. the patient that leave the Emergency Department if they have to wait excessively long for an available examination room. Unfortunately, this number is generally very hard to approximate in practice [18]. Nevertheless, when an indication of this number is known we can use the dimensioning scheme of the Erlang-A model.

To provide a concrete example, we reconsider the example of Section 3.2.3. By Little's law we can approximate the parameter θ by $P(\text{Ab})/E(W)$. We assume that once a patient is in the examination room, she does not abandon until her service is completed. We would like to determine the number of examination rooms needed such that no more than 1% leaves without being seen. By the techniques explained in the previous section, we find $\beta = 0.5834$, $\beta = 1.1371$ and $\beta = 0.6991$ in case we consider all patients, only the Fast-Track or only the High-Care patients respectively. This yields a recommended number of examination rooms of ten in total, or one for the Fast-Track patients and ten for the High-Care patients.

However, the classical patient flow models miss the prevalent feature of patients reentering for service. Instead we will propose a new model in Part II, and find a staffing policy in the QED regime for this system.

Part II

Advanced patient flow models

Chapter 4

Advanced models in the literature

Currently, the most commonly used queueing models to support workforce management are the classical models introduced in Part I. However, more recent work in the literature proposes other models designed specifically for modeling purposes in hospital settings. We will consider these models in this chapter.

4.1 The closed ward model

A typical feature of the classical models is that they are open in the sense that the patient arrivals originate from an external (infinite) source. In contrast, Jennings and De Véricourt propose a closed queueing model to determine an efficient nurse staffing policy within a nursing unit [26, 27]. In addition, this model will prove to be of interest in Chapter 5 when we consider the overloaded regime of the Semi-Open Erlang-R model with Waiting (SERW).

Jennings and De Véricourt were interested in suggesting a staffing policy subject to a constraint on the probability of (excessive) delay for a service-requesting patient under the QED regime. We are also interested in the workload of the nurses under the proposed QED staffing rule and therefore we extend their results by considering the utilization level of the nurses.

4.1.1 Model description

Suppose that there are n patients in the medical ward with s nurses. Patients can reside in two possible states: content and needy. Stable patients become needy after an exponentially distributed activation time with rate δ. Needy patients are served by the nurses in accordance with a FCFS service discipline, and require an exponentially distributed service time with rate μ. Then the patient returns to her stable state until she needs other care procedures. A structural representation of the closed ward model is given in Figure 4.1. This model corresponds to a closed $M/M/s//n$ queue and to exclude trivial situations, we assume $s < n$.

4.1.2 Stationary distribution

Before we consider any performance measures of the closed ward model, we first look at the stationary distribution of the number of patients in the service queue. Since there are n patients in the ward, there are $n+1$ possible states. Figure 4.2 illustrates the transition rates between the different states.

Let $\rho = \delta/(\mu s)$ denote the load intensity and $R = \delta/\mu$ the offered load. Let π_i denote the stationary probability of the queue length being equal to i, i.e. the number of patients in need of service. Then the stationary distribution of the queue length is given by the following proposition [15]

Proposition 4.1.1. *The stationary distribution of the queue length of the closed ward model is given by*

$$\pi_i = \begin{cases} \pi_0 \binom{n}{i} \rho^i s^i & \text{for } 0 \le i \le s, \\ \pi_0 \binom{n}{i} \frac{i!}{s!} \rho^i s^s & \text{for } s+1 \le i \le n \end{cases} \tag{4.1}$$

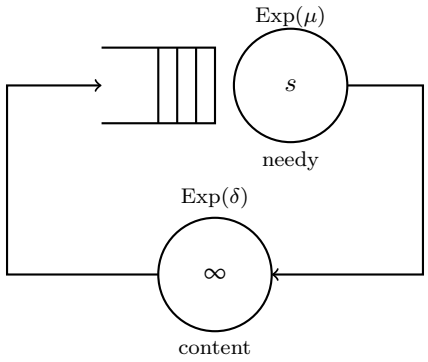


Figure 4.1: The closed ward model.

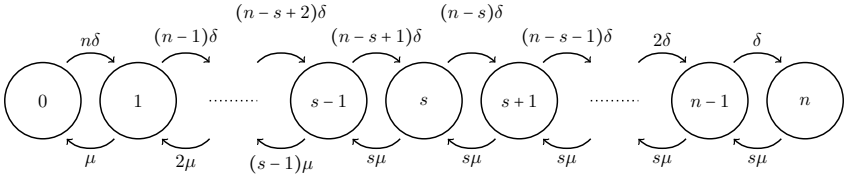


Figure 4.2: Transition diagram of the queue length in the closed ward model.

with

$$\pi_0 = \left[\sum_{i=0}^{s}\binom{n}{i}\rho^i s^i + \sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}\rho^i s^s\right]^{-1}. \tag{4.2}$$

We included the proof of Proposition 4.1.1 in Appendix C.

4.1.3 Performance measures

A needy patient must wait for service when it finds all s nurses occupied upon activation. Hence, the performance of this system can be measured by the probability of excessive delay. This is defined as the probability that an arriving patient has a waiting time that exceeds a certain threshold $t \geq 0$. In particular, if $t = 0$, it coincides with the probability that an arriving patient must wait. In addition, we are interested in the expected waiting time and the utilization level of the nurses.

Proposition 4.1.2. *Let ρ_J denote the utilization level of the closed ward model and let $\pi^{(n)}$ denote the stationary probability distribution of the number of needy patients in a ward of size n. The following properties hold:*

$$\begin{aligned}
P(W > 0) &= \sum_{k=s}^{n-1}\pi_k^{(n-1)},\\
P(W > t) &= \sum_{k=s}^{n-1}\pi_k^{(n-1)}\sum_{j=0}^{k-s}\frac{(\mu s t)^j}{j!}e^{-\mu s t},\\
E(W) &= \frac{1}{\mu s}\sum_{k=s}^{n-1}\pi_k^{(n-1)}(k-s+1),\\
\rho_J &= \sum_{k=0}^{s-1}\frac{k}{s}\pi_k^{(n)} + \sum_{k=s}^{n}\pi_k^{(n)}.
\end{aligned}$$

4.1.4 QED behavior

Normally, we consider the stationary probability that all nurses are occupied to determine the probability of delay. However, in this case the activation process is not Poisson and therefore the standard PASTA argument cannot be used. Nevertheless, Jennings and De Véricourt prove that the PASTA property holds asymptotically as $n \to \infty$. In other words, in the limit the probability that all nurses are occupied and the probability of waiting upon activation coincide. Therefore, to derive heavy-traffic approximations for the probability of delay we focus on the stationary probability that all nurses are occupied instead.

We introduce the quantity $r = \delta/(\delta + \mu)$, representing the long-run fraction of time that a patient would spend in service if her service requirement would always be addressed immediately. Moreover, a subscript is inserted for the number of servers s_n, emphasizing the dependency on the number of patients n.

Proposition 4.1.3. *The probability that all servers are occupied has a nondegenerate limit*

$$f(\beta) := \lim_{n\to\infty} P(Q_n \geq s_n) = \left(1 + e^{-\beta^2/(2r^2)}\sqrt{r}\frac{\Phi\left(\beta/\sqrt{r(1-r)}\right)}{\Phi\left(-\beta/(r\sqrt{1-r})\right)}\right)^{-1} \tag{4.3}$$

if and only if

$$\beta_n = \left(\tfrac{s_n}{n} - r\right)\sqrt{n} \to \beta \quad \text{as } n \to \infty$$

for some $\beta \in \mathbb{R}$.

Jennings and De Véricourt provide a proof of this proposition [26], which we included in Appendix D. Then Proposition 4.1.3 gives rise to the staffing rule

$$s = rn + \beta\sqrt{n}. \tag{4.4}$$

Jennings and De Véricourt extended their analysis when they considered the probability of excessive delay [27].

Proposition 4.1.4. *For any given* $t > 0$*, the probability of excessive delay has a nondegenerate limit*

$$f_t(\beta) := \Phi\left(-\frac{\beta}{\hat{r}_t}\sqrt{\frac{1 + r\mu t}{(1-r) + r\mu t}}\right), \tag{4.5}$$

if and only if

$$(s_n - \hat{r}_t)\sqrt{n} \to \beta, \quad \text{as } n \to \infty,$$

where

$$\hat{r}_t = \frac{r}{1 + r\mu t},$$

for some $\beta \in (-\infty, \infty)$.

The proof of Proposition 4.1.4 can be found in [27]. Say one wants to design the system such that the probability of excessive delay does not exceed α. Then Proposition 4.1.4 gives rise to a staffing rule for all $t > 0$, namely

$$s_n = \hat{r}_t n + \beta_t\sqrt{n},$$

where β_t is found by (4.5). Equivalently,

$$\beta_t = -\hat{r}_t\Phi^{-1}(\alpha)\sqrt{\frac{(1-r) + r\mu t}{1 + r\mu t}}.$$

Analogously to the classical patient flow models, these staffing rules provide a powerful tool in capacity management within a nursing unit. We will demonstrate this in Section 4.1.5.

The results in [27] take the patients waiting probability as the main performance criterion. This model will prove to be of interest when we introduce the Semi-Open Erlang-R model with Waiting, for which we are also interested in the utilization level of the nurses under the QED-scaling (4.4). Therefore we extend the results in [27] by providing the asymptotic utilization level in the QED regime.

Proposition 4.1.5. *Under the QED staffing rule (4.4), the utilization level ρ_J converges to one for any $\beta \in \mathbb{R}$ as $s \to \infty$.*

Proof. By Proposition 4.1.3 we know that $\sum_{k=s}^{n} \pi_k \longrightarrow f(\beta)$ under (4.4). What is left to show is

$$\sum_{k=0}^{s-1} \frac{k}{s} \pi_k \longrightarrow 1 - f(\beta).$$

We adopt the notation used in the proof of Propostion 4.1.3 (see Appendix D). Note that

$$\pi_0 = \frac{1}{A_n + B_n} = \frac{1}{B_n} \frac{B_n}{A_n + B_n}$$

with

$$B_n = \frac{(1+R)^n}{C_n} P(Y_n \leq n - s).$$

Hence

$$\frac{1}{B_n} \sim \frac{g(\beta)}{(1+R)^n},$$

where

$$g(\beta) = \frac{\sqrt{r} e^{\frac{-\beta^2}{2r^2}}}{\Phi\left(-\beta/(r\sqrt{1-r})\right)}.$$

We have

$$\begin{aligned}
\sum_{k=0}^{s-1} \frac{k}{s} \pi_k &= \frac{\pi_0}{s} \sum_{k=1}^{s-1} k \binom{n}{k} R^k \\
&= \frac{\pi_0 n R}{s} \sum_{k=0}^{s-2} \binom{n-1}{k} R^k \\
&= \frac{\pi_0 n R}{s} (1+R)^{n-1} \sum_{k=0}^{s-2} \binom{n-1}{k} r^k (1-r)^{n-1-k} \\
&= \frac{\pi_0 n R}{s} (1+R)^{n-1} P(Z_n \leq s-2),
\end{aligned}$$

where Z_n is a binomially distributed random variable with parameters $n-1$ and r. By the CLT we have that $P(Z_n \leq s-2) \to \Phi\left(\beta/(\sqrt{r(1-r)})\right)$, since

$$\frac{s - 2 - (n-1)r}{\sqrt{(n-1)r(1-r)}} = \frac{\beta + (r-2)/\sqrt{n}}{\sqrt{(1-1/n)r(1-r)}} \to \frac{\beta}{\sqrt{r(1-r)}},$$

as $n \to \infty$.

Since $n/s \to 1/r$, we obtain

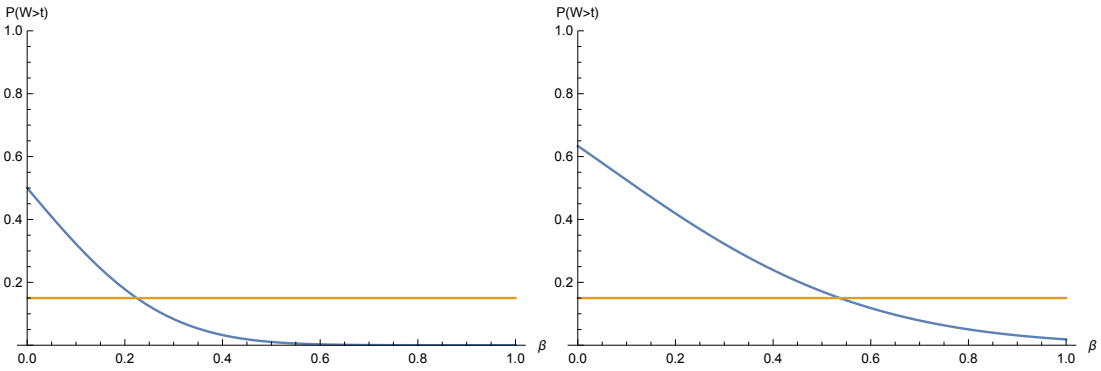
$$\begin{aligned}
\sum_{k=0}^{s-1} \frac{k}{s} \pi_k &\sim \frac{1}{r} \frac{R}{1+R} g(\beta) f(\beta) \Phi\left(\beta/(\sqrt{r(1-r)})\right) \\
&= g(\beta) f(\beta) \Phi\left(\beta/(\sqrt{r(1-r)})\right) \\
&= 1 - f(\beta).
\end{aligned}$$

In conclusion, under the QED regime we find that $\rho_J \to 1$. □

Summarizing, under the QED staffing rule (4.4), the approximate probability of waiting converges to a number strictly between zero and one. In contrast, the fraction of occupied servers converges to one. Remarkably, the fraction of time that some servers are busy while other servers are idle is not negligible, but tends to $1 - f(\beta)$. In practical terms, the probability for a patient to wait for a nurse converges to $f(\beta)$ as the number of examination rooms n grows large, while at the same time the fraction of time a nurse spends with a patient tends to one.

4.1.5 Dimensioning a nursing unit

To illustrate how to use the results in Section 4.1.4 for dimensioning a nursing unit, we will give a constraint satisfaction example next. Consider a nursing unit with $n = 15$ patients, on average requiring service every ten minutes each, with an expected service time of five minutes. Suppose a patient is considered to wait excessively if she has to wait for more than five minutes. Let one time unit correspond to five minutes, then we obtain the parameters $\delta = 1/2$, $\mu = 1$, $r = 1/3$, $t = 1$ and $n = 15$. Suppose we want that no more than 15% of the patients have to wait more than five minutes for service. The approximation for the probability of excessive delay is displayed in Figure 4.3a. We find that for $\beta_t = 0.2244$ the approximation for the probability of excessive delay intersects the performance level 0.15. Then the recommended staffing level is $s = \lceil \hat{r}_t n + \beta_t \sqrt{n} \rceil = 5$ servers in order to meet the desired performance level.



(a) Case $t = 1$.

(b) Case $t = 0$.

Figure 4.3: Excessive probability approximation for $\delta = 1/2$ and $\mu = 1$ for the closed ward model.

Similarly, we can find the required number of nurses when no more than 15% of the patients should wait at all for service. Figure 4.3b displays the probability of waiting as a function of β. We find that for $\beta = 0.3013$ the performance level is met, yielding a recommended staffing level of $s = 7$.

4.2 The Erlang-R model

In this section we consider the Erlang-R model, where the R stands for re-entering or returning, introduced by Yom-Tov and Mandelbaum [33]. Classical patient flow models do not accommodate for patients that return for service, a prevalent feature when considering healthcare environments and particularly the Emergency Department. Moreover, we will see in Chapter 5 that the Erlang-R model corresponds to the Semi-Open Erlang-R model with Waiting when we relax the constraint on the number of patients that can be served simultaneously.

4.2.1 Model description

Patients arrive to an Emergency Department according to a Poisson process with rate λ. Patients are served by s nurses with exponentially distributed service times with rate μ. After service completion, they leave the Emergency Department with probability $1-p$ and with probability p they return to service after an exponentially distributed delay time with rate δ. See Figure 4.4 for an illustration of this process. Following Jennings and de Véricourt [27], a patient in the Emergency Department can be in two possible states: the needy state when she is in need of service and the content state when she is in the delay phase.

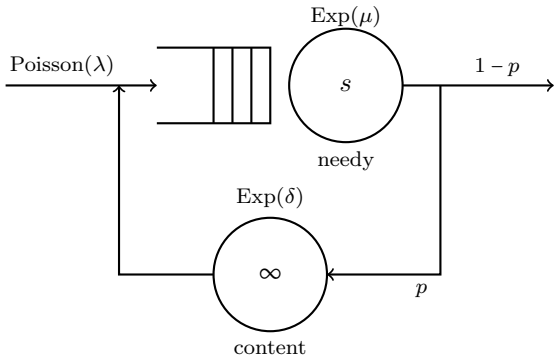


Figure 4.4: The Erlang-R model.

4.2.2 Stability

For stability in the Erlang-R model it is necessary and sufficient that the number of patients in the service station does not explode. We note that the Erlang-R model is an open two-station Jackson network with a product-form stationary distribution, see Proposition 4.2.2. Then we find that the marginal distribution on the number of needy patients is the same as the stationary distribution of an Erlang-C model with an adapted offered load. That is, there are two naturally corresponding Erlang-C models for the Erlang-R model in stationarity: either the arrival rate or the service rate can be modified to account for the returning patients, see Figure 4.5. From this observation, the stability condition follows directly.

(a) Case 1. (b) Case 2.

Figure 4.5: Corresponding Erlang-C models for an Erlang-R model.

Lemma 4.2.1. *The Erlang-R model is stable if and only*

$$\frac{\lambda}{(1-p)\mu s} < 1.$$

4.2.3 Stationary distribution

Consider the two-dimensional Markov process $Q^R(t) = (Q_1^R(t), Q_2^R(t))$, where $Q_1^R(t)$ represents the number of patients in need of service at time t and $Q_2^R(t)$ represents the number of patients in the content state at time t. Define the offered loads R and R_2 of the service queue and the content queue respectively. We note that $R = \lambda/\mu + R_2\delta/\mu$ and $R_2 = pR\mu/\delta$. Hence, the offered loads are given by $R = \lambda/((1-p)\mu)$ and $R_2 = p\lambda/((1-p)\delta)$. The Erlang-R model is an open Jackson network, with the following stationary distribution.

Proposition 4.2.2. *The stationary distribution of the queue length in the Erlang-R model is given by*

$$\pi_{ij} := P(Q_1(\infty) = i, Q_2(\infty) = j) = \begin{cases} \frac{(R)^i}{i!} c_1 \frac{(R_2)^j}{j!} c_2 & \text{if } i \le s, \\ \frac{R^i}{s!s^{i-s}} c_1 \frac{(R_2)^j}{j!} c_2 & \text{if } i \ge s, \end{cases}$$

where

$$c_1 = \left[\frac{R^s}{s!(1-R/s)} + \sum_{i=0}^{s-1} \frac{R^i}{i!} \right]^{-1}$$

and

$$c_2 = \left[\sum_{j=0}^{\infty} \frac{R_2^j}{j!}\right]^{-1} = e^{-R_2}.$$

The distribution is a direct result of the Erlang-R model being an open (product-form) Jackson network with offered loads R and R_2 for the two stations.

4.2.4 Performance measures

The next results can be found in [33].

Proposition 4.2.3. *Let $\rho_R = R/s$. The following properties hold for the Erlang-R model:*

$$\begin{aligned} P(W > 0) &= \frac{R^s}{s!(1 - R/s)} c_1, \\ W|W > 0 &\stackrel{d}{=} \text{Exp}(\mu s(1 - \rho_R)), \\ P(W > t) &= P(W > 0)e^{-\mu s(1-\rho_R)t}, \\ E(W) &= \frac{P(W > 0)}{\mu s(1 - \rho_R)}, \\ \rho_R &= \frac{R}{s} = \frac{\lambda}{(1-p)\mu s}. \end{aligned}$$

We included the proof in Appendix C.

4.2.5 QED behavior

Observe that the stationary distribution of the Erlang-R model for the nurse station is the same as that of the Erlang-C model with offered load $\lambda/((1-p)\mu)$. Therefore, the QED results for the Erlang-C model apply directly when using the QED staffing rule

$$s = R + \beta\sqrt{R}. \tag{4.6}$$

Proposition 4.2.4. *Under QED staffing (4.6) we have the following asymptotic behavior of the performance measures:*

$$\lim_{R\to\infty} P(W > 0) = \left(1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right)^{-1}, \tag{4.7}$$

$$\lim_{R\to\infty} \sqrt{s}E(W) = \frac{1}{\beta\mu}\left(1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right)^{-1}, \tag{4.8}$$

$$\lim_{R\to\infty} \rho_R = 1. \tag{4.9}$$

4.3 The Semi-Open Erlang-R model with Blocking

As an extension of the Erlang-R model, Yom-Tov [32] considered a semi-closed model that is illustrated in Figure 4.6 in her thesis [32]. We will refer to this model as the Semi-Open Erlang-R model with Blocking (SERB), and include the main results here. This model describes the referral from the Emergency Department to another medical unit. When all beds in the internal ward are occupied, the patient is directed to another hospital or another medical unit. Once the patient is in the nursing unit, she requires a nurse's service (possibly) multiple times until she leaves the Emergency Department.

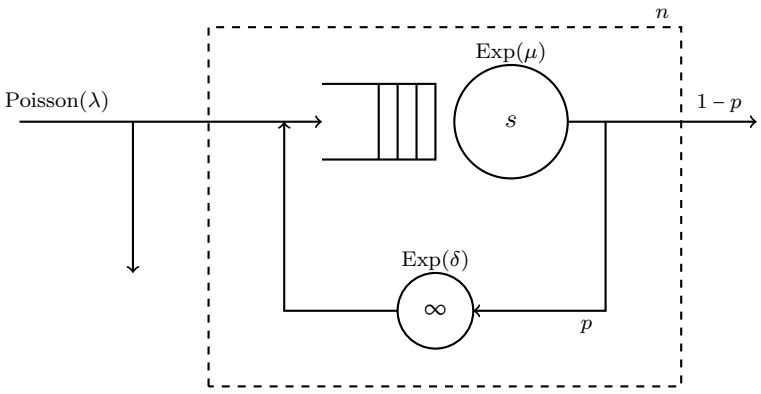


Figure 4.6: The Semi-Open Erlang-R model with Blocking.

4.3.1 Model description

Patients arrive to the service facility according to a Poisson process with rate λ. An arriving patient finding n patients being served in the service facility is blocked. Once a patient is in the service facility, the patient is served by one of the s nurses with an exponentially distributed service time with rate μ. After service completion, the patient leaves the system with probability $1-p$. With probability p the patient returns for service after an exponentially distributed time with mean $1/\delta$. Yom-Tov [32] includes a third station to represent a cleaning phase. For our research purposes, we excluded this cleaning station, which yields some simplified expressions for the distribution and performance measures.

4.3.2 Stationary distribution

The analysis of the SERB model can be reduced to a closed Jackson network consisting of three nodes. Two nodes correspond to the service station and the 'content' station of the SERB model. The patients that arrive when the system is full are blocked, hence these patients will not contribute to the number of patients within the system. Therefore, let the third node represent the number of available positions in the system. Once positions are available, the patients that arrive according a Poisson process with rate λ are accepted. In other words, as long as the third queue is nonempty, patients arrive successively following a Poisson process with rate λ. By the memoryless property, the third station can therefore be modeled as a single server queue with exponentially distributed service times with mean $1/\lambda$. An illustration of this closed network is shown in Figure 4.7.

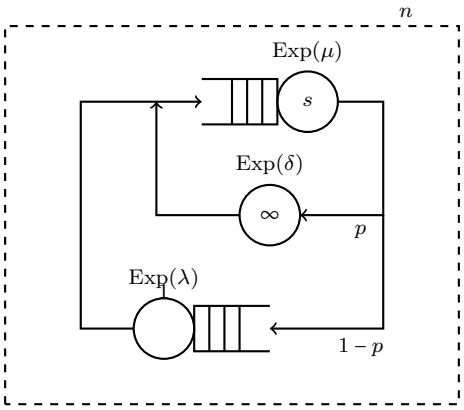


Figure 4.7: The corresponding closed Jackson of the SERB model.

Let $Q = \{Q^B(t), t \geq 0\}$ be a two-dimensional stochastic process with $Q^B(t) = (Q_1^B(t), Q_2^B(t))$, where Q_1^B represents the number of patients in need of service and Q_2^B represents the number of patients in the

content state. Let R and R_2 be the offered loads when there is no constraint on the number of patients, i.e. $R = \lambda/((1-p)\mu)$ and $R_2 = p\lambda/((1-p)\mu)$.

The corresponding closed Jackson network yields the following result on the distribution of the queue length in the SERB model [32] and the proof can be found in Appendix C.

Proposition 4.3.1. *The stationary distribution of the queue length in the SERB model is given by*

$$\pi_{ij} := P(Q_1^B(\infty) = i, Q_2^B(\infty) = j) = \begin{cases} \pi_0 \frac{(R)^i}{i!} \frac{(R_2)^j}{j!} & \text{if } i \le s, 0 \le i+j \le n, \\ \pi_0 \frac{R^i}{s! s^{i-s}} \frac{(R_2)^j}{j!} & \text{if } i \ge s, , 0 \le i+j \le n, \end{cases}$$

where

$$\pi_0^{-1} = \sum_{l=0}^{n} \frac{(R+R_2)^l}{l!} + \sum_{l=s+1}^{n} \sum_{i=s+1}^{l} \left(\frac{1}{s!} s^{s-i} - \frac{1}{i!} \right) \frac{1}{(l-i)!} (R)^i (R_2)^{l-i}.$$

From this, we can determine the distribution of the total number of patients in the system.

Corollary 4.3.2. *The distribution of the total number of patients in the SERB model is given by*

$$\pi_l := P(Q_1(\infty) + Q_2(\infty) = l) = \begin{cases} \pi_0 \frac{(R+R_2)^l}{l!} & \text{if } 0 \le l \le s, \\ \pi_0 \left(\frac{(R+R_2)^l}{l!} + \sum_{i=s+1}^{l} \left(\frac{1}{s!} s^{s-i} - \frac{1}{i!} \right) \frac{1}{(l-i)!} (R)^i (R_2)^{l-i} \right) & \text{if } s < l \le n. \end{cases},$$

The proof follows directly from the joint distribution, conditioned on the number of patients at the nurse station.

4.3.3 Performance measures

Performance measures of interest are the waiting time, blocking probability and the utilization level of the nurses. We summarize the results in Proposition 4.3.3.

Proposition 4.3.3. *Let ρ_B denote the utilization level of the SERB model and let $\pi^{(n)}$ denote the probability distribution of a system that can handle n patients simultaneously. The following properties hold for the SERB model:*

$$P(\text{Block}) = \pi_0 \left(\frac{(R+R_2)^n}{n!} + \sum_{i=s+1}^{n} \left(\frac{1}{s!} s^{s-i} - \frac{1}{i!} \right) \frac{1}{(n-i)!} (R)^i (R_2)^{n-i} \right), \tag{4.10}$$

$$P(W > 0) = \sum_{l=s}^{n-1} \sum_{i=s}^{l} \pi_{i,l-i}^{(n-1)}, \tag{4.11}$$

$$P(W > t) = \sum_{l=s}^{n-1} \sum_{i=s}^{l} \pi_{i,l-i}^{(n-1)} \sum_{k=0}^{i-s} \frac{(\mu s t)^k}{k!} e^{-\mu s t}, \tag{4.12}$$

$$E(W) = \frac{1}{\mu s} \sum_{l=s}^{n-1} \sum_{i=s}^{l} \pi_{i,l-i}^{(n-1)} (i-s+1), \tag{4.13}$$

$$\rho_B = \sum_{l=0}^{n} \sum_{i=0}^{l} i \pi_{i,l-i}^{(n)}. \tag{4.14}$$

4.3.4 QED behavior

The proposed QED staffing rules for this model read

$$n = \frac{R}{r} + \gamma \sqrt{\frac{R}{r}}, \tag{4.15}$$

$$s = R + \beta \sqrt{R}, \tag{4.16}$$

where $\gamma = \beta\sqrt{r}+\eta\sqrt{1-r}$ for some $\beta, \eta \in \mathbb{R}$. Equivalently, we can formulate the conditions (4.15) and (4.16) as

$$\begin{cases} \lim_{R\to\infty} \frac{n-s-R_2}{\sqrt{R_2}} = \eta, & -\infty < \eta < \infty, \\ \lim_{R\to\infty} \sqrt{s}\left(1-\frac{R}{s}\right) = \beta, & -\infty < \beta < \infty. \end{cases}$$

When we consider the Semi-Open Erlang-R model with Waiting in Chapter 5, we will explain heuristically how these staffing rules are established.

