Asymptotic dimensioning of stochastic service systems

Asymptotic dimensioning of stochastic service systems

This work is part of the Free Competition grant no. 613.001.213, which is financed by the Netherlands Organisation for Scientific Research (NWO).

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-4255-0

This thesis is number D 205 of the thesis series of the Beta Research School for Operations Management and Logistics.

Cover design by Kim Mathijsen.

Asymptotic dimensioning of stochastic service systems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 18 mei 2017 om 16.00 uur

door

Britt Walthera Johanna Mathijsen

geboren te 's-Hertogenbosch

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. B. Koren
1e promotor:	prof.dr. J.S.H. van Leeuwaarden
2e promotor:	prof.dr. A.P. Zwart
leden:	prof.dr. S. Bhulai (Vrije Universiteit Amsterdam)
	dr. J.E. Reed (New York University Stern)
	prof.dr. A.G. de Kok
	prof.dr.ir. O.J. Boxma
adviseur(s):	dr. G.B. Yom-Tov (Technion)

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Acknowledgments

The completion of this thesis would not have been possible without the many people who have supported and encouraged me along the way. I take this opportunity to express my gratitude towards them.

First and foremost, I am greatly indebted to my supervisors Johan van Leeuwaarden and Bert Zwart. Johan, I think I can safely say that you have been my guiding light throughout this academic journey. Ever since my bachelor project, your endless optimism, patience and sense of perspective have been exactly what I needed to keep me motivated through the ups and downs that come with research. Thank you for teaching me to be persistent, critical and positive. Bert, thank you for the continuous flow of ingenious ideas that came during our many discussions. I truly admire your drive, enthusiasm and passion for mathematics.

Much of work presented in this thesis is the result of fruitful collaborations with some great researchers, to whom I wish to express my appreciation as well. First of all, Onno Boxma, your inspiring lectures on stochastic processes may well have marked the starting point of the path that led to this PhD, and I am very happy that you have been willing to be a member of my defense committee. It has been delightful to work with you, together with Shaul Bar-Lev and David Perry, on the subject of Chapter 7. Guido Janssen, thank you for getting me acquainted with the mathematical theory behind asymptotics. Your thoroughness, preciseness and willingness to share your expertise are much appreciated. Galit Yom-Tov, thank you for the nice and intensive collaboration that resulted in Chapter 5, and for serving on my committee. Also, I hold nice memories on our joint effort to organize the YEQT workshop together with Jan-Pieter.

I am also thankful to Sandjai Bhulai, Ton de Kok and Josh Reed for being part of my defense committee, for their time to read this thesis and for providing me with valuable feedback.

During my visits to the Technion, I have been warmly welcomed by Avi Mandelbaum and the people at the SEELab. It has been fascinating to get a glimpse of the goldmine of service system data that they have collected and I am grateful to them for showing me the way. Avi, thank you for kindly introducing me to the

Israeli culture, and for your many advices on research and academic life.

I owe many thanks to my office mates, Fabio, Fiona and Thomas, for creating the most pleasant and comforting work environment. I am pleasantly surprised you have been able to tolerate my ever-fluctuating stress levels, certainly in the last couple of months. The fact that I have remained (reasonably) sane throughout these years is in large part thanks to you.

Outside my office door, I have been lucky to be part of an amazing research group. I want to thank my fellow PhDs of the Stochastics section, including the former generations, for creating and sustaining such an amiable atmosphere. Further, I am grateful to Remco van der Hofstad and Marko Boon for granting me the opportunity to develop my teaching and lecturing skills, which has brought refreshing variation into my job as a PhD student.

Having been a member of the departmental PhD Council, I got the chance to meet many bright young minds, and it has been great fun to collaborate with them to strengthen the PhD community. I wish all of them the best for the future.

When you stick around at the same university for almost nine years, you run the risk of befriending some mathematicians. I am nevertheless very happy to have met Christine, Jorg, Jorn, Laura, Mark, Thomas and many others at this place. Thank you all for making my time at TU/e very enjoyable.

Rik, thank you for holding my hand throughout these years. You have been my rock.

Finally, to my parents Jan and Willemien and my sister Kim, I owe my deepest gratitude. Your unconditional support and comfort have kept me grounded and realistic. I could never have done this without you.

Britt Mathijsen
March 2017

Contents

1 Introduction

Stochastic service systems describe situations in which customers compete for service from scarce resources. Think of check-in lines at airports, waiting rooms in hospitals or queues in supermarkets, where the scarce resource is human manpower. Next to these traditional settings, resource sharing is also important in large-scale service systems such as the internet, wireless networks and cloud computing facilities. In these virtual environments, geographical location does not play a restricting role on the system size, paving the way for the emergence of large-scale resource sharing networks. This thesis investigates how to design large-scale systems in order to achieve economies-of-scale, by which we mean that the system is highly occupied and hence utilizes efficiently the expensive resources, while at the same time, the offered service levels remain high. In this introductory chapter, we give an overview of the available machinery that supports such principles and explain how this thesis contributes to the existing study of large-scale service systems. A crucial concept behind most of the results discussed in the chapter is the Central Limit Theorem (CLT) – arguably one of the most important theorems in mathematics and science.

1.1 Service systems & queueing theory

1.1.1 Quality vs. Efficiency

Large-scale service systems take many shapes and forms. Classical examples of large-scale service systems include call centers [72, 175, 220, 80, 43, 49, 229, 30, 135] and communication systems [142, 14, 133, 143, 207]. More recently, congestion-related issues in health care facilities and cloud-computing facilities have received much attention [17, 94, 225, 95, 205]. In all settings, one can think of service systems as being composed of *customers* and *servers*. In call centers, customers typically call to request help from one of the agents (servers). In communication networks, the data packets are the customers and the communication channels are the servers. In health care facilities, patients are the customers, and nurses/physicians are the servers. The system scale may refer to the size of the client base it caters to, or the magnitude of its capacity, or both. Next to the central notions of customers and servers, we emphasize that service systems are inherently stochastic, that is, subject to uncertainty. Although arrival volumes can be anticipated to some extent over a certain planning horizon, for instance through historical data and forecasting methods, one cannot predict with certainty future arrival patterns. Moreover, service requirements are typically random as well, adding more uncertainty. This intrinsic stochastic variability is a predominant cause of delay experienced by customers in the system.

Due to the inherent randomness in both their arrival and service processes, stochastic models have proved instrumental in both quantifying and improving the operational performance of service systems. Queueing theory and stochastics provide the mathematical tools to describe and evaluate these service systems. Queueing models are often able to capture and explain fundamental phenomena that are common across applications.

A standard model for service systems is the $M/GI/s$ queue, which we will refer to as the *many-server* queue. This model assumes that customers arrive to the queue according to a Poisson process with rate λ, and customer service times are mutually independent and identically distributed (i.i.d.) samples from the distribution of a non-negative random variable B. The parameter s denotes the number of servers in the system, and hence restricts the number of simultaneous services. The case $s = 1$ corresponds to a single-server queue.

First principles say that the queueing process is stable, that is, the number of customers does not explode as time evolves, if and only if the expected workload $R := \lambda\mathbb{E}[B]$ brought into the system per time unit is strictly less than the system capacity. In other words, the *utilization* of the queue, defined as $\rho := \lambda\mathbb{E}[B]/s$ should remain strictly below one. Naturally, a system manager prefers to operate at a utilization level close to one, so that resources are used efficiently. However, it is known that pushing the occupation levels to 100% leads to an explosive increase in congestion. That is, the expected queue length and customer waiting time increase indefinitely, thereby reducing the quality-of-service (QoS) and also customer satisfaction. These seemingly conflicting objectives give rise to a classical trade-off between customer satisfaction and costs of resources.

1.1.2 Economies-of-scale

Under the assumption that service times are exponentially distributed with mean $1/\mu$, the many-server queue reduces to the well-studied $M/M/s$ queue. Despite its simplicity, the analysis of the $M/M/s$ queue explains mathematically the distinctive traits of queues in general, such as the non-linear effect of utilization on the queue size, and pooling effects.

Let $W^{(s)}$ denote the waiting time of a customer and $Q^{(s)}$ the queue length (including the customers in service) in the steady-state $M/M/s$ queue. Without loss of generality, we fix $\mu = 1$, so that $\rho = \lambda/s$. A straightforward balance argument gives the stationary distribution:

$$\pi_k := \mathbb{P}(Q^{(s)} = k) = \begin{cases} \pi_0 \frac{\lambda^k}{k!}, & \text{if } k < s, \\ \pi_0 \frac{\lambda^s}{s!} \rho^{k-s} & \text{if } k \geq s, \end{cases} \tag{1.1}$$

where

$$\pi_0 := \Big(\sum_{k=0}^{s-1} \frac{\lambda^k}{k!} + \frac{1}{1-\rho} \frac{\lambda^s}{s!} \Big)^{-1}.$$

Natural QoS indicators include the expected waiting time $\mathbb{E}[W^{(s)}]$ and the delay probability $\mathbb{P}(W^{(s)} > 0)$. Invoking the PASTA (Poisson arrivals see time averages) property [224], we know that the delay probability equals the probability of the queue length being greater or equal to the number of servers s. Thus,

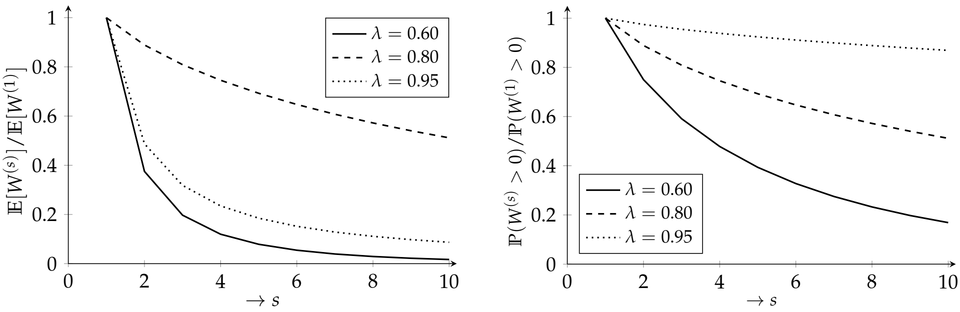
$$\mathbb{P}(W^{(s)} > 0) = \mathbb{P}(Q^{(s)} \geq s) = \frac{\lambda^s}{s!} \Big((1-\rho) \sum_{k=0}^{s-1} \frac{\lambda^k}{k!} + \frac{\lambda^s}{s!} \Big)^{-1}. \tag{1.2}$$

By Little's law, which says that $\mathbb{E}[(Q^{(s)} - s)^+] = \lambda \mathbb{E}[W^{(s)}]$, we furthermore have

$$\mathbb{E}[W^{(s)}] = \mathbb{P}(W^{(s)} > 0) \, \frac{1/s}{1-\rho}. \tag{1.3}$$

From these formulae, it is readily seen that $\mathbb{P}(W^{(s)} > 0) \to 1$ and $\mathbb{E}[W^{(s)}] \to \infty$ as $\rho \uparrow 1$. That is, increasing λ to s, while keeping the latter fixed, leads to a system in which all customers are delayed before service, and the expected delay before reaching a server increases to infinity.

The $M/M/s$ queue also reveals the effect of *resource pooling*. To illustrate the operational benefits of sharing resources, we compare a system of s separate $M/M/1$ queues, each serving a Poisson arrival stream with rate $\lambda < 1$, against one $M/M/s$ queue facing arrival rate λs. The two systems thus experience the same total workload and utilization, namely $\rho = \lambda$. We fix the value of λ and vary s. Obviously, the waiting time and queue length distribution in the first scenario are unaffected



(a) Expected waiting time (b) Probability of delay

Figure 1.1: Effects of resource pooling in the $M/M/s$ queue.

by the parameter s, since there is no interaction between the single-server queues. This lack of coordination tolerates a scenario of having an idle server, while the total number of customers in the system exceeds s, therefore wasting resource capacity. Such an event cannot happen in the many-server scenario, due to the central queue. This central coordination improves QoS. Indeed Figure 1.1 shows that the reduction in expected waiting time can be substantial.

So pooling kills two birds with one stone: QoS for customers improves and the system efficiency increases.

1.1.3 Many-server scaling regimes

Now that we know that economies-of-scale can be achieved, it is relevant to ask how to match capacity s to a demand λ in the setting where both s and λ become large. The expressions in (1.2) and (1.3) provide a starting point for finding such demand-capacity relations, particularly when we apply asymptotic analysis for $s \to \infty$, [101, 43, 184]. Asymptotic theory of many-server systems relies on the prerequisite that the limiting behavior of the service system is determined by the way in which capacity s is adjusted to demand, assuming demand grows large. We illustrate this idea by investigating typical sample paths of the queue length process $Q = \{Q(t)\}_{t\geq 0}$ of an $M/M/s$ queue for increasing values of λ.

Figure 1.2 depicts a sample path for $\lambda = 3$ and $s = 4$. The number of customers queueing at time t is given by $(Q(t) - s)^+$ with $(\cdot)^+ := \max\{0, \cdot\}$. The number of idle servers is given by $(s - Q(t))^+$. In Figure 1.2, the red and green area hence represent the cumulative queue length and cumulative number of idle servers, respectively, over the given time period. Bearing in mind the dual goal of QoS and efficiency, we want to minimize both of these areas simultaneously.

Next, we conduct a similar sample path experiment for increasing values of λ. Since $s > \lambda$ is required for stability, the value of s needs to be adjusted accordingly.

We propose three scaling rules:

$$s_\lambda^{(1)} = [\lambda + \beta], \qquad s_\lambda^{(2)} = \left[\lambda + \beta\sqrt{\lambda}\right], \qquad s_\lambda^{(3)} = [\lambda + \beta\,\lambda], \tag{1.4}$$

for some $\beta > 0$, where $[\cdot]$ denotes the rounding operator. Note that these three rules differ in terms of overcapacity $s - \lambda$. Figure 1.3 depicts typical sample paths of the queue length process for increasing values of λ for the three scaling rules with $\beta = 0.5$.

Observe that for all scaling rules, the stochastic fluctuations of the queue length processes relative to λ decrease with the system size. Moreover, the paths in Figure 1.3 appear to become smoother with increasing λ. Of course, the actual sample path always consists of upward and downward jumps of size 1, but we will show how proper centering and scaling of the queue length process indeed gives rise to a *diffusion process* in the limit as $\lambda \to \infty$. Although the difference in performance of the three regimes is not yet evident for relatively small λ, clear distinctive behavior occurs for large λ.

Under ${s_\lambda}^{(1)}$, the majority of customers is delayed and server idle time is low, since $\rho = (1 + \beta/\lambda)^{-1} \to 1$ as $\lambda \to \infty$. Systems dimensioned according to this rule value server efficiency over customer satisfaction and therefore this regime is in the literature also known as the *efficiency-driven* (ED) regime [229].

In contrast, the third scaling rule $s_\lambda^{(3)}$ yields a constant utilization level $\rho = 1/(1 + \beta)$, which stays away from 1, even for large λ. Queues operating in this regime exhibit significant server idle times. Moreover, for the particular realization of the queueing processes for $\lambda = 50$ and $\lambda = 100$ none of the customers waits. This customer-centered regime is known as the *quality-driven* (QD) regime [229].

The scaling rule $s_\lambda^{(2)}$ is in some ways a combination of the other two regimes. First, we have $\rho = (1 + \beta/\sqrt{\lambda})^{-1} \to 1$ as $\lambda \to \infty$, which indicates efficient usage of resources as the system grows. The sample paths, however, indicate that only a fraction of the customers is delayed, and only small queues arise, which suggest good QoS. This regime is therefore called *quality-and-efficiency driven* (QED) regime.

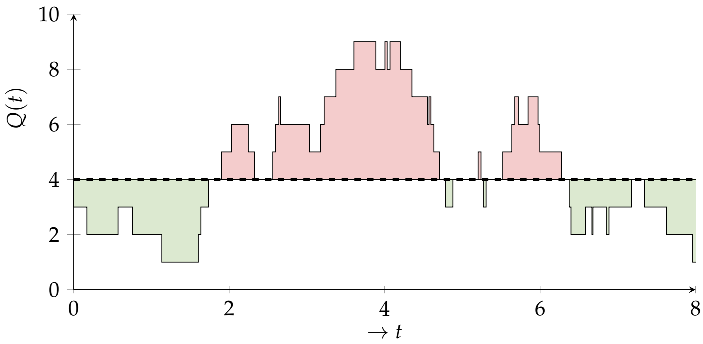


Figure 1.2: Sample path of the $M/M/s$ queue with $\lambda = 3$ and $s = 4$.

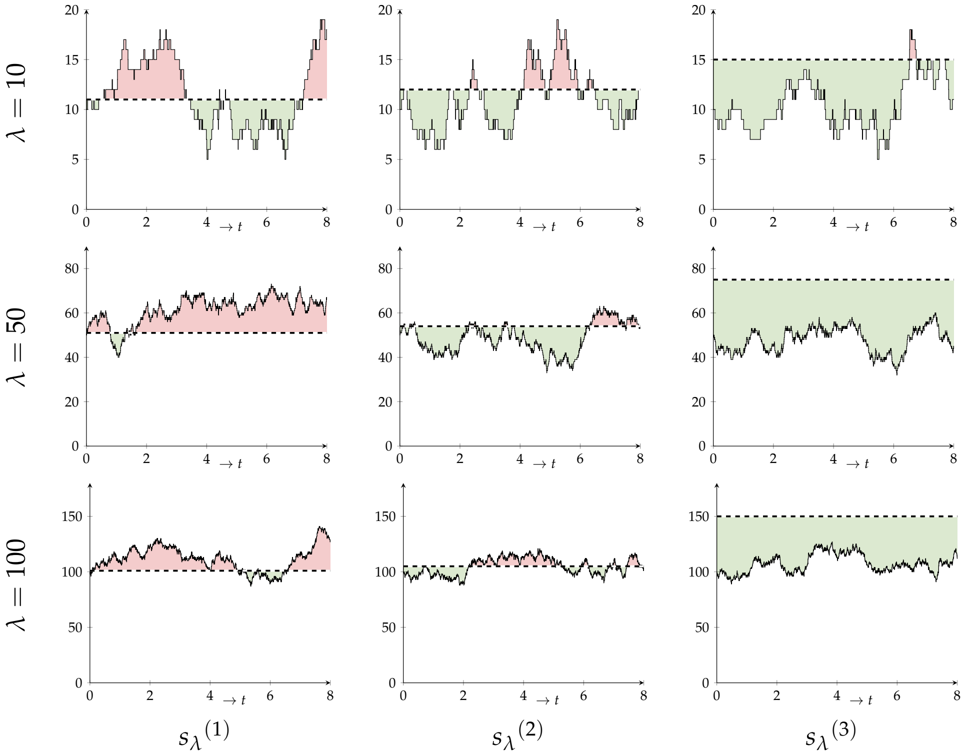


Figure 1.3: Sample paths of the $M/M/s$ queue with $\lambda = 10$, 50 and 100 and s set according to the three scaling rules in (1.4).