Proposition 4.3.4. *For the QED staffing rules in* (4.15) *and* (4.16)*, we find the following asymptotic behavior on the performance measures:*

$$\lim_{R\to\infty} P(W>0) = \begin{cases} \left(1+\frac{\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)}\right)^{-1}, & \beta\neq 0 \\ \left(1+\frac{\int_{-\infty}^{0}\Phi(\eta-t\sqrt{\frac{r}{1-r}})d\Phi(t)}{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-r}{r}}(\eta\Phi(\eta)+\phi(\eta))}\right)^{-1}, & \beta=0, \end{cases} \tag{4.17}$$

$$\lim_{R\to\infty} \sqrt{s}P(\text{Block}) = \begin{cases} \frac{\sqrt{r}\phi(\gamma\sqrt{r})\Phi(\beta\sqrt{1-r}-\eta\sqrt{r})+\phi(\sqrt{\eta^2+\beta^2})e^{\eta_1^2/2}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)}, & \beta\neq 0, \\ \frac{\sqrt{r}\phi(\gamma\sqrt{r})\Phi(\beta\sqrt{1-r}-\eta\sqrt{r})+\frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0}\Phi(\eta-t\sqrt{\frac{r}{1-r}})d\Phi(t)+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-r}{r}}(\eta\Phi(\eta)+\phi(\eta))}, & \beta=0, \end{cases} \tag{4.18}$$

$$\lim_{R\to\infty} \sqrt{s}E(W) = \begin{cases} \frac{\phi(\beta)\Phi(\eta)+\phi(\sqrt{\eta^2+\beta^2})e^{\eta_1^2/2}\Phi(\eta_1)\left(\frac{1-r}{r}\beta^2-\eta\beta\sqrt{\frac{1-r}{r}}-1\right)}{\mu\beta^2\left(\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)\right)}, & \beta\neq 0, \\ \frac{\frac{1-r}{r}\left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{2\mu\sqrt{2\pi}\left(\int_{-\infty}^{0}\Phi(\eta-t\sqrt{\frac{r}{1-r}})d\Phi(t)+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1-r}{r}}(\eta\Phi(\eta)+\phi(\eta))\right)}, & \beta=0, \end{cases} \tag{4.19}$$

(4.20)

where $\eta_1 = \eta - \beta\sqrt{\frac{1-r}{r}}$.

The proofs of these results can be found in [32], where we exclude the extra cleaning station in this case.

Chapter 5

The Semi-Open Erlang-R model with Waiting

In this chapter we return to our original problem of modeling an Emergency Department of a hospital. We introduce a new queueing model that is an extension of the Erlang-R model introduced by Yom-Tov and Mandelbaum [33] with the following essential feature: The total number of patients that can reside simultaneously in the Emergency Department is bounded by n, and all other patients wait for admission in a holding room. In practice, the parameter n can be interpreted as the number of beds. When this constraint is relaxed, i.e. $n = \infty$, we obtain the Erlang-R model again.

The model is a three-station open queueing network, with the additional requirement of maximally n patients present simultaneously in two of the three stations. While the Erlang-R model is a two-station Jackson network with a product-form solution, our model presents more mathematical difficulties due to a maximum of n patients. We therefore use matrix-geometric methods to solve a system of equations to obtain the stationary distribution.

Since is not easy to derive results for the (limit) behavior of this new model, we make stochastic comparisons with the Erlang-R model and the closed ward model by Jennings and De Véricourt [26]. We show that in order to obtain QED behavior, one needs to not only dimension the number of nurses according to square-root staffing rule, but also to scale the number of beds. We propose a two-fold scaling and show that nurses can work effectively, even in prevalent scenarios where the patient resides in the content state most of the time. The remedy for inefficient usage of nurses turns out to allow for an increase of n in a certain way. Indeed, letting more patients enter the Emergency Department creates more work for nurses, and we identify the exact way in which this expansion of operations should be designed in order to preserve the advantages of the QED regime.

5.1 Model description

In this section we will give a formal description of our model, which we call the Semi-Open Erlang-R model with Waiting (SERW), see Figure 5.1.

Patients arrive at the Emergency Department according to a Poisson process with rate λ. There are n beds, and in case these are all occupied, the arriving patient needs to wait until a bed becomes available. We assume that the service discipline is First Come First Served (FCFS). Once the patient is assigned to an available bed, she will be served by a nurse for a duration that is exponentially distributed with parameter μ. With probability p the patient remains in the Emergency Department and will return for service after an exponentially distributed delay time with rate δ. With probability $1 - p$ the patient leaves the Emergency Department. In this case, the patient can be in three possible states, namely the holding state when she is waiting for an available bed, in the needy state when she is in need of service (or being served), or in the content state when she is in the delay phase.

This process can be modelled as a quasi-birth-death (QBD) process, from which we can derive the stability condition and the stationary distribution. A quasi-birth-death (QBD) process is a generalization

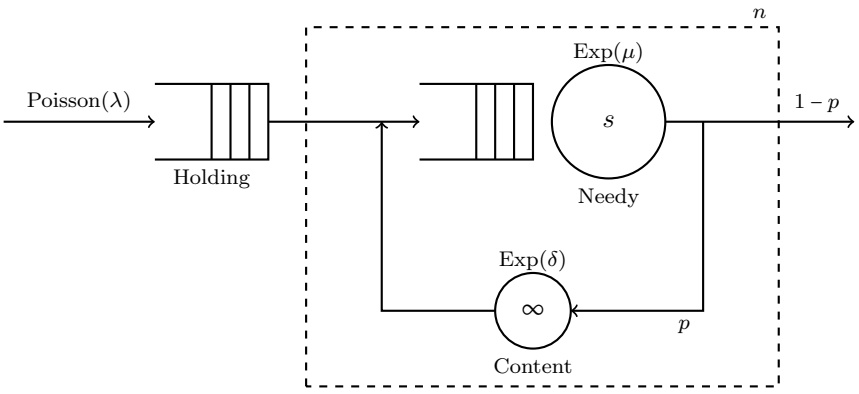


Figure 5.1: The Semi-Open Erlang-R Model with Waiting.

of the standard birth-death process.

Definition 5.1.1. *A quasi-birth-death (QBD) process is a continuous-time Markov process with a two-dimensional state space that consists of two parts: A finite number of boundary states and a semi-infinite strip. A level i consists of all (finite) states (i, j) of the two-dimensional state space. The only possible transitions in a QBD-process are within the same level or between two adjacent levels.*

For an illustration of the state space of a QBD-process, see Figure 5.2. The states within the grey box are the (finite number of) boundary states.

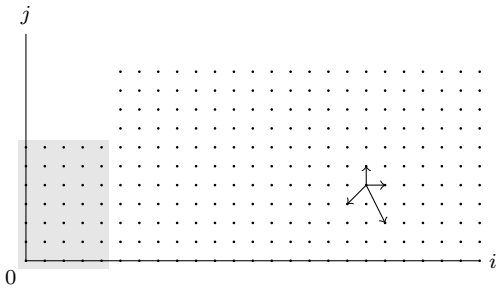


Figure 5.2: Illustration of a QBD-Process.

Note that there is a countable infinite number of states that can be ordered by ascending level. Moreover, the state space can be partitioned according to these levels. This gives rise to an infinite-sized transition matrix with a tridiagonal block structure, where each block is a square matrix of a finite size.

For the SERW model, we want to describe the process $\{Q(t), t \geq 0\}$ with $Q(t) = (Q_1(t), Q_2(t), Q_3(t))$ and $Q_i(t), i \in \{1, 2, 3\}$ is the number of patients at queue i at time t. Queue 1 corresponds to the number of needy patients, queue 2 corresponds to the number of content patients and queue 3 corresponds to the number of patients that are holding for an available bed.

Note that this process can be completely described by the two processes Q_1 and $N = Q_1 + Q_2 + Q_3$, the total number of patients in the Emergency Department. For example, suppose there are $n = 15$ beds and $s = 8$ nurses. If at time t there are in total $N(t) = i$ patients in the Emergency Department and $Q_1(t) = j \leq \min\{i, n\}$ in need of service, then $Q(t) = (j, \min\{n, i\} - j, (i - n)^+)$, where $(x)^+ = \max\{0, x\}$.

Therefore, we will look at the QBD-process $\{N(t), Q_1(t)\}$. Observe that $Q_1(t)$ is bounded by $\min\{n, N(t)\}$ and in total there can be an infinite number of patients in the Emergency Department. So level k of the QBD-process corresponds to the situation that there are k patients in the Emergency Department in total.

Define $\nu(i) = \min\{i, s\}\mu$. To determine the (outgoing) transition rates we distinguish between the following cases:

- *Transitions from* $(0,0)$*:* In this case there are no patients in the Emergency Department. The only possible occurrence is when a new patient arrives, resulting in a transition to $(1,1)$, which occurs with rate λ.
- *Transitions from* $(i,0), 1 \leq i < n$*:* There are exactly i patients assigned to a bed of which none are seen by a nurse. Then either one of those patients becomes needy, or a new patient arrives at the Emergency Department that can immediately be seen by a nurse. The first occurrence results in a transition to $(i,1)$ which happens at rate $i\delta$, and the second occurrence results in a transition to $(i+1,1)$ which happens with rate λ.
- *Transitions from* $(i,0), i \geq n$*:* Again the only possible transitions come from a newly arrived patient and a patient assigned to a bed becoming needy. However, a newly arrived patient finds all beds occupied and needs to wait. Thus, this event only increases N by one and Q_1 remains zero. So with rate λ we have a transition to $(i+1,0)$ and with rate $n\delta$ a transition to $(i,1)$.
- *Transitions from* $(i,i), i < n$*:* In this case all patients assigned to a bed are in need of service. With rate λ a new patient arrives at the Emergency Department. She joins the (possible) queue to be seen by a nurse immediately, so this results in a transition to $(i+1,i+1)$. Moreover, since there are only $s < n$ nurses, a service completion occurs with rate $\nu(i)$. With probability p the patient turns to queue 3, so in total we still have i patients with one patient less in queue for a nurse. With probability $1-p$ the patient leaves the Emergency Department, decreasing both N and Q_1 by one. In other words, with rate $p\nu(i)$ we have a transition to $(i,i-1)$ and with rate $(1-p)\nu(i)$ we have a transition to $(i-1,i-1)$.
- *Transitions from* (n,n)*:* Similar to the previous case, we have a transition to $(n,n-1)$ with rate $ps\mu$ and with rate $(1-p)s\mu$ we have a transition to $(n-1,n-1)$. In this case however, a newly arrived patient finds all beds occupied. So this results in a transition to $(n+1,n)$ with rate λ.
- *Transitions from* $(i,n), i > n$*:* Similar to the previous case we have a transition to $(i+1,n)$ with rate λ and a transition to $(i,n-1)$ with rate $ps\mu$. However, in case that the patient leaves the Emergency Department (after service completion) there are $i-n>0$ patients in the holding room waiting for an available bed. Thus after service completion, one of the $i-n$ patients in the holding room is assigned to the available bed, in need of service. That is, with rate $(1-p)s\mu$ we have a transition to $(i-1,n)$.
- *Transitions from* $(i,j), 1 \leq j < i < n$*:* There are four possible transitions. Firstly, with rate λ there is a new arrival, which results in a transition to $(i+1,j+1)$. Secondly, with rate $(i-j)\delta$ a patient in one of the beds becomes needy, which results in a transition to $(i,j+1)$. Thirdly, with rate $p\nu(j)$ a patient turns to the content state after service completion, which results in a transition to $(i,j-1)$. Lastly, with rate $(1-p)\nu(j)$ a patient leaves the Emergency Department after service completion, which results in a transition to $(i-1,j-1)$.
- *Transitions from* $(n,j), 1 \leq j < n$*:* This case is similar to the previous case. The only difference arises when a new patient arrives, since all n beds are already occupied. Thus, with rate λ we have a transition to $(n+1,j)$.
- *Transitions from* $(i,j), i > n, 1 \leq j \leq n$*:* This case is similar to the previous case, except that a patient leaves the Emergency Department after service completion. Then one of the $(i-n)$ patients in the holding room will be assigned to a bed in need of service. This results in a transition to $(i-1,j)$ with rate $(1-p)\nu(j)$.

The state space and transition rates of the SERW Model are displayed in Figure 5.3.

Recall that the state space can be partitioned according to its levels. In this case, level i corresponds to the states $\{(i,0),...,(i,\min\{i,n\})\}$, the states with in total i patients in the Emergency Department. This results in an infinite-sized matrix consisting of blocks, where each block corresponds to the transition flow from one level to another. Since the only transitions allowed are within the same level or between two adjacent levels in a QBD-process, we obtain a tridiagonal block structure. Each block consists of elements representing the transition rate of one state to another, and therefore each block is a matrix of size at most $(n+1)\times(n+1)$.

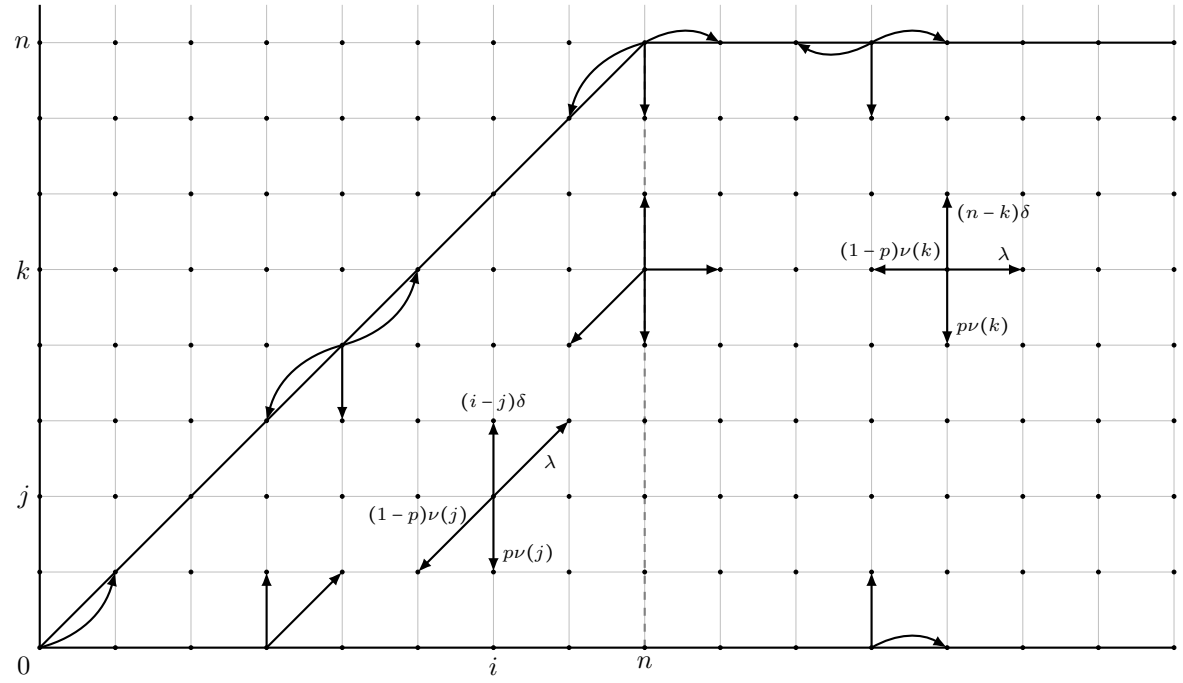


Figure 5.3: QBD-Process states space and transitions of the SERW model.

For the SERW model this gives the following result. Let P denote the transition matrix of the process $\{N(t), Q_1(t)\}$. We have the boundary levels $\{1, 2, ..., n\}$ and P is of the form

$$P = \begin{pmatrix} B_{00} & B_{01} & & & & & & & \\ B_{10} & B_{11} & B_{12} & & & & & & \\ & B_{21} & B_{22} & B_{23} & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & & & B_{n\,n-1} & & & \\ & & & & B_{n-1\,n} & B_{nn} & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & A_2 & A_1 & A_0 \\ & & & & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where $B_{ii} \in \mathbb{R}^{(i+1)\times(i+1)}$, $B_{i\,i-1} \in \mathbb{R}^{(i+1)\times i}$, $B_{i-1\,i} \in \mathbb{R}^{i\times(i+1)}$, and $A_0, A_1, A_2 \in \mathbb{R}^{(n+1)\times(n+1)}$. The matrices of transition rates for the boundary states are given by

$$B_{00} = (-\lambda), \quad B_{i-1\,i} = \begin{pmatrix} 0 & \lambda & & \\ & \ddots & \lambda & \\ & & \ddots & \ddots \\ & & & 0 & \lambda \end{pmatrix}, \quad B_{i\,i-1} = \begin{pmatrix} 0 & & & \\ (1-p)\mu & 0 & & \\ & (1-p)\nu(2) & \ddots & \\ & & \ddots & 0 \\ & & & (1-p)\nu(i) \end{pmatrix},$$

and

$$B_{ii} = \begin{pmatrix} -(\lambda + i\delta) & i\delta & & & & \\ p\mu & -(\lambda + \mu + (i-1)\delta) & (i-1)\delta & & & \\ & \ddots & \ddots & \ddots & & \\ & & p\nu(i-1) & -(\lambda + \nu(i-1) + \delta) & \delta & \\ & & & & p\nu(i) & -(\lambda + \nu(i)) \end{pmatrix}.$$

Moreover, the transition rates for the semi-infinite strip are given by

$$A_0 = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & & & & & & \\ & (1-p)\mu & & & & & \\ & & 2(1-p)\mu & & & & \\ & & & \ddots & & & \\ & & & & s(1-p)\mu & & \\ & & & & & \ddots & \\ & & & & & & s(1-p)\mu \end{pmatrix},$$

and

$$A_1 = \begin{pmatrix} -(\lambda+n\delta) & n\delta & & & & & & \\ p\mu & -(\lambda+\mu+(n-1)\delta) & (n-1)\delta & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & sp\mu & -(\lambda+s\mu+(n-s)\delta) & (n-s)\delta & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & sp\mu & -(\lambda+s\mu+\delta) & \delta & \\ & & & & & sp\mu & -(\lambda+s\mu) & \end{pmatrix}.$$

5.2 Stability

From the general theory of QBD-processes, we have the following result [29].

Lemma 5.2.1. *The Markov process $\{N(t), Q_1(t)\}$ is ergodic (stable) if and only if*

$$\pi A_0 e < \pi A_2 e, \tag{5.1}$$

where e is the all one column vector and $\pi = (\pi_0, ..., \pi_n)$ is the equilibrium distribution of the Markov process with generator $A_0 + A_1 + A_2$. In other words, π is such that

$$\pi(A_0 + A_1 + A_2) = 0, \quad \pi e = 1. \tag{5.2}$$

Hence, when the stability condition (5.1) is satisfied, the Markov process $\{N(t), Q_1(t)\}$ has a unique stationary distribution. Before presenting an algorithm to calculate this stationary distribution, we first take a closer look at the stability condition.

We have that

$$A_0+A_1+A_2 = \begin{pmatrix} -n\delta & n\delta & & & & & & \\ p\mu & -(p\mu+(n-1)\delta) & (n-1)\delta & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & sp\mu & -(ps\mu+(n-s)\delta) & (n-s)\delta & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & ps\mu & -(ps\mu+\delta) & \delta & \\ & & & & & ps\mu & -ps\mu & \end{pmatrix}.$$

Lemma 5.2.2. *Let π be defined as in (5.2). Then,*

$$\pi_i = \begin{cases} \pi_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \le i \le s, \\ \pi_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1 \le i \le n \end{cases} \tag{5.3}$$

with

$$\pi_0 = \left(\sum_{i=0}^{s} \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n} \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i \right)^{-1}.$$

Proof of Lemma 5.2.2. We have the balance equations

$$\begin{aligned} -n\delta\pi_0 + p\mu\pi_1 &= 0, \\ (n-j+1)\delta\pi_{j-1} - (p\nu(j) + (n-j)\delta)\pi_j + p\nu(j+1)\pi_{j+1} &= 0, \\ \delta\pi_{n-1} - ps\mu\pi_n &= 0, \end{aligned}$$

where $\nu(j) = \min\{j, s\}\mu$, and the normalization condition

$$\sum_{i=0}^{n} \pi_i = 1.$$

It is readily verified that the solution in (5.3) satisfies the balance equations and the normalization condition. □

Corollary 5.2.3. *The stability condition (5.1) of the SERW model can alternatively be expressed as $R < s\rho_J$, where ρ_J is the occupation rate in a closed $M/M/s//n$ system with load intensity $\rho = \delta/(p\mu s)$.*

Proof. It follows from Lemma 5.2.1 that the system is stable if and only if $\pi A_0 e < \pi A_2 e$. That is,

$$\begin{aligned} &\pi A_0 e < \pi A_2 e, \\ &\lambda \sum_{i=0}^{n} \pi_i < \sum_{i=0}^{n} (1-p)\mu \min\{i, s\}\pi_k, \\ &\lambda < (1-p)\mu s \pi_0 \left(\sum_{i=0}^{s-1} \frac{i}{s} \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i + \sum_{k=s}^{n} \binom{n}{k} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i \right), \\ &\lambda < (1-p)\mu s \rho_J, \end{aligned}$$

where ρ_J is the utilization level of a closed $M/M/s//n$ model with load intensity $\delta/(p\mu)$. □

Corollary 5.2.3 can be understood as follows. For stability we need that the 'effective' arrival rate should not exceed the total amount of work per time unit that can be done in the inner box when there are always the maximal number of patients present. In this case, the inner box has two naturally corresponding closed $M/M/s//n$ queues, which are illustrated in Figure 5.4. We will elaborate on this notion in Section 5.6. The effective arrival rate corresponds to the actual arrival rate adapted with the rate of patients leaving the system. That is, the rate of patients leaving the system is $(1-p)\mu s$ and the arrival rate is λ, which yields an effective arrival rate of $\lambda/((1-p)\mu s)$. Indeed, the stability condition is $\lambda < (1-p)\mu s \cdot \rho_J$.

(a) Case 1. (b) Case 2.

Figure 5.4: Corresponding $M/M/s//n$ models for inner box (when it is always fully occupied).

5.3 Stationary distribution

To obtain the distribution of the QBD-process, we use the matrix-geometric method [29]. Let $\pi_i \in \mathbb{R}^{\min\{i,n\}+1}$ denote the stationary probability vector of level i, i.e.

$$\pi_i = (\pi(i,0), ..., \pi(i, \min\{i,n\})),$$

where $\pi(i,j)$ denotes the stationary probability of being in state (i,j). Assuming that the stability condition is satisfied, we know from the general theory of QBD-processes that the vector π_i can be written as $\pi_{n+i} = \pi_n G^{i-n}$ for $i = 0, 1, ...$, where G is the minimal nonnegative solution of the non-linear matrix equation

$$A_0 + GA_1 + G^2 A_2 = 0. \tag{5.4}$$

The balance equations can be written as

$$\pi_{i-1}A_0 + \pi_i A_1 + \pi_{i+1}A_2 = 0, \quad i = n+1, n+2, ...$$

and when using $\pi_{n+i} = \pi_n G^{i-n}$ for $i = 0, 1, ...$, this gives

$$\pi_n G^{i-n-1}\left(A_0 + GA_1 + GA_2\right) = 0, \quad i = n+1, n+2, ...$$

Moreover, we have the boundary equations

$$\begin{aligned}
\pi_0 B_{00} + \pi_1 B_{10} &= 0 \\
\pi_0 B_{01} + \pi_1 B_{11} + \pi_2 B_{21} &= 0 \\
\pi_1 B_{12} + \pi_1 B_{22} + \pi_2 B_{32} &= 0 \\
&\vdots \\
\pi_{n-2} B_{n-2\,n-1} + \pi_{n-1} B_{n-1\,n-1} + \pi_n B_{n\,n-1} &= 0 \\
\pi_{n-1} B_{n-1\,n} + \pi_n B_{nn} + \pi_{n+1} A_2 &= 0,
\end{aligned}$$

along with the normalization equation

$$1 = \sum_{i=0}^{\infty} \pi_i e = \sum_{i=0}^{n-1} \pi_i e + \pi_n (I - G)^{-1} e,$$

where we slightly abuse notation by using e as the all ones vector of appropriate size. We note that the matrix G has a spectral radius less than one and therefore $(I - G)$ is invertible.

These equations provide the tools for finding the equilibrium probabilities. Although it is hard to solve G analytically from Equation (5.4), it is easy to solve numerically by using the following algorithm. Rewriting (5.4) gives

$$G = -(A_0 + G^2 A_2)A_1^{-1},$$

where A_1 is invertible, since A_1 is a transient generator matrix. We can use this equation to solve for G. Let

$$G_{k+1} = -(A_0 + G_k^2 A_2)A_1^{-1},$$

starting with $G_k = 0$. It can be shown that $G_k \uparrow G$ as $k \to \infty$. Once $\|G_{k+1} - G_k\|_2$ is below a certain preset threshold, we approximate G by G_{k+1}.

A concrete method to obtain G is the following. Rewriting (5.4) gives

$$G = -(A_0 + G^2 A_2)A_1^{-1},$$

where A_1 is invertible since A_1 is a transient generator matrix. Using this equation we can use this equation to solve for G. Let

$$G_{k+1} = -(A_0 + G_k^2 A_2)A_1^{-1},$$

starting with $G_k = 0$. It can be shown that $G_k \uparrow 0$ as $k \to \infty$. Once $\|G_{k+1} - G_k\|_2$ is below a certain preset threshold, set $G = G_{k+1}$. This algorithm is referred to as the matrix-geometric method.

5.4 Non-waiting distribution

To determine the patient's LOS there are three components of interest: the holding time in the holding room if all beds are occupied upon arrival, the waiting times for service if all nurses are occupied and the time that a patient is not waiting. That is, patients are not waiting when they are being served by a nurse or in the content state.

Lemma 5.4.1. *Let B denote the time of a patient's LOS she is not waiting. The probability density function of B is given by*

$$f_B(t) = (1-p)\mu e^{-\frac{1}{2}(\mu+\delta)t}\left(\cosh\left(\frac{1}{2}\sqrt{a}t\right) - \frac{\mu-\delta}{\sqrt{a}}\sinh\left(\frac{1}{2}\sqrt{a}t\right)\right) \tag{5.5}$$

for $t \geq 0$, where

$$a = (\mu-\delta)^2 + 4p\mu\delta = (\mu+\delta)^2 - 4(1-p)\mu\delta.$$

Proof. Let K denote the number of times that the patient reenters for service. That is, the patient needs service $K+1$ times and is K times in the content state. Then K is geometrically distributed with parameter p. Hence,

$$B = X_0 + \sum_{i=1}^{K}(X_i + Y_i),$$

where $X_i, i = 0,1,2,...$and $Y_j, j = 1,2,...$ are (independent) exponentially distributed with rates μ and δ respectively. Observe that X_0 represents the first time the patient needs service. Conditioning on K gives

$$\begin{aligned}\tilde{B}(x) &= \sum_{k=0}^{\infty}(1-p)p^n\left(\frac{\mu}{\mu+x}\right)^{n+1}\left(\frac{\delta}{\delta+x}\right)^n \\ &= \frac{(1-p)\left(\frac{\mu}{\mu+x}\right)}{1-p\left(\frac{\mu}{\mu+x}\right)\left(\frac{\delta}{\delta+x}\right)}.\end{aligned} \tag{5.6}$$

Next we will show that a random variable Z with probability density function given in (5.5) has Laplace Stieltjes transform (5.6), which concludes our proof. After some rewriting we find

$$\begin{aligned}\tilde{Z}(x) = E\left(e^{-Zx}\right) &= \int_0^\infty e^{-xt}f_B(t)dt \\ &= \frac{1}{2}(1-p)\mu\left(\int_0^\infty e^{-(x+\frac{1}{2}(\mu+\delta)-\frac{1}{2}\sqrt{a})t} + e^{-(x+\frac{1}{2}(\mu+\delta)+\frac{1}{2}\sqrt{a})t}dt\right. \\ &\qquad \left. -\frac{\mu-\delta}{\sqrt{a}}\int_0^\infty e^{-(x+\frac{1}{2}(\mu+\delta)-\frac{1}{2}\sqrt{a})t} - e^{-(x+\frac{1}{2}(\mu+\delta)+\frac{1}{2}\sqrt{a})t}dt\right).\end{aligned}$$

Then,

$$\begin{aligned}&\int_0^\infty e^{-(x+\frac{1}{2}(\mu+\delta)-\frac{1}{2}\sqrt{a})t} + e^{-(x+\frac{1}{2}(\mu+\delta)+\frac{1}{2}\sqrt{a})t}dt \\ &\qquad = \frac{1}{x+\frac{1}{2}(\mu+\delta)-\frac{1}{2}\sqrt{a}} + \frac{1}{x+\frac{1}{2}(\mu+\delta)+\frac{1}{2}\sqrt{a}} \\ &\qquad = \frac{2x+\mu+\delta}{(x+\frac{1}{2}(\mu+\delta))^2-\frac{1}{4}a},\end{aligned}$$

and similarly

$$\int_0^\infty e^{-(x+\frac{1}{2}(\mu+\delta)-\frac{1}{2}\sqrt{a})t} - e^{-(x+\frac{1}{2}(\mu+\delta)+\frac{1}{2}\sqrt{a})t}dt = \frac{\sqrt{a}}{(x+\frac{1}{2}(\mu+\delta))^2-\frac{1}{4}a}.$$

Hence,

$$\begin{aligned}\tilde{Z}(x) &= \frac{1}{2}(1-p)\mu\frac{2x+\mu+\delta-\mu+\delta}{(x+\frac{1}{2}(\mu+\delta))^2-\frac{1}{4}a} = (1-p)\mu\frac{x+\delta}{x^2+(\mu+\delta)x+(1-p)\mu\delta} \\ &= (1-p)\mu\frac{x+\delta}{(x+\mu)(x+\delta)-p\mu\delta} = \frac{(1-p)\mu}{x+\mu-p\left(\frac{\delta}{x+\delta}\right)\mu}.\end{aligned}$$

which equals (5.6). □

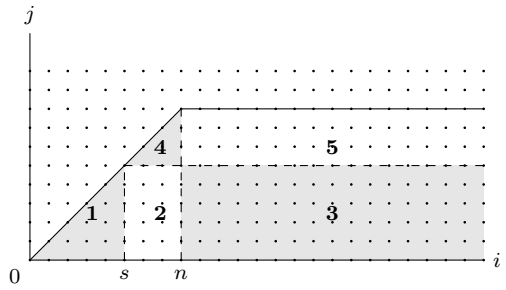


Figure 5.5: Partitioned state space of the SERW model.

5.5 Performance measures

Denote by Q_1, Q_2, Q_3 and N the stationary counterparts of $Q_1(t), Q_2(t), Q_3(t)$ and $N(t)$ respectively. The probability of waiting upon arrival is due to the PASTA property equal to $P(N \geq n)$.

Proposition 5.5.1. *The probability of waiting upon arrival is given by*

$$P(N \geq n) = \pi_n (I - G)^{-1}.$$

Proof. The event $\{N \geq n\}$ corresponds to area 3 and 5 displayed in Figure 5.5. Then,

$$\sum_{i=n}^{\infty} \sum_{j=0}^{n} \pi(i,j) = \sum_{j=0}^{n} \sum_{i=n}^{\infty} \pi(i,j) = \pi_n (I + G + G^2 + ...) = \pi_n (I - G)^{-1}.$$

□

Denote the stationary waiting time for the nurse station by W. For random variable W we cannot make a standard PASTA argument, since the activation rate is not Poisson. Nevertheless, we approximate the probability of delay for service under the conjecture that PASTA holds asymptotically as $n \to \infty$. Then it corresponds to the stationary probability that all s nurses are occupied.

Proposition 5.5.2. *The stationary probability that all s nurses are occupied is given by*

$$P(Q_1 \geq s) = \sum_{i=s}^{n-1} \sum_{j=s}^{i} \pi(i,j) + \sum_{j=s}^{n} \left(\pi_n (I - G)^{-1} \right)_j .$$

Proof. Observe that

$$P(Q_1 \geq s) = \sum_{i=s}^{n-1} \sum_{j=s}^{i} \pi(i,j) + \sum_{i=n}^{\infty} \sum_{j=s}^{n} \pi(i,j) = \sum_{i=s}^{n-1} \sum_{j=s}^{i} \pi(i,j) + \sum_{j=s}^{n} \left(\pi_n (I - G)^{-1} \right)_j .$$

□

Finally, we consider the utilization level of the servers.

Proposition 5.5.3. *The utilization level, i.e. the fraction of time the nurse is with a patient, is equal to*

$$\sum_{i=1}^{s-1} \sum_{j=1}^{i} \frac{j}{s} \pi(i,j) + \sum_{i=s}^{n-1} \sum_{j=1}^{s-1} \frac{j}{s} \pi(i,j) + \sum_{j=1}^{s-1} \frac{j}{s} \left(\pi_n (I - G)^{-1} \right)_j + \sum_{i=s}^{n-1} \sum_{j=s}^{i} \pi(i,j) + \sum_{j=s}^{n} \left(\pi_n (I - G)^{-1} \right)_j ,$$

where $\left(\pi_n (I - G)^{-1} \right)_j$ *corresponds to the j'th coordinate of the vector* $\left(\pi_n (I - G)^{-1} \right)$.

Proof. This is a direct result of computing the utilization level. We sum over five terms, that corresponds to partitioning the state space into five parts; see Figure 5.5.

There are finitely many states in areas 1, 2 and 4, which correspond to the terms with two summations over finitely many states in the expression. For the third area we have

$$\sum_{i=n}^{\infty}\sum_{j=1}^{s-1}\frac{j}{s}\pi(i,j)=\sum_{j=1}^{s-1}\frac{j}{s}\sum_{i=n}^{\infty}\pi(i,j)=\sum_{j=1}^{s-1}\frac{j}{s}\left(\pi_n(I+G+G^2+...)\right)=\sum_{j=1}^{s-1}\frac{j}{s}\left(\pi_n(I-G)^{-1}\right),$$

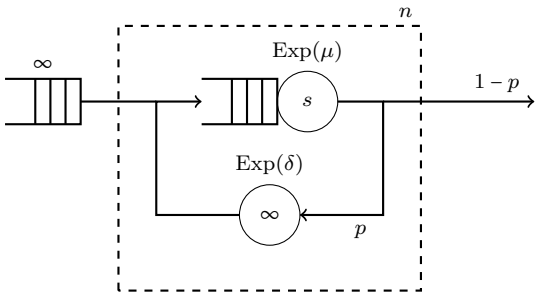
and similarly for the fifth term,

$$\sum_{i=n}^{\infty}\sum_{j=s}^{n}\pi(i,j)=\sum_{j=s}^{n}\sum_{i=n}^{\infty}\pi(i,j)=\sum_{j=s}^{n}\left(\pi_n(I-G)^{-1}\right)_j.$$

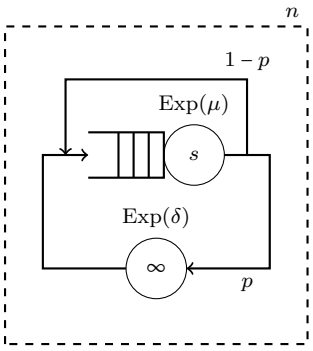
Adding all terms concludes the proof. □

5.6 The overloaded regime

In this section we consider the behavior of the SERW model when the system is in the overloaded regime, i.e. there is always some patient waiting for a bed. In other words, we consider the situation when the stability condition is violated, so that there will be infinitely many patients in the holding room and all n beds are occupied, see Figure 5.6a.



(a) The overloaded SERW model.



(b) The corresponding closed Jackson network.

Figure 5.6: The SERW model in the overloaded regime.

After service completion the patient turns to the content state with probability p. If the patient leaves the Emergency Department, which occurs with probability $1-p$, a new patient joins the service queue immediately. This corresponds to the closed two-node Jackson network in Figure 5.6b.

To determine the distribution of this closed Jackson network, we solve the balance equations. This leads to the following result.

Proposition 5.6.1. *The distribution of the closed two-node Jackson network illustrated in Figure 5.6b, is given by*

$$\pi_i=\begin{cases}\pi_0\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i & \text{for } 0\le i\le s,\\ \pi_0\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1\le i\le n\end{cases} \tag{5.7}$$

with

$$\pi_0=\left[\sum_{i=0}^{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i+\sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i\right]^{-1}.$$

Proof. We adopt the notation used in the proof of Proposition 4.3.1. We have a two-node closed Jackson network, with probability transition matrix

$$P = \begin{pmatrix} 1-p & p \\ 1 & 0 \end{pmatrix}.$$

Moreover, we have service parameters μ and δ, with rates $r_1(m) = \min\{m, s\}$ and $r_2(m) = m$ respectively. Let π_i denote the probability that there are i patients in need of service. The throughput must satisfy $p\gamma_1 = \gamma_2$, and we choose $\gamma = (p, 1)$. Let $\kappa = 1$, then

$$g_1(i) = \begin{cases} \frac{1}{i!\mu^i} & \text{for } 0 \le i \le s \\ \frac{1}{s!s^{i-s}\mu^i} & \text{for } s+1 \le i \le n \end{cases},$$
$$g_2(n-i) = \frac{1}{(n-i)!}\left(\frac{p}{\delta}\right)^n \left(\frac{\delta}{p}\right)^i.$$

We conclude that

$$\pi_i = \begin{cases} R^{-1}\left(\frac{p}{\delta}\right)^n \frac{1}{i!(n-i)!} & \text{for } 0 \le i \le s, \\ R^{-1}\left(\frac{p}{\delta}\right)^n \frac{1}{i!(n-i)!}\frac{i!}{s!}s^{s-i} & \text{for } s+1 \le i \le n. \end{cases},$$

which is equivalent to (5.7). □

We note that the stationary distribution (5.7) is the same as the distribution of the corresponding closed $M/M/s//n$ models with an offered load of $\delta/(p\mu)$. This confirms the intuitive explanation we provided for the stability condition for the SERW model. Hence, we see that in distribution the model of Figure 5.6b is equivalent to the models of Figure 5.4.

Remark 1. We obtained the stationary distribution of the model illustrated in Figure 5.6b by considering it as a closed two-node Jackson network. An alternative way to determine this distribution is as follows. Let B_1 denote the total service time at the service queue during a patient's LOS. In distribution, this is a geometric sum of exponential random variable with rate μ. Recall that the Laplace Stieltjes transform of an exponential random variable is given by $\psi(s) := \mu/(\mu + s)$. Then

$$E(e^{-sB_1}) = \sum_{k=0}^{\infty} p(1-p)^k \psi(s)^{k+1} = \frac{p\psi(s)}{1-(1-p)\psi(s)} = \frac{p\mu}{\mu + s - (1-p)\mu} = \frac{p\mu}{p\mu + s}.$$

Hence, B_1 is exponentially distributed with parameter $p\mu$.

Remark 2. Another view on how one can determine the distribution is by a memorylessness argument. After every possible service completion, one flips a coin. With probability p the service is actually done and with probability $1-p$ the patient returns for service. Then the effective service rate of the exponential is $p\mu$.

5.7 QED behavior

The goal is to provide a two-fold scaling rule for the SERW model on both the number of beds and the number of nurses, in order to design an efficient system in terms of a high utilization and short delays. If we would fix the staffing rule $s = R + \beta\sqrt{R}$ for the number of nurses and let both R and s become large, then how should we scale the number of beds n as a function of R? If we scale n too modestly, it will turn into a bottleneck as the number of patients in the holding room will grow with time and the system will explode. On the other hand, if we scale n too aggressively, the system will behave as if the constraint of n beds does not exist, i.e. it will behave as the Erlang-R model.

We will explain heuristically how to choose n in a balanced way, by determining the effective space at the service station and the content station. The service station should be able to handle the offered load, even when the effect of the constraint of n beds is negligible. Then the effective offered load at the service station is $R = \lambda/((1-p)\mu)$, which can then serve as the input for the square root staffing rule $s = R + \beta\sqrt{R}$ in the QED regime.

Moreover, we need the system to be able to handle the offered load, even when there are always more than n patients in the Emergency Department. In this case, there are s patients that can be handled at the service station, and hence $n-s$ patients reside at the content station. The effective offered load at the content station is $p\lambda/((1-p)\delta)$, which gives rise to

$$n - s = \frac{p\lambda}{(1-p)\mu\delta} + \eta\sqrt{\frac{p\lambda}{(1-p)\mu\delta}}$$

with $-\infty < \eta < \infty$. Substituting the server staffing rule in the above expression yields

$$n = s + \frac{p\lambda}{(1-p)\mu\delta} + \eta\sqrt{\frac{p\lambda}{(1-p)\mu\delta}} = R\left(1 + \frac{p\mu}{\delta}\right) + \left(\beta + \eta\sqrt{\frac{p\mu}{\delta}}\right)\sqrt{R}.$$

Then the given staffing rules are equivalent to

$$n = \frac{R}{r} + \gamma\sqrt{\frac{R}{r}}, \tag{5.8}$$

and

$$s = R + \beta\sqrt{R}, \tag{5.9}$$

where $\beta > 0$ and $\gamma > 0$ and $r = \delta/(\delta + p\mu)$.

We are interested in the (asymptotic) behavior of the SERW model under the two-fold staffing rule (5.8) and (5.9) and particularly, whether they lead to QED behavior. For the latter we use as a benchmark that the limiting probability of waiting is strictly between zero and one. Next we will explain heuristically why these staffing rules will result in a nondegenerate limit for the probability of waiting.

In the overloaded regime we assume that there is always a patient waiting in the holding room. In other words, all beds will be constantly occupied. Hence, in distribution there are always more patients at the service queue in the overloaded regime than in the original SERW model with the same parameters. That is, the corresponding $M/M/s//n$ model of the SERW model in the overloaded regime should provide an upper bound on the probability of waiting.

On the other hand, the SERB model blocks arriving patients when the beds are all occupied. Therefore, there are less patients in the inner box in the SERB model than in the SERW model with the same parameter setting. Consequently, there are also less patients at the service queue in the SERB model than in the SERW model (in distribution). Hence, the probability of waiting in the SERB model should serve as a lower bound on the probability of waiting in the SERW model.

Although the staffing rules (5.8) and (5.9) are given in terms of the offered load, this implicitly yields a dependency between n and s. Therefore, we can express the staffing level of the number of nurses in terms of n. That is,

$$\begin{aligned} s = R + \beta\sqrt{R} &= rn + \left(\beta\sqrt{r} - \gamma r\right)\sqrt{\frac{R}{r}} \\ &\sim rn + \left(\beta\sqrt{r} - \gamma r\right)\sqrt{n} = rn + \xi\sqrt{n}, \end{aligned}$$

with $\xi = \beta\sqrt{r} - \gamma r \in \mathbb{R}$. We observe that this is the QED staffing rule for the closed $M/M/s//n$ model [26]. Since this scaling yields a nondegenerate limit for the probability of waiting, we find that the probability of waiting in the SERW model cannot converge to one.

On the other hand, the SERB model also has a nondegenerate limit on the probability of waiting under the staffing rules (5.8) and (5.9) and hence the probability of waiting in the SERW model cannot converge to zero. More specifically, we find

$$\left(1 + e^{-\xi^2/(2r^2)}\sqrt{r}\frac{\Phi\left(\xi/\sqrt{r(1-r)}\right)}{\Phi\left(-\xi/(r\sqrt{1-r})\right)}\right)^{-1} \le P(W>0) \le \left(1 + \frac{\int_{-\infty}^{\beta}\Phi(\eta + (\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)}\right)^{-1} \tag{5.10}$$

with $\gamma = \beta\sqrt{r} + \eta\sqrt{1-r}$ and $\eta_1 = \eta - \beta\sqrt{\frac{1-r}{r}}$.

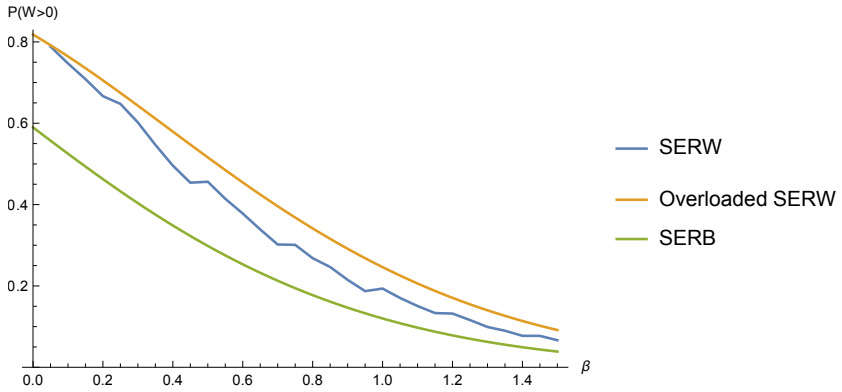


Figure 5.7: Probability of waiting for $r = 0.2$ and $\gamma = 0.5$.

This relation is illustrated in Figure 5.7, where the (asymptotic) probability of waiting for the SERW model is approximated via a simulation experiment with $R = 250$ (fluctuations are due to rounding).

In conclusion, we expect that the chosen staffing policy will result in stabilizing behavior for the probability of waiting. Although the matrix-geometric algorithm provides us the exact stationary distribution, it takes too much computational power when n grows large. Therefore we will use a simulation program to study the asymptotic behavior of the SERW model in Chapter 7. In the next chapter, we will first compare the SERW model to other models with respect to the total number of patients in the Emergency Department.

Chapter 6

Comparison between models

In the previous chapters we introduced the Erlang-R model, the Semi-Open Erlang-R model with Blocking (SERB) and the Semi-Open Erlang-R model with Waiting (SERW). The SERB model blocks patients from the system when the total number of patients in the Emergency Department exceeds a certain threshold n, while the SERW model lets these patients wait before entering the service loop. The Erlang-R model does not have such a constraint. But how do these systems relate to one another with respect to the total number of patients in the Emergency Department?

Since the SERW model assumes a maximum of n patients in the service loop, and hence limits the usage of resources, we expect that there are less patients in the Emergency Department in the Erlang-R model than the SERW model at any point in time. Moreover, we also expect that the SERB model contains less patients than the Erlang-R model, since the blocked patients will not receive service. As a reference model, we also consider the SERW model where we set the number of beds and the number of nurses equal to one another.

To formalize this idea, let N^B, N^R, N^W and $N^W_{n=s}$ denote the total number of patients in the system in stationarity in respectively the SERB model, the Erlang-R model, the SERW model and the SERW model with $n = s$. In this chapter we will study whether we can find the following stochastic ordering of these four models:

$$N^B \leq_{\text{st}} N^R \leq_{\text{st}} N^W \leq_{\text{st}} N^W_{n=s}. \tag{6.1}$$

6.1 Stochastic dominance

In probability theory, there exist several orders in which we say that a random variable stochastically dominates some other random variable.

Definition 6.1.1. *Let A and B be random variables. Random variable A is statewise stochastically dominated, or statewise stochastically less, by B if $A \leq B$ for every possible outcome. This is denoted as $A \leq_{(0)} B$.*

Definition 6.1.1 is stochastic dominance in the strongest sense. A typical example for which $A \leq_{(0)} B$ holds is $A = B - 1$. Usual stochastic dominance is defined in terms of its distributions.

Definition 6.1.2. *Let A and B be random variables. Random variable A is stochastically dominated in the usual sense, or stochastically less, by B if*

$$P(A > x) \leq P(B > x), \quad \forall x \in \mathbb{R}. \tag{6.2}$$

This is denoted as $A \leq_{st} B$.

Lemma 6.1.3. *Let A and B be random variables. Then $A \leq_{st} B$ if and only if*

$$E(f(A)) \leq E(f(B)) \tag{6.3}$$

for all non-decreasing functions f.

Statewise dominance is also referred to as dominance of the zeroth order, and dominance in the usual sense as dominance of the first order. Moreover, a random variable A is second order stochastically dominated by random variable B if $\int_{-\infty}^{x} F_B(t) - F_A(t)\,dt \geq 0$ for all $x \in \mathbb{R}$, where F_A and F_B denote the probability distribution function of A and B respectively, and A is third order stochastically dominated by B if $\int_{-\infty}^{x}\int_{-\infty}^{z} F_B(t) - F_A(t)\,dt\,dz \geq 0$ for all $x \in \mathbb{R}$. This iterative procedure results in the definitions of higher order stochastic dominance.

We note that stochastic dominance of some order implies stochastic dominance of higher orders. Therefore, we will first consider whether statewise stochastic dominance holds, since this implies the relation (6.1).

6.2 The SERB model vs the Erlang-R model

The SERB model behaves as the Erlang-R model, except that it refuses patients when the total number of patients in the Emergency Department exceeds a certain threshold n. This suggests that the total number of patients in the system is smaller in the SERB model than in the Erlang-R model. We will first show that $N^B(t) \preceq N^R(t)$ can not be implied by statewise stochastic dominance.

We provide an example of arriving patients, with fixed arrival times, number of loops (number of times the patient returns for service), service times and delay times and show that at some point in time $N^R(t) \leq N^B(t)$. Let $n = 2$ and $s = 1$ and consider the arrival stream of Table 6.1. Initially, the system is empty.

Customer	Arrival Time	Number of Loops	Service time(s)	Content time(s)
1	$t = 0$	0	(3)	-
2	$t = 1$	0	(1)	-
3	$t = 2$	0	(6)	-
4	$t = 5$	1	(1,3)	(5)
5	$t = 12$	0	(1)	-

Table 6.1: Example of five arriving patients.

Note that for this example customer 3 will be blocked from the system in the SERB model. This leads to the sample paths displayed in Figure 6.1.

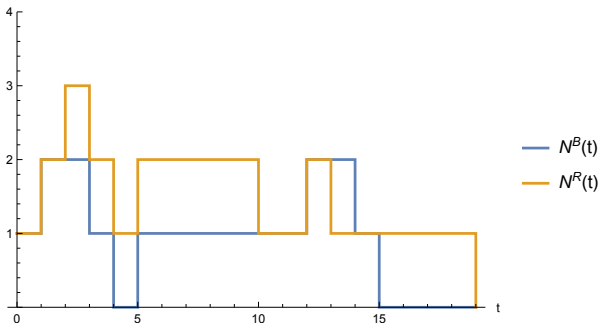
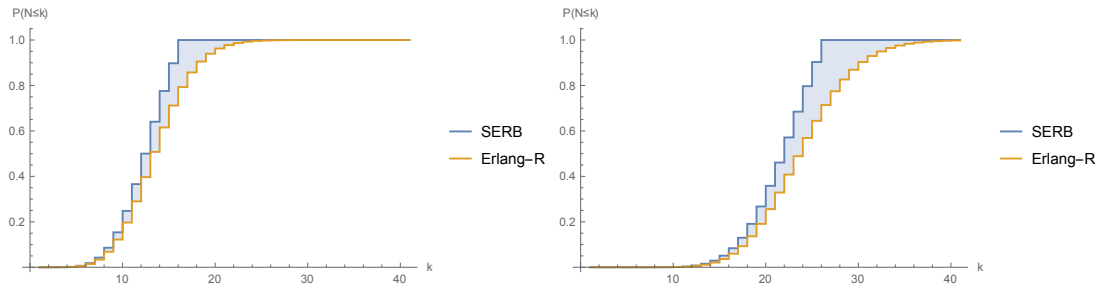


Figure 6.1: Total number of patients for the arrival stream in Table 6.1.

Observe that for $t \in [13, 14)$ we find $N^B(t) > N^R(t)$, and hence there is a time span where there are more patients in the Emergency Department in the SERB model than in the Erlang-R model.

Although this event can occur for some point in time, via simulations we observed that this is a rare event. Still, we are interested whether $N^B \preceq_{\text{st}} N^R$ holds. The total number of patients in the SERB model can never exceed threshold n, hence for $k > n$ we certainly have that $P(N^B \leq k) \leq P(N^R \leq k)$. Moreover, when n grows large the two models will coincide.

Recall that the stationary distribution are known for these models, which yields the probability distribution for the total number of patients in the Emergency Department. We considered these probability distributions via numerical experiments, where we chose our parameters such that threshold n has a considerable effect. We observed in all cases that $N^B \leq_{\text{st}} N^R$ and we display two outcomes of our numerical experiments with parameters $\mu = 10$, $\delta = 2$ and $p = 0.8$ in Figure 6.2. This supports our intuition that indeed there are more patients in the Erlang-R model than in the SERB model.



(a) Case $n = 15$, $s = 5$ and $\lambda = 5$.

(b) Case $n = 25$, $s = 7$ and $\lambda = 9$.

Figure 6.2: Probability distributions for the total number of patients SERB vs. Erlang-R.

6.3 The SERW model vs the Erlang-R model

The SERW model behaves as the Erlang-R model, except that patients need to wait for an available bed when the total number of patients in the Emergency Department exceeds a certain threshold n. Again, we will first show that $N^R \leq_{\text{st}} N^W$ cannot be implied by statewise dominance.

Let $n = 2$ and $s = 1$ and consider the arrival stream of Table 6.2. The total number of patients in the Emergency Department is given in Figure 6.3.

Customer	Arrival Time	Number of Loops	Service time(s)	Content time(s)
1	$t = 0$	1	(2,1)	(3)
2	$t = 1$	1	(1,2)	(4)
3	$t = 4$	0	(4)	-

Table 6.2: Example of three arriving patients.

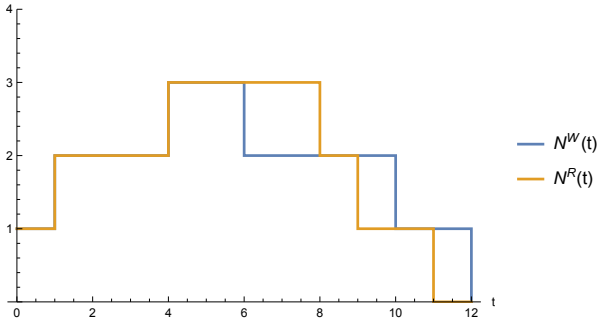
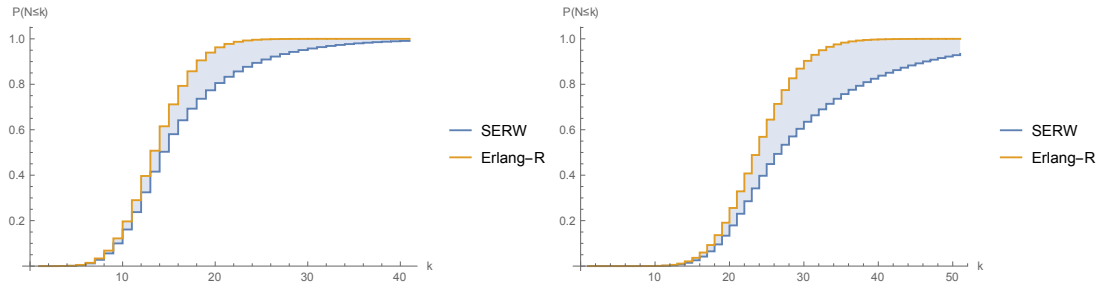


Figure 6.3: Total number of patients for arrival stream in Table 6.2.

We observe that for $t \in [6, 8)$ the policy of the Erlang-R model leads to more patients in the Emergency Department than the policy of the SERW model. In other words, the relation $N^R(t) \leq_{(0)} N^W(t)$ does not hold.

Nevertheless, intuitively $N^R \preceq_{\text{st}} N^W$ seems a reasonable relation, since nurses can already serve patients in the Erlang-R model that are waiting in the holding room in the SERW model. Hence, we expect that there are more patients in the Emergency Department than in the Erlang-R model. Our numerical experiments support this intuition, e.g. Figure 6.4, but the relation remains a conjecture.



(a) Case $n = 15$, $s = 5$ and $\lambda = 5$.