Since this scaling regime and the related *square-root staffing rule*

$$s_\lambda = \lambda + \beta\sqrt{\lambda} \tag{1.5}$$

strikes the right balance between the two profound objectives of capacity allocation in service systems, we discuss in the next section the mathematical foundations of the QED regime and quantify the favorable properties revealed by Figure 1.3.

1.2 The QED regime: two canonical examples

We saw in Figure 1.1 the advantageous effect of resource pooling and economies-of-scale in many-server systems. In this section, we will explain how this is related to the Central Limit Theorem (CLT).

Theorem 1.1 (Central Limit Theorem, e.g. [37, Thm. 27.1])**.** *Consider a sequence $X_1, X_2, \ldots, X_n$ of independent and identically distributed random variables having mean μ and positive variance σ^2. Then,*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \overset{d}{\Rightarrow} \mathcal{N}(0,1), \qquad \textit{for } n \to \infty.$$

where $\overset{d}{\Rightarrow}$ denotes convergence in distribution and $\mathcal{N}(0,1)$ is a random variable with standard normal distribution.

We shall now apply the CLT to the delay probability in the $M/M/s$ queue. Striking the proper balance between queueing delay and server efficiency asymptotically, i.e. balancing the green and red areas in Figure 1.3, in mathematical terms boils down to choosing a service level s_λ such that both the delay probability $\mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda)$ and $\mathbb{P}(Q^{(s_\lambda)} < s_\lambda)$ remain strictly smaller than 1 as $\lambda \to \infty$. In other words, one would like to see that $\mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda)$ converges to a non-degenerate limit $\alpha \in (0,1)$ as $\lambda \to \infty$.

To get a feel for the natural scale of the queue, we first examine the situation with unlimited capacity. More precisely, let $Q^{(\infty)}$ be the number of customers in a steady-state $M/GI/\infty$ queue with mean service requirement $\mathbb{E}[B] = 1$. Notice that in this infinite-server setting, $Q^{(\infty)}$ also represents the steady-state number of busy servers. It is well known that $Q^{(\infty)}$ follows a Poisson distribution with mean equal to the expected workload, in our case $R = \lambda$. Moreover, if we assume that λ is integer, then a Poisson random variable with rate λ can be viewed as the sum of λ i.i.d. Poisson random variables with rate 1. In other words, $Q^{(\infty)} = \sum_{i=1}^{\lambda} P_i$, where the P_i's, $i = 1, 2, \ldots, n$, have Poisson distribution with unit mean and variance, and are mutually independent. The CLT thus gives

$$\mathbb{P}(Q^{(\infty)} \geq x_\lambda) = \mathbb{P}\left(\frac{Q^{(\infty)} - \lambda}{\sqrt{\lambda}} \geq \frac{x_\lambda - \lambda}{\sqrt{\lambda}}\right) \approx 1 - \Phi\left(\frac{x_\lambda - \lambda}{\sqrt{\lambda}}\right), \tag{1.6}$$

where Φ denotes the cumulative distribution function of the standard normal distribution for large λ. Hence, the probability in (1.6) converges to a constant value away from both 0 and 1 if and only if $(x_\lambda - \lambda)/\sqrt{\lambda} \to x \in \mathbb{R}$, or equivalently $x_\lambda = \lambda + x\sqrt{\lambda} + o(\sqrt{\lambda})$, as $\lambda \to \infty$. Here, the relation $u(\lambda) = o(v(l))$ implies that $u(\lambda)/v(\lambda) \to 0$ as $\lambda \to \infty$. Equation (1.6) also shows that the leading order of the random variable describing the queue length is λ, while the stochastic fluctuations are of order $\sqrt{\lambda}$.

If we now pretend, for a moment, that the infinite-server queue serves as a good approximation for the many-server queue with s_λ servers, then (1.6) says that the steady-state probability of delay for $s_\lambda = \lambda + \beta\sqrt{\lambda}$ obeys the Gaussian approximation

$$\mathbb{P}(W^{(s_\lambda)} > 0) = \mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda) \approx 1 - \Phi(\beta), \tag{1.7}$$

where Φ denotes the cumulative distribution function (cdf) of the standard normal distribution. Of course, the infinite-server system ignores the one thing that makes a queueing system unique, namely that a queue is formed when all servers are busy. During these periods of congestion, customers will depart from a system with a finite number of servers at a slower pace than in its infinite-server counterpart. So the approximation in (1.7) is likely to overestimate the actual delay probability, and a more careful investigation of the queue length process in many-server settings is needed. Nevertheless, the infinite-server heuristic reveals that in a well-managed system, i.e. queues are of acceptable length, the size at which the system operates is of the order λ, with fluctuations of order $\sqrt{\lambda}$. We shall now demonstrate through two canonical examples how these guessed natural scalings can be turned into mathematically rigorous statements. Both examples which will play a key role in this thesis.

1.2.1 The $M/M/s$ queue

Converging delay probability. Let $Q^{(s)}$ denote the steady-state number of customers in an $M/M/s$ queue with arrival rate λ and mean service requirement 1, of which the probability distribution is given in (1.1). Halfin & Whitt [101] showed that, just as the tail probability in the infinite-server setting, the delay probability in the $M/M/s$ queue converges under scaling (1.5) to a value between 0 and 1. Moreover, they showed that this is in fact the only scaling regime in which such a non-degenerate limit exists and identified its value. Let φ denote the probability density function (pdf) of the standard normal distribution.

Proposition 1.1 ([101, Prop. 2.1]). *The probability of delay in the $M/M/s_\lambda$ queue has the non-degenerate limit*

$$\lim_{\lambda\to\infty} \mathbb{P}(W^{(s_\lambda)} > 0) = \alpha \in (0,1) \tag{1.8}$$

if and only if

$$\lim_{\lambda\to\infty} (1 - \rho_{s_\lambda})\sqrt{s_\lambda} \to \beta, \quad \beta > 0, \tag{1.9}$$

where α is given by

$$\alpha = \left(1 + \frac{\beta\,\Phi(\beta)}{\varphi(\beta)}\right)^{-1} =: g(\beta). \tag{1.10}$$

Proof. We first prove the sufficiency condition. Rewrite (1.2) as

$$\mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda) = \left(1 + (1-\rho_{s_\lambda})\frac{\mathbb{P}(\mathrm{Pois}(\lambda) < s_\lambda)}{\mathbb{P}(\mathrm{Pois}(\lambda) = s_\lambda)}\right)^{-1}. \tag{1.11}$$

Similar to (1.6) we find

$$\begin{aligned}
\mathbb{P}(\mathrm{Pois}(\lambda) < s_\lambda) &= \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < \frac{s_\lambda - \lambda}{\sqrt{\lambda}}\right)\\
&= \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < (1-\rho_{s_\lambda})\,\frac{s_\lambda}{\sqrt{\lambda}}\right)\\
&= \mathbb{P}\left(\frac{\mathrm{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} < (1-\rho_{s_\lambda})\,\sqrt{s_\lambda}\,(1+o(1))\right) \to \Phi(\beta), \qquad (1.12)
\end{aligned}$$

for $\lambda \to \infty$. Using Stirling's approximation, we get

$$\mathbb{P}(\mathrm{Pois}(\lambda) = s) = \mathrm{e}^{-\lambda}\frac{\lambda^{s_\lambda}}{s_\lambda!} \sim \mathrm{e}^{-\lambda}\lambda^{s_\lambda}\cdot\frac{1}{\sqrt{2\pi\,s_\lambda}}\left(\frac{\mathrm{e}}{s_\lambda}\right)^{s_\lambda} = \frac{1}{\sqrt{2\pi s_\lambda}}\,\mathrm{e}^{s_\lambda - \lambda - s_\lambda \ln(\rho_{s_\lambda})}.$$

Since $\ln(\rho_{s_\lambda}) = -(1-\rho_{s_\lambda}) - \frac{1}{2}(1-\rho_{s_\lambda})^2 + o((1-\rho_{s_\lambda})^2)$ we find that

$$\frac{\mathbb{P}(\mathrm{Pois}(\lambda) = s)}{1-\rho_{s_\lambda}} = \frac{1}{(1-\rho_{s_\lambda})\sqrt{s_\lambda}}\,\frac{\mathrm{e}^{-\frac{1}{2}(1-\rho_{s_\lambda})^2 s_\lambda + o\left((1-\rho_{s_\lambda})^2 s_\lambda\right)}}{\sqrt{2\pi}} \to \frac{1}{\beta}\,\frac{\mathrm{e}^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} = \frac{\varphi(\beta)}{\beta}. \tag{1.13}$$

Substituting (1.12) and (1.13) into (1.11) gives (1.10). The necessary condition follows directly by the characterization of $\mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda)$ as in (1.11) by observing, through (1.12) and (1.13), that the term

$$(1-\rho_{s_\lambda})\frac{\mathbb{P}(\mathrm{Pois}(\lambda) < s_\lambda)}{\mathbb{P}(\mathrm{Pois}(\lambda) = s_\lambda)}$$

has a limiting value in $(0,\infty)$ only if $(1-\rho_{s_\lambda})\sqrt{s_\lambda} \to \beta$ for some $\beta > 0$. □

Observe that $g(\beta)$ is a strictly decreasing function on $(0,\infty)$ with $g(\beta) \to 1$ as $\beta \to 0$ and $g(\beta) \to 0$ for $\beta \to \infty$. Thus all possible delay probabilities are achievable in the QED regime, which will prove useful for the dimensioning of systems (see Section 1.3).

Although Proposition 1.1 is an asymptotic result for $\lambda \to \infty$, Figure 1.4 shows that $g(\beta)$ can serve as an accurate approximation for the delay probability for relatively small λ.

From Proposition 1.1, it also follows that under (1.9),

$$\sqrt{s_\lambda}\,\mathbb{E}[W^{(s_\lambda)}] = \frac{\mathbb{P}(W^{(s_\lambda)} > 0)}{(1-\rho_{s_\lambda})\sqrt{s_\lambda}} \to \frac{g(\beta)}{\beta} =: h(\beta), \qquad \text{as } \lambda \to \infty, \tag{1.14}$$

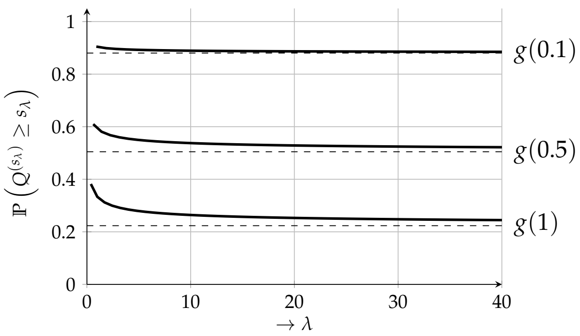


Figure 1.4: The delay probability $\mathbb{P}(Q^{(s_\lambda)} \geq s_\lambda)$ with $s_\lambda = [\lambda + \beta\sqrt{\lambda}]$ for $\beta = 0.1,\ 0.5,$ and 1 as a function of λ.

where we have used the characterization of $\mathbb{E}[W^{(s_\lambda)}]$ in (1.3). This implies that in the QED regime, the expected waiting time vanishes at rate $1/\sqrt{s_\lambda}$ as $\lambda \to \infty$. By Little's law this implies that the expected queue length is $O(\sqrt{s_\lambda})$. By the relation $u(\lambda) = O(v(\lambda))$ we mean that $\limsup_{\lambda\to\infty} u(\lambda)/v(\lambda) < \infty$. While these are all steady-state results, similar statements can be made for the entire queue-length process, as shown next.

The theoretical results of the QED regime we presented here are based on steady-state queueing analysis. But at the heart of the QED theory lies a much deeper result in which the entire queue-length process, over all points in time, converges to some other limiting process.

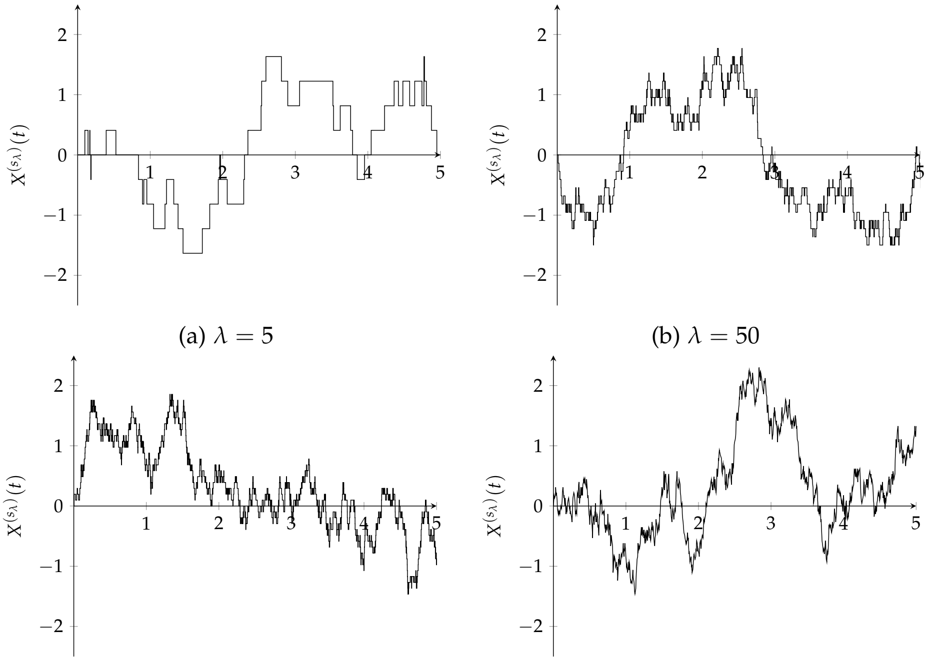
Process-level convergence. Obtaining rigorous statements about stochastic-process limits poses considerable mathematical challenges. Rather than presenting the deep technical details of the convergence results, we give a heuristic explanation of how the limiting process arises and what it should look like.

The queue-length process $Q^{(s_\lambda)}$ in Figure 1.3 with scaling rule $s_\lambda = [\lambda + \beta\sqrt{\lambda}]$ appears to concentrate around the level s_λ. As argued before, the stochastic fluctuations are of order $\sqrt{\lambda}$, or equivalently $\sqrt{s_\lambda}$. For that reason, we consider the centered and scaled process

$$X^{(s_\lambda)}(t) := \frac{Q^{(s_\lambda)}(t) - s_\lambda}{\sqrt{s_\lambda}}, \qquad \text{for all } t \geq 0, \tag{1.15}$$

and ask what happens to this process as $\lambda \to \infty$. First, we consider the expected drift conditioned on $X^{(s_\lambda)}(t) = x$. When $x > 0$, this corresponds to a state in which $Q^{(s_\lambda)} > s_\lambda$ and hence all servers are occupied. Therefore, the expected rate at which customers leave the system is s_λ, while the arrival rate remains λ, so that the expected drift of $X^{(s_\lambda)}(t)$ in $x > 0$ satisfies

$$\frac{\lambda - s_\lambda}{\sqrt{s_\lambda}} \to -\beta, \qquad \text{as } \lambda \to \infty,$$



(a) $\lambda = 5$ (b) $\lambda = 50$

(c) $\lambda = 100$ (d) $\lambda = 500$

Figure 1.5: Sample paths of the normalized queue length process $X^{(s_\lambda)}(t)$ with $\lambda = 5$, $\lambda = 5$ and $\lambda = 500$ and $s_\lambda = [\lambda + 0.5\sqrt{\lambda}]$.

under scaling $\sqrt{s_\lambda}(1 - \rho_{s_\lambda}) \to \beta$ in (1.9). When $x \leq 0$, only $s_\lambda + x\sqrt{s_\lambda}$ servers are working, so that the net drift is

$$\frac{\lambda - (s_\lambda + x\sqrt{s_\lambda})}{\sqrt{s_\lambda}} \to -\beta - x, \qquad \text{as } \lambda \to \infty.$$

Now, imagine what happens to the sample paths of $\{X^{(s_\lambda)}(t)\}_{t \geq 0}$ as we increase λ. Within a fixed time interval, larger λ and s_λ will trigger more and more events, both arrivals and departures. Also, the jump size at each event epoch decreases as $1/\sqrt{s_\lambda}$ as a consequence of the scaling in (1.15). Hence, there will be more events, each with a smaller impact, and in the limit as $\lambda \to \infty$, there will be infinitely many events of infinitesimally small impact. This heuristic explanation suggests that the process $X^{(s_\lambda)}(t)$ converges to a stochastic-process limit, which is continuous, and has infinitesimal drift $-\beta$ above zero and $-\beta - x$ below zero. Figure 1.5 visualizes the appearance of the suggested scaling limit as λ and s_λ increase.

The following theorem by Halfin & Whitt [101] characterizes this scaling limit more formally.

Theorem 1.2. *Let $X^{(s_\lambda)}(0) \stackrel{d}{\Rightarrow} X(0) \in \mathbb{R}$ and $\sqrt{s_\lambda}(1-\rho_{s_\lambda}) \to \beta$. Then for all $t \geq 0$,*

$$X^{(s_\lambda)}(t) \stackrel{d}{\Rightarrow} X(t), \qquad \text{as } \lambda \to \infty,$$

where $X(t)$ is the diffusion process with infinitesimal drift $m(x)$ given by

$$m(x) = \begin{cases} -\beta, & \text{if } x > 0, \\ -\beta - x, & \text{if } x \leq 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2$.

The limiting diffusion process $\{X(t)\}_{t\geq 0}$ in Theorem 1.2 is a combination of a negative-drift Brownian motion in the upper half plane and an Ornstein-Uhlenbeck (OU) process in the lower half plane. We refer to this hybrid diffusion process as the Halfin-Whitt diffusion. Much is known for such diffusion processes with piecewise linear drift coefficient, see [210, 75]. Its stationary distribution can for instance be derived, see e.g. [51].