(b) Case $n = 25$, $s = 7$ and $\lambda = 9$.

Figure 6.4: Probability distributions for the total number of patients SERW vs. Erlang-R

6.4 The SERW model vs the rescaled SERW model

As a reference model, we also consider the SERW model where we adapt the scaling. There are two ways to rescale the SERW model on a service level, namely modifying the number of beds or modifying the number of servers. That is, we will consider the relation

$$N^W_{s=n} \preceq_{\text{st}} N^W \preceq_{\text{st}} N^W_{n=s}, \tag{6.4}$$

where $N^W_{s=n}$ represents the total number of patients in a rescaled SERW model with an increased number of servers, and $N^W_{n=s}$ represents the total number of patients in a rescaled SERW model with a decreased number of beds.

First, we consider the case where the number of servers is increased. We will refer to this system as the Increased SERW model.

Proposition 6.4.1. *The Increased SERW model is statewise stochastically dominant over the SERW model, i.e.*

$$N^W_{s=n}(t) \preceq_{(0)} N^W(t) \tag{6.5}$$

for each point in time t.

Proof. Without loss of generality we can assume we have an empty system at $t = 0$. Let LOS and LOS$'$ denote the patient's length of stay in the SERW model and the Increased SERW model respectively. Let customer k be the first customer for which $\text{LOS}(k) < \text{LOS}'(k)$. In addition, let A and A' represent the time the patient is in the holding state.

Observe that a patient's length of stay in the SERW model consists of the time in the holding queue, the time(s) being served, the time(s) that the patient is in the content state and the time(s) that a patient needs to wait for service. In the Increased SERW model there is always a server available to serve the patient in a bed, and hence a patient will never need to wait for service. Thus the patient's LOS boils down to the time in the holding queue, the time(s) being served and the time(s) that the patient is in the content state. Hence, patient k is the first patient for which the time in the holding queue and the time(s) waiting for service in the SERW model is strictly less than the time in the holding queue for the Increased SERW model. In particular, we have that $A(k) < A'(k)$.

Then there is a time t such that $N^W(t) < N^W_{s=n}(t)$, in particular when patient k is assigned to a bed in the SERW model and not yet in the Increased SERW model. Hence, there must be at least one patient that left the Emergency Department before t in the SERW model than in the Increased SERW model. This contradicts that customer k is the first patient for which $\text{LOS}(k) < \text{LOS}'(k)$.

In conclusion, for every patient we have $\text{LOS}(k) \geq \text{LOS}'(k)$ and hence $N^W_{s=n}(t) \leq N^W(t)$ for each $t \geq 0$. □

Particularly, Proposition 6.4.1 implies that $N^W_{s=n} \leq_{\text{st}} N^W$ holds.

Alternatively, we can decrease the number of beds to the number of servers. We will refer to this model as the Decreased SERW model. Note that the stability condition of the Decreased SERW model is more restrictive than for the original SERW model. That is, if the Decreased SERW model is stable, then so is the original SERW model, but the converse does not necessarily hold.

In this case, $N^W \leq_{\text{st}} N^W_{n=s}$ can also not be shown via a higher stochastic ordering. Let $n = 2$ and $s = 1$ and consider the arrival stream of Table 6.3. Initially, we start with an empty system.

Customer	Arrival Time	Number of Loops	Service time(s)	Content time(s)
1	$t = 0$	1	(1,2)	(3)
2	$t = 2$	1	(1,1)	(4)
3	$t = 5$	0	(7)	-

Table 6.3: Example of three arriving patients.

This example results to the sample paths displayed in Figure 6.5.

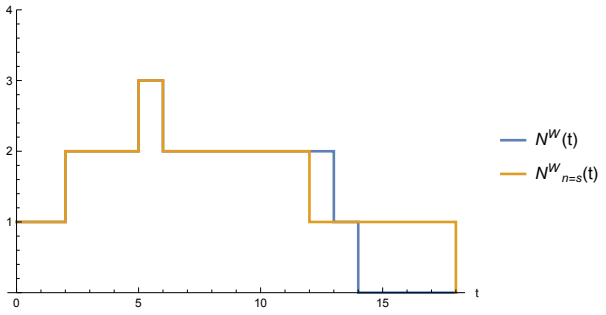
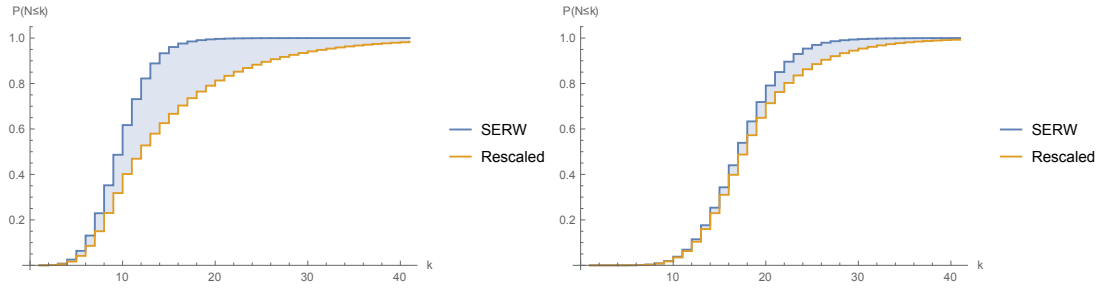


Figure 6.5: Total number of patients for arrival stream in Table 6.3.

Then, for $t \in [12, 13)$ we have $N^W(t) > N^W_{n=s}(t)$. Nevertheless, in this case our numerical results also support our intuition that stochastic dominance holds in the usual sense, see Figure 6.6, but a formal proof is still required.



(a) Case II. (b) Case IV.

Figure 6.6: Probability distributions for the total number of patients in SERW vs. Decreased SERW model.

Summarizing, we have shown that (6.1) can not be implied by statewise stochastic ordering. Nevertheless, our numerical experiments support that this relation does hold. To prove this formally provides a nice mathematical challenge for further research!

Part III

Data-driven staffing procedures in the Emergency Department

Chapter 7

System behavior of the SERW model

In this chapter we will use our simulation program for the Semi-Open Erlang-R model with Waiting (SERW) to obtain insights in various performance measures under the staffing rules (5.8) and (5.9). We will verify that these staffing rules indeed prescribe a QED policy. That is, the probability of waiting for a nurse and also the probability of holding for a bed have nondegenerate limits and the utilization levels will converge to one. In addition, we will see that compared to the Erlang-R model the constraint on the number of beds has a significant effect on the performance measures.

Although so far we considered all models in stationarity, in Emergency Departments the arrival rates are typically not stationary. We will propose a method to modify our staffing policy for the SERW model to account for time-varying arrival rates. Since the arrival stream in an Emergency Department typically resembles somewhat of a sinusoidal shape, and moreover, any periodic time-varying arrival rate can be approximated by a finite linear combination of sine functions, we will test the performance of this staffing policy for a sinusoidal arrival rate.

7.1 Stationary behavior

We simulated scenarios for many different parameters to support our findings we pose in this chapter. When not specified otherwise, we display the results of the parameter setting $\mu = 10$, $\delta = 2$ and $p = 0.8$, and hence $r = 0.2$. This parameter setting is motivated by data and based on experience of experts working in our cooperating Emergency Department. In our experiments we let R grow till $R = 250$, yielding systems with large numbers of nurses and beds such that we can consider the asymptotic behavior of the SERW model. In the time-varying arrival setting we will also consider smaller system sizes that might represent an Emergency Department better. We observe some irregular fluctuations in all figures, which are due to the rounding effect.

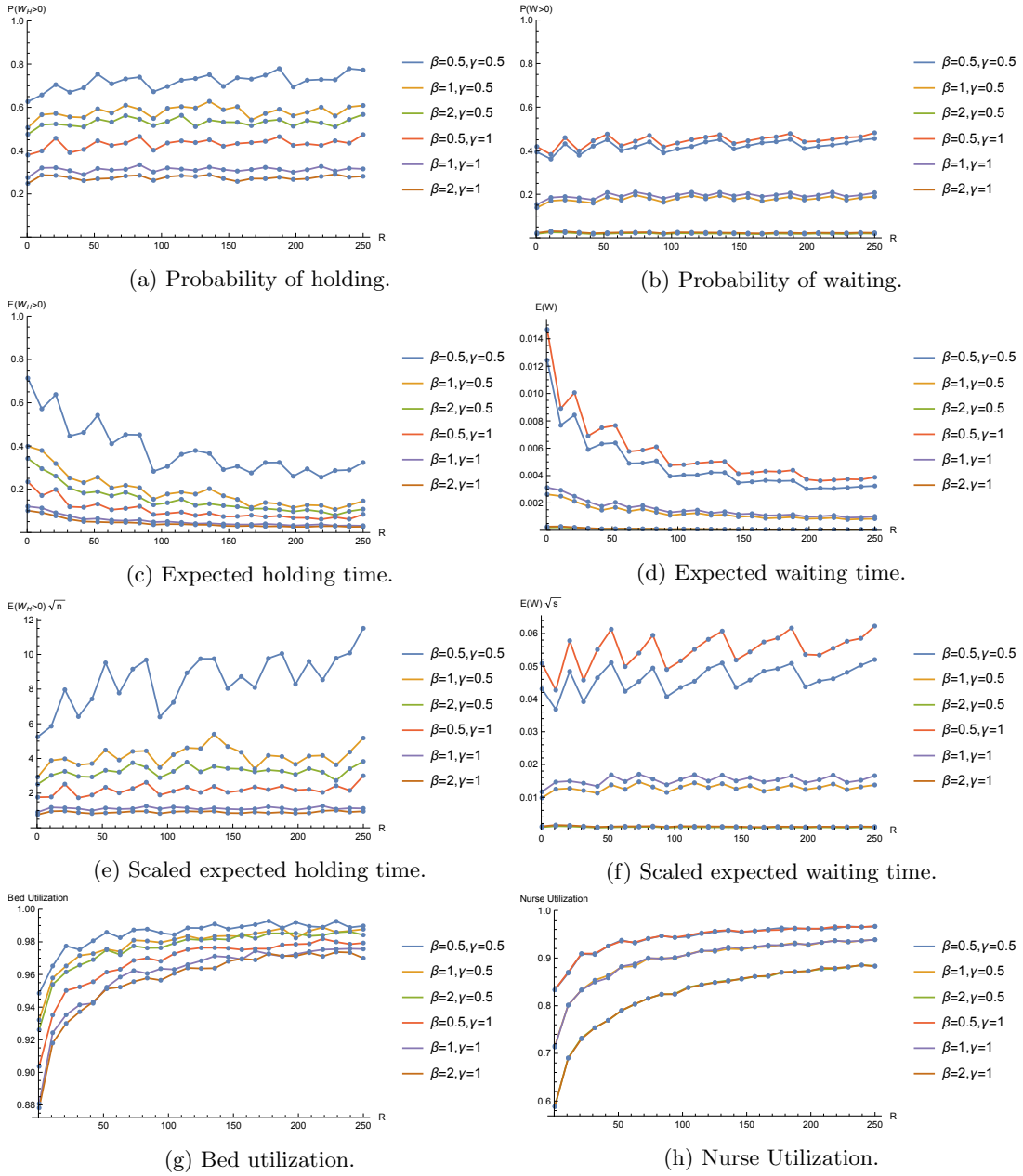
7.1.1 Performance measures

Next, we will consider the behavior of all performance measures of interest separately under the staffing rules (5.8) and (5.9). The behavior of different performance measures is displayed in Figure 7.1.

Probability of Holding:
We observe stabilizing behavior for the probability of holding. This probability is determined by r, β and γ and as these values grow, the probability of holding will tend to zero. Naturally, the parameter γ has the most impact on this performance measure, but we find that β has a significant effect as well. In other words, increasing the number of nurses will ensure that patients need to wait less for service, and therefore the beds will be occupied for a shorter amount of time.

Probability of Waiting:
For a QED policy the probability of waiting must have a nondegenerate limit. Our numerical experiments show that this probability indeed stabilizes to a value strictly between zero and one, e.g. Figure 7.1b.

The probability of waiting for a nurse depends on r, β and γ. As β grows to a value such that $s \geq n$, the probability of waiting is zero and moreover, as β and γ grow simultaneously, this probability will tend to zero as well. Moreover, it seems that the impact of γ is rather insignificant compared to β in the QED regime. Naturally, in a QED regime where the utilization rate of beds is high, the probability of waiting is most influenced by the number of nurses itself. However, for small values of γ the number of beds might prove a bottleneck and impact this performance measure considerably. We believe more research is needed to have a deep understanding of the interplay between these parameters.



(a) Probability of holding.

(b) Probability of waiting.

(c) Expected holding time.

(d) Expected waiting time.

(e) Scaled expected holding time.

(f) Scaled expected waiting time.

(g) Bed utilization.

(h) Nurse Utilization.

Figure 7.1: Various performance measures of the SERW model for $r = 0.2$.

Expected Holding Time for a Bed:
When we consider Figure 7.1c and Figure 7.1e, we see that the expected holding time for a bed will converge to zero with rate $1/\sqrt{n}$. However, this performance measure behaves rather capricious due to the rounding effect.

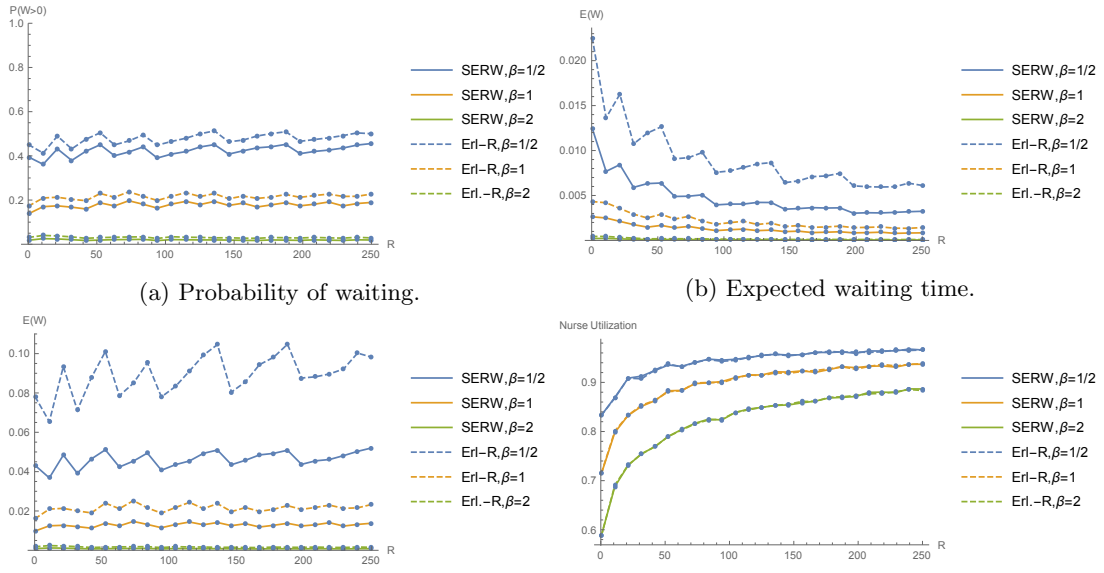
Expected Waiting Time for a Nurse:
We expect that the expected waiting time for a nurse will converge to zero at rate $1/\sqrt{s}$, which is the same order of magnitude as the rate of convergence for the expected holding time for a bed. Our simulation results seem to support this notion, see Figure 7.1d and Figure 7.1f. However, we find rather capricious behavior due to the rounding effect and therefore our simulation results do not provide conclusive support for our intuition.

Expected Length of Stay:
Since the LOS consists of a sum of service times, content times, holding times and waiting times, this performance measure will obviously not stabilize. However, as we let R grow to infinity the expected waiting time and the expected holding time tend to zero with rate $1/\sqrt{R}$, which is of the same order as $1/\sqrt{s}$ and $1/\sqrt{n}$. Therefore, the expected LOS will tend to $p/((1-p)\delta) + 1/((1-p)\mu)$ with rate $1/\sqrt{R}$.

Utilization levels:
Under the staffing policy (5.8) and (5.9) we see that the bed utilization is very high and tends to one as R grows large, e.g. Figure 7.1g. Both β and γ have a significant impact on this performance measure. Our simulation results also support that the nurse utilization level grows to one as the offered load increases, see Figure 7.1h.

7.1.2 Comparison with the Erlang-R model

Figure 7.2 demonstrates that the constraint on the number of beds has a significant impact on the performance measures with respect to the Erlang-R model. Intuitively, we expect that this impact will decrease as β and particularly γ increase, and our simulation results support this intuition.



(a) Probability of waiting.

(b) Expected waiting time.

(c) Scaled expected waiting time.

(d) Nurse Utilization.

Figure 7.2: Comparison performance measures SERW model vs. Erlang-R model for $\gamma = 0.5$.

Note that the expected waiting time for a nurse converges to zero with the same rate as in the Erlang-R model. Moreover, when we compare the nurse utilization level in the SERW model to the Erlang-R model, we observe that these measures coincide in our simulation results, see Figure 7.2d. This can be expected, since the amount of work for the nurses is equal in both. Hence the expected amount of work each nurse executes per time unit is the same in both systems. Therefore this measure should only depend on the choice of r and β, and not on the choice of γ.

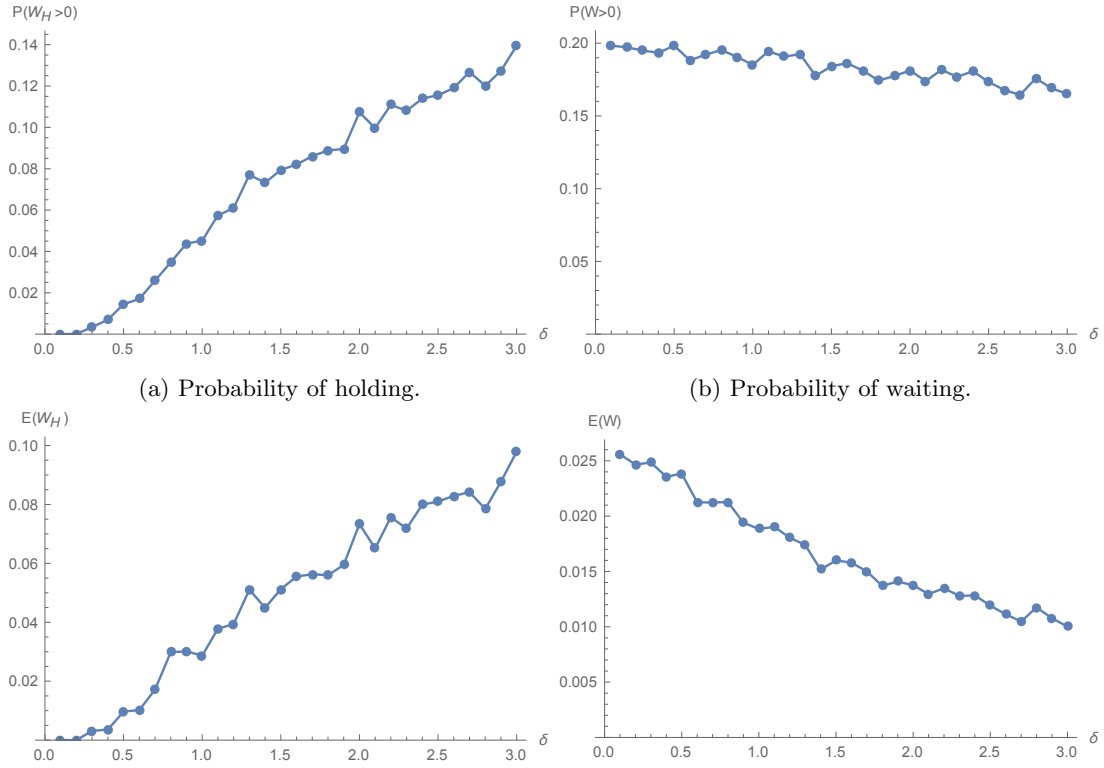
7.1.3 Influence of r

Recall that r corresponds to the fraction of the time patients in beds would be served when there is always a nurse available. It is a measure that determines the intensity of service requirement of a patient. In the Emergency Department this value is typically rather low, since the time a patient waits for lab results to return is much longer compared to the service times. To obtain a general understanding of the influence of r on the performance measures, we fixed $\lambda = 10$, $\mu = 1$, $p = 0.8$, $\beta = 1$, $\gamma = 1$ and varied δ between 0.1 and 3. Note that this setting yields a fixed number of nurses of $s = 58$, an increasing r as a function of δ and and hence decreasing number of beds as a function of δ.

As the (average) expected time patients are in the content state decreases, the total duration the patient spends in a bed will also decrease. As long as the expected holding time does not increase too fast, the LOS will decrease as well as δ increases. However, this measure cannot be captured by r itself, since it depends heavily on the magnitude of μ and δ.

In our simulation results we observed that the waiting probability for a nurse and the expected waiting time are increasing function with respect to δ, e.g. Figure 7.3b and Figure 7.3d. This suggests that these performance measures decrease as r increases. This seems counterintuitive, since we increase the ratio between the time a patient spends in a bed in the needy state and the content state. However, we simultaneously decrease the number of beds by order $1/r$ due to staffing rule (5.8). Consequently, there are less patients per nurse per time unit and hence the probability of waiting and also the expected waiting time decrease.

Thus, for larger values of r we staff fewer beds in (5.8). Although the time patients spend in a bed decreases, our simulation results suggest this is not proportional to the decrease of beds. Therefore the probability of holding and the expected holding time will increase, see Figure 7.3a and Figure 7.3c.



(a) Probability of holding.

(b) Probability of waiting.

(c) Expected holding time.

(d) Expected waiting time.

Figure 7.3: Influence of r.

7.2 Time-varying environment

As Figure 7.4 illustrates, the arrival rate differs every hour where most patients arrive during office hours. Therefore we need a staffing procedure for a non-stationary system such that every patient encounters the same service level at each moment in time. This has typically been addressed via two approaches. The first approach concerns stationary approximations, as in the PSA (Pointwise Stationary Approximation), SIPP (Stationary Independent Period by Period) or the lag-SIPP method. In the PSA method one fixes the arrival rate at the start of each time interval and uses this in a stationary model. In the SIPP-method the arrival rate is first averaged over each staffing interval before using it in a stationary model [19]. This method has been improved by the lag-SIPP method, where one slacks the arrival rate by the mean service time before applying the stationary-model method. As a result, systems with longer service times might be handled as well, yet it still requires that the steady state is quickly reached. The second approach approximates the time-varying offered load via a corresponding system with ample servers, and uses the time-varying adaption of the square root staffing rule. This method is referred to as MOL staffing method [19].

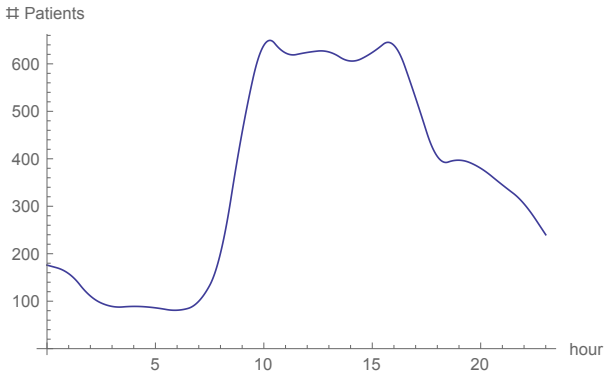


Figure 7.4: Number of arrivals in six months for an Emergency Department in the Netherlands.

Following Yom-Tov and Mandelbaum [33], we will consider the performance of the PSA and the MOL staffing method for a sinusoidal arrival rate:

$$\lambda(t) = \bar{\lambda} + \bar{\lambda}\kappa \sin(2\pi t/f) = \bar{\lambda} + \bar{\lambda}\kappa \sin(\omega t), \tag{7.1}$$

where $\bar{\lambda}$ is the average arrival rate, κ is the relative amplitude, f is the period and ω the frequency.

7.2.1 The PSA staffing procedure

The Pointwise Stationary Approximation (PSA) is a standard way to cope with time-dependent arrivals [19]. At the start of each time interval, the distribution of the number of patients is approximated by a stationary model with the same parameters and a fixed arrival rate set to $\lambda = \lambda(t)$, where t is the start time of the time interval.

The SIPP-method differs from the PSA method in the sense that the arrival rate is first averaged over each staffing interval before using it in a stationary model. In other words, for each time interval $[t_1, t_2]$ the arrival rate is set to $\lambda = 1/(t_2 - t_1) \int_{t_1}^{t_2} \lambda(t)\, dt$ before applying it to the corresponding stationary model. In the lag-SIPP method, the arrival rate is additionally slacked by the mean service time.

7.2.2 The MOL staffing procedure

The approach in the MOL staffing method is based on staffing for a time-varying offered load in a related infinite server system where all patients can be served immediately. We will approximate the offered load of the service station by the offered load in a related system where the number of beds and the number of nurses are set to infinity. Therefore, the modified offered load coincides with the modified offered load

in case of an Erlang-R model, for which the results are already established in [33]. For the SERB model, which also deals with an additional constraint of n, Yom-Tov already showed that this works well for the probability of waiting and blocking (only) in the QED regime [32].

Introduce the time-varying offered load $\{R(t), R_2(t)\}$, where $R(t)$ and $R_2(t)$ are the offered loads in the related Erlang-R model with an infinite number of servers and beds. In other words, $R(t)$ is the least number of servers required so no arriving patient needs to wait for service.

Since the Erlang-R model has exponentially distributed service times with rate μ and exponentially distributed content times with rate δ, the modified offered loads are given by the unique solution of the following ordinary differential equation [33]:

$$\frac{d}{dt}R(t) = \lambda(t) + \delta R_2(t) - \mu R(t), \tag{7.2}$$

$$\frac{d}{dt}R_2(t) = p\mu R(t) - \delta R_2(t), \tag{7.3}$$

for $t \geq 0$ and initial condition $R(0) = 0$ and $R_2(0) = 0$.

This system of equations can be understood as follows. At time t the service station finds an offered load from arriving patients and patients from the content station. With rate $\lambda(t)$ patients arrive to the service station and the content station processes a load of $R_2(t)$ with rate δ. Moreover, the service station processes a load of $R(t)$ with rate μ, and hence the change of offered load is captured by (7.2). Similarly, the service station processes with rate μ and turns to the content station with probability p. On the other hand, the content station processes load with rate δ, which yields (7.3). The formal proof of this system of differential equations is given in [32]. Moreover, Yom-Tov shows that this can be extended for general arrival times and general service times.

For the sinusoidal arrival rate in (7.1) the system of differential equations (7.2) and (7.3) can be solved explicitly [33].

Lemma 7.2.1. *The solution of* (7.2) *for the sinusoidal arrival rate* (7.1) *is given by*

$$R(t) = \frac{\bar{\lambda}}{(1-p)\mu} + \bar{\lambda}\kappa \cdot \cos(\omega t + \pi + \tan^{-1}(\theta)) \cdot \sqrt{\frac{\delta - i\omega}{(\mu - i\omega)(\delta - i\omega) - p\mu\delta} \cdot \frac{\delta + i\omega}{(\mu + i\omega)(\delta + i\omega) - p\mu\delta}}, \tag{7.4}$$

where $\theta = \mu(\delta^2 - p\delta^2 + \omega^2)/(\omega(\delta^2 + \omega^2 + p\mu\delta))$.

Using the modified offered load of (7.4), we dimension the SERW model by

$$n(t) = \frac{R(t)}{r} + \gamma\sqrt{\frac{R(t)}{r}}, \tag{7.5}$$

$$s(t) = R(t) + \beta\sqrt{R(t)}. \tag{7.6}$$

7.2.3 Performance

We are interested in whether the staffing procedure (7.5) and (7.6) via the MOL staffing method works well for a sinusoidal arrival stream. Moreover, we will compare its performance with the PSA staffing method and consider the influence of r.

7.2.3.1 Moderate system

First, we reconstruct the experiment of case study 1 in [32], thus we have the parameter setting $\bar{\lambda} = 30$, $\kappa = 0.2$, $\mu = 1$, $\delta = 0.5$, $p = 2/3$, $\eta = 1$, $\gamma = \beta\sqrt{r} + \eta\sqrt{1-r}$ and vary $\beta \in \{0.1, 0.3, 0.5, 0.7, 1, 1.5\}$. We will describe the behavior of all performance measures, and particularly consider the probability of holding and the probability of waiting, see Figure 7.6.