Proposition 1.2. *Let $X(t) \stackrel{d}{\Rightarrow} X(\infty)$ as $t \to \infty$ for a random variable $X(\infty)$ and $(1-\rho_{s_\lambda})\sqrt{s_\lambda} \to \beta$ for $\lambda \to \infty$. Then*

$$\mathbb{P}(X(\infty) > 0) = g(\beta), \tag{1.16}$$

$$\mathbb{P}(X(\infty) \geq x | X(\infty) > 0) = \mathrm{e}^{-\beta x}, \quad \text{for } x > 0, \tag{1.17}$$

$$\mathbb{P}(X(\infty) \leq x | X(\infty) \leq 0) = \frac{\Phi(\beta + x)}{\Phi(\beta)}, \quad \text{for } x \leq 0. \tag{1.18}$$

This result shows that as the system grows large, the $Q^{(s_\lambda)}(t)$ concentrates around s_λ, and the fluctuations are of order $\sqrt{s_\lambda}$. Moreover, Proposition 1.2 iterates the limiting values for the delay probability and scaled expected delay. Namely,

$$\mathbb{P}\big(W^{(s_\lambda)} > 0\big) \to \mathbb{P}(X(\infty) > 0) = g(\beta)$$

and

$$\sqrt{s_\lambda}\mathbb{E}[W^{(s_\lambda)}] \approx \frac{\mathbb{E}[Q^{(s_\lambda)}]}{\sqrt{s_\lambda}} \to \mathbb{E}[X(\infty)] = \int_0^\infty g(\beta)\mathrm{e}^{-\beta x}\mathrm{d}x = \frac{g(\beta)}{\beta},$$

For obvious reasons, the QED regime is also referred to as the Halfin-Whitt regime, and both these names are used interchangeably in this thesis.

1.2.2 The $M/D/s$ queue

We next consider a many-server queue with deterministic service requirements equal to one, a Poisson arrival process of rate λ and s_λ servers. We let $Q^{(s_\lambda)}(t)$ be the process describing the number of customers in the system and only examine the process at discrete time epochs $t = 0, 1, 2, \ldots$. In our analysis, we focus on the queue length process $Z^{(s_\lambda)}(t) := (Q^{(s_\lambda)}(t) - s_\lambda)^+$.

Since we discretize time, the number of new arrivals per time period is given by the sequence of i.i.d. random variables $\{A_k\}_{k\geq 1}$, which has a Poisson distribution with mean λ. At the start of the k^{th} period, $Z^{(s_\lambda)}(k)$ customers are waiting. Because the service time of a customer is equal to the period length, all $\min\{Q^{(s_\lambda)}(k), s_\lambda\}$ customers in service at the beginning of the period will have left the system by time $t = k+1$. This implies that $\min\{Z^{(s_\lambda)}(k), s_\lambda\}$ of the waiting customers are taken into service during period k, but could not possibly have departed before its end, due to the deterministic service times. If $Z^{(s_\lambda)}(k) < s_\lambda$, then additionally $\min\{A_k, s_\lambda - Z^{(s_\lambda)}(k)\}$ of the new arrivals are taken into service. This yields a total of A_k arrivals, and $\min\{Z^{(s_\lambda)}(k) + A_k, s_\lambda\}$ departures from the queue during period k, which gives the Lindley type recursion [148], with $Z^{(s_\lambda)}(0) = 0$,

$$Z^{(s_\lambda)}(k+1) = Z^{(s_\lambda)}(k) + A_k - \min\{Z^{(s_\lambda)}(k) + A_k, s_\lambda\} = \max\{0, Z^{(s_\lambda)}(k) + A_k - s_\lambda\}. \tag{1.19}$$

The queue length process thus gives rise to a random walk with i.i.d. steps of size $(A^{(s_\lambda)} - s_\lambda)$, with a reflecting barrier at zero. We can iterate the recursion in (1.19) to find

$$\begin{aligned} Z^{(s_\lambda)}(k+1) &= \max\Big\{0, Z^{(s_\lambda)}(k) + A_k - s_\lambda\Big\} \\ &= \max\Big\{0, \max\{0, Z^{(s_\lambda)}(k-1) + (A_{k-1} - s_\lambda)\} + (A_k - s_\lambda)\}\Big\} \\ &= \max\Big\{0, (A_k - s_\lambda), Z^{(s_\lambda)}(k-1) + (A_k - s_\lambda) + (A_{k-1} - s_\lambda)\Big\} \\ &= \max_{0\leq j\leq k}\Big\{\sum_{i=1}^{j}(A_{k-i} - s_\lambda)\Big\} \stackrel{d}{=} \max_{0\leq j\leq k}\Big\{\sum_{i=1}^{j}(A_i - s_\lambda)\Big\}, \end{aligned} \tag{1.20}$$

where the last equality in distribution holds due to the duality principle for random walks, see e.g. [188, Sec. 7.1]. For stability, the expected step size satisfies $\mathbb{E}[A_k - s_\lambda] = \lambda - s_\lambda < 0$. We use the shorthand notation for the partial sum $S_k := \sum_{i=1}^{k}(A_i - s_\lambda)$. Let $Z^{(s_\lambda)}(\infty) := \lim_{k\to\infty} Z^{(s_\lambda)}(k)$ denote the stationary queue length in this $M/D/s$ queue, which can be shown to exist under our assumptions. The probability generating function (pgf) of $Z^{(s_\lambda)}(\infty)$ can then be expressed in terms of the pgf of the positive parts of the partial sum:

$$\mathbb{E}[w^{Z^{(s_\lambda)}(\infty)}] = \exp\Big\{-\sum_{k=1}^{\infty}\frac{1}{k}\,(1 - \mathbb{E}[w^{S_k^+}])\Big\}, \qquad |w| \leq 1. \tag{1.21}$$

From (1.21), which is a special case of Spitzer's identity [197], we obtain for the mean queue length and empty-queue probability the expressions

$$\begin{aligned} \mathbb{E}[Z^{(s_\lambda)}(\infty)] &= \sum_{k=1}^{\infty}\frac{1}{k}\,\mathbb{E}[S_k^+], \\ \mathbb{P}(Z^{(s_\lambda)}(\infty) = 0) &= \exp\Big\{-\sum_{k=1}^{\infty}\frac{1}{k}\,\mathbb{P}(S_k^+ > 0)\Big\}. \end{aligned} \tag{1.22}$$

Although explicit, the expressions in (1.22) reveal little of the structure of the queue length process. Hence, we again turn to asymptotics.

Gaussian random walk. We take another look at the identity in (1.20), and ask ourselves what happens if λ grows large. Since $\mathbb{E}[A_k - s_\lambda] = \lambda - s_\lambda = -\beta\sqrt{\lambda} + o(\sqrt{\lambda})$ under the QED scaling (1.5), it makes sense to consider the scaled queue length process $X^{(s_\lambda)}(k) := Z^{(s_\lambda)}(k)/\sqrt{\lambda}$ for all $k \geq 0$, with scaled steps $Y_k^{(s_\lambda)} := (A_k - s_\lambda)/\sqrt{\lambda}$. Dividing both sides of (1.20) by $\sqrt{\lambda}$ then gives

$$X^{(s_\lambda)}(k+1) = \max_{0 \leq j \leq k} \Big\{ \sum_{i=1}^{j} Y_i^{(s_\lambda)} \Big\}. \tag{1.23}$$

Observe that $A_k \stackrel{d}{=} \mathrm{Pois}(\lambda)$ with $\mathrm{Pois}(\lambda)$ a random variable with mean λ. Hence by the CLT

$$Y_k^{(s_\lambda)} = \frac{A_k - s_\lambda}{\sqrt{\lambda}} = \frac{A_k - \lambda}{\sqrt{\lambda}} - \beta \stackrel{d}{\Rightarrow} Y_k \stackrel{d}{=} \mathcal{N}(-\beta, 1),$$

for $\lambda \to \infty$, where $\mathcal{N}(-\beta, 1)$ denotes a normally distributed random variable with mean $-\beta$ and standard deviation 1. So we expect the scaled queue length process to converge in distribution to a reflected random walk with normally distributed increments, i.e. a reflected *Gaussian random walk*. Indeed, it is easily verified that [121],

$$X^{(s_\lambda)}(k) \stackrel{d}{\Rightarrow} M_\beta(k) := \max_{0 \leq j \leq k} \Big\{ \sum_{i=1}^{j} Y_j \Big\}, \qquad \lambda \to \infty. \tag{1.24}$$

Let $M_\beta := \lim_{k\to\infty} M_\beta(k)$ denote the all-time maximum of a Gaussian random walk. It can be shown that M_β almost surely exists and that

$$X^{(s_\lambda)}(\infty) := \lim_{k\to\infty} X^{(s_\lambda)}(k) \stackrel{d}{\Rightarrow} M_\beta,$$

for instance by [197, Prop. 19.2] and [20, Thm. X6.1]. The following theorem can be proved using a similar approach as in [122]. (We prove this result in a more general setting in Chapter 3.)

Theorem 1.3. *Let $X^{(s_\lambda)}(\infty)$ be the scaled queue length in steady-state. If $(1 - \rho_{s_\lambda})\sqrt{\lambda} \to \beta$, then as $\lambda \to \infty$,*

(i) $X^{(s_\lambda)}(\infty) \stackrel{d}{\Rightarrow} M_\beta$,

(ii) $\mathbb{P}(X^{(s_\lambda)}(\infty) = 0) \to \mathbb{P}(M_\beta = 0)$,

(iii) $\mathbb{E}[X^{(s_\lambda)}(\infty)^k] \to \mathbb{E}[M_\beta^k]$, *for any $k > 0$.*

The Gaussian random walk is well studied [191, 57, 115, 38, 115] and there is an intimate connection with Brownian motion. The only difference, one could say, is

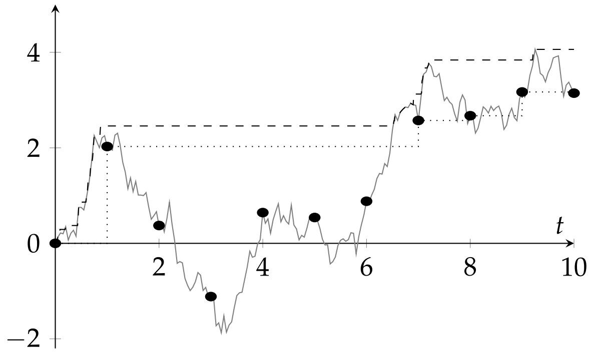


Figure 1.6: Brownian motion (gray) and embedded Gaussian random walk (marked) with their respective running maxima (dashed and dotted, respectively).

that Brownian motion is a continuous-time process, whereas the Gaussian random walk only changes at discrete points in time. If $\{B(t)\}_{t\geq 0}$ is a Brownian motion with drift $-\mu < 0$ and infinitesimal variance σ^2 and $\{W(t)\}_{t\geq 0}$ is a random walk with $\mathcal{N}(-\mu, \sigma^2)$ steps and $B(0) = W(0)$, then W can be regarded as the process B embedded at equidistant time epochs. That is, $W(t) \stackrel{d}{=} B(t)$ for all $t \in \mathbb{N}^+$. For the maximum of both processes this coupling implies

$$\max_{k\in\mathbb{N}^+} W(k) = \max_{k\in\mathbb{N}^+} B(k) \leq_{\mathrm{st}} \max_{t\in\mathbb{R}^+} B(t), \tag{1.25}$$

where $\leq_{\mathrm{st}}$ denotes stochastic dominance. This difference in maximum is visualized in Figure 1.6. It is known that the all-time maximum of Brownian motion with negative drift $-\mu$ and infinitesimal variable σ^2 has an exponential distribution with mean $\sigma/2\mu$ [104]. Hence, (1.25) implies that M_β is stochastically upper bounded by an exponential random variable with mean $1/2\beta$.

Despite this easy bound, precise results for M_β are more involved. Let ζ denote the Riemann zeta function, which is defined as, see e.g. [174, Eq. 25.2.1],

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{1.26}$$

Theorem 1.4 ([57, Thm. 1] & [115, Thm. 2 & 3]). *For* $0 < \beta < 2\sqrt{\pi}$,

$$\mathbb{P}(M_\beta = 0) = \sqrt{2}\beta \exp\left\{ \frac{\beta}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \frac{\zeta(1/2-l)}{l!(2l+1)} \left(\frac{-\beta^2}{2}\right)^l \right\}, \tag{1.27}$$

$$\mathbb{E}[M_\beta] = \frac{1}{2\beta} + \frac{\zeta(1/2)}{\sqrt{2\pi}} + \frac{\beta}{4} + \frac{\beta^2}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \frac{\zeta(-1/2-l)}{l!(2l+1)(2l+2)} \left(\frac{-\beta^2}{2}\right)^l, \tag{1.28}$$

$$\begin{aligned}\operatorname{Var} M_\beta = {} & \frac{1}{4\beta^2} - \frac{1}{4} - \frac{2\zeta(-1/2)}{\sqrt{2\pi}}\beta - \frac{\beta^2}{24} \\ & - \frac{2\beta^3}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \frac{\zeta(-3/2-l)}{l!(2l+1)(2l+2)(2l+3)} \left(\frac{-\beta^2}{2}\right)^l.\end{aligned} \tag{1.29}$$

1.2.3 Characteristics of the QED regime

Now that we have seen how the square-root staffing rule (1.5) yields non-degenerate limiting behavior in two classical queueing models, we shall elaborate on how the QED regime gives rise to (at least) three desirable properties. The first property relates to the efficient usage of resources, expressed as

$$\rho_{s_\lambda} = \frac{\lambda}{s_\lambda} = 1 - \frac{\beta}{\sqrt{s_\lambda}} + O(1/\lambda), \tag{Efficiency}$$

where we have used that $s_\lambda = O(\lambda)$. The second distinctive property is the balance between QoS and efficiency:

$$\mathbb{P}(W^{(s_\lambda)} > 0) \to g(\beta), \qquad \text{and} \qquad \mathbb{P}(W^{(s_\lambda)} > 0) \to 1 - \mathbb{P}(M_\beta = 0), \tag{Balance}$$

as $s_\lambda \to \infty$, in the $M/M/s$ queue and $M/D/s$ queue, respectively. The third property relates to good QoS:

$$\mathbb{E}[W^{(s_\lambda)}] = \frac{h(\beta)}{\sqrt{s_\lambda}} + o(1/\sqrt{s_\lambda}) \qquad \text{and} \qquad \mathbb{E}[W^{(s_\lambda)}] = \frac{\mathbb{E}[M_\beta]}{\sqrt{s_\lambda}} + O(1/\sqrt{s_\lambda}), \tag{QoS}$$

in the $M/M/s$ queue and $M/D/s$ queue, respectively. Hence the expected waiting time vanishes at rate $1/\sqrt{s_\lambda}$.

Both limiting functions $g(\beta)$ and $1 - \mathbb{P}(M_\beta = 0)$ can take all values in $(0,1)$ by tuning the parameter β.

Since the mathematical underpinning of these properties comes from the CLT, we can expect the properties to hold for a much larger class of models. These models should then be members of the same universality class (to which the CLT applies). Let us again show this by example.

Consider a stochastic system in which demand per period is given by some random variable A, with mean μ_A and variance $\sigma_A^2 < \infty$. For systems facing large demand we propose to set the capacity according to the more general rule

$$s = \mu_A + \beta\sigma_A,$$

which consists of a minimally required part μ_A and a variability hedge $\beta\sigma_A$. Assume that the workload brought into the system is generated by n stochastically identical and independent sources. Each source i generates $A_{i,j}$ work in the j^{th} period, with $\mathbb{E}[A_{i,j}] = \mu$ and Var $A_{i,j} = \sigma^2$. Then the total amount of work arriving to the system during one period is $A_j^{(n)} = \sum_{i=1}^n A_{i,j}$ with mean $n\mu$ and variance $n\sigma^2$. Assume that the system is able to process a deterministic amount of work s_n per period and denote by $U^{(n)}(j)$ the amount of work left over at the end of period j. Then,

$$U^{(n)}(j+1) = \left(U^{(n)}(j) + A_j^{(n)} - s_n\right)^+. \tag{1.30}$$

Given that $s_n > \mathbb{E}[A_1^{(n)}] = n\mu$, the steady-state limit $U^{(n)} := \lim_{j\to\infty} U^{(n)}(j)$ exists and satisfies

$$U^{(n)} \stackrel{d}{=} \left(U^{(n)} + A_1^{(n)} - s_n\right)^+. \tag{1.31}$$

This framework is also known as the bulk service queue or the Anick-Mitra-Sondhi model [14, 113, 117]. In this scenario, increasing the system size is done by increasing n, the number of input flows. As we have seen before, it requires a rescaling of the process $U^{(n)}$ by an increasing function $c(n)$, in order to obtain a non-degenerate scaling limit $U := \lim_{n\to\infty} U^{(n)}/c(n)$. (We omit the technical details needed to justify the interchange of limits.) From (1.31) it becomes clear that the scaled increment

$$\frac{A_j^{(n)} - s_n}{c(n)} = \frac{\sum_{i=1}^n A_{i,j} - n\mu}{c(n)} + \frac{n\mu - s_n}{c(n)} \tag{1.32}$$

only admits a proper limit if $c(n)$ is of the form $c(n) = O(\sqrt{n})$, by the virtue of the CLT, and $(s_n - n\mu)/c(n) \to \beta > 0$ as $n \to \infty$. Especially for $c(n) = \sigma\sqrt{n}$, this reveals that U has a non-degenerate limit, which is equal in distribution to the maximum of a Gaussian random walk with drift $-\beta$ and variance 1, if

$$s_n = n\mu + \beta\sqrt{n}\sigma + o(\sqrt{n}).$$

Moreover, the results on the Gaussian random walk presented in Section 1.2.2 are applicable to this model and the key features of the QED scaling carry over to this more general setting as well. In conclusion, the many-sources framework shows that the QED scaling finds much wider applications than queueing models with Poisson input only.

1.2.4 Related literature

We now provide a partial overview on the literature on heavy-traffic analysis in queueing theory and the QED regime in particular.

Conventional heavy-traffic. Before the formal introduction of the Halfin-Whitt scaling regime in 1981, see [101], the existing literature on the asymptotic analysis of queues mostly evolved around two types of scaling regimes: single-server and infinite-server regimes.

The idea of studying a sequence of queues in which the utilization approaches 100%, i.e. heavy-traffic, was first laid out by Kingman in the 1960s. In [140, 141] he showed how in the $GI/G/1$ queue, under mild conditions on the arrival and service processes, the scaled steady-state waiting time $(1-\rho)W^{(1)}$ converges to an exponentially distributed random variable. The notion that heavily loaded systems admit a scaling limit that is remarkably simple compared to the otherwise intractable pre-limit queueing systems triggered a surge of research within the field of queueing theory in the 1960s and 1970s, see [42, 111, 52, 171, 145, 146, 217] among others. These works conduct their asymptotic analysis in what we now call conventional heavy-traffic. That is, the service times and number of servers are held fixed, while the arrival rate approaches the critical value from below. A noteworthy result of these efforts is the extension of Kingman's findings to the $GI/G/s$, which finds that the scaled queue length $(1-\rho)Q^{(s)}$ converges in distribution to an exponential random variable with mean $(c_a^2+c_s^2)/2$, where c_a and c_s denote the coefficient of variation of the interarrival and service time distribution, respectively. We remark that this limiting result is the key ingredient to the widely applied Kingman formula

$$\mathbb{E}[W^{(1)}] \approx \frac{\rho}{1-\rho} \cdot \frac{c_a^2+c_s^2}{2} \cdot \mathbb{E}[B], \tag{1.33}$$

which serves as an approximation to the expected waiting time in the single-server queue. The limit (1.33) reveals that in the conventional heavy-traffic regime, the expected waiting time explodes as $\rho \to 1$. Hence, efficient usage of resources is achieved, at the expense of poor QoS.