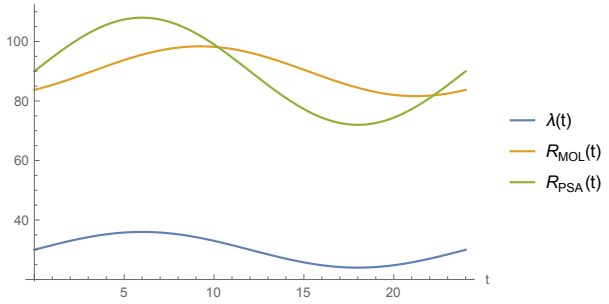
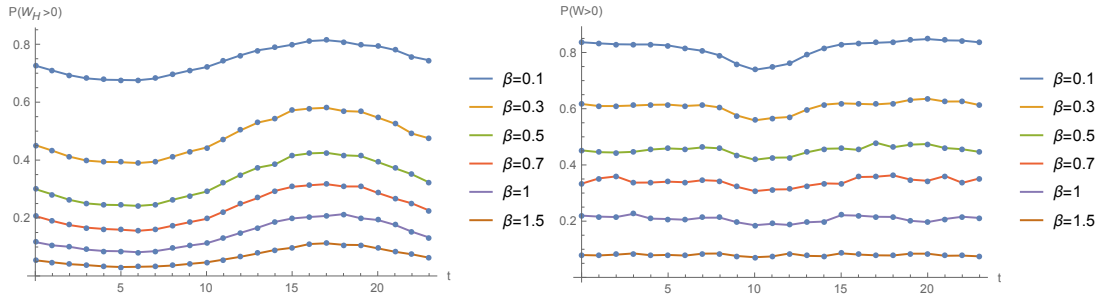


Figure 7.5: Arrival rate and offered loads.



(a) Probability of holding.

(b) Probability of waiting.

Figure 7.6: Performance of the PSA method for a sinusoidal arrival rate (moderate system).

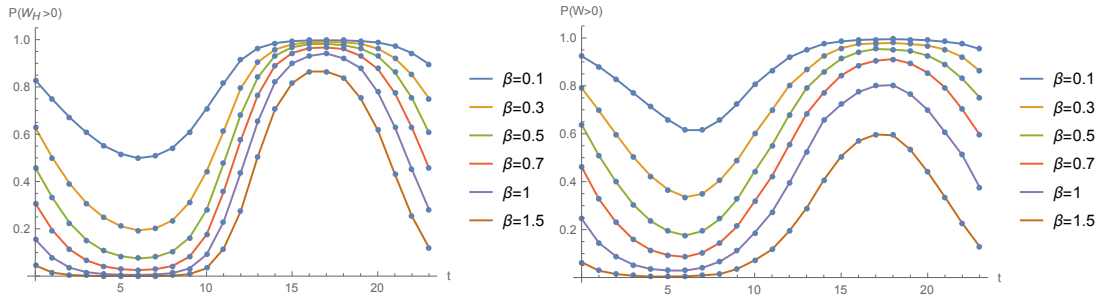
Our simulation results indicate that this staffing procedure has the desirable result of stabilizing the probability of waiting, as well as the bed utilization level and the nurse utilization level. We do observe a minor sinusoidal shape with respect to the probability of holding, which is asymmetric with respect to the offered load in the MOL staffing, see Figure 7.5. We note that this phenomenon is present even more clearly for the expected waiting time and the expected LOS. The expected length of stay at time t corresponds to patients leaving the Emergency Department at hour t. Hence, patients that leave when the system is highly congested experienced less waiting times during their stay on average. Thus, when the offered load increases the provided MOL staffing procedure, it seems that we tend to overstaff the number of nurses slightly, yielding a decrease in the expected waiting time. This can also indirectly provide the slight decrease in the holding probability.

We also considered the performance of the PSA method and the lag-PSA with a lag of $1/((1-p)\mu) = 3$ and $1/\mu = 1$ hours. We displayed the results for the PSA method in 7.7. We observe that the MOL staffing method stabilizes the waiting probabilities considerably more than the PSA. The lag-PSA method shows similar results, and hence we did not include these figures here. Thus, we find that the MOL method is the preferable staffing method for our model and therefore we will henceforth use this method for our staffing policies.

7.2.3.2 Small system

Next we consider a smaller system with a more realistic parameter setting for an Emergency Department. In this setting we have $\bar{\lambda} = 2.5$, $\kappa = 0.6$, $\mu = 1$, $\delta = 0.2$, $p = 2/3$, $\eta = 1$, $\gamma = \beta\sqrt{r} + \eta\sqrt{1-r}$ and vary $\beta \in \{0.1, 0.3, 0.5, 0.7, 1, 1.5\}$. This yields a system with $r = 0.2$ and $65 \leq n \leq 80$ and $12 \leq s \leq 20$.

The results for the probability of waiting and the probability of holding are included in Figure 7.8. In this example we observe that the probability of holding does not stabilize under this staffing procedure. Our MOL staffing rule (7.5) uses the modified offered load of the service station. We observe that the probability of holding for a bed does not stabilize to a single level, this shows that we cannot simply adopt the modified offered load for the service station to determine the number of beds. Further research



(a) Probability of holding.

(b) Probability of waiting.

Figure 7.7: Performance of the PSA method for a sinusoidal arrival rate (moderate system).

is necessary to obtain a deeper understanding of this relation.

7.2.3.3 Influence of r

To study the influence of r we fixed the parameters $\lambda = 2.5$, $\kappa = 0.6$, $f = 24$, $\mu = 1$, $p = 0.8$, and varied δ. Observe that r is an increasing function of δ, s is fixed and n will decrease as δ grows. The results for the probability of holding for a bed and waiting for a nurse for $r \in \{0.2, 0.5, 0.8\}$ are given in Figure 7.8.

As r increases the probability of waiting will fluctuate a little more, but this effect is quite modest. Interestingly, we find that the probability of holding for a bed displays more stabilizing behavior for larger r. Note that here we increase r such that n decreases while s is fixed. Therefore n grows to s, which yields improving results for the modified offered load approximation for the bed holding queue and hence more stabilizing behavior for the probability of holding.

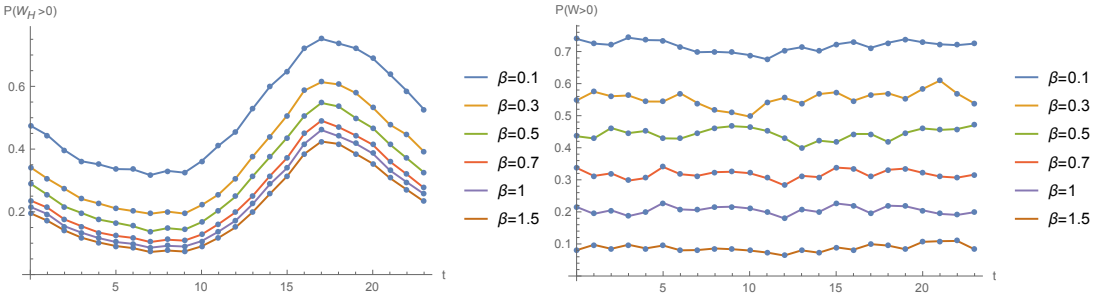
7.3 Rounding effect

Note that the rounding procedure of the number of nurses and the number of beds has a greater effect in smaller systems and hence the probability of waiting can fluctuate more. In our simulations we observed many fluctuations due to both rounding up the number of nurses and rounding up the number of beds with respect to a chosen R in smaller systems. Moreover, these fluctuations are more due to rounding up the number of nurses than the number of beds. In order to decrease these fluctuations, we can alternatively scale the system with respect to s. That is, asymptotically the staffing rules (5.8) and (5.9) are equivalent to

$$R = s - \beta\sqrt{s} \tag{7.7}$$

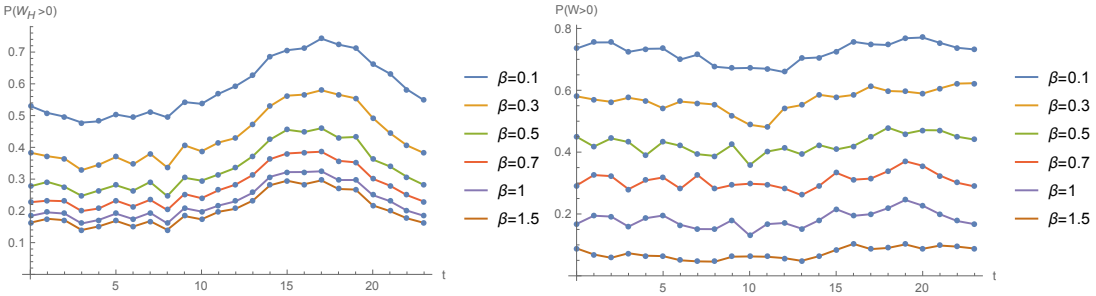
$$n = \frac{s}{r} + \left(\gamma - \frac{\beta}{\sqrt{r}}\sqrt{\frac{s}{r}}\right). \tag{7.8}$$

In the next Chapter we will take a closer look at the probability of waiting and its behavior. We will approximate it by means of a heuristic, and to validate the approximation it is necessary to reduce the impact of the rounding up. For the stationary case, we perform our experiments with the equivalent staffing rules (7.7) and (7.8).



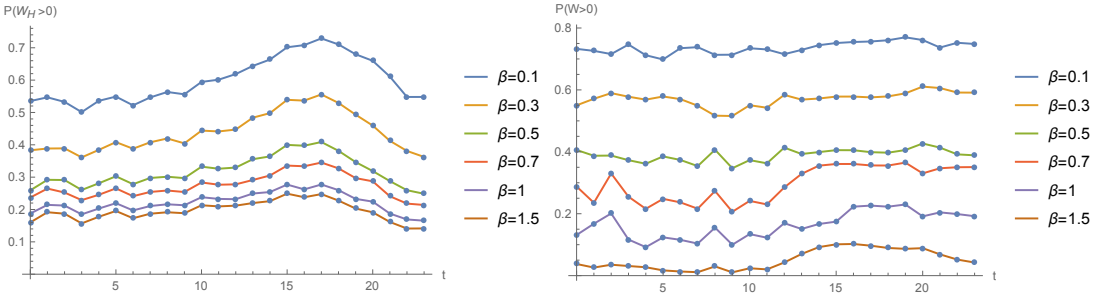
(a) Probability of holding for $r = 0.2$.

(b) Probability of waiting for $r = 0.2$.



(c) Probability of holding for $r = 0.5$.

(d) Probability of waiting for $r = 0.5$.



(e) Probability of holding for $r = 0.8$.

(f) Probability of waiting for $r = 0.8$.

Figure 7.8: Performance of the MOL staffing method for a sinusoidal rate (small system).

Chapter 8

Dimensioning scheme

In this chapter we will provide a heuristic that approximates the behavior of the Semi-Open Erlang-R model with Waiting (SERW) with that of the Semi-Open Erlang-R Model with Blocking (SERB). Since the stationary distribution of the SERB model does have a closed form representation, this heuristic method makes available closed-form, yet approximative, expressions for performance measures like the probability of waiting in the SERW model. These closed-form approximations will turn out to be useful with the MOL staffing method.

8.1 Inspiration

Our heuristic method is inspired by the recent work on the Erlang-B model with slow retrials [3], where slow refers to the feature that blocked arrivals will return to the system after a relatively long time compared to the service time. Cohen [9] showed that for a (stable) Erlang-B model with retrials, with an exponential retrial rate ξ, arrival rate λ and service rate μ, the steady state distribution of the number of busy servers converges to the corresponding distribution of the standard Erlang-B model as $\xi \downarrow 0$, but with an increased arrival rate $\lambda + \Omega$, where Ω is the unique positive root of the equation

$$\Omega = (\lambda + \Omega) B\left(s, \frac{\lambda + \Omega}{\mu}\right) \tag{8.1}$$

with $B(\cdot, \cdot)$ the Erlang-B formula in Equation (2.9). Equation (8.1) can be understood as follows. In case of slow retrials, the retrial rate becomes roughly Poisson and hence the system converges to a standard Erlang-B model with an increased arrival rate, namely the original arrival rate plus an additional arrival rate due to retrials. Then the additional arrival rate Ω equals the blocked arrival rate in the corresponding standard Erlang-B model.

The key difference between the SERW model and the SERB model is that patients finding all beds occupied upon arrival, wait indefinitely till there is a bed available or are blocked respectively. Our goal is to approximate the SERW model with an appropriate version of the SERB model. For this, we extend the SERB with the feature of retrials: blocked patients in the SERB model will retry to enter the system after some time. If we assume that this time until retrial is rather long (slow retrials), we find that the retrial process becomes roughly Poisson again. Then we obtain a corresponding SERB model with an increased offered arrival rate or equivalently, an increased offered load. The key question then is how to determine the limiting retrial rate, i.e. the counterpart of Ω in (8.1) in this case.

8.2 A heuristic method to connect SERW and SERB

First we will intuitively explain how one comes up with a counterpart of the Cohen's equation for the SERB model. It follows from Proposition 4.3.4 that for a standard SERB model, when n and s are

scaled according to the QED staffing rules (5.8) and (5.9),

$$\lim_{R\to\infty} \sqrt{R}P(\text{Block}) = f(\beta,\gamma),$$

where $\beta,\gamma > 0$ and

$$f(\beta,\gamma) := \frac{\sqrt{r}\phi(\gamma\sqrt{r})\Phi(\beta\sqrt{1-r}-\eta\sqrt{r}) + \phi(\sqrt{\eta^2+\beta^2})e^{\eta_1^2/2}\Phi(\eta_1)}{\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)} \tag{8.2}$$

and $\eta = \frac{\gamma}{\sqrt{1-r}} - \beta\sqrt{\frac{r}{1-r}}$ and $\eta_1 = \eta - \beta\sqrt{\frac{1-r}{r}}$. Thus the probability of blocking is of the order $\mathcal{O}(R^{-1/2})$, yielding a rate of blocked load of roughly $R\cdot\mathcal{O}(R^{-1/2}) = \mathcal{O}(\sqrt{R})$ per time unit. Therefore, we approximate the SERB model with retrials by a standard SERB model with an additional offered load of the order $\mathcal{O}(\sqrt{R})$.

We assume that the additional load takes the form $a\sqrt{R}$ with a some positive constant that is yet to be determined. This additional offered load should be equal to the the fraction of blocked total offered load in the corresponding (standard) SERB model. Hence, we obtain the following counterpart of Cohen's equation in this case:

$$(R + a\sqrt{R})P(\text{Block}) = a\sqrt{R}. \tag{8.3}$$

Next we will explain heuristically how we can use this result to approximate the behavior of a SERW model that is dimensioned according to (5.8) and (5.9). We relate this system with a SERB model with retrials with the same parameters. When we assume that the additional load takes the form $a\sqrt{R}$ with $a > 0$, then

$$R + a\sqrt{R} = s - (\beta - a)\sqrt{R} \sim s - (\beta - a)\sqrt{R + a\sqrt{R}},$$

and similarly

$$R + a\sqrt{R} = rn - (\gamma r - a\sqrt{r})\sqrt{\frac{R}{r}} \sim rn - (\gamma r - a\sqrt{r})\sqrt{\frac{R + a\sqrt{r}}{r}}.$$

In other words, we find the staffing rules

$$s = \hat{R} + (\beta - a)\sqrt{\hat{R}} \tag{8.4}$$

and

$$n = \frac{\hat{R}}{r} + \left(\gamma - \frac{a}{\sqrt{r}}\right)\sqrt{\frac{\hat{R}}{r}}. \tag{8.5}$$

for the corresponding standard SERB model with an offered load of $\hat{R} = R + a\sqrt{R}$. Then it follows from Proposition 4.3.4 and the counterpart of Cohen's Equation (8.3) that

$$f\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right) = \sqrt{R + a\sqrt{R}}P(\text{Block}) = \frac{a\sqrt{R}}{\sqrt{R + a\sqrt{R}}} \sim a$$

with f as in (8.2).

Let us now summarize the approximative scheme we have just constructed heuristically. We want to approximate the SERW model dimensioned according to the QED staffing rules (5.8) and (5.9). Therefore we approximate the behavior of the service station by a SERB model with an adapted offered load $\hat{R} = R + a\sqrt{R}$ and dimensioned according to (8.4) and (8.5), where a can be found by solving

$$f\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right) = a. \tag{8.6}$$

with f as in (8.2). Then the probability of waiting is approximately $g(\beta - a, \gamma - \frac{a}{\sqrt{r}})$, where

$$g(\beta,\gamma) := \left(1 + \frac{\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\frac{r}{1-r}})d\Phi(t)}{\frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\eta_1^2/2}\Phi(\eta_1)}\right)^{-1} \tag{8.7}$$

with $\eta = \frac{\gamma}{\sqrt{1-r}} - \beta\sqrt{\frac{r}{1-r}}$ and $\eta_1 = \eta - \beta\sqrt{\frac{1-r}{r}}$.

8.3 Open technical problem

When applying the heuristic we implicitly assume that (8.6) has a unique positive solution a. Although f defined in (8.2) looks rather complicated, this function is smooth, convex and strictly decreasing with respect to β and γ for all r. Based on our numerical findings, we state the following conjecture.

Conjecture 1. Assume that $0 < \beta < \sqrt{R}$ and $0 < \gamma < \sqrt{\frac{R}{r}}$. Then f defined in (8.2) is continuous and satisfies

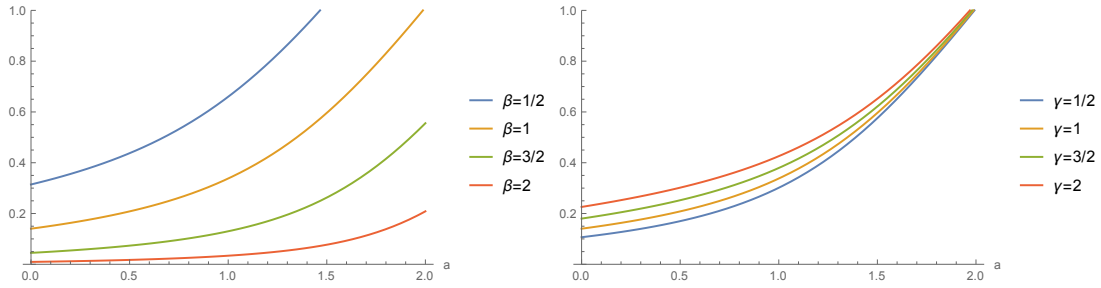
$$f(\beta, \gamma) > 0, \tag{8.8}$$

$$0 < \frac{df\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right)}{da} < 1 \tag{8.9}$$

and for $a > 0$ large enough

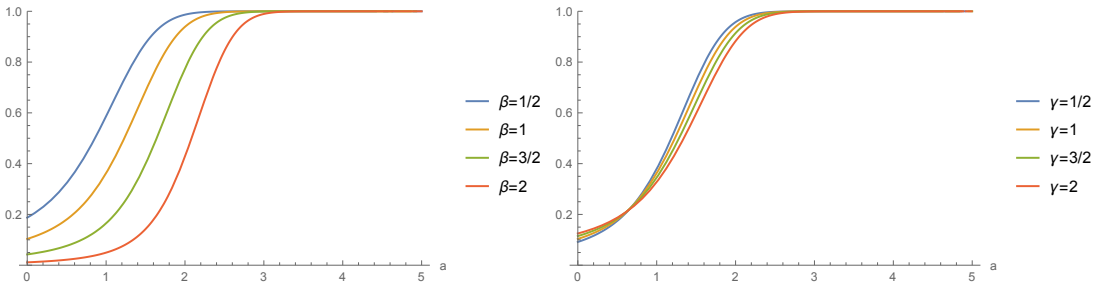
$$f\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right) < a. \tag{8.10}$$

Under Conjecture 1, (8.6) has a unique positive solution for a by the following reasoning. From (8.8) we have $f(\beta - a, \gamma - \frac{a}{\sqrt{r}}) > 0$ at $a = 0$. Combining this notion with (8.10) and the continuity of f, we see that (8.6) has at least one solution, e.g. Figure 8.1. Moreover, this is uniquely determined since $0 < \frac{df(\beta-a,\gamma-a/\sqrt{r})}{da} < 1$, see Figure 8.2.



(a) Case $\gamma = 1$. (b) Case $\beta = 1$.

Figure 8.1: Function $f(\beta - a, \gamma - a/\sqrt{r})$ for $r = 0.2$.



(a) Case $\gamma = 1$. (b) Case $\beta = 1$.

Figure 8.2: Derivative of $f(\beta - a, \gamma - a/\sqrt{r})$ w.r.t. a for $r = 0.2$.

8.4 Performance

We shall now apply the heuristic to approximate the probability of waiting in the SERW model in terms of the SERB model with slow retrials. Moreover, we will consider the significance of the retrial

feature with respect to the SERB model without retrials, i.e. the SERB model that is dimensioned as the original SERW model. We fix $\gamma = 1$ and consider the staffing rules (7.7) and (7.8) to decrease the effect of rounding in our analysis. For the shown results in this section we set $\mu = 1$, $p = 0.8$ and $\delta \in \{0.2, 0.8, 3.2\}$, which yields $r \in \{0.2, 0.5, 0.8\}$.

When patients wait for an available bed, this results in a higher bed utilization than in the case that patients are blocked when all beds are occupied. Consequently, there will be more patients at the service station on average in the SERW model and hence it is more likely that a patient waits for service. Therefore, the probability of waiting in the SERW model is typically underestimated by the probability of waiting in the standard SERB model without retrials. This effect can be quite severe for smaller β, as is illustrated in Figure 8.3. Here the straight line is the probability of waiting derived from our simulations for the SERW model, the dashed line is the heuristic approximation and the dotted line the asymptotic probability of waiting in the SERB model without retrials. We observe that the heuristic improves the approximation dramatically.

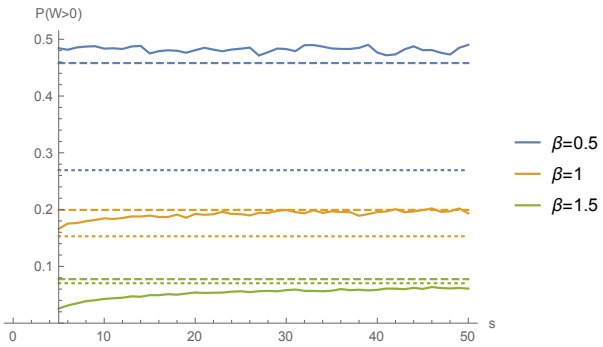


Figure 8.3: Probability of waiting for $r = 0.2$.

Important to note is that the heuristic is an approximation for the asymptotic probability of waiting of the SERW model. In other words, the probability of waiting in the SERW model will not converge to the value provided by the heuristic. This effect is illustrated in Table 8.1.

	$r = 0.2$			$r = 0.5$			$r = 0.8$		
β	Heuristic	QED	Gap	Heuristic	QED	Error	Heuristic	QED	Gap
0.1	0.8472	0.8391	0.0081	0.7971	0.8409	0.0439	0.6897	0.8644	0.1747
0.3	0.6243	0.6439	0.0197	0.5424	0.6288	0.0864	0.3759	0.6559	0.2800
0.5	0.4581	0.4804	0.0223	0.3761	0.4607	0.0846	0.2225	0.4711	0.2486
0.7	0.3322	0.3504	0.0182	0.2638	0.3226	0.0588	0.1487	0.3157	0.1670
1	0.1994	0.2020	0.0026	0.1579	0.1760	0.0181	0.1000	0.1498	0.0499
1.5	0.0775	0.0681	0.0094	0.0670	0.0505	0.0164	0.0567	0.0225	0.0343

Table 8.1: Performance of heuristic w.r.t. asymptotic $P(W > 0)$ in SERW model.

We observe that for small r the heuristic approximates the asymptotic probability of waiting rather well. As r increases the gap between the asymptotic probability of waiting and the heuristic approximation increases as well. This error is modest enough for small r, but as r increases the gap can exceed 25%. From our simulation results we find that the heuristic can only be used for $r < 0.5$. Since the value of r is typically low for an Emergency Department, we can use this machinery for dimensioning purposes in this case.

From our simulation results we observed that the probability of waiting converges quickly to its asymptotic value, and therefore the heuristic already approximates the true probability of waiting rather well for small systems (in case of small r). This is illustrated for $r = 0.2$, $\beta = 0.5$ and $\gamma = 1$ in Table 8.2. In theory, as s grows the gap between the probability of waiting and the QED value will converge to zero and the error of the heuristic will converge to the gap between the asymptotic probability of waiting

and the heuristic approximation (0.0233 in this case). Since the errors are very small in this case, this convergence cannot be verified from this table. Nevertheless, it does show that the SERB model without retrials would provide a poor approximation, as expected, and that the heuristic provides a good approximation with a small error of around 2%.

s	n	$P(W>0)$	Gap QED	Error Heuristic	Error SERB
5	25	0.4841	0.0037	0.0260	0.2147
6	30	0.4816	0.0011	0.0234	0.2121
7	35	0.4856	0.0052	0.0275	0.2162
8	40	0.4871	0.0067	0.0290	0.2177
9	45	0.4877	0.0072	0.0296	0.2182
10	50	0.4835	0.0031	0.0254	0.2141
15	74	0.4750	0.0054	0.0169	0.2056
20	99	0.4807	0.0003	0.0226	0.2113
25	124	0.4833	0.0029	0.0252	0.2139
30	149	0.4828	0.0023	0.0246	0.2133
40	199	0.4766	0.0038	0.0185	0.2072
50	249	0.4901	0.0097	0.0320	0.2207

Table 8.2: Influence of s.

Summarizing, for sufficiently small r the heuristic provides good approximations for the probability of waiting in the stationary case. Next, we will explain intuitively why larger values of r will yield poor approximations.

Note that large values of δ with respect to μ correspond to high values of r. Moreover, the patient returns faster to service as δ increases and thus the frequency that patients depart the Emergency Department will increase. In the SERW model a holding patient will be immediately assigned to a bed. However, we assume a slow retrial rate in our heuristic and consequently, it can take a relatively long time for a patient to retry for a bed. That is, the retrial rate is too small to approximate the actual behavior of the SERW model where the patient immediately moves to the service queue. This effect will be more significant when the frequency of departures increases and hence the heuristic will provide poorer estimations.

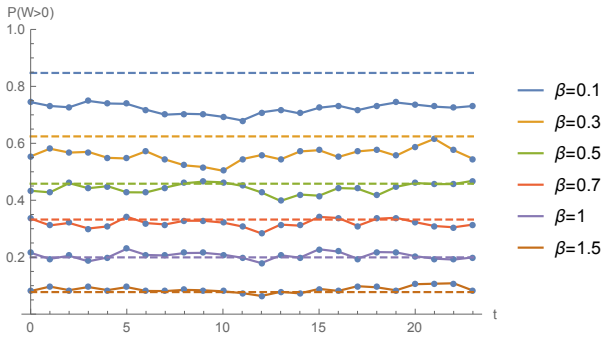


Figure 8.4: Heuristic approximation on the probability of waiting for a sinusoidal arrival stream.

Lastly, we consider whether the heuristic works well for a sinusoidal arrival rate in case of small r. We will illustrate our results for the case described in Section 7.2.3.2, where we scaled again by the staffing rules (7.5) and (7.6) in this time-varying setting. Observe that in this case we have a large relative amplitude and decreasing the relative amplitude will lead to more stationary behavior. In other words, there will be less fluctuations as the relative amplitude decreases.