An alternative regime that received much attention, see e.g. [108, 42, 109, 110, 218], fixes the service time distribution while increasing both the arrival rate λ and the number of servers to infinity simultaneously, such that the ratio λ/s remains constant. It has been shown that the sequence of queues under this scaling start resembling the behavior of infinite-server queues as λ and s grow. That is, the probability of a customer finding a queue on arrival is negligible. The sample paths in Figure 1.3 are illustrative for this regime. Since the utilization level ρ remains strictly away from one in the limit, this setting is in the literature typically not recognized as heavy-traffic.

As Halfin & Whitt indicate themselves, the QED regime in which service times are held fixed, and λ and s tend to infinite while satisfying $(1-\rho)\sqrt{s} \to \beta$, is a hybrid between the two aforementioned regimes. Namely, it considers the efficiency property of the conventional heavy-traffic scaling, and the good QoS levels from infinite-server queues.

The $G/G/s$ queue in the QED regime. We have demonstrated in Section 1.2 how to obtain QED limits for the $M/M/s$ queue and the $M/D/s$ queue. When one moves beyond the exponential and deterministic assumptions, establishing QED behavior becomes mathematically more challenging.

The heavy-traffic analysis of the $G/G/s$ queue requires fundamentally different approaches than for Markovian queues. Most of the research conducted on the $G/G/s$ in the Halfin-Whitt regime evolves around the characterization of the stochastic process limit of the appropriately centered and scaled queueing process in terms of diffusion processes, under various assumptions on the model primitives. Puhalskii & Reiman [181] analyzed the multi-class queue with phase-type service times in the Halfin-Whitt regime. Heavy-traffic limits for queues in which service time distributions are lattice-based and/or have finite support are studied by Mandelbaum & Momčilović [155] and Gamarnik & Momčilović [78]. Approaches through measure-valued processes are taken by Kang, Kaspi & Ramanan [129, 128, 130]. The most general class of distributions is considered by Reed [184] and Puhalskii & Reed [180], who impose no assumption on the service time distribution except for the existence of the first moment. For a survey on the techniques required for the analysis of process limits of $G/G/s$ queues, we refer the reader to [176] and references therein.

Considerably less is known for the corresponding steady-state distribution of the $G/G/s$ queue in the QED regime. Namely, under the assumption of general service time distributions, truly infinite-dimensional limits arise, since the Markovian nature of the service time and 'age' process can no longer be exploited. Works that have been able to characterize limiting behavior for the specific service time distribution classes include Jelenkovic et al. [122], who assume deterministic service times, and Whitt [222], who identifies the heavy-traffic limit in the case of hyper-exponentially distributed service times. Progress in the understanding of steady-state behavior of $G/G/s$ queues in the Halfin-Whitt regime has been facilitated by Gamarnik & Goldberg [86, 77], who perform their analysis under the mild assumption that the service time distribution has finite $(2+\varepsilon)$ moment. A significant advance has been made by Aghanjani & Ramanan [9], who identify the limit as the steady-state distribution of infinite-dimensional Markov process, given that the service time distribution has finite $(3+\varepsilon)$ moment, while drawing upon previous results by Kang, Kaspi & Ramanan [129, 128, 130].

Model extensions. Many extensions to the standard many-server queue can be considered. A feature ubiquitous to service systems involving humans is customer abandonment [80, 49, 229, 159]. The $M/M/s+M$ queue introduced by Palm [175], also known as the Erlang-A model [82, 210], acknowledges this feature by assigning every customer an exponentially distributed *patience time* upon his arrival (denoted by $+M$ in the model definition). If a customer has not yet started receiving service by the expiration of his patience, he leaves the system. Note that abandonments render queues stable under any load. Under QED scaling, the more general $G/G/s+G$ queue has received much attention under various modeling assumptions, see e.g. [82, 80, 223, 158, 229, 156, 128, 64, 185, 126, 231]. Noteworthy findings include the vanishing abandonment probability [82] and insensitivity of the patience time distribution as long as its density at 0, i.e. the probability of abandoning immediately upon arrival, is fixed, as the system grows large under QED scaling.

Overviews of queues with abandonment and their asymptotic counterpart are given by Zeltyn & Mandelbaum [229] and Dai & He [65] and Ward [214].

Other features that have been studied in the QED regime include multiple customer classes, see e.g. [105, 23, 97, 100, 206], or heterogeneous servers [16, 19, 157, 202]. These models are all interesting in their own respect and are fairly well-understood. Therefore, we choose to focus in this thesis on a different set of extensions, which will be discussed in Section 1.4.

1.3 Dimensioning

We adopt the term *dimensioning* used by Borst, Mandelbaum & Reiman [43] to say that the capacity of a service system is adapted to the load in order to reach certain performance levels. In [43] dimensioning refers to the staffing problem in a large-scale call center and key ingredients are the square-root staffing rule in (1.5) and the QED regime. We now revisit the results in [43] and its follow-up works to explain this connection to the QED regime.

1.3.1 Constraint satisfaction

Consider the $M/M/s$ queue with arrival rate λ and service rate μ. A classical dimensioning problem is to determine the minimum number of servers s necessary to achieve a certain target level of service, say in terms of waiting time.

Suppose we want to determine the minimum number of servers such that the fraction of customers who are delayed in the queue is at most $\varepsilon \in (0,1)$. Hence we should find

$$s_\lambda^*(\varepsilon) := \min\left\{s \geq \lambda \mid \mathbb{P}(W^{(s)} > 0) \leq \varepsilon\right\}. \tag{1.34}$$

But alternatively, we can use the QED framework, which says that under (1.9), $\lim_{s\to\infty} \mathbb{P}(W^{(s_\lambda)} > 0) = g(\beta)$ (see Proposition 1.1). Then by (1.34), $s_\lambda^*(\varepsilon)$ can be replaced by

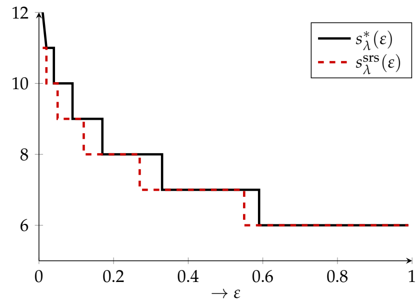
$$s_\lambda^{\mathrm{srs}}(\varepsilon) = \lceil \lambda + \beta^*(\varepsilon)\sqrt{\lambda}\rceil, \tag{1.35}$$

where $\beta^*(\varepsilon)$ solves

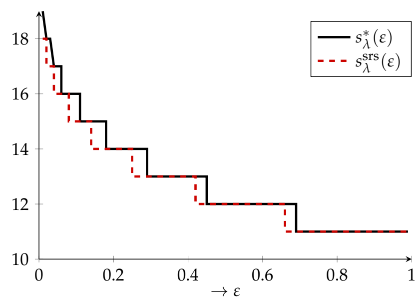
$$g(\beta^*) = \varepsilon. \tag{1.36}$$

In Figure 1.7 we plot the exact staffing level $s_\lambda^*(\varepsilon)$ and the heuristically obtained staffing level $s_\lambda^{\mathrm{srs}}(\varepsilon)$ as functions of ε for several loads λ.

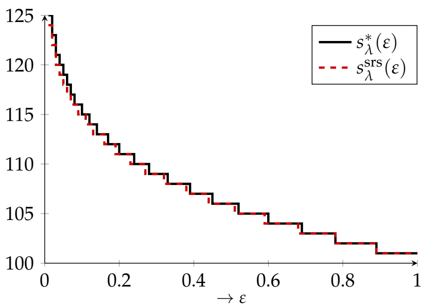
Observe that even for very small values of λ, the staffing function $s^{\mathrm{srs}}(\varepsilon)$ coincides with the exact solution for almost all $\varepsilon \in (0,1)$ and differs no more than by one server for all ε. Borst et al. [43] recognized this in their numerical experiments too, and Janssen, van Leeuwaarden & Zwart [120] later confirmed this theoretically. One can easily formulate other constraint satisfaction problems and reformulate them in the QED regime. For instance, constraints on the mean waiting time or



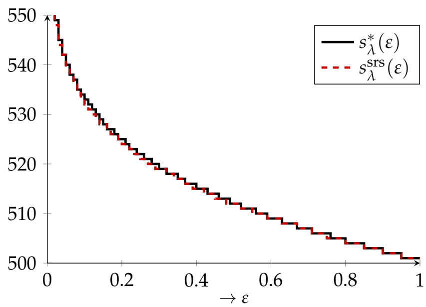
(a) $\lambda = 5$



(b) $\lambda = 10$



(c) $\lambda = 100$



(d) $\lambda = 500$

Figure 1.7: Staffing levels as a function of the delay probability targets ε.

the tail probability of the waiting time, e.g. $\mathbb{P}(W^{(s)} > T)$, which are asymptotically approximated by $h(\beta)/\sqrt{\lambda}$ and $g(\beta)\mathrm{e}^{-\beta\sqrt{\lambda}T}$, respectively. See [43] for more examples.

1.3.2 Optimization

One can also consider optimization problems, for instance to strike the right balance between the costs for servers and costs incurred by customer dissatisfaction. More specifically, assume a salary cost of a per server per unit time, and a penalty cost of q per waiting customer per unit time, yielding the total cost function

$$\bar{C}_\lambda(s) := a\, s + q\, \lambda \mathbb{E}[W^{(s)}]$$

and then ask for the staffing level s that minimizes $\bar{C}_\lambda(s)$. Since $s > \lambda$, we have $\bar{C}_\lambda(s) > a\,\lambda$ for all feasible solutions s. Moreover, the minimizing value of $\bar{C}_\lambda$ is invariant with respect to scalar multiplication of the objective function. Hence we have to optimize

$$C_\lambda(s) = r\,(s-\lambda) + \lambda \mathbb{E}[W^{(s)}], \qquad r = a/q. \tag{1.37}$$

Denote by $s^*_\lambda(r) := \arg\min_{s>\lambda} C_\lambda(s)$ the true optimal staffing level. With $s_\lambda = \lambda + \beta\sqrt{\lambda}$ and the QED limit in (1.14), we can replace (1.37) by its asymptotic counterpart:

$$\frac{C_\lambda(s_\lambda)}{\sqrt{\lambda}} = r\,\beta + \sqrt{\lambda}\mathbb{E}[W^{(s)}] \to r\,\beta + \frac{g(\beta)}{\beta} =: \hat{C}(\beta), \qquad \lambda \to \infty.$$

Once again we obtain a limiting objective function that is easier to work with than its exact pre-limit counterpart. Hence, in the spirit of the asymptotic staffing procedure in the previous subsection, we propose the following method to determine the staffing level that minimizes overall costs. First, (numerically) compute the value $\beta^*(r) = \arg\min_{\beta>0} \hat{C}(\beta)$, which is well-defined, because the function $\hat{C}(\beta)$ is strictly convex for $\beta > 0$. Then, set $s^{\mathrm{srs}}_\lambda(r) = [\lambda + \beta^*(r)\sqrt{\lambda}]$. In Figure 1.8 we compare the outcomes of this asymptotic staffing procedure against the true optima as a function of $r \in (0,\infty)$, for several values of λ. The staffing levels $s^{\mathrm{srs}}_\lambda(r)$ and $s^*_\lambda(r)$ are aligned for almost all r, and differ no more than one server for all instances.

1.3.3 Time-varying dimensioning

So far we have only considered queues in which the model primitives are constant over time. In practice, though, the arrival rate can fluctuate and depends on the time of day, the day of the week, season or even larger time scales. It is therefore more realistic to describe these mostly predictable fluctuations through $\lambda(t)$, which represents the instantaneous arrival rate of the arrival process at time $t \in \mathbb{R}$. The existence of time-varying demand requires a re-evaluation of staffing levels throughout the planning horizon as well. That is, the number of servers $s(t)$ becomes a

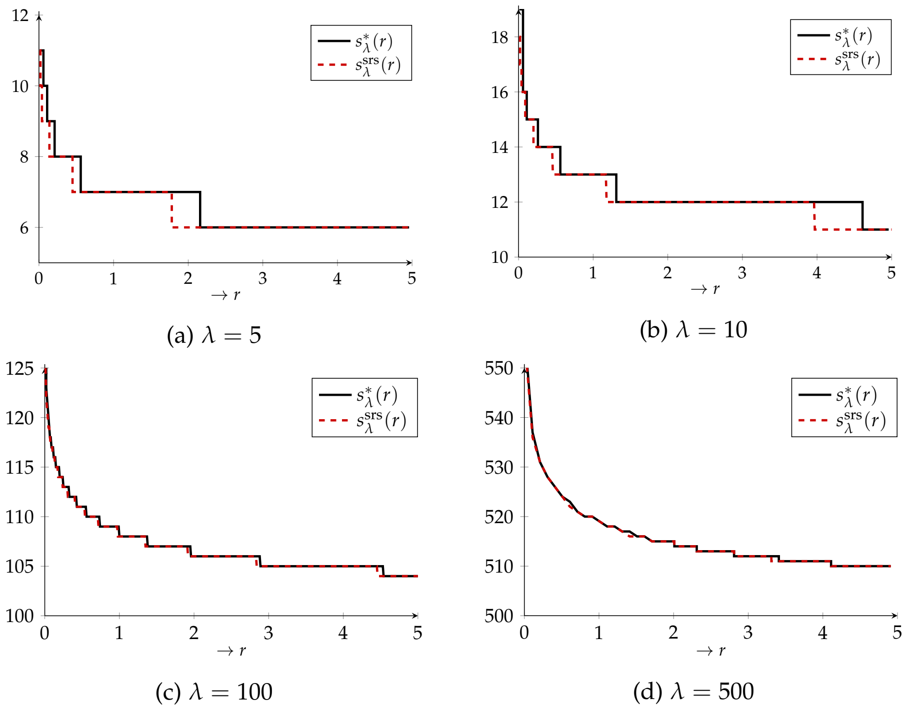


Figure 1.8: Optimal staffing levels as a function of $r = a/q$.

function of time, rather than a constant and this clearly asks for an adaptation of the dimensioning procedures in Sections 1.3.1 and 1.3.2.

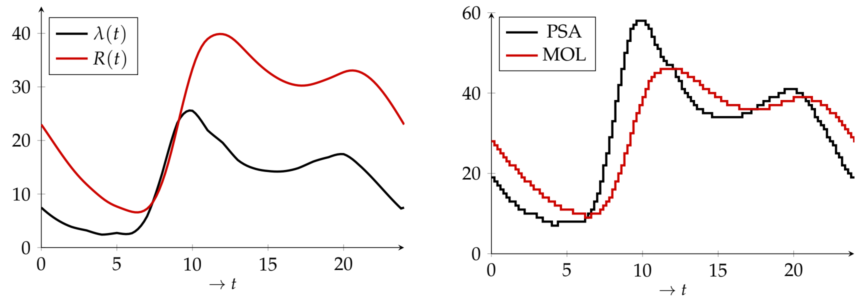
We explain the concept of time-varying staffing and the connection with the QED regime through the time-varying extension of the $M/M/s$ queue known as the $M_t/M/s_t$ queue, where the subscript t refers to the time-varying nature of both the arrival process and the staffing level. In this setting, customers arrive according to a non-homogeneous Poisson process with rate function $\lambda(t)$ and customers have exponentially distributed service times with mean $1/\mu$. Under a constraint satisfaction strategy, we aim to find the staffing function $s(t)$ such that the delay probability is at most $\varepsilon \in (0,1)$ for all t. The analysis and optimization of time-varying many-server queueing systems is known to be intrinsically hard, but many approximation techniques and heuristic methods have been proposed throughout the years [91, 125].

A natural but naive approach is the *pointwise-stationary approximation* (PSA) [91], which evaluates the system at time t as if it were in steady-state with instantaneous parameters $\lambda = \lambda(t)$, μ and $s = s(t)$. Consequently, the analysis and optimization of queues is performed on steady-state performance metrics. Variants of the PSA method include the *simple-stationary approximation* (SSA) [92], which uses the long-term (moving) average arrival rate instead of the instantaneous arrival rate, and the *stationary-independent-period-by-period approximation* (SIPP) [92], which splits the time-horizon into multiple intervals and performs steady-state analysis with the averaged parameters in each of these intervals, among others. PSA performs well in slowly varying environments with relatively short service times [91, 219]. However, when the model parameters fluctuate significantly, as is often the case in real-life systems, the accuracy of PSA can be poor, as we will see in the numerical experiment at the end of this section.

The main reason why PSA, SSA and SIPP can fail is that these methods neglect that customers are actually residing in the system (being in service or waiting in the queue) for some time. In contrast, staffing decisions should be based on the number of customers present in the system rather than the arrival rate at that particular time. Jennings et al. [125] introduced a more sophisticated method that exploits the relation with infinite-server queues. We explain their idea in the context of the $M_t/M/s_t$ queue. By Eick et al. [71], the number of customers in the $M_t/M/\infty$ queue at time t is Poisson distributed with mean

$$R(t) = \mathbb{E}\left[\lambda(t - B_e)\right]\mathbb{E}[B] = \int_0^\infty \lambda(t-u)\,\mathbb{P}(B > u)\,\mathrm{d}u = \int_0^\infty \lambda(t-u)\,\mathrm{e}^{-\mu u}\,\mathrm{d}u. \tag{1.38}$$

We remark that this result holds for more general service time distributions. Now, recall that in large systems in the QED regime, the expected delay is negligible. Therefore, under these conditions, the many-server system may be approximated by the infinite-server approximation with offered load as in (1.38). Accordingly, we can determine the staffing levels $s(t)$ for each t based on steady-state $M/M/s$ measures with offered load $R = R(t)$. Jennings et al. [125] proceed by exploiting the heavy-traffic results of Halfin-Whitt (1.14). In conjunction with the dimensioning



(a) Arrival rate and offered load functions

(b) Staffing functions.

Figure 1.9: Time-varying parameters of a real-world emergency department.

scheme in Section 1.3.1, the authors propose to set

$$s(t) = \left\lceil R(t) + \beta^*(\varepsilon)\sqrt{R(t)} \right\rceil, \tag{1.39}$$

where $\beta^*(\varepsilon)$ solves $g(\beta^*(\varepsilon)) = \varepsilon$. Remark that the number of servers is rounded up to ensure that the achieved delay probability is indeed below ε. This method was called in [125, 161] the *modified-offered-load* (MOL) approximation, and we adopt this term in this thesis.

Let us demonstrate that this approximation scheme works. Figure 1.9(a) shows an arrival rate pattern $\lambda(t)$ and corresponding offered load function $R(t)$ for $\mu = 1/2$. This arrival rate stems from a real-world emergency department [194]. The resulting staffing level functions based on the PSA and MOL approximations with $\varepsilon = 0.3$ are plotted in Figure 1.9(b).

Through simulation, we evaluate the delay probability as a function of time for $\varepsilon = 0.1$, 0.3 and 0.5. In Figure 1.10 we see how the PSA approach fails to stabilize the performance of the queue, whereas the MOL method does stabilize around the target performance. The erratic nature of the delay probability as a function of time can be explained by rounding effects of the staffing level. Since this rather simple but elegant technique to address time-varying dimensioning is provably effective, we will adopt the underlying idea of the MOL method in various different settings in this thesis.

1.4 Contributions

We have explained how the QED regime can be used to dimension and staff large-scale service systems. The basic concepts, however, were explained for the relatively simple $M/M/s$ and $M_t/M/s_t$ queue. Many real-world service systems have essential features that are not captured by these elementary models. We will now discuss some of these features and address the need to consider more involved models and extend the existing QED theory.

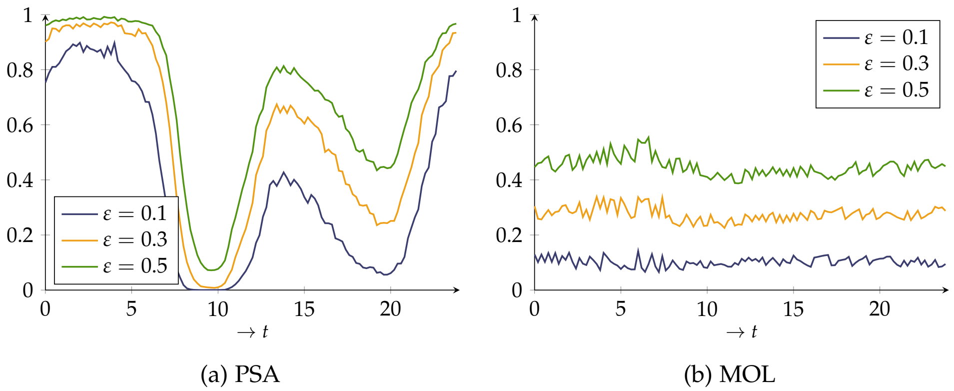


Figure 1.10: Probability of delay under staffing functions obtained through PSA and MOL approximations.

1.4.1 Non-classical scaling regimes and pre-limit behavior

The QED theory is centered around the scaling relation $\sqrt{\lambda}(1-\rho_\lambda) \to \beta$, or equivalently $s_\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda})$, for $\lambda \to \infty$.

It is worthwhile to study how pre-limit behavior of many-server queues is affected when one deviates from the square-root staffing rule. Consider a novel family of heavy-traffic scaling regimes, described in terms of the parameter η for which we assume that

$$\lambda^\eta(1-\rho_\lambda) \to \beta, \qquad \text{as } \lambda \to \infty,\ \beta > 0. \tag{1.40}$$

The parameter $\eta \geq 0$ defines a whole range of possible scaling regimes, including the classic case $\eta = 1/2$, as well as the cases $\eta = 0$ and $\eta = 1$ investigated in Subsection 1.1.3. In terms of a capacity sizing rule, the condition (1.40) is tantamount to $s_\lambda = \lambda + \beta\lambda^{1-\eta}$. This framework thus bridges the gap between the QD and QED regime if $\eta \in (0, 1/2)$ and the QED and ED regime if $\eta \in (1/2, 1)$, in the $M/M/s$ model. Similar capacity sizing rules have been considered in [29, 150] for many-server systems with uncertain arrival rates. Hence, for $\eta \in (0, 1/2)$ the variability hedge is relatively large, so that the regime parameterized by $\eta \in (0, 1/2)$ can be seen as *moderate* heavy traffic: heavy-traffic conditions in which the full occupancy is reached more slowly, as a function of λ, than for classical heavy traffic. See [56, 182, 179, 21, 23, 24, 22] for more details. For opposite reasons the range $\eta \in (1/2, \infty)$ corresponds to *extreme* heavy traffic due to a relatively small variability hedge.

We use the insights of Section 1.2 and the connection of the QED scaling to the CLT to argue intuitively that the following trichotomy in the qualitative system behavior as $\lambda \to \infty$ holds under scaling (1.40). For $\eta \in (0, 1/2)$ the empty-system probability converges to 1, because the order of the variability hedge $\beta\lambda^{1-\eta}$ is greater than strictly necessary to accommodate the stochastic fluctuations in demand. Scalings in which $\eta \in (1/2, \infty)$, have adverse behavior, since stochastic fluctuations are not accounted for sufficiently, so that the probability of delay converges to 1. The value $\eta = 1/2$ is therefore the tipping point, at which the delay probability converges to a limit between 0 and 1. Above and below this critical value, the asymptotic performance of the queue flips to either one of the extremes.

In Chapter 2, we formalize this heuristic argument and conduct an asymptotic analysis to reveal the rate at which the limit of performance metrics is attained, depending on the parameters η and β and the system size λ, s_λ.

1.4.2 Overdispersed arrivals

Until now we have considered queueing systems with perfect knowledge on the model primitives, including the mean demand per time period. For large-scale service systems, the dominant assumption in the literature is that demand arrives according to a non-homogeneous Poisson process, which in practice translates to the assumption that arrival rates are known for each basic time period (second, hour or day).

Although natural and convenient from a mathematical viewpoint, the Poisson assumption often fails to be confirmed in practice. A deterministic arrival rate implies that the demand over any given period is a Poisson random variable, whose variance equals its expectation. A growing number of empirical studies of service systems shows that the variance of demand typically exceeds the mean significantly, see [26, 29, 30, 49, 58, 80, 99, 127, 138, 150, 165, 187, 200, 228]. The feature that variability is higher than one expects from the Poisson assumption is referred to as *overdispersion*.

Due to its inherent connection with the CLT, the dimensioning rule in (1.5) relies heavily on the premise that the variance of the number of customers entering the system over a period of time is of the same order as the mean. Subsequently, when stochastic models do not take into account overdispersion, resulting performance estimates are likely to be overoptimistic. The system then ends up being under-provisioned, which possibly causes severe performance problems, particularly in critical loading.

To deal with overdispersion, existing capacity sizing rules like the square-root staffing rule need to be modified in order to incorporate a correct hedge against (increased) variability. Following our findings in Section 1.2.3, we propose a capacity allocation rule similar to (1.5) in which the original variability hedge is replaced by an amount that is proportional to the square-root of the variance of the arrival process.

In Chapter 3, we elaborate on this idea and show how to adapt the scaling of the queueing process appropriately to achieve QED-type behavior in the presence of overdispersion.

1.4.3 Finite-size constraints

The canonical examples in Section 1.2 assume an infinite amount of waiting space. Physical service systems, however, are sometimes limited in the number of customers that can be held in the system simultaneously. For instance in a call center, the maximum number of clients in service or queueing is restricted by the number of available trunk lines [135], while in the emergency department of a hospital, the number of beds constrains the number of patients that can be admitted [225]. Depending on the practical setting and admission policy, if the maximum capacity, say n, is reached, newly arriving customers either leave the system immediately (blocking), reattempt getting access later (retrials) or queue outside the facility (holding). In any case, expectations are that the queueing dynamics within the service facility are affected considerably in the presence of such additional capacity constraints.

We illustrate these implications through the $M/M/s/n$ queue, that is, the standard $M/M/s$ queue with additional property that a customer who finds upon arrival n customers already present in the system, is deferred and considered lost. To avoid trivialities, let $n \geq s$. Since the expected workload reaching the servers is less than in the unconstrained scenario, one expects less congestion and resource utilization.

Consider the $M/M/s_\lambda/n_\lambda$ in the QED regime. So, let λ increase while s_λ scales as $s_\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda})$. We then ask how n_λ should scale along with λ and s_λ to maintain the non-degenerate behavior as seen in Section 1.2. We provide a heuristic answer. Let $Q^{(s_\lambda,n_\lambda)}$ and $W^{(s_\lambda,n_\lambda)}$ denote the number of customers in the system and the waiting time in the $M/M/s_\lambda/n_\lambda$ queue in steady state. Note through Proposition 1.2 that if there were no finite-size constraints, we would have, for λ large,

$$\mathbb{P}(Q^{(s_\lambda)} \geq n_\lambda) = \mathbb{P}\left(\frac{Q^{(s_\lambda)} - s_\lambda}{\sqrt{s_\lambda}} \geq \frac{n_\lambda - s_\lambda}{\sqrt{s_\lambda}}\right)$$
$$\to \begin{cases} g(\beta), & \text{if } n_\lambda = s_\lambda + o(s_\lambda), \\ g(\beta)\,\mathrm{e}^{-\beta\gamma}, & \text{if } n_\lambda = s_\lambda + \gamma\sqrt{s_\lambda} + o(\sqrt{s_\lambda}), \\ 0, & \text{if } n_\lambda = s_\lambda + \Omega(\sqrt{s_\lambda}), \end{cases} \tag{1.41}$$

as $\lambda \to \infty$ for some $\gamma > 0$. Here, the relation $u(\lambda) = \Omega(v(\lambda))$ implies $u(\lambda)/v(\lambda) > 1$ for $\lambda \to \infty$. Hence, asymptotically the finite-size effects only play a role if the extra variability hedge of n_λ is of order $\sqrt{s_\lambda}$ (or equivalently $o(\sqrt{\lambda})$). Furthermore, if the variability hedge is $o(\sqrt{\lambda})$, then we argue that asymptotically, all customers who do enter the system have probability of delay equal to zero. More formally, under the *two-fold scaling rule*

$$\begin{cases} s_\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda}), \\ n_\lambda = s_\lambda + \gamma\sqrt{s_\lambda} + o(\sqrt{\lambda}), \end{cases} \tag{1.42}$$

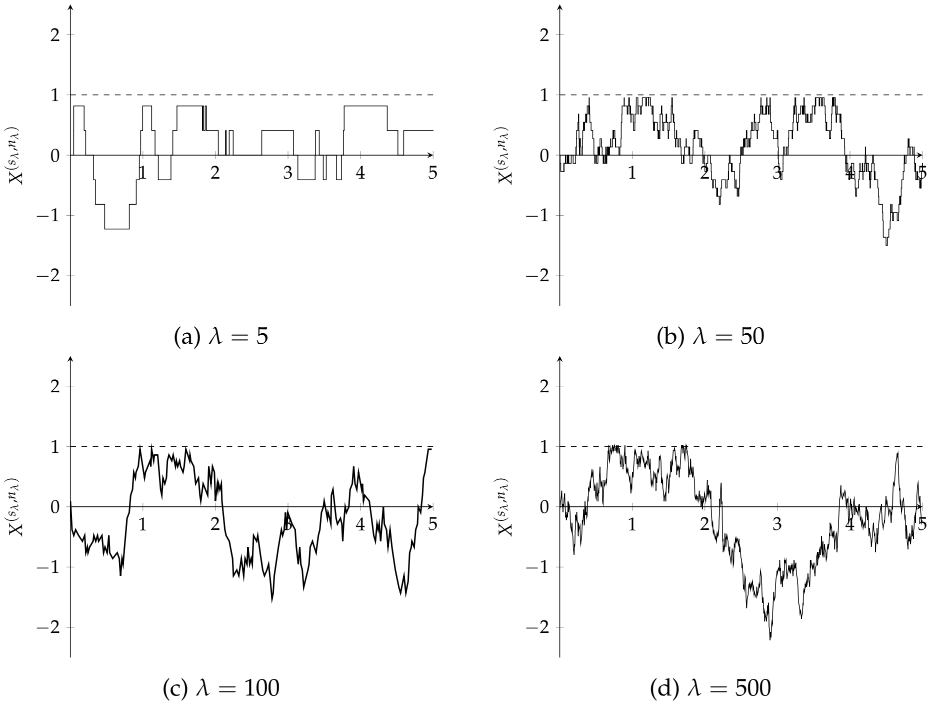


Figure 1.11: Sample paths of the normalized queue length process $X^{(s_\lambda,n_\lambda)}(t)$ with $\lambda = 5$, 50, 100 and 500 under scaling (1.42) with $\beta = 0.5$ and $\gamma = 1$.

it is not difficult to deduce that, see e.g. [160],

$$\mathbb{P}(W^{(s_\lambda,n_\lambda)} > 0) \to \left(1 + \frac{\beta\,\Phi(\beta)}{(1-\mathrm{e}^{-\beta\gamma})\varphi(\beta)}\right)^{-1}, \quad \text{as } \lambda \to \infty, \tag{1.43}$$

which is strictly smaller than $g(\beta)$ in (1.4), but still bounded away from both 0 and 1. Furthermore, the buffer size of the queue is $n_\lambda - s_\lambda = \gamma\sqrt{s_\lambda}$, so that by Little's law, the expected waiting time of an admitted customer is $O(1/\sqrt{s_\lambda})$. Even though resource utilization in the $M/M/s_\lambda/n_\lambda$ is less efficient than in the queue with unlimited waiting space, it can easily be shown that $\rho \to 1$ as $\lambda \to \infty$. Hence, all three key characteristics of the QED regime are carried over to the finite-size setting if adhered to scaling (1.42).

On a process level, adding a capacity constraint translates to adding a reflection barrier to the normalized queue length process $X^{(s_\lambda,n_\lambda)} = (Q^{(s_\lambda,n_\lambda)} - s_\lambda)/\sqrt{s_\lambda}$, at γ, as is illustrated by the sample paths of $X^{(s_\lambda,n_\lambda)}$ for three values of λ in Figure 1.11.

It has been shown by [160] that under (1.42)

$$\sqrt{s_\lambda}\,\mathbb{P}(\text{block}) = \sqrt{s_\lambda}\,\mathbb{P}(Q^{(s_\lambda,n_\lambda)} = n_\lambda) \to f(\beta,\gamma), \quad \text{as } \lambda \to \infty, \tag{1.44}$$

for a non-negative function f.

The idea of the two-fold scaling in (1.42) can be extended to settings in which the interior is in fact a network of queues, rather than the single-station setting discussed here, see [135, 225, 205] for examples of such *semi-open* queueing networks.

When customers retry getting access after being blocked initially, the QED analysis becomes much more difficult, and no explicit limiting results are known. Nevertheless, observe that the volume of blocked arrivals is by (1.44) of order $\sqrt{\lambda}$, the exact same magnitude as the variability hedge of both s_λ and n_λ. Therefore, retrials and holding customers have a non-negligible effect on the service levels within the facility in the QED regime. This will be the topic of Chapters 4 and 5.

1.4.4 Pre-limit behavior

The results on queues in the QED regime discussed in Section 1.2 are in two ways of an asymptotic nature. First, the heavy-traffic limits prescribe the queueing dynamics for $\lambda, s_\lambda \to \infty$. Real-world systems obviously do not experience infinite demand nor have infinite capacity, and hence the heavy-traffic limits only form an approximation for such finite-sized systems. Although these approximations are qualitatively insightful, the asymptotic analyses do not reveal much about their accuracy with respect to actual performance. For instance, we would like to know how fast the convergence takes place, and how inaccuracies in asymptotic approximations percolate into inaccuracies in pre-limit systems. To answer such questions, it would be helpful to have an asymptotic estimate for the difference between the (scaled) queueing process and its limiting counterpart, to be able to judge the error made by relying on asymptotic as opposed to actual performance evaluation. Characterization of the error term gives rise to so-called *corrected diffusion approximations* [191, 38, 117], which are refinements to heavy-traffic limits for finite systems, and are also related to *universal approximations* [98, 107, 47, 48]. We will derive such corrected diffusion approximations in the context of the novel scaling regimes mentioned in Section 1.4.1 in Chapter 2.

Second, the bulk of queueing literature is concerned with the performance analysis and optimization of steady-state systems. However, in practice, service systems certainty do not run infinitely long, which renders this assumption questionable. Validation of the steady-state assumption is related to the *relaxation time* of a queueing process [5, 6, 114, 209, 210, 76], which prescribes the time it takes a system starting out of equilibrium to approximate its stationary distribution. In case the relaxation time is small, stationary performance evaluation is likely to be accurate. On the contrary, if the relaxation time is large, a time-dependent analysis of the queueing system is required in order to capture realistic behavior. Subsequently, we can investigate the implications of applying staffing principles that are based on steady-state performance metrics in settings which are inherently transient over the planning period. We will touch upon this topic in Chapter 6.

1.5 Outline of the thesis

The remainder of this thesis builds upon the ideas behind the QED scaling regime exhibited in this introductory chapter, and is organized as follows.

Chapter 2 is concerned with the analysis of the limiting behavior of queues in case one deviates from the square-root staffing principle as demand grows large. Using the bulk-service queue together with the many-sources paradigm as a vehicle, we derive corrected diffusion approximations for the performance metrics of pre-limit systems in these alternative scaling regimes. The work presented in Chapter 2 is based on [118].

In Chapter 3, we also analyze the bulk-service queueing model, but with many correlated sources, so that demand becomes overdispersed. As we alluded to in Section 1.4.2, this requires an alternative scaling of the queue length process and associated staffing rule. This chapter exhibits the ideas of [163].

In Chapter 4, we discuss how QED-type behavior prevails in simple settings in which the system size is finite, given appropriate capacity-sizing rules. More specifically, we show how customer retrials can be incorporated heuristically into the performance analysis of finite-size systems in the QED regime. The content of this chapter is based on [211] and [212].

Building upon the insights gained in Chapter 4, we show in Chapter 5 how the approximation methods carry over to a more complex finite-size queueing system, inspired by delay analysis in a health care facility. We show how the QED scaling limits for this model offer surprisingly accurate approximations for realistic model parameters in systems of small to moderate size, and develop a staffing algorithm for dimensioning such systems. Chapter 5 is based on the ideas of [213].

Chapter 6 investigates the validity of a capacity allocation rule based on steady-state performance metrics in practical settings. Namely, in realistic scenarios, the parameters of a queueing model are typically subject to change over the planning period. This asks for a more elaborate transient analysis of the queue dynamics, and an adaptation of the staffing level. In this chapter, we present how to do so appropriately in a single-server queueing model facing a Lévy input process by prescribing a correction to the steady-state optimum, which has a square-root form. This chapter is based on [164].

Chapter 7 presents the analysis of an inventory model with backlogs, perishable goods and consumer impatience. This model resembles the inventory level of a blood bank, and can be regarded as a shot-noise model with both positive and negative jumps and exponential decay rates above and below zero. Besides the derivation of the stationary distribution of the inventory level, we show how under appropriate scaling the process converges to an Ornstein-Uhlenbeck process. The latter allows for a more tractable approximate analysis of the model in case the number of blood deliveries and demand is large. Chapter 7 is based on [28].