Even in this time-varying setting, it seems that the heuristic works quite well in case that the probability of waiting is not too high, which is illustrated in Figure 8.4 and Table 8.3. That is, in case that

the probability of waiting for a nurse we want to achieve is not too high, the heuristic can be used for dimensioning the Emergency Department. Nevertheless, more simulations and further research is needed to obtain a deeper understanding on the performance of the heuristic approximation, particularly in a time-varying setting.

t	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 1$	$\beta = 1.5$
1	0.1025	0.0701	0.0248	0.0036	0.0159	0.0036
2	0.1160	0.0431	0.0304	0.0206	0.0052	0.0193
3	0.1206	0.0569	0.0019	0.0119	0.0056	0.0067
4	0.0981	0.0544	0.0158	0.0314	0.0138	0.0176
5	0.1072	0.0752	0.0086	0.0240	0.0004	0.0069
6	0.1086	0.0776	0.0301	0.0100	0.0301	0.0174
7	0.1291	0.0521	0.0305	0.0124	0.0084	0.0040
8	0.1466	0.0811	0.0136	0.0171	0.0067	0.0035
9	0.1434	0.1016	0.0015	0.0042	0.0152	0.0079
10	0.1451	0.1076	0.0075	0.0043	0.0160	0.0062
11	0.1541	0.1207	0.0049	0.0106	0.0096	0.0027
12	0.1669	0.0799	0.0077	0.0243	0.0014	0.0035
13	0.1384	0.0655	0.0306	0.0485	0.0201	0.0136
14	0.1289	0.0807	0.0605	0.0177	0.0074	0.0009
15	0.1422	0.0530	0.0380	0.0214	0.0010	0.0046
16	0.1223	0.0483	0.0421	0.0093	0.0285	0.0112
17	0.1149	0.0724	0.0152	0.0055	0.0203	0.0041
18	0.1307	0.0527	0.0173	0.0221	0.0044	0.0206
19	0.1154	0.0469	0.0415	0.0028	0.0186	0.0168
20	0.1037	0.0678	0.0111	0.0062	0.0177	0.0062
21	0.1112	0.0357	0.0026	0.0092	0.0041	0.0276
22	0.1183	0.0085	0.0015	0.0226	0.0045	0.0294
23	0.1221	0.0481	0.0013	0.0272	0.0078	0.0309
24	0.1161	0.0786	0.0104	0.0177	0.0001	0.0053

Table 8.3: Error values of heuristic.

8.5 Dimensioning of an Emergency Department

We now illustrate how this heuristic can help in dimensioning an Emergency Department using a constraint based example. When there is a fixed number of beds available, solve (5.8) for γ. In case that on average no more than q% of the patients should wait for service, (numerically) solve

$$f\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right) = a,$$
$$g\left(\beta - a, \gamma - \frac{a}{\sqrt{r}}\right) = q,$$

for the values of β and a, where f is as in (8.2) and g as in (8.7). Then set the number of nurses to $s = R + \beta\sqrt{R}$.

We note that this dimensioning scheme can also be applied for a time-varying setting. Let t represent the start of a time interval. At the start of each time interval we solve (5.8) for γ_t and the system of equations for β_t and a_t. Then we set the number of nurses to $s = R(t) + \beta\sqrt{R(t)}$, where $R(t)$ is the modified offered load of the MOL staffing procedure.

Example 1. Suppose there are $n = 30$ beds available in the Emergency Department of a hospital with $\mu = 5$, $\delta = 2$ and $p = 0.8$. How many nurses are needed such that no more than 30% of the patients need to wait for service?

When there are five patients arriving per hour, this yields an offered load of $R = 5$ and hence $\gamma = 1$. Then numerically solving

$$f\left(\beta - a, 1 - \frac{a}{\sqrt{0.2}}\right) = a,$$
$$g\left(\beta - a, 1 - \frac{a}{\sqrt{0.2}}\right) = 0.4,$$

yields $a \approx 0.3$ and $\beta \approx 0.75$. Hence we would need $s = \lceil 5 + 0.75\sqrt{5} \rceil = 7$ nurses.

Chapter 9

Conclusions

In this thesis we introduced a large-scale service system facing patients returning for service in order to determine dimensioning rules for the Emergency Department in the QED regime, in terms of both the number of beds and the number of nurses.

9.1 Our contributions

The main research objective of this thesis was to analyze and improve the patient flow in Emergency Departments. Our study was based on a new stochastic model that we called the Semi-Open Erlang-R model with Waiting. This model builds on recently developed models and adds the essential restriction that the total number of patients that can reside in the Emergency Department simultaneously is bounded. Our study has led to three main contributions:

1. We explained how the classical queueing models in the QED regime can be used in hospital settings. Moreover, we identified the prevalent features that need to be accounted for in order to dimension an Emergency Department.

2. We introduced and analyzed the Semi-Open Erlang-R Model with Waiting (SERW) and

   - Showed how the stationary distribution of the number of patients at each station can be determined by matrix-geometric methods and identified the stability condition of this system;
   - Made stochastic comparisons with other models to obtain insights in the behavior of the SERW model;
   - Proposed a two-fold scaling rule that leads to QED behavior in the SERW model.

3. Using simulations, we validated our proposed staffing policy for both stationary and time-varying environments, and identified the conditions for which the proposed staffing policy stabilizes performance. We provided a practical algorithm that can be used for dimensioning the Emergency Department in real-life.

9.2 Our results

Most delays experienced in the Emergency Department are due to a lack of beds or nurses. Since these two resources are highly interdependent, it is essential to find a suitable dimensioning rule for both beds and nurses simultaneously. To include the important feature of patients returning multiple times for service by a nurse we adopted the Erlang-R model, but extending it with the restriction that the number of patients that can reside in the Emergency Department simultaneously is bounded. We called our model the Semi-Open Erlang-R Model with Waiting and identified the staffing policy that stabilizes

performance in the QED regime. In the stationary case, the number of beds n and the number of servers s should be dimensioned by the following square-root staffing rules:

$$n = \frac{R}{r} + \gamma\sqrt{\frac{R}{r}},$$
$$s = R + \beta\sqrt{R}.$$

Via our simulations we verified that these staffing rules indeed lead to QED behavior, e.g. a limiting probability of waiting strictly between zero and one. However, since the arrival process into an Emergency Department is typically non-stationary, we translated our two-fold staffing rule into

$$n(t) = \frac{R(t)}{r} + \gamma\sqrt{\frac{R(t)}{r}},$$
$$s(t) = R(t) + \beta\sqrt{R(t)}.$$

We use the MOL (Modified Offered Load) staffing method to approximate the time-varying offered load $R(t)$. That is, we approximate the offered load via a corresponding system with infinitely many servers, which gives an explicit expression in case of a sinusoidal arrival rate. Our simulations indicated that this staffing procedure again stabilizes the probability of waiting for a nurse in the QED regime.

Since the stationary distribution of the SERW model does not have a closed-form expression, we designed a heuristic method to approximate the probability of waiting by a corresponding SERB model with retrials. Using this heuristic method, we designed a practical algorithm that determines the number of nurses for a predetermined number of beds and a target probability of waiting. Moreover, we verified via simulations this heuristic works well in the QED regime.

9.3 Open problems

Naturally, our study gives rise to various extensions that might lead to a more realistic model for the Emergency Department. Moreover, our mathematical model raises several open problems that are worth investigation. We now present an overview of these extensions and open problems.

Multi-class patients:
Before patients are treated by any care provider, the severity of their condition is established during triage. This process provides an urgency ordering and the patients are served accordingly. When assumed that every class of patients has the same mean service time, this scenario corresponds to the SERW model illustrated in Figure 9.1 for two types of patients. This can be extended for many types of patients, or even with different service rates for each patient.

To be able to provide immediate care to the prioritized patients, it seems natural to reserve some servers for these prioritized patients. In the queueing literature, this concept is also known as trunk reservation. The main objective for this model is finding the number of beds and nurses that should be reserved such that certain predetermined performance levels are achieved. Moreover, we are interested in how this affects our original staffing policy.

For the Erlang-B model, the answers of these types of questions are already known (see e.g. [30, 31]). An answer in this more complex queueing system provides important and relevant challenges for further research. The simulation program we have created can be easily extended for example to investigate these more complex situations.

Multi-type servers:
Patients are typically seen by both nurses and doctors during their stay. When assumed that service of doctors and nurses are required separately and independently, we can include an additional service station in the inner box of the SERW model representing the doctors. However, there is a cost difference between nurses and doctors, and hence it can be preferable to staff the doctors in the Efficiency-Driven regime with very high utilization and the nurses in the QED regime. Moreover, we note that the number

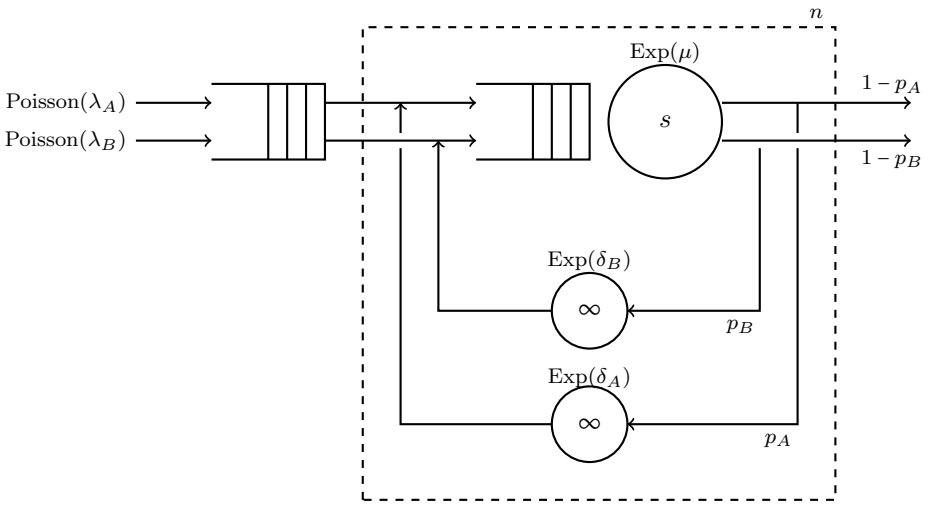


Figure 9.1: The SERW model with two types of patients.

of doctors is typically low in an Emergency Department and hence heavy-traffic approximations might not work well. It would be interesting to see if one can still find appropriate staffing policies in this case.

Bed staffing rule:
In our simulation results we observed that the probability of waiting for a bed does not stabilize by staffing the number of beds in a time-varying environment. Note that in (7.5) we use the same offered load approximation of the service queue to approximate the offered load for the bed queue, which might cause the less stabilizing behavior for the probability of holding.

It would be interesting to find an alternative method to approximate the offered load for the bed queue such that the probability of holding for a bed stabilizes. One possible method is by neglecting the waiting time of a nurse, since this converges to zero as $\mathcal{O}\left(s^{-1/2}\right)$ as the offered load and s grow to infinity. Then, the inner box can be considered as an $M_t/G/n$ queue with Poisson arrivals with rate $\lambda(t)$ and a service distribution given in (5.5). Following [12], the offered load of the bed queue can be approximated by $R_{\text{bed}}(t) := \int_{-\infty}^{t} (1 - F_B(t-u))\,\lambda(u)\,du$, where F_B is the probability distribution function of B. Then set the number of beds to $n(t) = R_{\text{bed}}(t) + \gamma\sqrt{R_{\text{bed}}(t)}$ with $\gamma > 0$. We note that this staffing rule is equivalent to (5.8) in the stationary case. Nevertheless, further research is needed to obtain an appropriate time-varying diffusion approximation for the offered load of the bed queue and understand why such an approximation would result in stabilizing behavior.

Rounding effect:
In our simulation results we found that the effect of rounding up the number of servers and the number of beds is quite significant. In Section 7.3 we already proposed an equivalent scaling to reduce this effect for our performance analysis. Nevertheless, in order to have a clean performance analysis, we would like to design a method that can account for noninteger values of s and n. One way to do that would be to perform the asymptotic analysis on 'analytical continuations' of the performance measures that hold for non-integer n and s, in the same spirit as writing the Erlang-B formula as an incomplete gamma function [23].

Proving stochastic ordering:
Rigorously proving the stochastic ordering we discovered in Chapter 6 provides a further mathematical challenge. Moreover, how can these results be used to bound the performance measures of interest? This will require delicate reasoning and a deep understanding of the relation in behavior of the systems. Nevertheless, if it is possible to bound the performance measures, we might be able to rigorously prove that the performance of SERW model stabilizes.

Moreover, in Section 5.7 we reasoned that the probability of waiting can be 'sandwiched', i.e. bounded from below and above, by the probability of waiting of the corresponding SERB model and the $M/M/s//n$

model, respectively. This reasoning is implicitly based on a stochastic ordering of these three models for the number of patients at the service queue. Our simulations support this reasoning, but this relation still requires a proof. Moreover, how tight are these bounds and can we improve them?

Prove that Equation (8.6) *has a unique solution:*
For the proposed dimensioning scheme based on the heuristic in Chapter 8 we conjectured that (8.6) has a unique positive solution. To ensure that the heuristic provides a unique approximation, it needs to be proven that this is indeed the case. Moreover, following the idea of the heuristic, the SERB models with retrials should also provide an approximation for the expected waiting time in the SERW model or any other stationary performance measure, but more simulations are needed to verify this.

Data analysis:
We believe that one of the most important challenges for future research is testing our model and staffing algorithms to real hospital data. For instance, we assumed exponential service times and content times in our models for mathematical tractability, which is typically not the case [1]. How robust are our results when it is tested to detailed hospital data? What features arise from the data and how can we account for this in our model?

Moreover, when real traces of arrival times at Emergency Departments are made available, our time-varying staffing policy can be tested and hopefully leads to stabilized performance, just as for the sinusoidal arrival rate. This would be a great avenue for further research.

In this thesis we introduced a simple model to capture a complex reality. Probably the most challenging opportunity for further research is to extend this model to account for essential features that arise in an Emergency Department, such as multi-class patients and multi-type servers. However, there are many more factors that should be accounted for and we believe that collaboration between the different fields is essential to identify these features.

Appendices

Appendix A

Preliminaries

In this appendix several basic mathematical concepts will be explained that are used throughout this thesis.

A.1 Basic concepts from probability theory

To describe the arrival and service process of a queueing model, we will use certain notions from probability theory. Therefore, we will provide some basic concepts in this section.

Within probability theory, we represent a quantity subject to variation due to chance by a random variable X. For example, X can represent the number of patients within the ICU, but also the interarrival time of two patients within the ICU. We denote its mean by $E(X)$ and its variation by $\sigma^2(X)$, where $\sigma(X)$ is the standard deviation of X. A measure of the variability of X is the coefficient of variation, defined as

$$c_X = \frac{\sigma(X)}{E(X)}. \tag{A.1}$$

Alternatively, a continuous nonnegative random variables can be expressed by its Laplace-Stieltjes transform. For a random variable X with distribution function $F(\cdot)$, this is defined as

$$\tilde{X}(s) = E(e^{-sX}) = \int_0^\infty e^{-sx}\, dF(x), \quad s \geq 0.$$

Observe that $|\tilde{X}(s)| \leq 1$ for all $s \geq 0$ and

$$\tilde{X}(0) = 1 \quad \tilde{X}'(0) = -E(X) \quad \tilde{X}^{(k)}(0) = (-1)^k E(X^k).$$

Every random variable has a unique Laplace-Stieltjes transform and vice versa, and therefore it is a characterization of the random variable. Often it is more convenient to work with the Laplace-Stieltjes transform of a random variable instead of its distribution for its nice properties. For example, let X, Y be two independent random variables. If $Z = X + Y$, then the Laplace-Stieltjes transform of Z is given by $\tilde{Z}(s) = \tilde{X}(s)\tilde{Y}(s)$. Moreover, if Z is equal to X with probability p and equal to Y with probability $1 - p$, then $\tilde{Z}(s) = p\tilde{X}(s) + (1 - p)\tilde{Y}(s)$.

Next, we will describe a few commonly used distributions.

A.1.1 Normal distribution

One of the most commonly encountered continuous probability distributions is the normal (or Gaussian) distribution. It is fully described by its mean μ and and variance σ^2 and is denoted by $N(\mu, \sigma)$. The probability density function of a random variable X that is $N(\mu, \sigma)$ distributed, is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

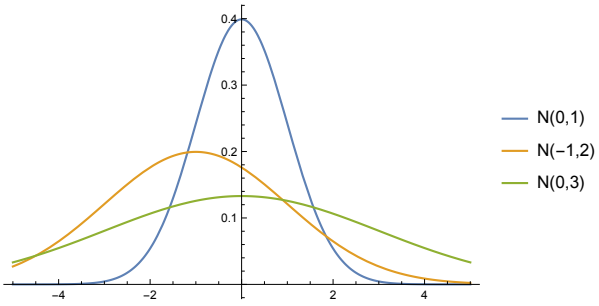


Figure A.1: Probability density function of $N(-1,2)$, $N(0,1)$ and $N(0,3)$.

In particular, we have the standard normal distribution with $\mu = 0$ and $\sigma = 1$. Then the probability density function is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

In this thesis, we will denote the density function of the standard normal distribution as ϕ and the cumulative distribution function by Φ.

The normal distribution is among other things extremely relevant due to its key role in the central limit theorem (CLT), which we will use multiple times in this thesis. Informally, this theorem states that the arithmetic mean of a sufficiently large number of independent random variables will be approximately normally distributed. Many formulations of the CLT exist, such as the following.

Theorem A.1.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with finite mean μ and finite standard deviation σ. Then*

$$\lim_{n\to\infty} P\left(\frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x),$$

where Φ denotes the standard cumulative normal distribution function.

Equivalently, if $S_n = \sum_{k=1}^{n} X_k$, then the random variable

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to $N(0,1)$ as $n \to \infty$.

A.1.2 Geometric distribution

A (discrete) random variable X with a geometric distribution with parameter p has probability distribution

$$P(X = k) = (1-p)p^k \quad k = 0, 1, 2, \ldots$$

Its mean, variance and coefficient of variance are given by

$$E(X) = \tfrac{p}{1-p} \quad \sigma^2(X) = \tfrac{p}{(1-p)^2} \quad c_X^2 = \tfrac{1}{p}.$$

A.1.3 Exponential distribution

The exponential distribution is a continuous probability distribution often used to model the times between two events, e.g. the interarrival times. If X denotes the exponentially distributed random variable with rate λ, for example representing the time for a next patient to arrive, then the probability that the patient arrives within t time units is given by

$$P(X \le t) = 1 - e^{-\lambda t}.$$

We have

$$E(X) = \frac{1}{\lambda}, \quad \sigma^2(X) = \frac{1}{\lambda^2} \quad c_x = 1,$$

and its Laplace-Stieltjes transform is given by

$$\tilde{X}(s) = \frac{\lambda}{\lambda+s}, \quad s \geq 0.$$

A key property of an exponential distribution is the memoryless property. Informally, this property states that the time spent in some state only depends on the state where it is presently. In other words, the time that the process will remain in the state that it is currently at will not be prolonged or shortened by anything that happened in the past. Mathematically, we can write this as

$$P(X > x + t | X > t) = P(X > x) = e^{-\lambda x},$$

for all $x \geq 0$ and $t \geq 0$, where X is an exponentially distributed random variable with rate λ. To be concrete, let X represent the arrival time for a new patient to arrive at the Emergency Department. The memoryless property states that the remaining interarrival time, given that the patient has not arrived before time t, is again exponentially distributed with the same rate λ. That is, the remaining arrival time is independent of the time past since the previous arrived patient.

If $X_1, ..., X_n$ are independent exponentially distributed random variables with parameters $\lambda_1, ..., \lambda_n$, then $\min\{X_1, ..., X_n\}$ is again exponentially distributed with rate $\lambda_1 + ... + \lambda_n$. The probability that X_i is the smallest one is given by the fraction $\lambda_i/(\lambda_1 + ... + \lambda_n)$, $i = 1, ..., n$.

A.1.4 Erlang distribution

A random variable X has an Erlang-k distribution, $k \in \mathbb{N}$, if X is the sum of k independent exponentials with the same mean. That is, if $X_1, ..., X_n$ are all exponentially distributed random variables with rate λ, then $X = X_1 + ... + X_k$ is Erlang-k distributed. A phase diagram of the Erlang-k distribution is shown in Figure A.2.

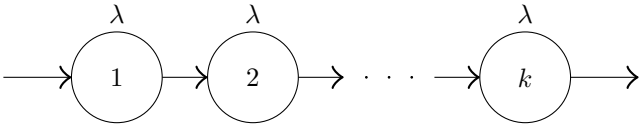


Figure A.2: Phase-independent diagram of the Erlang-k distribution.

The density is given by

$$f(t) = \lambda \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}, \quad t > 0,$$

and its distribution function equals

$$F(t) = 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t}, \quad t \geq 0.$$

The distribution function shows strong similarities with the Poisson distribution, which we consider in Section A.1.5. Its mean, variance and coefficient of variation are given by

$$E(X) = \frac{k}{\lambda}, \quad \sigma^2(X) = \frac{k}{\lambda^2}, \quad c_X^2 = \frac{1}{k}.$$

Since X is the sum of k independent exponentially distributed random variables, its Laplace-Stieltjes transform is given by

$$\tilde{X}(s) = \left(\frac{\lambda}{\lambda+s}\right)^k, \quad s \geq 0.$$

A.1.5 Poisson distribution

For now we will only give a formal description of the Poisson distribution and mention some properties. In Section A.2 the Poisson *process* will be explained. A Poisson random variable X with parameter λ has probability distribution

$$P(X = n) = \frac{\lambda^n}{n!}e^{-\lambda}, \quad n = 0, 1, 2, ...$$

Its mean, variance and coefficient of variation are given by

$$E(X) = \sigma^2(X) = \lambda, \quad c_X = \frac{1}{\sqrt{\lambda}}.$$

A.2 Poisson process

The Poisson process is the most commonly used assumption to model the arrival process of patients to the Emergency Department [21, p. 206]. Let $Q(t)$ be a random variable, denoting the number of arrivals in the interval $[0, t]$, which is described by a Poisson process with rate λ. That is, the time between successive arrivals is exponentially distributed with parameter λ. Then $Q(t)$ has a Poisson distribution with parameter λt, i.e.

$$P(Q(t) = n) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \quad n = 0, 1, 2, ...$$

The rate λ represents the expected number of arrivals per unit of time. For example, if $Q(t)$ describes the number of arriving patients within the Emergency Department per hour, then rate $\lambda = 3$ corresponds to an expected number of three arrivals per hour.

The Poisson process has the property that anywhere in time the occurrence of an arrival is equally likely, which can be derived from the memoryless property of the exponential distribution. To provide a concrete example, the likelihood that a new patient arrives to the Emergency Department is independent of the number of patients that already arrived that day. Moreover, it is independent of the arrival times of the preceding patients.

The following three properties define a Poisson process [22, p. 287]:

1. Patients arrive one at a time.
2. The probability that a patient arrives at any time is independent of when other patients arrived.
3. The probability that a patient arrives at a given time is independent of time.

To determine whether a Poisson process is reasonable for modeling purposes, it is useful to consider these properties. For instance, the Poisson process does not incorporate events such as major accidents, since such events trigger multiple simultaneous arrivals.

Next, we note two another important properties of the Poisson process.

(i) *Merging:* let $Q_1(t)$ and $Q_2(t)$ be two independent Poisson processes with respective rates λ_1 and λ_2. Then the sum $Q_1(t) + Q_2(t)$ is again a Poisson process with rate $\lambda_1 + \lambda_2$.

(ii) *Splitting:* Let $Q(t)$ be a Poisson process with rate λ. With probability p the arrival is of type one and with probability $1 - p$ of type two. Let $Q_1(t)$ and $Q_2(t)$ denote the arrival process of type one and type two patients respectively. Then $Q_1(t)$ and $Q_2(t)$ are independent Poisson processes with respective rates λp and $\lambda(1 - p)$.

To illustrate property (i), consider two clinical wards A en B within a hospital. Assume that the arrival process of the wards can be modeled by a Poisson process with rates $\lambda_1 + \lambda_2$. If clinical wards A en B would be merged to one big ward, according to property (i), its arrival process is described by a Poisson process with rate λ_1 and λ_2. In contrast, let the arrival process of a clinical ward be described by a Poisson process with rate λ. There are two types of patients in need of slightly different care. To be able to provide directed care, it is considered to separate the ward in two more specialized wards. Suppose patients of type one arrive with probability p, and so patients of type 2 arrive with probability $1 - p$. Let $Q_1(t)$ and $Q_2(t)$ denote the arrival process of the two specialized wards. Then the arrival

process of these specialized wards are described by independent Poisson processes with rates λp and $\lambda(1-p)$.

Last, we will mention a special property that holds for queueing systems with Poisson arrivals, namely the Poisson-Arrivals-See-Time-Averages (PASTA) property. This property states that on average an arriving patient finds the system in the same state as an outside random observer at an arbitrary point in time. More precisely, the fraction of patients finding the system in some state equals the fraction of time that the system is in this state. To provide an illustrative example, consider again the example of arriving patients to the Emergency Department. Then the PASTA property states that the probability that an arriving patient finds twelve patients in the ED equals the fraction of time that there actually are twelve patients in the ED. We emphasize that this property only holds for Poisson arrivals. To provide a counterexample, consider the D/D/1 queue with deterministic (fixed) interarrival times of two time units and a service time of one time unit. Then every arriving patient finds the system empty and can be served immediately. However, half of the time the system contains one patient.

A.3 Little's law

This principle in queueing theory provides an important relation between the mean number of patients in the system and the mean time that a patient spends in the system. Let λ denote the average number of patients entering the system per unit of time. We emphasize that the arrival process does not need to be Poisson. Let Q denote the number of patients in the system and S the time spent in the system. Then Little's law states that

$$E(Q) = \lambda E(S).$$

In words, the mean number of patients in the system equals the mean time spends in the system times the average number of arrivals per time unit. Intuitively, this law might seem reasonable and even trivial. However, it is a quite remarkable result, since it is not influenced by the arrival process distribution, the service distribution or the service discipline.

An equivalent definition can be given to describe the relation between the mean waiting time $E(W)$ and the mean queue length $E(Q^w)$ (number of patients that must wait for service). This is given by

$$E(Q^w) = \lambda E(W).$$

Let B denote the service time and ρ the occupation rate, i.e. the fraction of time the server is busy. In terms of the server, Little's law states

$$\rho = \lambda E(B).$$

Appendix B

Simulation

Since there is no analytical expression for the probability distribution of the Semi-Open Erlang-R model with Waiting (SERW), we resort to simulations to obtain insight in the limit behavior and compare these results with the Erlang-R model and the Semi-Open Erlang-R model with Blocking (SERB). In this appendix we will consider how we these three models can be simulated.

B.1 General setup

For all three models we make use of a discrete-event simulation. In a discrete-event simulation, one typically keeps track of the events and how these events affect the system state. In general, a number of different events occurs, each yielding its own corresponding activities. We create an event list that keeps track of the time points at which the events (of any type) are scheduled to occur. Then, the simulation consists of (iteratively) finding the next event with the smallest time point. The current time is set to this event time and the corresponding activities are executed. A typical construction of such a simulation program is given in Algorithm 1.

Input : Parameters (e.g. arrival/service distribution and parameters) and runlength T.
Output : Desired statistical measures.

Initialize: Simulation time $t = 0$, system state, statistical counters and event list.

while $t < T$ **do**

1. Determine next event type;
2. Update simulation time;
3. Update system state and statistical counters;
4. Generate future events and update event list;

end
return Desired statistical measures;

Algorithm 1: Typical construction of a discrete-event simulation.

In our patient flow models the changes in the system state are due to patient arrivals, service completions and the events that a patient assigned to a bed returns to service. We will refer to the latter as a content completion event. Moreover, we will refer to the station with patients in the needy state as the service queue, the station with patients in the content state as the content queue and the station with patients waiting for a bed as the holding queue.

Instead of setting a (maximum) runlength in our simulation, we choose to set a threshold K on the number of arriving patients. Each of these K arriving patients possesses some attributes, namely an

arrival time, an array of service times and an array of content times. Moreover, we (possibly) store the last time stamp that a patient starts waiting for an available nurse. In each iteration, we keep track of certain statistical counters that are of interest, such as the number of waiting patients, waiting times and utilization levels. These values depend on the parameters of the simulation, of which an overview is given in Table B.1.

Parameter	Description
n	Number of beds
s	Number of nurses
λ	Arrival rate
μ	Service rate
δ	Content rate
p	Probability of staying in the Emergency Department after service completion
K	Number of patients that arrive in the simulation

Table B.1: Parameters of simulation.