2

Novel heavy-traffic regimes

In this chapter, we introduce a family of heavy-traffic scalings for a large-scale service system meant to serve jobs coming from a large pool of customers. The scaling rules are inspired by the classical QED regime discussed in Chapter 1, but lead to a range of different system behaviors that includes the ED, QED and QD regime as special cases. To determine the scaling limits, we describe the performance measures in terms of Pollaczek integrals and use asymptotic techniques to evaluate these integrals in the large-system limit.

Based on
Novel heavy-traffic regimes for large-scale service systems
Guido Janssen, Johan van Leeuwaarden & Britt Mathijsen
In *SIAM Journal of Applied Mathematics, 75(2), 787-812 (2015)*

2.1 Introduction & motivation

We study the workload process of a system, experiencing stochastic demand and deterministic capacity s_n per period, at equidistant time epochs. Demand is assumed to be generated by n stochastically identical and independent sources. Let $A_{i,j}$ denote the workload brought into the system by source i in period j, for which $\mathbb{E}[A_{i,j}] = \mu$ and $\operatorname{Var} A_{i,j} = \sigma^2$. Then the total amount of demand arriving to the system in period j is $A_j^{(n)} = \sum_{i=1}^{n} A_{i,j}$ with $\mathbb{E}[A_j^{(n)}] = n\mu$ and $\operatorname{Var} A_j^{(n)} = n\sigma^2$. As explained in Chapter 1, a good capacity sizing rule for achieving economies-of-scale is $s_n = n\mu + \beta\sqrt{n}\sigma$ for some $\beta > 0$. If we denote the system utilization by $\rho_n := n\mu/s_n$, then this dimensioning rule in the bulk service queue with many sources is tantamount to the heavy-traffic scaling

$$\sqrt{n}(1-\rho_n) \to \gamma = \frac{\beta\sigma}{\mu}, \qquad \text{as } n \to \infty. \tag{2.1}$$

Starting from this setting, we introduce a novel family described in terms of a parameter η for which we assume that

$$n^{\eta}(1-\rho_n) \to \gamma, \quad \text{as } n \to \infty,\ \gamma > 0. \tag{2.2}$$

The parameter $\eta \geq 0$ defines a whole range of possible scaling regimes, including the classical case $\eta = 1/2$. In terms of a capacity sizing rule for systems with many customers, the condition (2.2) is tantamount to $s_n = n\mu + \beta\sigma n^{1-\eta}$. Similar capacity sizing rules have been considered in [29, 150] for many-server systems with uncertain arrival rates. Hence, for $\eta \in (0, 1/2)$ the variability hedge is relatively large, so that the regime parameterized by $\eta \in (0, 1/2)$ can be seen as *moderate* heavy traffic: heavy-traffic conditions in which the full occupancy is reached more slowly, as a function of n, than for classical heavy traffic. For opposite reasons the range $\eta \in (1/2, \infty)$ corresponds to *extreme* heavy traffic due to a relatively small variability hedge. Note that the case $\eta = 0$ does not lead to 100% system utilization when $n \to \infty$.

In this chapter we show that economies-of-scale can be achieved for a large range of η, although the nature of the benefits obtained by operating on large scale depends on the precise capacity sizing rule (hence the parameter η). We quantify performance in terms of stationary measures: The mean and variance of the congestion in the system, and the probability of an empty system. For these performance measures we derive heavy-traffic limits under the scalings (2.2) that are relatively simple functions of only the first two moments of the demand per period. Such parsimonious expressions are useful for quantifying and improving system behavior. The heavy-traffic limits, however, provide also qualitative insight into the system behavior. Our asymptotic analysis shows that mean congestion is $O(n^{\eta})$, which implies that delays experienced by the customers are negligible for all values of $\eta \in [0, 1)$, are roughly constant for $\eta = 1$, and grow without bound for $\eta > 1$. We expect this qualitative behavior to be universal for a wide range of stochastic models to which the regime (2.2) is applied. We further show the existence of the following trichotomy as $n \to \infty$ under (2.2): For $\eta \in (0, 1/2)$ the empty-system probability converges to 1, for $\eta \in (1/2, 1)$ it converges to 0, while only for $\eta = 1/2$ there is a limiting value in $(0, 1)$. Hence, as expected, the system performance deteriorates with η, in a rather crude way for the empty-system probability, and in only a mild way for mean congestion levels. The regime (2.2) thus presents a range of possible capacity sizing rules that all lead to economies-of-scale, and depending on what is the desired nature of performance for a particular service system, an appropriate η can be selected. From the quantitative perspective, our detailed asymptotic analysis leads to more precise asymptotic estimates for the performance measures in heavy traffic, which reveal the exact manner in which congestion is influenced by η and γ.

Motivating examples. The bulk service queueing model is one of the canonical models in queueing theory, having a wide range of applications in fields like digital communication, wireless networks, road traffic, reservation systems, health care and many more (see [53] and [207, Chap. 2] for an overview). In road traffic, the basic model for congestion at an intersection, known as the fixed-cycle traffic-light queue [170, 208], is related to our discrete bulk service queue. Then s_n represents the maximum number of delayed cars in front of a traffic light that can depart during one green period, while $A_j^{(n)}$ is the number of newly arriving cars during a consecutive green and red period.

An example from health care is panel sizing [227]. Say a general practitioner has a pool of n clients (typically in the order of 2,500 [93]), all of which are potential patients, and together require $A_j^{(n)}$ consults per day. Further assume that the practitioner can see a maximum number of s_n patients per day. What is then an appropriate patient panel size n, which strikes a reasonable balance between accessing medical care in a timely manner and restricting the time that the practitioner sits idle? The panel size application is one of many examples of an appointment book, referring to some schedule of appointments for a fixed period, with capacity s_n appointments per period and newly arriving number of appointments $A_j^{(n)}$ per period. See [66] for another recent example of an appointment book in a health care setting, again in terms of our bulk service queue, with $A_j^{(n)}$ the new patients per day and s_n the number of available beds.

For all examples above, and many more, our new class of heavy-traffic scalings (2.2) presents capacity sizing rules for which the expected performance can be quantified using the results in this chapter. This will be helpful in dimensioning the systems (How much capacity is needed to achieve a certain target performance?) while exploiting economies-of-scale. For appointment books, our model together with the capacity sizing rules (2.2) is particularly relevant for *advanced access* [93], a scheduling approach in health care designed to reduce delays by offering every patient a same-day appointment, regardless of the urgency of the problem. In that way, patients do not have to wait long for appointments, and practices do not waste capacity by holding appointments in anticipation of urgent cases.

Pollaczek's formula. Next to the freedom to model different situations, another advantage of our model is that it is mathematically tractable, in the sense that it can be subjected to powerful mathematical methods from complex and asymptotic analysis. In order to establish the heavy-traffic limits we start from Pollaczek's formula for the transform of the stationary queue length distribution in terms of a contour integral. From this famous transform representation, contour integrals for the empty-system probability and the mean and variance of the congestion immediately follow. Contour integrals are often amenable to asymptotic evaluation (see e.g. [61]), particularly for obtaining classical heavy-traffic asymptotics. We also subject the contour integral representations to asymptotic evaluation, but not under classical heavy-traffic scaling. This asymptotic analysis requires a *non-standard* saddle point method, tailored to the specific form of the integral expressions that arise under the capacity sizing rule (2.2).

Saddle point method. In complex analysis, the saddle point method in its standard form is a useful technique to estimate the asymptotic behavior of integrals of the form

$$I(n) = \int_C h(z)\,\mathrm{e}^{nf(z)}\,\mathrm{d}z, \tag{2.3}$$

as $n \to \infty$, where C is a contour in the complex plane, and $f(z)$ and $h(z)$ are functions that are analytic in some neighborhood of C. The main idea behind the saddle point method is that if the integrand in (2.3) exhibits a sharp peak along the contour, then one may naturally expect that a small neighborhood around this peak provides the dominant contribution to the integral. More specifically, for large values of n, the function f and its associated maximum $f(z^*)$ for $z^* \in C$ to a large extent determine the magnitude of the integrand (where z^* is well-defined due to analyticity of f. In the setting of this chapter, C is a closed curve, which implies that the value z^* must be a *saddle point* of f, i.e. $f'(z^*) = 0$. Subsequently, one can replace $f(z)$ in (2.3) by its Taylor expansion around z^* and deduce through the Laplace method, see e.g. [67], that

$$I(n) = \sqrt{2\pi}\,i\,\frac{h(z^*)\,\mathrm{e}^{nf(z^*)}}{\sqrt{n\,|f''(z^*)|}}\Big(1 + O(1/n)\Big),$$

as $n \to \infty$. In Section 2.3, we show how the contour integrals describing stationary measures for the queue length, derived through Pollaczek's formula, can be reformulated into the shape of (2.3). However, we will show that the saddle point method in its standard from cannot be applied to asymptotically characterize other stationary measures like the mean or mass at zero. Indeed, for our model the saddle point (the solution of (2.21)) converges to one (as $n \to \infty$), which is a singular point of the integrand, and renders the standard saddle point method useless. The non-standard saddle point method discussed in this chapter, originally proposed by [67], is made specifically to overcome this complication. This leads to asymptotic expansions for the performance measures, of which the limiting forms correspond to the heavy-traffic limits, and pre-limit forms present refined approximations for pre-limit systems ($n < \infty$) in heavy traffic. Such refinements to heavy-traffic limits are commonly referred to as *corrected diffusion approximations* [191, 38, 20].

Further connections to the literature. We now discuss two classes of stochastic systems for which the heavy-traffic regime (2.1) has been studied extensively, and for which our new family of regimes (2.2) is largely unexplored. We discuss these classes because, despite the fact that the Pollaczek formula does not hold, we believe the qualitative results that we reveal for our particular model should to a large extent carry over to these settings as well, presenting some interesting avenues for further research (see Section 2.6.2).

The first class concerns so-called *nearly-deterministic* systems [192, 193], denoted by $G_n/G_n/1$ system, where G_n stands for *cyclic thinning* of order n, indicating that some point process is thinned to contain only every nth point. As $n \to \infty$, the $G_n/G_n/1$ systems approach the deterministic $D/D/1$ system. For $G_n/G_n/1$ systems, [192] establishes stochastic-process limits, and [193] derives heavy-traffic limits for stationary waiting times. In the framework of [192, 193], our stochastic model corresponds to a $D/G_n/1$ queue, where the sequence of service times $\{A_j^{(n)}\}_{j\geq 1}$ follows from a cyclically thinned sequence of i.i.d. random variables $A_{i,j}$. It follows from [193, Theorem 3] that the rescaled stationary waiting time process converges under (2.1) to a reflected Gaussian random walk. Hence, the performance measures of the nearly deterministic system, under (2.4) and (2.1), should be well approximated by the performance measures of the reflected Gaussian random walk, giving rise to heavy-traffic approximations. This connection is discussed in detail in Section 2.4.2. It seems likely that results similar to those presented in this chapter can be obtained by applying the scaling (2.2) to the nearly-deterministic systems in [192, 193], and because Pollaczek's formula also applies to this setting, the non-standard saddle point method developed in this chapter can provide the appropriate methodology.

The second class concerns multi-server systems, and in particular the many-server regime. When we interpret s_n as the number of servers, instead of capacity per time slot or order of thinning, the scaling (2.1) is similar to the QED or Halfin-Whitt regime for the $M/M/s_n$ system. As we have reviewed in Chapter 1, the QED regime is characterized by a delay probability that converges to a non-degenerate limit away from both zero and one, and the mean delay is asymptotically negligible as the number of servers grows large. The QED regime (2.1) is naturally positioned in between the Quality-Driven (QD) regime and the Efficiency-Driven (ED) regime. In the QD regime, the load remains bounded away from 1, which corresponds to setting $\eta = 0$ in (2.2). Hence, the range $\eta \in (0, 1/2)$ bridges the gap between the QED regime and the QD regime. Likewise, the ED regime corresponds to setting $\eta = 1$ in (2.2), so that the range $\eta \in (1/2, 1]$ connects the QED regime and ED regime. For the birth-death process describing the $M/M/s_n$ system, Maman [150] introduced a scaling similar to (2.2), and called it the QED-c regime, also bridging the ED and QD regimes. Theorem 4.1 of [150] says that the expected waiting time under the scaling $s_n = n\mu + \beta\sigma n^{1-\eta}$ is of order $s_n^{1-\eta}$, which is equivalent to the expected queue length being of order n^η by Little's law. We should stress though that we expect the mathematical techniques that are needed to establish heavy-traffic results could be entirely different than in this chapter, because Pollaczek's formula does not apply to many-server settings.

The specific model assumptions will determine to a large extent the appropriate methodology. Under Markovian assumptions leading to the $M/M/s_n$ system, simple exact solutions are available for the stationary distribution. This makes it possible to describe performance measures like the mean congestion directly in terms of real integrals. Where the saddle point method is used for integrals in the complex plane, the Laplace method (see e.g. [74]) is used for real integrals. Hence, for the asymptotic evaluation of the $M/M/s_n$ system under the scaling (2.2), the Laplace method seems an appropriate methodology, although again one needs to deal with possible singularities in the integrand. For $G/D/s_n$ systems, which assume deterministic service times, it has been shown in [122] that using a decomposition property the dynamics of this multi-server systems can be captured in terms of a single-server system. Hence, for these systems, Pollaczek's formula applies, and our saddle point method can most likely be applied to obtain heavy-traffic results in the regimes (2.2). Under more general conditions, for instance leading to a $G/G/s_n$ system, it is simply unclear at this stage how to obtain precise heavy-traffic approximations for (2.2), because a tractable description of the performance measures is not available; see Section 1.2.4 for details.

Structure of the chapter. In Section 2.2 we present in detail the model and the family of heavy-traffic scalings. In Section 2.3 we introduce the saddle point method. In Section 2.4 we apply the saddle point method for the mean congestion level. Theorem 2.1 gives for all heavy-traffic scalings the limiting behavior in terms of an integral expression. As a consequence, we show in Proposition 2.1 that there are two types of heavy-traffic behavior, depending on whether $\eta \in (0, 1/2)$ or $\eta \geq 1/2$. In Section 2.4.2 we discuss for the case $\eta = 1/2$ the connection with the Gaussian random walk and the Riemann zeta function. In fact, we show that for all $\eta \geq 1/2$ there exists a connection between the integral expression in Theorem 2.1 and the Riemann zeta function. In Section 2.5 we apply the saddle point method to obtain several more heavy-traffic results, including refined heavy-traffic approximations for the mean congestion level, and the leading heavy-traffic behavior for the variance of the stationary congestion level and for the empty-system probability. Finally, in Section 2.6 we confirm through numerical experiments the accuracy of our heavy-traffic approximations, and moreover show that under (2.2), various multi-server systems behave similar to our discrete bulk service queue.

2.2 Model description & heavy-traffic regimes

We thus consider a discrete stochastic model in which time is divided into periods of equal length. At the beginning of each period $j = 1, 2, 3, ...$ new demand $A_j^{(n)}$ arrives to the system. The demands per period $A_1^{(n)}, A_2^{(n)}, \ldots$ are assumed independent and equal in distribution to some non-negative integer-valued random variable $A^{(n)}$. We will omit the superscript (n) if no ambiguity is possible. The system has a service capacity $s_n \in \mathbb{N}$ per period, so that the recursion

$$Q(j+1) = \max\{Q(j) + A_j^{(n)} - s_n, 0\}, \qquad j = 1, 2, ..., \tag{2.4}$$

assuming $Q(0) = 0$, gives rise to a Markov chain $\{Q(j)\}_{j \geq 1}$ that describes the congestion in the system over time. The probability generation function (pgf)

$$\tilde{A}(z) = \sum_{k=0}^{\infty} \mathbb{P}\big(A^{(n)} = k\big) z^k$$

of $A^{(n)}$ is assumed analytic in a disk $|z| < r$ with $r > 1$, which implies that all moments of $A^{(n)}$ exist. We also assume that

$$\tilde{A}'(1) = \mathbb{E}[A_j^{(n)}] = \mu_A < s_n. \tag{2.5}$$

Under the assumption (2.5) the function $z^{s_n} - \tilde{A}(z)$ has exactly s_n zeros in the closed unit disk, one of these being $z = 1$ (see [8]). We further assume that $\mathbb{P}(A^{(n)} = j) > 0$ for some $j > s_n$. Under this assumption the function $z^{s_n} - \tilde{A}(z)$ also has zeros outside $|z| \leq 1$, and we let r_0 be the minimum modulus of these zeros. The number r_0 is the unique zero of $z^{s_n} - \tilde{A}(z)$ with real $z > 1$; see e.g. [113]. Moreover, under assumption (2.5) the stationary distribution $\lim_{j \to \infty} \mathbb{P}\left(Q(j) = k\right) = \mathbb{P}(Q = k)$, $k = 0, 1, \ldots$ exists, with the random variable Q defined as having this stationary distribution.

We let

$$\tilde{Q}(w) = \sum_{j=0}^{\infty} \mathbb{P}(Q = j) w^j$$

be the pgf of the stationary distribution. $\tilde{Q}(w)$ is analytic in $|w| < r_0$, and given by Pollaczek's formula (see e.g. [2, 61]). In our discrete setting, we shall first derive a useful expression for $\tilde{Q}(w)$.

Lemma 2.1. *For any $\varepsilon > 0$ with $1 + \varepsilon < r_0$,*

$$\tilde{Q}(w) = \exp\Big(\frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \ln\Big(\frac{w-z}{1-z}\Big) \frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)} \,\mathrm{d}z\Big) \tag{2.6}$$

holds when $|w| < 1 + \varepsilon$.

Proof. We shall establish (2.6) for any $w \in (1, 1+\varepsilon)$, and then the full result follows from analyticity of $\tilde{Q}(w)$ and of

$$\ln\Big(\frac{w-z}{1-z}\Big) = \ln\Big(\frac{1-w/z}{1-1/z}\Big) = -\sum_{k=1}^{\infty} \frac{1}{k}\left(\left(\frac{w}{z}\right)^k - \left(\frac{1}{z}\right)^k\right)$$

in w, $|w| < 1+\varepsilon$ for any z with $|z| = 1+\varepsilon$.

Our starting point is the formula, see [45],

$$\tilde{Q}(w) = \frac{(s_n - \mu_A)(w-1)}{w^{s_n} - \tilde{A}(w)} \prod_{k=1}^{s_n-1} \frac{w - z_k}{1 - z_k} \tag{2.7}$$

that holds for all w, $|w| < r_0$, in which $z_1, \ldots, z_{s_n-1}$ are the $s_n - 1$ zeros of $z^{s_n} - \tilde{A}(z)$ in $|z| < 1$. Fix $w \in (1, 1+\varepsilon)$. Then $\ln[(w-z)/(1-z)]$ is analytic in $z \in \mathbb{C}\backslash[1, w]$. It follows that

$$\begin{aligned} I_C &= \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{w-z}{1-z}\Big)\frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)}\,\mathrm{d}z \\ &= \sum_{k=1}^{s_n-1} \ln\Big(\frac{w-z_k}{1-z_k}\Big) + \frac{1}{2\pi i}\int_C \ln\Big(\frac{w-z}{1-z}\Big)\frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)}\,\mathrm{d}z, \end{aligned} \tag{2.8}$$

where C is a contour encircling $[1, w]$ in the positive sense with none of the z_k's in its interior. We let $\delta \in (0, \frac{w-1}{2})$ and we take C the union of two line segments, from $1 + \delta - i0$ to $w - \delta - i0$ and from $w - \delta + i0$ to $1 + \delta - i0$, and two circles, of radius δ and encircling 1 and w in positive sense. A careful administration of the various contributions to the integral I_C in (2.8), taking account of the branch cut $[1, w]$, yields

$$I_C = \ln\left(\frac{(s_n - \mu_A)(w-1)}{w^s - \tilde{A}(w)}\right) + O(\delta \ln \delta).$$

Using this in (2.7) and letting $\delta \downarrow 0$, we get (2.6) for $w \in (1, 1+\varepsilon)$ and the proof is complete. □

Using $\mathbb{P}(Q = 0) = \tilde{Q}(0)$, $\mu_Q = \tilde{Q}'(1)$ and $\sigma_Q^2 = \tilde{Q}''(1) + \tilde{Q}'(1) - (\tilde{Q}'(1))^2$, it follows by straightforward manipulations that

$$\mathbb{P}(Q = 0) = \exp\Big[\frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{z}{z-1}\Big)\frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)}\,\mathrm{d}z\Big], \tag{2.9}$$

$$\mu_Q = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{1}{1-z}\,\frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)}\,\mathrm{d}z, \tag{2.10}$$

$$\sigma_Q^2 = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{-z}{(1-z)^2}\,\frac{(z^{s_n} - \tilde{A}(z))'}{z^{s_n} - \tilde{A}(z)}\,\mathrm{d}z. \tag{2.11}$$

Because s_n appears directly in expressions (2.9)-(2.11), we will be conducting our analysis with respect to s_n rather than n. Note that this has no consequences for our results on the convergence speed of the performance metrics, since $s_n = O(n)$. Furthermore, we will omit the index n when describing the capacity s_n in the remainder of the chapter for brevity.

We next discuss in more detail the family of heavy-traffic scalings considered in this chapter, which combines two features. First, we have assumed that $A_j^{(n)}$ is in distribution equal to the sum of work generated by all sources, $A_{1,j} + ... + A_{n,j}$, where the $A_{i,k}$ are for all i and k i.i.d. copies of a random variable X, of which the pgf $\tilde{X}(z) = \sum_{k=0}^{\infty} \mathbb{P}(X = k) z^k$ has radius of convergence $r > 1$, and

$$0 < \mathbb{E}[A^{(n)}] = n\mu = n\tilde{X}'(1) < s_n.$$

Hence

$$\vartheta := \frac{n}{s_n} \in (0, 1/\mu). \tag{2.12}$$

Second, we scale the system according to (2.2), for which we assume that

$$\rho_{s_n} = \vartheta\,\mu = 1 - \frac{\gamma}{s_n^\eta} \tag{2.13}$$

in which $\gamma > 0$ is bounded away from 0 and ∞ as $s_n \to \infty$. In the remainder of this chapter, we will omit the subscript in s_n. The condition that $\mathbb{P}(A^{(n)} = k) > 0$ for some $k > s$ holds when the degree d of $\tilde{X}(z)$ (with $d = \infty$ if $\tilde{X}(z)$ is not a polynomial) is such that $nd > s$.

To avoid certain complications when applying the saddle point method, we further assume that

$$|\tilde{X}(z)| < \tilde{X}(r_1), \qquad |z| = r_1\,, \;\; z \neq r_1, \tag{2.14}$$

for any $r_1 \in (0, r)$. This implies that r_0 is the unique zero of $z^s - \tilde{A}(z)$ on $|z| = r_0$. This condition is related to Cramér's condition, see [20, pp. 189 and 355], and it has also been used in [114]. Condition (2.14) holds when the set of all $j = 0, 1, \ldots$ such that $\mathbb{P}(X = k) > 0$ is not contained in an arithmetic progression with a ratio larger than one (see also [8]).

2.3 Non-standard saddle point method

We illustrate our saddle point method for μ_Q. As a first step, we bring (2.10) in a form which is amenable to saddle point analysis.

Lemma 2.2.

$$\mu_Q = \frac{s}{2\pi i} \int_{|z|=1+\varepsilon} \frac{g'(z)}{z-1} \frac{\exp(s\,g(z))}{1 - \exp(s\,g(z))}\,\mathrm{d}z \tag{2.15}$$

with

$$g(z) = -\ln z + \vartheta \ln(\tilde{X}(z)). \tag{2.16}$$

Proof. With $\tilde{A}(z) = \tilde{X}^n(z)$,

$$\frac{(z^s - \tilde{A}(z))'}{z^s - \tilde{A}(z)} = \frac{s\, z^{s-1} - n\, \tilde{X}'(z)\, \tilde{X}^{n-1}(z)}{z^s - \tilde{X}^n(z)}$$
$$= \frac{s}{z} - \frac{s}{z}\Big(\frac{n}{s}\, \frac{z\,\tilde{X}'(z)}{\tilde{X}(z)} - 1\Big)\, \frac{z^{-s}\,\tilde{X}^n(z)}{1 - z^{-s}\,\tilde{X}^n(z)}. \tag{2.17}$$

Write $z^{-s}\,\tilde{X}^n(z) = \exp(s\,g(z))$. Noting that

$$\frac{1}{2\pi i} \int_{|z|=1+\varepsilon} \frac{s}{z}\, \frac{1}{1-z}\, \mathrm{d}z = 0, \tag{2.18}$$

and that

$$g'(z) = \frac{1}{z}\Big(\vartheta\, \frac{z\,\tilde{X}'(z)}{\tilde{X}(z)} - 1\Big), \tag{2.19}$$

gives (2.15). □

Let us now explain how the standard saddle point method can be applied to (2.15). Since

$$g(1) = g(r_0) = 0\,; \qquad g(z) < 0\,,\ \ 1 < z < r_0, \tag{2.20}$$

and by strict convexity of

$$z^{-s}\,\tilde{X}^n(z) = z^{-s}\tilde{A}(z) = \sum_{k=0}^{\infty} a_k\, z^{k-s}, \qquad z \in (0, r),$$

$g(z)$ has a unique minimum on $[1, r_0]$. This minimum is found by solving $z \in [1, r_0]$ from $g'(z) = 0$, and this yields the equation

$$\tilde{X}(z) = \vartheta\, z\, \tilde{X}'(z). \tag{2.21}$$

Denote the solution $z \in (1, r_0)$ of (2.21) by z_{sp}, and observe that z_{sp} is a saddle point of $g(z)$, explaining the notation. Thus, the saddle point method can be used for the integral in (2.15) by taking $1 + \varepsilon = z_{\mathrm{sp}}$.

In the case that $\vartheta = n/s$ is bounded away from $1/\mu$ as $s \to \infty$, we have that the minimum value of $g(z)$, $1 \le z \le r_0$, is negative and bounded away from 0. Furthermore, z_{sp} is bounded away from 1, and the saddle point method can be applied in the classical way by replacing

$$\frac{\exp(s\,g(z))}{1 - \exp(s\,g(z))} \quad \text{by} \quad \exp(s\,g(z)),$$

at the expense of an exponentially small relative error, and performing an expansion of $g'(z)/(z_{\mathrm{sp}} - 1) = d_1(z - z_{\mathrm{sp}}) + O((z - z_{\mathrm{sp}})^2)$ with $d_1 = g''(z_{\mathrm{sp}})/(z_{\mathrm{sp}} - 1) \neq 0$. Using that $g(z^*) = (g(z))^*$, where the * denotes complex conjugation, it can be shown that

$$\mu_Q = \frac{\exp(s\,g(z_{\mathrm{sp}}))}{(z_{\mathrm{sp}} - 1)^2\,\sqrt{2\pi s\, g''(z_{\mathrm{sp}})}}\,(1 + O(s^{-1})). \tag{2.22}$$

We next explain why the standard saddle point method does not work for the heavy-traffic scaling considered in this chapter. Since we operate in (2.13), $\vartheta\mu \to 1$ as $s \to \infty$, and

$$z_{\rm sp} - 1 = \frac{\gamma}{a_2\, s^\eta} + O(s^{-2\eta}), \tag{2.23}$$

$$g(z_{\rm sp}) = \frac{-\gamma^2}{2a_2 s^{2\eta}} + O(s^{-3\eta}), \tag{2.24}$$

$$g''(z_{\rm sp}) = a_2 + O(s^{-\eta}), \tag{2.25}$$

where

$$a_2 = \frac{\sigma^2}{\mu} - \frac{\gamma}{s^\eta}\Big(\frac{\sigma^2}{\mu} - 1\Big). \tag{2.26}$$

Hence, $\exp(sg(z))$ near $z = z_{\rm sp}$ is (as $s \to \infty$): vanishingly small when $\eta \in (0, 1/2)$, bounded away from 1, but non-negligible when $\eta = 1/2$, and tending to 1 when $\eta \in (1/2, \infty)$. Furthermore, $(z-1)^{-1}$ in (2.15) is unbounded near $z = z_{\rm sp}$ as $s \to \infty$. Therefore, an adaptation of the standard saddle point method is required, and the resulting asymptotic form of μ_Q will deviate significantly from the standard case (2.22). In particular, since $z_{\rm sp} \to 1$, this asymptotic form will contain information from $X(z)$ at $z = 1$, rather than at a point away from 1 as is the case in (2.22).

The required adaptation of the saddle point method is modeled after a device developed in [67, Sec. 5.12]. We use a substitution $z = z(v)$ in (2.15) with real v and $z(0) = z_{\rm sp}$ such that for sufficiently small v,

$$g(z(v)) = g(z_{\rm sp}) - \tfrac{1}{2}\, v^2\, g''(z_{\rm sp}). \tag{2.27}$$

This is feasible, since

$$g(z) = g(z_{\rm sp}) + \tfrac{1}{2}\, g''(z_{\rm sp})(z - z_{\rm sp})^2 \Big(1 + \frac{g'''(z_{\rm sp})}{3g''(z_{\rm sp})}\,(z - z_{\rm sp}) + \ldots\Big) \tag{2.28}$$

with $g''(z_{\rm sp})$ positive and bounded away from 0 as $s \to \infty$. Hence, $z(v)$ can be found for small v by inverting the equation

$$(z - z_{\rm sp})\Big(1 + \frac{g'''(z_{\rm sp})}{3g''(z_{\rm sp})}\,(z - z_{\rm sp}) + \ldots\Big)^{1/2} = iv. \tag{2.29}$$

By Lagrange's inversion theorem [67], there is a $\delta > 0$ (independent of s) such that

$$z(v) = z_{\rm sp} + iv + \sum_{k=2}^{\infty} c_k (iv)^k, \qquad |v| < \delta, \tag{2.30}$$

with real coefficients c_k (since $g(z)$ is real for real z) and

$$c_2 = -\frac{g'''(z_{\rm sp})}{6g''(z_{\rm sp})}. \tag{2.31}$$

Thus

$$z(v) = z_{\rm sp} + iv - c_2\, v^2 + O(v^3), \qquad |v| \le \tfrac{1}{2}\,\delta, \tag{2.32}$$

where the order term holds uniformly in s. The uniformity statement follows from an inspection of the usual argument by which Lagrange's theorem is proved, noting that the inversion in (2.27) with g as in (2.16) is considered for $\vartheta \to 1/\mu$, $z_{\rm sp} \to 1$ with radius of convergence r away from 1.

By (2.14) we can restrict the integration in (2.15) to a fixed but arbitrarily small subset of $|z| = z_{\rm sp}$ near $z = z_{\rm sp}$, at the expense of an exponentially small error. Furthermore, by Cauchy's theorem and again at the expense of an exponentially small error, the integration path can be deformed in accordance with the transformation in (2.27)–(2.32). Set

$$q(v) = g(z_{\rm sp}) - \tfrac{1}{2}\, v^2\, g''(z_{\rm sp}) \tag{2.33}$$

and note that from (2.27),

$$g'(z(v))\, z'(v) = -v\, g''(z_{\rm sp}).$$

Then substituting $z = z(v)$ in (2.15), μ_Q is given with exponentially small error by

$$\frac{s}{2\pi i} \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{g'(z(v))}{z(v)-1}\, \frac{\exp(s\, g(z(v)))}{1-\exp(s\, g(z(v)))} z'(v)\, {\rm d}v,$$

which gives the following result.

Lemma 2.3. *The mean stationary congestion level is given with exponentially small error by*

$$\mu_Q = \frac{-s}{2\pi i}\, g''(z_{\rm sp}) \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{v}{z(v)-1}\, \frac{\exp(s\, q(v))}{1-\exp(s\, q(v))}\, {\rm d}v. \tag{2.34}$$

In a similar fashion we get that $\mathbb{P}(Q=0)$ and σ_Q^2, see (2.9) and (2.11), are given, both with exponentially small error, by

$$\frac{-s}{2\pi i}\, g''(z_{\rm sp}) \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} v \ln\Big(\frac{z(v)}{z(v)-1}\Big) \frac{\exp(s\, q(v))}{1-\exp(s\, q(v))}\, {\rm d}v \tag{2.35}$$

and

$$\frac{-s}{2\pi i}\, g''(z_{\rm sp}) \int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{v\, z(v)}{(z(v)-1)^2}\, \frac{\exp(s\, q(v))}{1-\exp(s\, q(v))}\, {\rm d}v, \tag{2.36}$$

respectively.

2.4 Heavy-traffic limits for the mean congestion level

In this section we apply the non-standard saddle point method explained in Section 2.3 to the Pollaczek integral representation for the mean stationary congestion level μ_Q. In Section 2.4.1 we first derive an integral representation for the leading order behavior of μ_Q with a relative error of order $O(s^{-1})$, which serves as a heavy-traffic approximation in the regime $\rho_s = 1 - \gamma/s^\eta$ with $\eta > 0$. We also consider separately the cases of moderate heavy traffic ($\eta \in (0, 1/2)$) and extreme heavy traffic ($\eta \in (1/2, \infty)$), for which the integral representation leads to vastly different alternative expressions. We find that $\mu_Q \to 0$ more rapidly than any power of $1/s$ when $\eta \in (0, 1/2)$. When $\eta \geq 1/2$ the saddle point method yields an integral representation with relative error $O(s^{-\min(1,\eta)})$. In Section 2.4.2 we specialize this general result to the CLT case $\eta = 1/2$, and make a connection with existing results.

2.4.1 Leading order behavior in integral form

Theorem 2.1. *The mean stationary congestion level is given by*

$$\mu_Q = \frac{2}{\pi}\sigma\sqrt{\frac{s}{2\mu}}\int_0^\infty \frac{t^2}{d^2(s)+t^2}\,\frac{\exp(-d^2(s)-t^2)}{1-\exp(-d^2(s)-t^2)}\,\mathrm{d}t\,\left(1+O(s^{-\min(1,\eta)})\right) \tag{2.37}$$

with $d^2(s) = s^{1-2\eta}\gamma^2\mu/(2\sigma^2)$.

Proof. According to Lemma 2.3, μ_Q is given with exponentially small error by (2.34) with $q(v)$ given in (2.33). Since $z(-v) = z^*(v)$ for real v, we have

$$\begin{aligned}\frac{v}{z(v)-1}+\frac{-v}{z(-v)-1} &= -2iv\,\frac{\mathrm{Im}(z(v))}{|z(v)-1|^2}\\ &= \frac{-2iv^2+O(v^4)}{(z_{\mathrm{sp}}-1)^2+v^2-2c_2(z_{\mathrm{sp}}-1)\,v^2+O(v^4)}\\ &= \frac{-2iv^2\left(1+O(v^2)\right)}{(z_{\mathrm{sp}}-1)^2+v^2-2c_2(z_{\mathrm{sp}}-1)\,v^2},\end{aligned} \tag{2.38}$$

for $-\frac{1}{2}\delta \leq v \leq \frac{1}{2}\delta$. where (2.32) and $c_k \in \mathbb{R}$ have been used. Using (2.38) in (2.34) and extending the integration range from $[-\frac{1}{2}\delta, \frac{1}{2}\delta]$ to $(-\infty, \infty)$ while using symmetry of $q(v)$, we get that μ_Q is given with exponentially small error by

$$\frac{s\,g''(z_{\mathrm{sp}})}{\pi}\int_0^\infty \frac{v^2\left(1+O(v^2)\right)}{(z_{\mathrm{sp}}-1)^2+v^2-2c_2(z_{\mathrm{sp}}-1)\,v^2}\,\frac{\exp(s\,q(v))}{1-\exp(s\,q(v))}\mathrm{d}v. \tag{2.39}$$

With

$$B = \exp(s\,g(z_{\mathrm{sp}})), \qquad \alpha = g''(z_{\mathrm{sp}}), \tag{2.40}$$

Equation (2.39) takes the form

$$\frac{s\alpha}{\pi}\int_0^\infty \frac{v^2\left(1+O(v^2)\right)}{(z_{\mathrm{sp}}-1)^2+v^2-2c_2(z_{\mathrm{sp}}-1)\,v^2}\cdot\frac{B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}{1-B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}\mathrm{d}v. \tag{2.41}$$

Since $(z_{\mathrm{sp}}-1)^2 = (\gamma/a_2)^2 s^{-2\eta} + O(s^{-4\eta})$, see (2.23), the integrand in (2.41) in leading order has the form

$$\frac{B\,v^2\,\exp(-s\,D\,v^2)}{(v^2+C\,s^{-2\eta})(1-B\exp(-s\,D\,v^2))},$$

and this is reminiscent of the integrand in [67, Eq. (5.12.3)] for the case $\kappa = 2\eta$. Proceeding as in [67, Sec. 5.12], the substitution $v = t\sqrt{2/(s\alpha)}$ brings (2.41) into the form

$$\frac{2}{\pi}\sqrt{\tfrac{1}{2}s\alpha}\int_0^\infty \frac{t^2(1+O(t^2/s))}{\frac{1}{2}s\alpha(z_{\rm sp}-1)^2+t^2-2c_2(z_{\rm sp}-1)t^2}\,\frac{B\,\exp(-t^2)}{1-B\,\exp(-t^2)}\mathrm{d}t. \tag{2.42}$$