In Figure B.1 we find the general construction of the simulation we used for all three models. At initialization we start with empty queues, the number of arrived patients is set to zero as well as all statistical counters and we schedule the first arrival event. At each iteration we update the current time to the time the next event occurs, and register the time the system is in the state between the previous and the current event. Then the event type is determined, which is either an arrival event, a service completion event or a content completion event. These event types result in different activities for the three models, which we will consider in the next sections. Finally, it is determined whether K patients arrived in the simulation and if not, we iterate to the next event.

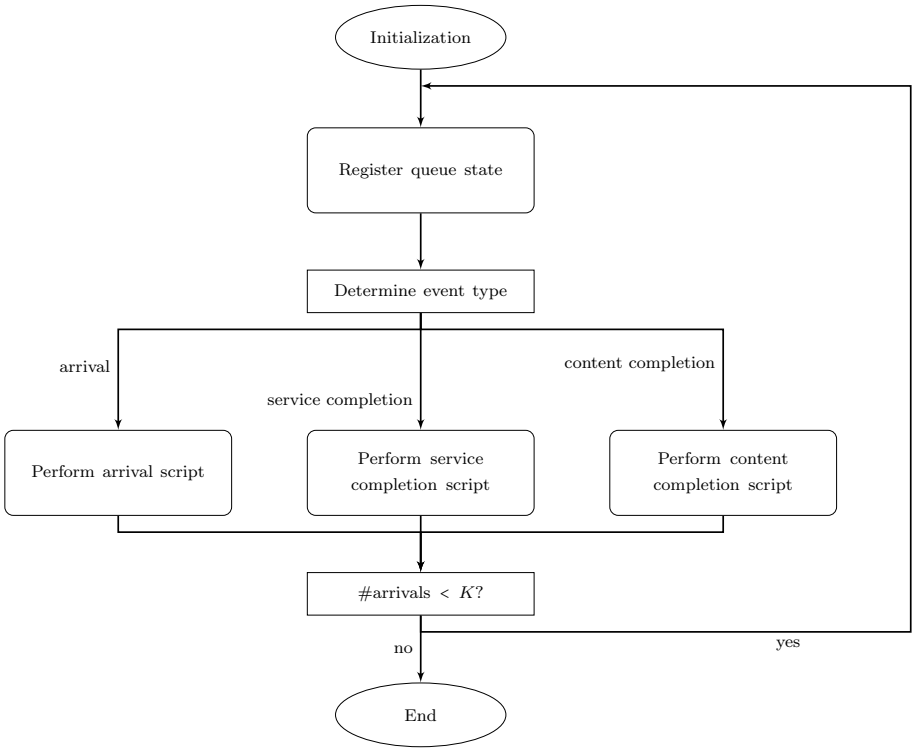


Figure B.1: General simulation construction for the three models.

B.2 Erlang-R model

There are three event types that influence the state of the system, namely an arrival event, a service completion event or a content completion event. In case of an arrival at time t, we first need to update the number of arrivals that entered the system, since we chose this counter to determine the runlength of the simulation. We immediately generate the amount of work that the patient needs. That is, we generate the number of returns N, which is geometrically distributed with parameter p, $N + 1$ exponentially distributed service time(s) with mean $1/\mu$ and N exponentially distributed content time(s) with mean $1/\delta$. After adding the patient to the service queue, we need to determine whether there are enough nurses available to serve this patient immediately. If so, we register that this patient has no waiting time and schedule a service completion event at time t plus the first generated service time of this patient. If not, we store t as the time that this patient starts to wait for service. Finally, we schedule the next arrival event after an exponential amount of time (with mean $1/\lambda$)

In case of a service completion event, we remove the patient from the service queue and determine whether the patient will return for service (from its generated service pattern). If so, the patient is added to the content queue and the next content completion event is scheduled. Otherwise, the patient leaves the system and her LOS is stored. In both cases, we determine whether there is another patient still waiting for service and if so, we register that patient's waiting time and schedule a service completion event.

In case of a content completion event, the patient moves from the content queue to the service queue. If there is a nurse available, we register that this patient does not need to wait and schedule a service completion event. Otherwise, t is stored as the time that this patient started to wait for service.

The activities of the three different event types are illustrated in Figure B.2. Here sq represents the number of patients at the service queue and cq the number of patients at the content queue.

B.3 SERW model

To simulate the SERW model, we need to take into account an additional queue for the beds. This results in a different script in case of an arrival and a service completion, but the content completion script given in Figure B.2 remains the same.

An overview of the arrival script and the case that a patient leaves the system is given in Figure B.3. Again, sq represents the number of patients at the service queue, cq the number of patients at the content queue and in this case hq represents the number of patients in the holding queue.

In case of an arrival, we need to check whether there is a bed available for this patient. If so, we register that this patient has no holding time (i.e. does not need to wait for a bed) and then the script is the same as for the Erlang-R model. If not, we add the patient to the holding queue and schedule the next arrival event after an exponentially distributed amount of time with mean $1/\lambda$.

Also, the script of the service completion event is different in case that a patient leaves the system. That is, we replace the node 'Store LOS' of B.2b with the scipt depicted in Figure B.3b. In case that the holding queue is empty, we execute the same activities as for the Erlang-R model. Otherwise, the next patient waiting for a bed is moved to the service queue and her holding time is registered. If a nurse is available, we register that the patient does not need to wait and schedule a next service completion event. Otherwise, there is another patient that was already waiting for an available nurse, and therefore the patient currently assigned to a bed needs to wait for service. We store the time this patient starts to wait for service and the nurse that completed service with the departed patient will start service with the next patient waiting. Hence, we register her waiting time and schedule a next service completion event.

B.4 SERB model

For the SERB model, only the arrival script differs from the Erlang-R model. In case of an arrival, we first check whether there is a bed available for this patient. Only if this is the case, we generate the

patient's service pattern, add the patient to the service queue and then execute the same activities as for the Erlang-R model. Otherwise, we register a refused patient and schedule the next arrival. This arrival script is illustrated in Figure B.4. Again, *sq* represents the number of patients at the service queue and *cq* the number of patients at the content queue.

B.5 Time-varying arrivals

These simulations can be extended to account for time-varying arrivals as well. First, this can be done by providing a data set with fixed arrival times. Then the next arrival is scheduled by finding the next arrival time of the data set.

Alternatively, the arrival rate can be changed over time. A next arrival is then generated after an exponentially distributed time with rate $\lambda(t)$, where t is the time stamp the next arrival event is generated.

In our simulation we assume the arrival rate to be constant throughout the period of an hour, after which it is allowed to change value. Accordingly, the number of beds and nurses only needs to be re-evaluated at the beginning of these time epochs. Note that in a time-varying setting there can be less nurses and less beds available in the previous hour for staffing rules (5.8) and (5.9). In reality, a nurse finishes her service to a patient and a patient will be redirected to the holding room once she is assigned to an examination room. Therefore, we decreased the number of nurses and the number of beds after a service completion and a departure respectively.

Remark 3. When we determine the bed utilization rate or the nurse utilization level, we average the mean number of occupied beds and the mean number of busy nurses over the number of beds or nurses that should be in the system according to the staffing policy (5.8) and (5.9). Therefore, our results might indicate that these utilization rates exceed one (in case a bed or a nurse finishes service). We should interpret this as an utilization rate nearly equal to one.

B.6 Our simulation settings

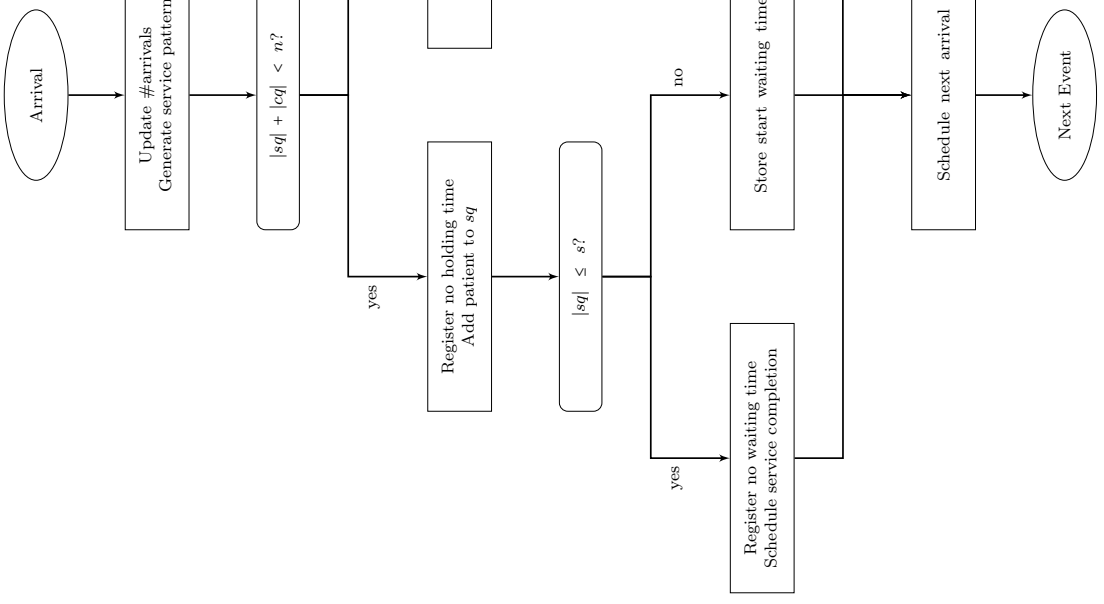
To determine the behavior of the models we simulated $K = 5000000$ patient arrivals for ten times and averaged the outcomes. Moreover, when we consider the asymptotic behavior we executed our simulation experiment five times for $K = 10000000$ and averaged the outcomes. Recall that in the QED regime the probability of waiting for a nurse and the probability of holding for a bed stabilizes for the SERW model in stationarity. To approximate the true limiting values in this case, we set $R = 250$ and called its outcome the true limiting probability. Lastly, for the time-varying setting we simulated 100000 days for ten times and averaged the outcomes per hour.

(a) Arrival script.

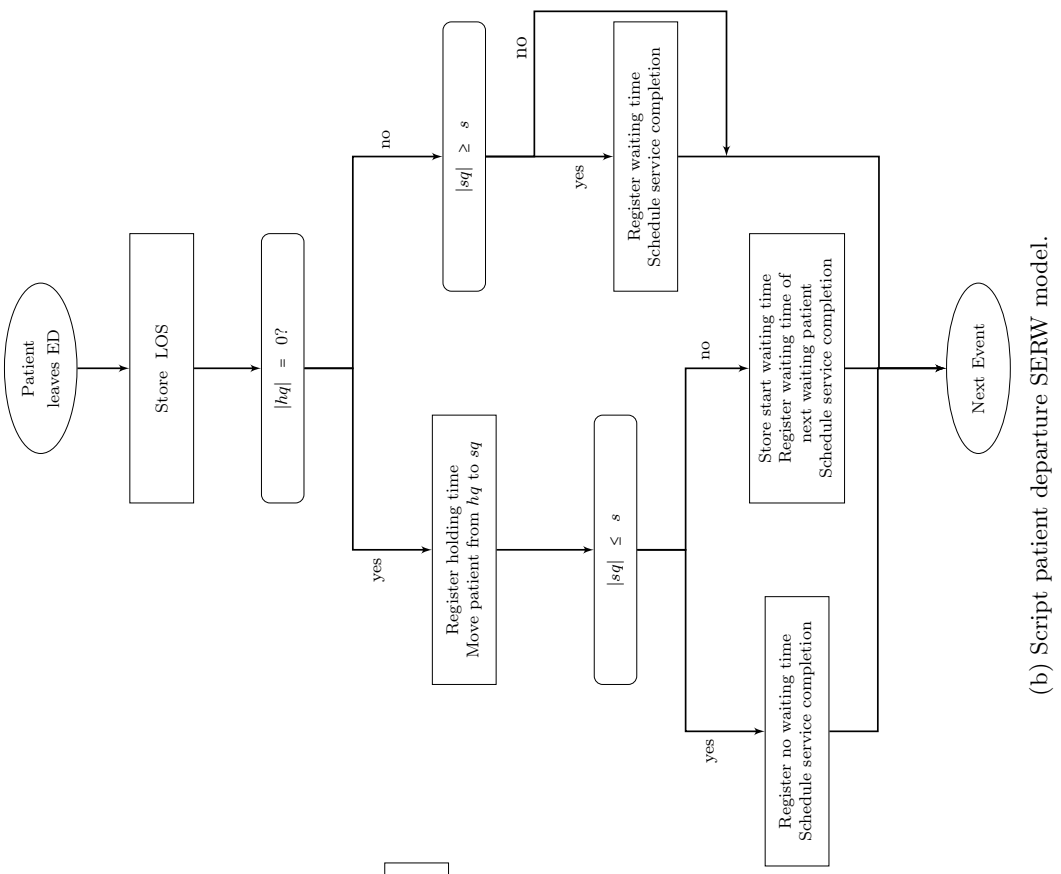
(b) Service completion

(c) Content completion.

Figure B.2: Script of the three event types of the Erlang-R model.



(a) Arrival script SERW model.



(b) Script patient departure SERW model.

Figure B.3: Scripts of two events of the SERW model.

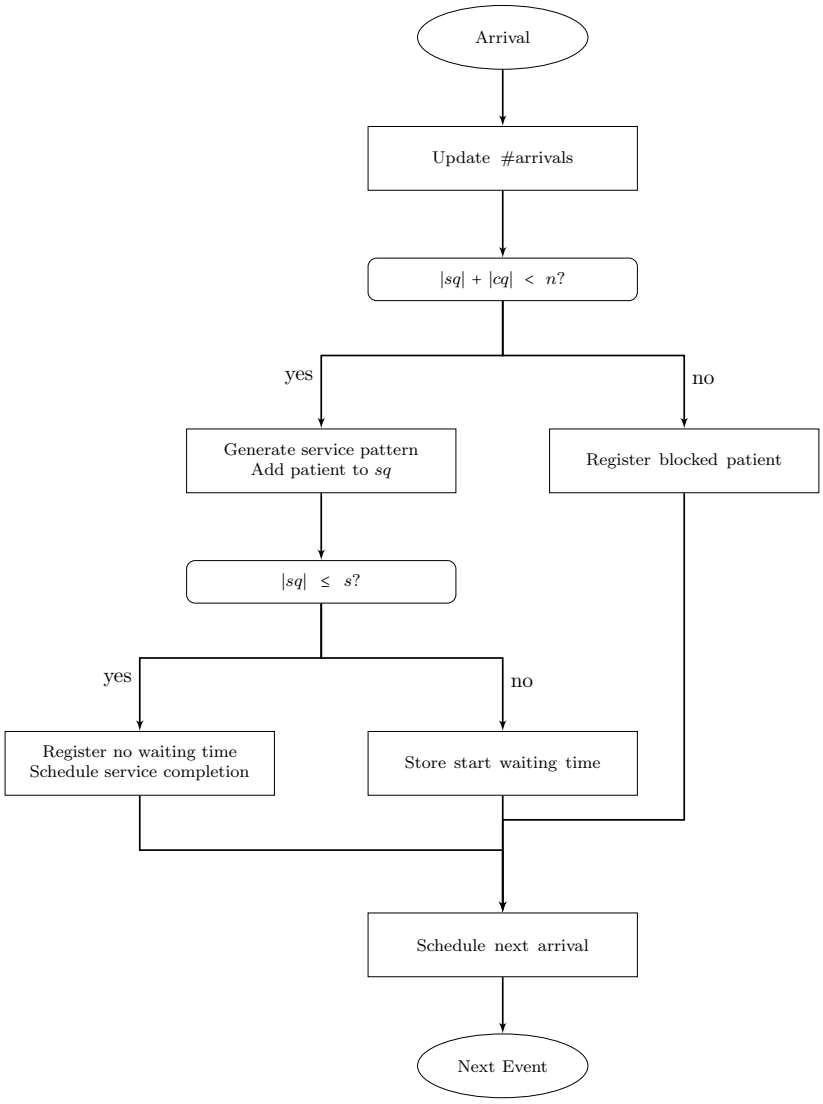


Figure B.4: Arrival script SERB model.

Appendix C

Proofs of analytical properties

Proof of Proposition 2.2.2. The balance equations of a birth-death process are given by

$$\lambda_i \pi_i = \mu_{i+1} \pi_{i+1}, \quad i = 0, 1, ..., I, \tag{C.1}$$

and we have the normalization condition

$$\sum_{i=0}^{I} \pi_i = 1. \tag{C.2}$$

Then (C.1) yields the recursion

$$\pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}, \quad i = 0, 1, ..., I, \tag{C.3}$$

and from the normalization condition we obtain

$$\pi_0 = \left[\sum_{i=0}^{I} \sum_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} \right]^{-1}. \tag{C.4}$$

For π_0 to be finite, we find the necessary and sufficient condition $\sum_{i=0}^{I} \sum_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} < \infty$, which ensures that a stationary distribution exists. □

Proof of Proposition 2.3.1. The queue length behaves like a birth-death process with $I = \infty$. The rates are given by $\lambda_i = \lambda$ for all $i \in \mathbb{N}$ and

$$\mu_i = \begin{cases} i\mu & \text{for } 0 \le i \le s-1, \\ s\mu & \text{for } i \ge s. \end{cases}$$

Then the proposition follows directly from Proposition 2.2.2. □

Proof of Proposition 2.4.1. The queue length behaves like a birth-death process with $I = s$. The rates are given by $\lambda_i = \lambda$ for all $0 \le i \le s-1$ and $\mu_i = i\mu$ for all $0 \le i \le s$. Then the stationary distribution follows directly from Proposition 2.2.2. □

Proof of Lemma 2.4.2. We have that

$$\begin{aligned} B(s,R)^{-1} &= \frac{\sum_{j=0}^{s} R^j/j!}{R^s/s!} \\ &= 1 + \frac{\sum_{j=0}^{s-1} R^j/j!}{R^{s-1}/(s-1)!} \cdot \frac{s}{R} \\ &= 1 + B(s-1,R)^{-1} \cdot \frac{s}{R} \\ &= \frac{R\,B(s-1,R) + s}{R\,B(s-1,R)}. \end{aligned}$$

We can conclude that

$$B(s+1,R) = \frac{R\,B(s,R)}{R\,B(s,R)+s+1}.$$

□

Proof of Lemma 2.4.3. We will prove this by considering its distribution in terms of R. We have that

$$\begin{aligned}
C(s,R)^{-1} &= \frac{s-R}{s}\cdot\frac{s!}{R^s}\cdot\left[\sum_{j=0}^{s-1} R^j/j! + \frac{s}{s-R}\frac{R^s}{s!}\right] \\
&= (1-\rho)\cdot\frac{s!}{R^s}\cdot\left[\sum_{j=0}^{s} R^j/j! + \left(\frac{s}{s-R}-1\right)\frac{R^s}{s!}\right] \\
&= (1-\rho)B(s,\rho)^{-1} + \frac{s-R}{s}\cdot\frac{r}{s-R} \\
&= \rho + (1-\rho)B(s,\rho)^{-1}.
\end{aligned}$$

□

Alternative proof of Lemma 2.4.3. The Erlang delay formula in its basic form is defined for integer values of s, but we can extend this formula for continuous values as well. Using the continuous extensions of the Erlang-B and the Erlang-C formula, one can provide an alternative proof, see [25]. We include this proof here for completeness.

For all real $s > R$ the continuous extensions of the Erlang-B formula and the Erlang-C formula read

$$B(s,R) = \left(R\int_0^\infty e^{-Rt}(1+t)^s dt\right)^{-1}, \tag{C.5}$$

$$C(s,R) = \left(R\int_0^\infty te^{-Rt}(1+t)^{s-1} dt\right)^{-1}. \tag{C.6}$$

Note that

$$\begin{aligned}
B(s,R)^{-1} &= R\int_0^\infty e^{-Rt}(1+t)^s dt \\
&= R\int_0^\infty e^{-Rt}(1+t)^{s-1}(1+t) dt \\
&= R\int_0^\infty e^{-Rt}(1+t)^{s-1} dt + R\int_0^\infty te^{-Rt}(1+t)^{s-1} dt \\
&= R\int_0^\infty e^{-Rt}(1+t)^{s-1} dt + C(s,R)^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
C(s,R)^{-1} &= B(s,R)^{-1} - R\int_0^\infty e^{-Rt}(1+t)^{s-1} dt \\
&= B(s,R)^{-1} - R\left(\left[\frac{(1+t)^s}{s}e^{-Rt}\right]_0^\infty - \int_0^\infty -Re^{-Rt}\frac{(1+t)^s}{s} dt\right) \\
&= B(s,R)^{-1} + \frac{R}{s} - \frac{R}{s}\int_0^\infty Re^{-Rt}(1+t)^s dt \\
&= \rho + (1-\rho)B(s,R)^{-1}.
\end{aligned}$$

□

Proof of Lemma 2.3.2. This can be proven by applying Lemmas 2.4.2 and 2.4.3. Now,

$$\begin{aligned}
C(s+1,R)^{-1} &= \frac{R}{s+1} + (1-\frac{R}{s+1})B(s+1,R)^{-1} \\
&= \frac{R}{s+1} + \frac{s+1-R}{s+1}(1+\frac{s+1}{R}B(s,R)^{-1}) \\
&= 1 + \frac{s+1-R}{R}B(s,R)^{-1} \\
&= 1 + \frac{s+1-R}{R}\frac{s}{s-R}\left(C(s,r)^{-1} - \frac{R}{s}\right) \\
&= 1 + \frac{(s+1-R)s}{R(s-R)C(s,r)} - \frac{s+1-R}{s-R} \\
&= \frac{(s+1-R)s}{R(s-R)C(s,r)} - \frac{1}{s-R} \\
&= \frac{(s+1-R)s - RC(s,r)}{R(s-R)C(s,r)}.
\end{aligned}$$

We can conclude that recursion (2.6) holds. □

Proof of Proposition 4.1.1. The queue length behaves like a birth-death process with $I = n$. The rates are given by $\lambda_i = \lambda(n-i)$ as the rate from the finite source for all $0 \le i \le n$. The death rates are $\mu_i = \min\{i, s\}\mu$ for all $0 \le i \le n$. Then the stationary distribution follows directly from Proposition 2.2.2. □

Proof of Proposition 4.1.2. We note that the closed ward model is a two-node closed Jackson network. The arrival theorem for a closed Jackson network states that the activation at any node observes time averages excluding the patient itself. That is, the probability that a patient finds k needy patients upon activation equals the fraction of time that there are a k needy patients in the same closed Jackson network with $n-1$ patients. Then the probability of waiting follows directly.

Suppose an activated patient finds $k \ge s$ patients in the system. Then the patient must wait for $(k-s+1)$ stages of exponentials with rate $s\mu$. By definition, this is an Erlang distributed random variable. Let W denote the stationary in-queue waiting time for a new arriving patient and Q the number of patients in need of service. Thus, the probability the patient must wait longer than threshold t is

$$P(W > t|Q = k) = e^{-s\mu t}\sum_{j=0}^{k-s}\frac{(s\mu t)^j}{j!}.$$

By conditioning on the number of needy patients upon activation, we obtain the result for the probability of excessive delay.

The expected waiting time can be determined by using the probability of excessive delay. That is,

$$\begin{aligned}
E(W) &= \int_0^\infty P(W > t)dt = \sum_{k=s}^{n-1}\pi_k^{(n-1)}\sum_{j=0}^{k-s}\int_0^\infty \frac{(\mu st)^j}{j!}e^{-\mu st}dt = \sum_{k=s}^{n-1}\pi_k^{(n-1)}\sum_{j=0}^{k-s}\frac{1}{\mu s} \\
&= \frac{1}{\mu s}\sum_{k=s}^{n-1}\pi_k^{(n-1)}(k-s+1).
\end{aligned}$$

The utilization level is a direct consequence of its definition. □

Proof of Proposition 4.2.3. The first property is quite straightforward. By the PASTA property, we have

that

$$\begin{aligned}\alpha := P(W>0) &= \sum_{j=0}^{\infty}\sum_{i=s}^{\infty}\frac{R^i}{s!s^{i-s}}c_1\frac{R_2^j}{j!}c_2\\ &= \sum_{i=s}^{\infty}\frac{R^i}{s!s^{i-s}}c_1\\ &= \frac{R^s}{s!(1-R/s)}c_1.\end{aligned}$$

For the second property we will condition on the number of patients that are already waiting or being seen by a nurse in the examination room. Suppose that at the moment a new patient arrives to the Emergency Department, there are already k patients. So the arriving patient must wait for $(i-s+1)^+$ stages, each exponentially distributed with rate μs. This corresponds to an Erlang distribution with $(i-s+1)^+$ stages, thus the probability that the arriving patient's waiting time exceeds t equals

$$\int_t^{\infty}\frac{s\mu(s\mu x)^{i-s}}{(i-s)!}e^{-\mu s x}dx.$$

Then we have

$$P(Q_1=i)=\sum_{j=0}^{\infty}\pi_{(i+s)j}=\frac{(R)^{i+s}}{s!s^i}c_1=\alpha\left(1-\frac{R}{s}\right)\left(\frac{R}{s}\right)^i=\alpha(1-\rho_R)\rho_R^i.$$

Equivalently, $N_1|W>0$ has a geometric distribution with parameter ρ_R.

Let E_k denote an Erlang-k distribution with rate μs. Then,

$$\begin{aligned}P(W>t|W>0) &= \sum_{i=0}^{\infty}P(N_1=i|W>0)P(E_{i-s+1}>t)\\ &= \sum_{i=0}^{\infty}(1-\rho_R)\rho_R^i\int_t^{\infty}\frac{s\mu(s\mu x)^i}{i!}e^{-\mu s x}\,dx\\ &= \mu s(1-\rho_R)\int_t^{\infty}\sum_{i=0}^{\infty}\frac{(R\mu x)^i}{i!}e^{-\mu s x}\,dx\\ &= \mu s(1-\rho_R)\int_t^{\infty}e^{-\mu s(1-\rho_R)x}\,dx\\ &= e^{-\mu s(1-\rho_R)t}.\end{aligned}$$

Thus indeed, $W|W>0$ is exponentially distributed with rate $\mu s(1-\rho_R)$.

The last two properties follow from the second property. We have

$$P(W>t)=P(W>t|W>0)P(W>0)=\alpha e^{-\mu s(1-\rho_R)t},$$

and

$$E(W)=\int_0^{\infty}P(W>t)\,dt=\int_0^{\infty}\alpha e^{-\mu s(1-\rho_R)t}\,dt=\frac{\alpha}{\mu s(1-\rho_R)}.$$

We will give a direct proof for the utilization level of the Erlang-R model. The probability that there are i patients in front of the service queue is

$$\begin{cases}\frac{R^i}{i!}c_1 & \text{if } i\le s,\\ \frac{R^i}{s!s^{i-s}}c_1 & \text{if } i\ge s,\end{cases}$$

where

$$c_1=\left[\frac{R^s}{s!(1-R/s)}+\sum_{i=0}^{s-1}\frac{R^i}{i!}\right]^{-1}.$$

Therefore the occupation rate is equal to

$$\rho_R = c_1 \left[\sum_{i=0}^{s} \frac{i}{s} \frac{R^i}{i!} + \sum_{i=s+1}^{\infty} \frac{R^i}{s! s^{i-s}} \right] \tag{C.7}$$

$$= c_1 \frac{R}{s} \left[\sum_{i=1}^{s} \frac{R^{i-1}}{(i-1)!} + \sum_{i=s+1}^{\infty} \frac{R^{i-1}}{s! s^{i-1-s}} \right] \tag{C.8}$$

$$= c_1 \frac{R}{s} \left[\sum_{i=0}^{s-1} \frac{R^i}{i!} + \sum_{i=s}^{\infty} \frac{R^i}{s! s^{i-s}} \right] \tag{C.9}$$

$$= c_1 \frac{R}{s} c_1^{-1} = \frac{R}{s}. \tag{C.10}$$

□

Proof of Proposition 4.3.1. We consider the closed Jackson network given in Figure 4.7. The distribution follows by the theory of closed Jackson networks, see e.g. [15].

Consider a general closed Jackson network, with M nodes, n customers and $N(t) = (N_1(t), ..., N_M(t))$ representing the number of customers present at each queue. A customer at station i requires an exponentially distributed service duration with parameter μ_i and $r_i(m)$ denotes the rate of service when there are m customers at queue i. Let $P \in \mathbb{R}^{M \times M}$, with p_{ij} representing the probability that a customer after service completion at queue i proceed to queue j. Then the throughputs $\gamma = (\gamma_1, ..., \gamma_M)$ can be found by solving $\gamma = \gamma P$ (up to a scaling factor). Moreover, the stationary distribution is given by

$$\pi(n_1, ..., n_M) = G^{-1} \prod_{i=1}^{M} g_i(n_i),$$

with

$$g_i(n_i) = \frac{(\kappa \gamma_i / \mu_i)^{n_i}}{\prod_{m=1}^{n_i} r_i(m)}$$

and normalization constant

$$G = \sum_{n_1 + ... + n_M = n} \prod_{i=1}^{M} g_i(n_i).$$

The value $\kappa > 0$ is a scaling factor that can be chosen arbitrarily. We return to our particular closed Jackson network.

The probability transition matrix for the SERB model is given by

$$P = \begin{pmatrix} 0 & p & 1-p \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

The throughput vector $\gamma = \left(\frac{\lambda}{1-p}, \frac{p\lambda}{1-p}, \lambda \right)$ satisfies the equation $\gamma = \gamma P$. Note that the rates of the stations are given by $r_1(m) = \min\{m, s\}$, $r_2(m) = m$ and $r_3(m) = 1$. Let $\kappa = 1$, so that

$$g_1(i) = \begin{cases} \frac{1}{i!} \left(\frac{\lambda}{(1-p)\mu} \right)^i & \text{for } 0 \le i \le s, \\ \frac{1}{s! s^{i-s}} \left(\frac{\lambda}{(1-p)\mu} \right)^i & \text{for } s+1 \le i \le n, \end{cases}$$

$$g_2(j) = \frac{1}{j!} \left(\frac{p\lambda}{(1-p)\delta} \right)^j \quad \text{for } 0 \le j \le n-i,$$

$$g_3(n-i-j) = 1.$$

This yields

$$\pi_{ij} := P(Q_1(\infty) = i, Q_2(\infty) = j) = \begin{cases} \pi_0 \frac{R_1^i}{i!} \frac{R_2^j}{j!} & \text{if } i \le s, 0 \le i + j \le n, \\ \pi_0 \frac{R^i}{s! s^{i-s}} \frac{R_2^j}{j!} & \text{if } i \ge s, , 0 \le i + j \le n \end{cases}$$

with

$$\begin{aligned}\pi_0^{(}-1) &= \sum_{l=0}^{s}\sum_{i=0}^{l}\frac{R^i}{i!}\frac{R_2^{l-i}}{(l-i)!} + \sum_{l=s+1}^{n}\left(\sum_{i=0}^{s}\frac{R^i}{i!}\frac{R_2^{l-i}}{(l-i)!} + \sum_{i=s+1}^{l}\frac{R^i}{s!s^{i-s}}\frac{R_2^{l-i}}{(l-i)!}\right)\\ &= \sum_{l=0}^{n}\sum_{i=0}^{l}\frac{R^i}{i!}\frac{R_2^{l-i}}{(l-i)!} + \sum_{l=s+1}^{n}\sum_{i=s+1}^{l}\left(\frac{1}{s!}s^{s-i}-\frac{1}{i!}\right)R^i\frac{R_2^{l-i}}{(l-i)!}\\ &= \sum_{l=0}^{n}\frac{(R+R_2)^l}{l!} + \sum_{l=s+1}^{n}\sum_{i=s+1}^{l}\left(\frac{1}{s!}s^{s-i}-\frac{1}{i!}\right)\frac{1}{(l-i)!}R^iR_2^{l-i}.\end{aligned}$$

□

Proof of 4.3.3. The blocking probability follows directly from the PASTA property and Corollary 4.3.2. For the performance measures in terms of the waiting time, we exploit the arrival theorem for a closed Jackson network. In particular, the probability that a patient finds k patients upon arrival at the nursing station equals the average time that there are k patients in a closed system with $n-1$ patients. Then (4.11) follows directly.

Next we consider the probability that patients need to wait more than t time units. Suppose a patient finds i patients in the service queue upon arrival. The waiting time then follows an Erlang distribution with $(i-s+1)^+$ stages, where $(x)^+ = \max\{0,x\}$, each with rate $s\mu$. Thus, if the patient needs to wait, then the probability that she needs to wait longer than t time units is $e^{-\mu st}\sum_{j=0}^{i-s}(s\mu t)^j/j!$. Conditioning on the number of patients in the service queue upon arrival gives (4.12).

From this, we can determine the expected waiting time.

$$\begin{aligned}E(W) &= \int_0^\infty P(W>t)dt = \sum_{l=s}^{n-1}\sum_{i=s}^{l}\pi_{i\,l-i}^{(n-1)}\sum_{k=0}^{i-s}\int_0^\infty\frac{\mu st}{k!}e^{-\mu st}dt\\ &= \sum_{l=s}^{n-1}\sum_{i=s}^{l}\pi_{i\,l-i}^{(n-1)}\frac{i-s+1}{\mu s},\end{aligned}$$

which equals (4.13).

The utilization level follows from the definition. □

Appendix D

Proofs of QED properties

Proof of Proposition 3.2.1. We will first prove direction ($\Leftarrow$). So we assume that $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta$ for some $\beta > 0$. The Erlang loss formula is given by 2.5. Then,

$$\begin{aligned}
P(Q_n \geq n) &= \frac{(n\rho_n)^n}{n!(1-\rho_n)}\left(\sum_{j=0}^{n-1}\frac{(\rho_n n)^j}{j!} + \frac{(\rho_n n)^n}{(1-\rho_n)n!}\right)^{-1} \\
&= \left(\sum_{j=0}^{n-1}\frac{(\rho_n n)^j}{j!} \Big/ \left(\frac{(n\rho_n)^n}{n!(1-\rho_n)}\right) + 1\right)^{-1} \\
&= \left(1 + \sum_{j=0}^{n-1}\frac{(\rho_n n)^j}{j!}e^{-n\rho_n} \Big/ \left(\frac{(n\rho_n)^n}{n!(1-\rho_n)}e^{-n\rho_n}\right)\right)^{-1} \\
&= \left(1 + \frac{P(X_n \leq n-1)}{(1-\rho_n)P(X_n = n)}\right)^{-1},
\end{aligned}$$

where X_n is a Poisson distributed random variable with parameter $n\rho_n$. By the properties of the Poisson distribution, X_n is the sum of $n\rho_n$ independent Poisson random variables of rate one. From the Central Limit Theorem follows that

$$\frac{X_n - n\rho_n}{\sqrt{n\rho_n}}$$

converges in distribution to the standard normal distribution. Let

$$z_n = \frac{(n-1) - n\rho_n}{\sqrt{n\rho_n}} = \frac{n(1-\rho_n) - 1}{\sqrt{n\rho_n}}.$$

For $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta > 0$ to hold, we must have that $\rho_n \to 1$. Then

$$\lim_{n\to\infty} z_n = \frac{\sqrt{n}(1-\rho_n)}{\sqrt{\rho_n}} - 0 = \beta.$$

Thus,

$$P(X_n \leq n-1) = P\left(\frac{X_n - n\rho_n}{\sqrt{n\rho_n}} \leq z_n\right) \to \Phi(\beta). \tag{D.1}$$

To approximate $P(X_n = n)$ we will apply Stirling's formula. Stirling's formula states that

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + ...\right)$$

So we can approximate $n!$ by $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$. Then,

$$\begin{aligned}(1-\rho_n)P(X_n=n) &= \frac{(n\rho_n)^n}{n!(1-\rho_n)}e^{-n\rho_n}\\ &= \frac{\rho_n^n}{(1-\rho_n)\sqrt{2\pi n}}e^n n^{-n}e^{-n\rho_n}\\ &= \frac{\rho_n^n}{(1-\rho_n)\sqrt{2\pi n}}e^{n(1-\rho_n)}\\ &= \frac{1}{(1-\rho_n)\sqrt{2\pi n}}e^{n(1-\rho_n+\log\rho_n)}.\end{aligned}$$

Applying Taylor series to the logarithm function we derive

$$\log(x) = -\sum_{j=0}^{\infty}\frac{(1-x)^j}{j} = -(1-x) - \frac{(1-x)^2}{2} + o(1-x)^2.$$

Consequently,

$$\begin{aligned}(1-\rho_n)\lim_{n\to\infty}P(X_n=n) &= \lim_{n\to\infty}\frac{1}{(1-\rho_n)\sqrt{2\pi n}}e^{n(1-\rho_n+\log\rho_n)}\\ &= \lim_{n\to\infty}\frac{1}{(1-\rho_n)\sqrt{2\pi n}}e^{n(1-\rho_n)^2/2}\\ &= \frac{1}{\beta\sqrt{2\pi}}e^{\beta^2/2}\\ &= \frac{\phi(\beta)}{\beta}.\end{aligned}$$

So both $P(X_n \le n)$ and $(1-\rho_n)P(X_n=n)$ converge to a finite limit. Hence we conclude that

$$\begin{aligned}\lim_{n\to\infty}P(Q_n \ge n) &= \lim_{n\to\infty}\left(1+\frac{P(X_n\le n-1)}{(1-\rho_n)P(X_n=n)}\right)^{-1}\\ &= \left(1+\frac{\Phi(\beta)}{\phi(\beta)/\beta}\right)^{-1}\\ &= \left(1+\frac{\beta\Phi(\beta)}{\phi(\beta)}\right)^{-1}\end{aligned}$$

if $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta$.

For the converse direction ($\Rightarrow$), assume $\lim_{n\to\infty}P(Q_n \ge n) \in (0,1)$ and that $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} \ne \beta$ for any $\beta > 0$. Then we have three possibilities:

(a) $(1-\rho_n)\sqrt{n} \to 0$ as $n \to \infty$,

(b) $(1-\rho_n)\sqrt{n} \to \infty$ as $n \to \infty$,

(c) $(1-\rho_n)\sqrt{n}$ does not converge to any limit as $n \to \infty$.

We will show that all three cases lead to a contradiction.

For case (a) we have that $z_n \to 0$, where z_n is as before. Then $P(X_n \le n-1) \to \Phi(0) = 1/2$ and $(1-\rho_n)P(X_n=n) = \phi(\beta)/\beta \to \infty$ as $n \to \infty$. So $\lim_{n\to\infty}P(Q_n \ge n)$ will converge to one, which contradicts our assumption.

For case (b) we have that $z_n \to \infty$, hence $P(X_n \le n-1) \to \Phi(\infty) = 1$. Moreover, $(1-\rho_n)P(X_n=n) = \phi(\beta)/\beta \to 0$ which leads to $P(Q_n \ge n) \to 0$, contradicting our assumption.

Lastly, assume $(1-\rho_n)\sqrt{n}$ does not converge to any limit. Then the sequence will have two subsequences converging to two different limits (possibly infinite), for which the converse reasoning still holds. Note that (3.1) is a decreasing function of β, and hence we find that $\epsilon(\beta)$ strictly decreases in β. Then the two subsequences of $(1-\rho_n)\sqrt{n}$ gives rise to two subsequences of $P(Q_n \geq n)$ converging to different limits, which contradicts our assumption. □

Proof of Theorem 3.3.1. The proof is analogous to the proof of Theorem 3.2.1, we will only show ($\Leftarrow$) here. We have

$$\begin{aligned}
P(Q_n \geq n) &= \sqrt{n}\frac{(\rho_n n)^n}{n!}\left[\sum_{j=0}^{n}\frac{(\rho_n n)^j}{j!}\right]^{-1} \\
&= \sqrt{n}\left[1+\sum_{j=0}^{n-1}\frac{(\rho_n n)^j}{j!}e^{-n\rho_n}\Big/\left(\frac{(\rho_n n)^n}{n!}e^{-n\rho_n}\right)\right]^{-1} \\
&= \left[\frac{1}{\sqrt{n}}+\frac{P(X_n \leq n-1)}{\sqrt{n}P(X_n = n)}\right]^{-1},
\end{aligned}$$

where X_n is again a Poisson distributed random variable with parameter $n\rho_n$. Similar as in the proof of Theorem 3.2.1, we have that

$$P(X_n \leq n-1) \to \Phi(\beta).$$

To approximate $\sqrt{n}P(X_n = n)$ we apply Stirling's formula again and obtain

$$\begin{aligned}
\sqrt{n}P(X_n = n) &= \sqrt{n}\frac{(n\rho_n)^n}{n!}e^{-n\rho_n} \\
&\sim \frac{\rho_n^n\sqrt{n}}{\sqrt{2\pi n}}e^n n^{-n}e^{-n\rho_n} \\
&= \frac{1}{\sqrt{2\pi}}e^{n(1-\rho_n+\log\rho_n)} \\
&\to \frac{1}{\sqrt{2\pi}}e^{\beta^2/2} \\
&= \phi(\beta).
\end{aligned}$$

So both $P(X_n \leq n)$ and $\sqrt{n}P(X_n = n)$ converge to a finite limit. Hence,

$$\begin{aligned}
\lim_{n\to\infty} P(Q_n \geq n) &= \lim_{n\to\infty}\left[\frac{1}{\sqrt{n}}+\frac{P(X_n \leq n-1)}{\sqrt{n}P(X_n = n)}\right]^{-1} \\
&= \left[0+\frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1} \\
&= \frac{\phi(\beta)}{\Phi(\beta)}.
\end{aligned}$$

if $\lim_{n\to\infty}(1-\rho_n)\sqrt{n} = \beta$. □

Proof of Proposition 4.1.1. Recall the distribution of a $M/M/s//n$ queue

$$\pi_i = \begin{cases} \pi_0\binom{n}{k}R^k & \text{for } 0 \leq k < s, \\ \pi_0\binom{n}{k}\frac{k!}{s!}s^{s-k}R^k & \text{for } s \leq k \leq n. \end{cases}$$

with

$$\pi_0 = \left[\sum_{i=0}^{s}\binom{n}{k}R^k + \sum_{i=s+1}^{n}\binom{n}{k}\frac{k!}{s!}s^{s-k}R^k\right]^{-1}.$$

In this proof, for convenience we leave out the subscript in s_n although the dependency $s_n = rn+\beta\sqrt{n}$ still holds.

The approximate probability of delay can be written as

$$P(Q_n \geq s) = \left(1 + \frac{A_n}{B_n}\right)^{-1},$$

with

$$A_n = \sum_{k=0}^{s-1} \binom{n}{k} R^k, \quad B_n = \sum_{k=s}^{n} \binom{n}{k} \frac{k!}{s!} s^{s-k} R^k.$$

Note that $R = \lambda/\mu = r/(1-r)$, thus

$$\begin{aligned} A_n &= \sum_{k=0}^{s-1} \binom{n}{k} R^k \\ &= \sum_{k=0}^{s-1} \binom{n}{k} r^k (1-r)^{-k} \\ &= (1+R)^n \sum_{k=0}^{s-1} \binom{n}{k} r^k (1-r)^{n-k} \\ &= (1+R)^n P(X_n \leq s-1), \end{aligned}$$

where X_n is a binomial random variable with parameters n and r. Moreover,

$$\begin{aligned} B_n &= \sum_{k=s}^{n} \binom{n}{k} \frac{k!}{s!} s^{s-k} R^k \\ &= \frac{n!}{s!} s^s \sum_{k=s}^{n} \frac{1}{(n-k)!} s^{-k} R^{k-n} \\ &= \frac{n!}{s!} \frac{R^n}{s^{n-s}} \sum_{k=s}^{n} \frac{1}{(n-k)!} s^{n-k} R^{k-n} \\ &= \frac{n!}{s!} \frac{R^n}{s^{n-s}} e^{s/R} \sum_{k=s}^{n} \frac{1}{(n-k)!} \left(\frac{R}{s}\right)^{n-k} e^{-s/R} \\ &= \frac{n!}{s!} \frac{R^n}{s^{n-s}} e^{s/R} P(Y_n \leq n-s), \end{aligned}$$

where Y is a Poisson distributed random variable with parameter s/R. Thus the approximate probability of delay can be written as

$$P(Q_n \geq s) = \left(1 + C_n \frac{P(X_n \leq s-1)}{P(Y_n \leq n-s)}\right)^{-1},$$

where

$$C_n = \frac{s!}{n!} \left(\frac{s}{R}\right)^n \frac{(1+R)^n}{s^s} e^{-s/R}.$$

Recall $\beta_n = (s/n - r)\sqrt{n}$. Then by applying the CLT to X_n we obtain

$$\begin{aligned} P(X_n \leq s-1) &= P\left(\frac{X - rn}{\sqrt{nr(1-r)}} \leq \frac{s - rn - 1}{\sqrt{nr(1-r)}}\right) \\ &\to \Phi\left(\frac{\beta}{\sqrt{r(1-r)}}\right), \end{aligned}$$

since

$$\frac{s - rn - 1}{\sqrt{nr(1-r)}} = \frac{b\sqrt{n} - 1}{\sqrt{nr(1-r)}} \to \frac{\beta}{\sqrt{r(1-r)}}$$

as $n \to \infty$. Similarly,

$$P(Y_n \le n - s) = P\left(\frac{Y_n - s/R}{\sqrt{s/R}} \le \frac{n - s - s/R}{\sqrt{s/R}}\right) \to \Phi\left(\frac{-\beta}{r\sqrt{1-r}}\right),$$

since

$$\frac{n - s - s/R}{\sqrt{s/R}} = \frac{rn - rs - s(1-r)}{\sqrt{s(1-r)r}} = \frac{rn - s}{\sqrt{s(1-r)r}} = \frac{-\beta}{\sqrt{s/n(1-r)r}} \to \frac{-\beta}{\sqrt{(1-r)r^2}} = \frac{-\beta}{r\sqrt{1-r}}$$

as $n \to \infty$.

Using Stirling's approximation $n! \sim (2\pi n)^{1/2} n^n e^{-n}$ for C_n, we find

$$\begin{aligned} C_n &\sim \frac{(2\pi s)^{1/2} s^s e^{-s}}{(2\pi n)^{1/2} n^n e^{-n}} \left(\frac{s}{R}\right)^n \frac{(1+R)^n}{s^s} e^{-s/R} \\ &= \sqrt{r}\left(\frac{s}{rn}\right)^{n+1/2} e^{n - s(1-r)/r - s} \\ &= \sqrt{r}\left(1 + \frac{\beta}{r\sqrt{n}}\right)^{n+1/2} e^{n - s/r}. \end{aligned}$$

Note that by Taylor expansion

$$n \log\left(1 + \frac{\beta}{r\sqrt{n}}\right) = n\left(\frac{\beta}{r\sqrt{n}} - \frac{1}{2}\frac{\beta^2}{r^2 n} + \mathcal{O}(n^{-3/2})\right),$$

so that

$$\left(1 + \frac{\beta}{r\sqrt{n}}\right)^n = e^{n \log\left(1 + \frac{\beta}{r\sqrt{n}}\right)} \sim e^{\frac{\beta\sqrt{n}}{r} - \frac{\beta^2}{2r^2}}.$$

So this yields

$$C_n \sim \sqrt{r}\left(1 + \frac{\beta}{r\sqrt{n}}\right)^{1/2} e^{-\frac{\beta^2}{2r^2}} \to \sqrt{r} e^{-\frac{\beta^2}{2r^2}}.$$

We can conclude that

$$P(Q_n \ge s) \to f(\beta) = \left(1 + e^{-\beta^2/(2r^2)} \sqrt{r} \frac{\Phi\left(\beta/\sqrt{r(1-r)}\right)}{\Phi\left(-\beta/(r\sqrt{1-r})\right)}\right)^{-1},$$

as $n \to \infty$.

For the other direction, suppose that $\alpha_n := P(Q_n \ge s)$ has a limit $\alpha \in (0,1)$ and that $\beta \ne f^{-1}(\alpha)$ is a (possibly infinite) limit point of $\{(s_n - r)\sqrt{n}\}$. We note that $f : \mathbb{R} \to (0,1)$ is a strictly decreasing function.

Assume that $\beta > f^{-1}(\alpha)$. Then we can construct a sequence $\{s'_n\}$ such that $\{(s'_n - r)\sqrt{n}\} \to \min\{\frac{\beta + f^{-1}(\alpha)}{2}, f^{-1}(\alpha) + 1\}$ and (by taking a subsequence) such that $s'_n \le s_n$. Then, by definition $f^{-1}(\alpha) < \beta' < \infty$ and since f is a strictly decreasing function we find $f(\beta') < \alpha$. Let Q'_n denote the number of customers in the n'th system with s'_n servers. Since $s'_n \le s_n$, we have that $P(Q_n \ge s_n) \le P(Q'_n \ge s'_n)$, but taking the limit on both sides yields $\alpha \le f(\beta')$, a contradiction.

Assume that $\beta < f^{-1}(\alpha)$. Then we can construct a sequence $\{s'_n\}$ such that $\{(s'_n - r)\sqrt{n}\} \to \max\{\frac{\beta + f^{-1}(\alpha)}{2}, f^{-1}(\alpha) - 1\}$ and $s'_n \ge s_n$. Then $-\infty < \beta' < f^{-1}(\alpha)$ and thus $f(\beta') > \alpha$. Let Q'_n

denote the number of customers in the n'th system with s'_n servers. Since $s'_n \geq s_n$, we have that $P(Q_n \geq s_n) \geq P(Q'_n \geq s'_n)$, but taking the limit on both yields the contradiction $\alpha \geq f(\beta')$.

Hence the convergence $\alpha_n \to \alpha \in (0,1)$ implies that $(s_n - r)\sqrt{n}$ converges to a unique finite limit as well. □

Proof of Proposition 4.2.4. The asymptotic results for the probability of waiting and utilization level are direct consequences of the QED results we have for the Erlang-C model. For the expected waiting time, consider

$$\begin{aligned}\sqrt{s}E(W) &= \sqrt{s}E(W|W>0)P(W>0)\\ &= \frac{\sqrt{s}P(W>0)}{\mu s(1-\rho_R)} = \frac{P(W>0)}{\mu\beta\sqrt{R/s}}\end{aligned}$$

which tends to $\left(1+\frac{\beta\Phi(\beta)}{\phi(\beta)}\right)^{-1}/(\beta\mu)$ as $R \to \infty$. □

Bibliography

[1] M. Armony et al. *On patient flow in hospitals: A data-based queueing-science perspective*. 2011. URL: http://iew3.technion.ac.il/serveng/References/Short_Patient%20flow%20main_010114.pdf.

[2] American Hospital Association. *Survey of hospital leaders*. 2007.

[3] F. Avram, A.J.E.M. Janssen and J.S.H. van Leeuwaarden. 'Loss systems with slow retrials in the Halfin-Whitt regime'. In: *Advanced Applied Probability* 45 (2013), pp. 274–294.

[4] R. Bekker and A.M. de Bruin. 'Time-dependent analysis for refused admissions in clinical wards'. In: *Annals of Operations Research* (2009). DOI: 10.1007/s10479-009-0570-z. URL: https://www.vumc.nl/afdelingen-themas/239911/7351533/7359194/time-dependent_analysis_for1.pdf.

[5] R. Bekker, G. Koole and D. Roubos. *Bed reservation, earmarking and merging of clinical wards*. 2014. URL: http://www.math.vu.nl/~koole/publications/2009report1/art.pdf.

[6] S. Borst, A. Mandelbaum and M.I. Reiman. 'Dimensioning large call centers'. In: *Operations Research* 52(1) (2004), pp. 17–34.

[7] A.M. de Bruin et al. 'Dimensioning hospital wards using the Erlang loss Model'. In: *Annals of Operations Research* (2009), pp. 23–43. DOI: DOI10.1007/s10479-009-0647-8.

[8] A.M. de Bruin et al. 'Modeling the emergency cardiac in-patient Flow: an application of queuing theory'. In: *Health Care Management Science* (2007), pp. 125–137.

[9] J.W. Cohen. 'Basic problems of telephone traffic theory and the influence of repeated calls'. In: *Philips Telecommunication Rev.* 18 (1957), pp. 49–100.

[10] R.B. Cooper. *Introduction to queueing theory*. Springer, 1981.

[11] B.T. Denton, ed. *Handbook of healthcare operations management: Methods and applications*. 2nd. New York: North Holland, 2013.

[12] S.G. Eick, W.A. Massey and W. Whitt. 'The physics of the $M_t/G/\infty$ queue'. In: *Operations Research* 41 (1993), pp. 731–742.

[13] A.K. Erlang. 'The life and works of A.K. Erlang'. In: ed. by E. Brockmeyer, H.L. Halstrom and A. Jensen. The Copenhagen Telephone Company, Copenhagen, Denmark, 1948. Chap. On the rational determination of the number of circuits.

[14] O. Garnett, A. Mandelbaum and M.I. Reiman. 'Designing a call center with impatient customers'. In: *Manufacturing & Service Operations Management* 4(3) (2002), pp. 208–227.

[15] W.J. Gordon and G.F. Newell. 'Cyclic queueing networks with exponential servers'. In: *Operations Research* 15(2) (1967), pp. 254–265.

[16] L.V. Green. 'Queueing analysis in healthcare'. In: ed. by R.W. Hall. Springer, 2006. Chap. 10, pp. 281–308.

[17] L.V. Green. 'Using operations research to reduce delays for healthcare'. In: *Tutorials in Operations Research* (2008), pp. 1–16.

[18] L.V. Green. 'Using queueing theory to increase the effectiveness of physician staffing in the Emergency Department'. In: *Academic Emergency Medicine* (2006), pp. 61–68.

[19] L.V. Green, P.J. Kolesar and W. Whitt. 'Coping with time-varying demand when setting staffing requirements for a service system'. In: *Production and Operations Management* 16.1 (2007), pp. 13–39.

[20] S. Halfin and W. Whitt. 'Heavy-traffic limits for queues with many exponential servers'. In: *Operations Research* 29 (1981), pp. 567–588.

[21] R.W. Hall, ed. *Handbook of healthcare system scheduling*. Springer, 2012.

[22] R.W. Hall, ed. *Patient flow: Reducing delay in healthcare delivery*. Springer, 2006.

[23] A.A. Jagers and E.A. van Doorn. 'On the continued erlang loss function'. In: *Operations Research Letters* 5(1) (1986), pp. 43–46.

[24] A.J.E.M. Janssen, J.S.H. van Leeuwaarden and B. Zwart. 'Gaussian expansions and bounds for the Poisson distribution applied to the Erlang-B formula'. In: *Advances in Applied Probability* 40 (2008), pp. 122–143.

[25] A.J.E.M. Janssen, J.S.H. van Leeuwaarden and B. Zwart. 'Refining square-root safety staffing by expanding Erlang C'. In: *Operations Research* 59(6) (2011), pp. 1512–1522.

[26] O. Jennings and F. de Véricourt. 'Dimensioning large-scale memberschip services'. In: *Operations Research* 55(1) (2008), pp. 173–187.

[27] O. Jennings and F. de Véricourt. 'Nurse staffing in medical units: A queueing perspective'. In: *Operations Research* 59(6) (2011), pp. 1320–1331.

[28] A. Mandelbaum and S. Zeltyn. *The Palm/Erlang-A Queue, with applications to call centers*. Haifa, Israel, 2005. URL: http://ie.technion.ac.il/serveng/References/Erlang_A.pdf.

[29] M.F. Neuts. *Matrix-geometric solutions in stochastic models*. Baltimore: The John Hopkins University Press, 1981.

[30] M.I. Reiman. 'Asymptotically optimal trunk reservation for large trunk groups'. In: *Proceedings of 28th Conference on Decision and Control, IEEE* (1989), pp. 2536–2541.

[31] M.I. Reiman. 'Optimal trunk reservation for a critically loaded link'. In: *Teletraffic and Datatraffic: In a Period of Change* (1991). Ed. by A. Jensen and V.B. Iverson, pp. 247–252.

[32] G. Yom-Tov. 'Queues in hospitals: Queueing networks with ReEntering customers in the QED regime'. PhD thesis. Technion, 2010.

[33] G.B Yom-Tov and A. Mandelbaum. 'Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing'. In: *Manufacturing & Service Operations Management* 16(2) (2014), pp. 283–299.