From (2.23)–(2.26) and (2.40),

$$\frac{2}{\pi}\sqrt{\frac{s\alpha}{2}} = \frac{2}{\pi}\,\sigma_X\sqrt{\frac{s}{2\,\mu}}\,(1+O(s^{-\eta})), \tag{2.43}$$

$$\tfrac{1}{2}\,s\,\alpha\,(z_{\rm sp}-1)^2 = d^2(s)+O(s^{1-3\eta}), \tag{2.44}$$

$$2\,c_2(z_{\rm sp}-1) = O(s^{-\eta}), \tag{2.45}$$

$$s\,g(z_{\rm sp}) = -d^2(s)+O(s^{1-3\eta}), \tag{2.46}$$

where

$$d^2(s) = \frac{b_0^2}{s^{2\eta-1}},\quad b_0^2 := \frac{\gamma^2\mu}{2\,\sigma^2}. \tag{2.47}$$

In the case that $2\eta-1<0$, we have that $\frac{1}{2}\,s\,\alpha\,(z_{\rm sp}-1)^2\to\infty$ and that

$$B = \exp(s\,g(z_{\rm sp})) = O(\exp(-b^2s^{1-2\eta})) \tag{2.48}$$

for any $b\in(0,b_0)$. From (2.42) it then follows that $\mu_Q = O(\exp(-b^2\,s^{1-2\eta}))$ for any $b\in(0,b_0)$. In the case that $2\,\eta-1\geq 0$, we have that $d^2(s)$ is bounded, and using that $1/s^{3\eta-1} = O(d^2(s)/s^\eta)$, we get

$$\begin{aligned}\tfrac{1}{2}\,s\,\alpha\,(z_{\rm sp}-1)^2+t^2-2\,c_2\,(z_{\rm sp}-1)\,t^2 &= d^2(s)+t^2+O\left(s^{-\eta}\,(d^2(s)+t^2)\right)\\ &= \left(d^2(s)+t^2\right)\left(1+O(s^{-\eta})\right).\end{aligned}$$

Hence, in this case,

$$\frac{t^2(1+O(t^2/s))}{\frac{1}{2}s\,\alpha(z_{\rm sp}-1)^2+t^2-2c_2(z_{\rm sp}-1)t^2} = \frac{t^2}{d^2(s)+t^2}\left(1+O(s^{-\eta})+O(t^2/s)\right). \tag{2.49}$$

Furthermore,

$$\begin{aligned}1-B\,\exp(-t^2) &= 1-\exp(-d^2(s)-t^2)\,\left(1+d^2(s)\,O(s^{-\eta})\right)\\ &= (1-\exp(-d^2(s)-t^2))\left(1+\frac{d^2(s)}{\exp(d^2(s)+t^2)-1}O(s^{-\eta})\right)\\ &= (1-\exp(-d^2(s)-t^2))\,(1+O(s^{-\eta})),\end{aligned}$$

It follows therefore that

$$\frac{B\,\exp(-t^2)}{1-B\,\exp(-t^2)} = \frac{\exp(-d^2(s)-t^2)}{1-\exp(-d^2(s)-t^2)}\,(1+O(s^{-\eta})). \tag{2.50}$$

Combining the three items (2.43), (2.49) and (2.50), we obtain for (2.42) the result

$$\frac{2}{\pi}\,\sigma\sqrt{\frac{s}{2\,\mu}}\int_0^\infty \frac{t^2}{d^2(s)+t^2}\cdot\frac{\exp(-d^2(s)-t^2)}{1-\exp(-d^2(s)-t^2)}\mathrm{d}t\left(1+O(s^{-\eta})+O(s^{-1})\right),$$

and this gives (2.37). □

Theorem 2.1 gives the leading-order behavior of μ_Q as $s\to\infty$ with a relative error of $O(s^{-\min(1,\eta)})$. By considering in more detail the integral expressions, we obtain the following result, describing two different heavy-traffic behaviors.

Proposition 2.1. *If $\eta\in(0,1/2)$ the mean congestion level satisfies*

$$\mu_Q = O\left(\exp(-b^2 s^{1-2\eta})\right),$$

for any $b\in(0,b_0)$. If $\eta\in[1/2,\infty)$ the mean congestion level is given by

$$\mu_Q = s^\eta\,\frac{\sigma^2}{2\mu\gamma}\left(1+O(s^{\max(1/2-\eta,-1)})\right).$$

The first assertion in Proposition 2.1 follows from the observation in (2.48), together with (2.42). The second assertion is based on a connection between the integral in Theorem 2.1 and the Riemann zeta function, which is explained in the next subsection.

2.4.2 Classical heavy traffic and the Gaussian random walk

We now build on Theorem 2.1 to obtain further results for the classical heavy traffic case $\eta=1/2$, for which we know from [193, Thm. 3] that the rescaled congestion process converges under (2.1) to a reflected Gaussian random walk. The latter is defined as $(S_\beta(k))_{k\geq 0}$ with $S_\beta(0)=0$ and

$$S_\beta(j) = Y_1+\ldots+Y_j$$

with $Y_1,Y_2,\ldots$ i.i.d. copies of a normal random variable with mean $-\beta$ and variance 1. Assume $\beta>0$ (negative drift), and denote the all-time maximum of this random walk by M_β.

Denote by $Q_\infty^{(s)}$ the stationary congestion level for a fixed s (that arises from taking $j\to\infty$ in (2.4)), and remember that we have assumed $\vartheta=n/s$ fixed. Then, using $\rho_s=1-\gamma/\sqrt{s}$, with

$$\gamma = \frac{\beta\sigma}{\mu\sqrt{\vartheta}}, \tag{2.51}$$

the spatially-scaled stationary congestion levels reach the limit $Q_\infty^{(s)}/(\sigma\sqrt{n}) \stackrel{d}{\Rightarrow} M_\beta$ as $s,n\to\infty$ (see [122, 192, 193]). From [193, Thm. 4] we then know that under the standard heavy-traffic scaling (2.1)

$$\frac{\mathbb{E}[Q_\infty^{(s)}]}{\sigma\sqrt{n}} \to \mathbb{E}[M_\beta],\quad \text{as } s,n\to\infty, \tag{2.52}$$

from which it follows that

$$\mu_Q \approx \sigma\sqrt{n}\,\mathbb{E}[M_\beta]. \tag{2.53}$$

The random variable M_β was studied in [57, 115]. In particular, [115, Thm. 2] yields, for $\beta < 2\sqrt{\pi}$,

$$\mathbb{E}[M_\beta] = \frac{1}{2\beta} + \frac{\zeta(1/2)}{\sqrt{2\pi}} + \frac{\beta}{4} + \frac{\beta^2}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-1/2-r)}{r!(2r+1)(2r+2)} \left(\frac{-\beta^2}{2}\right)^r,$$

where ζ denotes the Riemann zeta function, which is defined as, see (1.26). Hence, for small values of β,

$$\mu_Q \approx \sigma\sqrt{n}\,\mathbb{E}[M_\beta] \approx \frac{\sigma\sqrt{n}}{2\beta} = \sqrt{s}\,\frac{\sigma^2}{2\mu\gamma}. \tag{2.54}$$

We will now show how the approximation (2.54) follows from Theorem 2.1, and also how similar steps give rise to Proposition 2.1.

Consider the integral

$$G_0(b) = G_1(b) - G_2(b) = \int_0^\infty \frac{t^2}{b^2+t^2}\,\frac{\exp(-b^2-t^2)}{1-\exp(-b^2-t^2)}\,\mathrm{d}t, \tag{2.55}$$

where $b > 0$ and

$$G_1(b) = \int_0^\infty \frac{\exp(-b^2-t^2)}{1-\exp(-b^2-t^2)}\,\mathrm{d}t, \quad G_2(b) = \int_0^\infty \frac{b^2}{b^2+t^2}\,\frac{\exp(-b^2-t^2)}{1-\exp(-b^2-t^2)}\mathrm{d}t. \tag{2.56}$$

We have, as in [115, Sec. 2],

$$\begin{aligned} G_1(b) &= \sum_{k=0}^{\infty} \int_0^\infty \exp(-(k+1)(b^2+t^2))\,\mathrm{d}t \\ &= \frac{\sqrt{\pi}}{2} \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-(k+1)b^2}}{\sqrt{k+1}} = \frac{\sqrt{\pi}}{2}\,\mathrm{e}^{-b^2}\,\Phi(\mathrm{e}^{-b^2}, 1/2, 1) \\ &= \frac{\pi}{2b} + \frac{\sqrt{\pi}}{2} \sum_{r=0}^{\infty} \zeta(\tfrac{1}{2}-r)\,\frac{(-1)^r\,b^{2r}}{r!}, \end{aligned} \tag{2.57}$$

where the last identity holds when $0 < b < \sqrt{2\pi}$ and $\Phi(z,s,v)$ is Lerch's transcendent, which is defined as, see [174, Eq. 25.14.1],

$$\Phi(z,s,v) = \sum_{n=0}^{\infty} \frac{z^n}{(v+n)^s}, \qquad \text{for } v \neq 0,-1,-2,\ldots,\ |z|<1;\ \Re s > 1,\ |z| = 1.$$

As to $G_2(b)$, we make a connection with the complementary error function

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{2}{\pi}\,\mathrm{e}^{-z^2} \int_0^\infty \frac{\mathrm{e}^{-z^2t^2}}{1+t^2}\mathrm{d}t,$$

see [174, Secs. 7.2 and 7.7.1]. We thus compute

$$\begin{aligned} G_2(b) &= \sum_{k=0}^{\infty} \mathrm{e}^{-(k+1)b^2} \int_0^{\infty} \frac{b^2}{b^2+t^2}\, \mathrm{e}^{-(k+1)t^2} \mathrm{d}t \\ &= \frac{\pi}{2}\, b \sum_{k=0}^{\infty} \operatorname{erfc}(b\sqrt{k+1}). \end{aligned} \tag{2.58}$$

From [115, Eq. (4.3) & (4.23)],

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{2\pi}} \int_{\beta\sqrt{n}}^{\infty} \mathrm{e}^{-x^2/2}\, dx = \frac{1}{2\beta^2} - \frac{1}{4} - \frac{1}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-1/2-r)(-1/2)^r}{r!\,(2r+1)}\, \beta^{2r+1} \tag{2.59}$$

in which $0 < \beta < 2\sqrt{\pi}$. Taking $\beta = b\sqrt{2}$ in (2.59), we get

$$G_2(b) = \frac{\pi}{4b} - \frac{\pi}{4}\, b - \sqrt{\pi} \sum_{r=0}^{\infty} \frac{\zeta(-1/2-r)(-1)^r\, b^{2r+2}}{r!\,(2r+1)} \tag{2.60}$$

when $0 < b < \sqrt{2\pi}$. The two results in (2.57) and (2.60) can be combined, as in [115, Sec. 2.5.2], and this yields

$$G_0(b) = \frac{\pi}{4b} + \frac{\pi}{4}\, b + \frac{\sqrt{\pi}}{2}\, \zeta(1/2) + \sqrt{\pi} \sum_{r=0}^{\infty} \frac{\zeta(-1/2-r)(-1)^r\, b^{2r+2}}{r!\,(2r+1)(2r+2)} \tag{2.61}$$

when $0 < b < \sqrt{2\pi}$.

Using (2.61) in (2.53), we find that the leading order behavior of μ_Q is given as

$$\sigma_X \sqrt{\frac{s}{2\mu}} \left[\frac{1}{2b_0} + \frac{b_0}{2} + \frac{\zeta(1/2)}{\sqrt{\pi}} + \frac{2}{\sqrt{\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-1/2-r)(-1)^r b_0^{2r+2}}{r!\,(2r+1)(2r+2)} \right] \tag{2.62}$$

with relative error of $O(s^{-1/2})$ in which b_0 is given by (2.47). The expression (2.62) is exactly equal to the right-hand side of [115, Eq. (4.25)] times $\sqrt{s}$ when we take there $\sigma = \mu = 1$ and $\beta = b_0\sqrt{2}$. Notice that, with γ as in (2.51),

$$\sigma \sqrt{\frac{s}{2\mu}} \frac{1}{2b_0} = \frac{\sigma\sqrt{n}}{2\beta},$$

which confirms the approximation (2.54).

According to Theorem 2.1, we have for $\eta \geq 1/2$,

$$\mu_Q = \frac{2}{\pi}\, \sigma \sqrt{\frac{s}{2\,\mu}} G_0(d(s)) \left(1 + O(s^{-\min(1,\eta)})\right).$$

When $\eta = 1/2$, so that $d(s) = b_0$ is independent of s, the series representation for G_0 in (2.61) can be used, as long as $b_0 \in (0, \sqrt{2\pi})$. When $\eta > 1/2$, we have that $d(s) = b_0/s^{\eta-1/2} \to 0$ as $s \to \infty$, and so this series representation can be used when

s is large enough. We then have from (2.61) and $b_0^2 = \gamma^2\mu/2\,\sigma^2$, while replacing the whole series at the right-hand side by $O(b^2)$, for μ_Q the leading order behavior

$$s^\eta \left[\frac{\sigma^2}{2\,\gamma\,\mu} + \frac{\sigma\,\zeta(1/2)}{\sqrt{2\,\pi\,\mu}}\,\frac{1}{s^{\eta-1/2}} + \frac{1}{4}\,\gamma\,\frac{1}{s^{2\eta-1}} + O(s^{3/2-3\eta}) \right] \tag{2.63}$$

with relative error $O(s^{-\min(1,\eta)})$. Retaining the constant term $\sigma^2/(2\gamma\mu)$ and estimating the other terms between the brackets in (2.63) as $O(s^{1/2-\eta})$, we get Proposition 2.1.

2.5 More heavy-traffic results

In this section we apply the non-standard saddle point method to obtain several more heavy-traffic results. In Section 2.5.1 we derive refined heavy-traffic approximations for the mean congestion level by considering higher-order correction terms. In Section 2.5.2 we derive the leading heavy-traffic behavior for the variance of the stationary congestion level, and in Section 2.5.3 for the empty-system probability. To keep the developments tractable, we restrict Section 2.5.1 to $\eta = 1/2$, and Section 2.5.2 and Section 2.5.3 to $\eta \in (0,1]$, although the same technique will work for all values $\eta > 0$.

2.5.1 Correction term for the mean congestion level for $\eta = 1/2$

Our saddle point method not only establishes the leading-order heavy-traffic approximations, but also allows to derive refinements to these approximations. In this section we demonstrate how this works for the mean congestion level in the case $\eta = 1/2$.

To obtain a refinement or correction term from (2.42), we must be more precise about the $O(s^{-\eta})$ terms that occur in the approximations in Section 2.4.1 for $\frac{1}{2}\,s\,\alpha(z_{\text{sp}} - 1)^2$, B and $\sqrt{s\,\alpha/2}$. When higher-order corrections are required, we should include higher-order terms in the approximations of these quantities, and be more specific about the $O(t^2/s)$ and $O(t^4/s)$ in the integrand in (2.42).

Let $g^{(i)}$, $i = 1, 2, \ldots$ denote the i^{th} derivative of g and define, see (2.12) and (2.16) with $\vartheta = (1 - \gamma/s^\eta)\,\mu^{-1}$,

$$a_i = g^{(i)}(1); \qquad g(z) = -\ln z + \vartheta \ln \tilde{X}(z).$$

Dropping the X from μ and σ^2 for brevity, we have

$$a_1 = -\frac{\gamma}{s^\eta}, \qquad a_2 = \frac{\sigma^2}{\mu} - \frac{\gamma}{s^\eta}\Big(\frac{\sigma^2}{\mu} - 1\Big),$$

$$a_3 = -2 + \Big(1 - \frac{\gamma}{s^\eta}\Big)\Big(\frac{\tilde{X}'''(1)}{\tilde{X}'(1)} - 3\tilde{X}''(1) + 2(\tilde{X}'(1))^2\Big).$$

For the purpose of finding a first-order correction term, we note that

$$\begin{aligned}
\alpha = g''(z_{\rm sp}) &= a_2 + (z_{\rm sp} - 1)\, a_3 + O(s^{-1}),\\
z_{\rm sp} - 1 &= -\,\frac{a_1}{a_2} - \frac{a_3}{2a_2}\Big(\frac{a_1}{a_2}\Big)^2 + O(s^{-3/2}),\\
c_2 = -\,\frac{g'''(z_{\rm sp})}{6g''(z_{\rm sp})} &= -\,\frac{a_3}{6a_2} + O(s^{-1/2}),\\
g(z_{\rm sp}) &= -\,\frac{a_1^2}{2a_2} - \frac{a_3}{6a_2^3}\, a_1^3 + O(s^{-2}).
\end{aligned}$$

This gives rise to

$$\sqrt{\tfrac{1}{2}\, s\, \alpha} = \sigma \sqrt{\frac{s}{2\mu}}\,\Big(1 + \frac{C_1}{\sqrt{s}} + O(s^{-1})\Big), \tag{2.64}$$

$$\tfrac{1}{2}\, s\, \alpha (z_{\rm sp} - 1)^2 = \frac{\gamma^2 \mu}{2\sigma^2} + \frac{C_2}{\sqrt{s}} + O(s^{-1}), \tag{2.65}$$

$$2c_2(z_{\rm sp} - 1) = \frac{C_3}{\sqrt{s}} + O(s^{-1}), \tag{2.66}$$

$$B = \exp(s\, g(z_{\rm sp})) = \exp\Big(-\,\frac{\gamma^2 \mu}{2\sigma^2}\Big)\Big(1 + \frac{C_4}{\sqrt{s}} + O(s^{-1})\Big), \tag{2.67}$$

with explicitly computable constants C_1, C_2, C_3, C_4. Remembering that $b_0^2 = \gamma^2\mu/2\sigma^2$, see (2.47), we then get with errors of order $1/s$

$$\begin{aligned}
&\frac{t^2(1 + O(t^2/s))}{\frac{1}{2}\, s\, \alpha(z_{\rm sp} - 1)^2 + t^2 - 2c_2(z_{\rm sp} - 1)\, t^2}\\
&\qquad = \frac{t^2}{b_0^2 + t^2} - \frac{1}{\sqrt{s}}\Big((C_2 + b_0^2\, C_3)\,\frac{t^2}{(b_0^2 + t^2)^2} - C_3\,\frac{t^2}{b_0^2 + t^2}\Big),
\end{aligned} \tag{2.68}$$

and

$$\frac{B\,\exp(-t^2)}{1 - B\,\exp(-t^2)} = \frac{\exp(-b_0^2 - t^2)}{1 - \exp(-b_0^2 - t^2)} + \frac{C_4}{\sqrt{s}}\,\frac{\exp(-b_0^2 - t^2)}{(1 - \exp(-b_0^2 - t^2))^2}. \tag{2.69}$$

Using (2.64), (2.68) and (2.69) in (2.42) we get with an absolute error of order $1/\sqrt{s}$

$$\begin{aligned}\mu_Q = {} & \frac{2}{\pi}\,\sigma\,\sqrt{\frac{s}{2\mu}}\,\Big(1+\frac{C_1}{\sqrt{s}}\Big)\\ & \cdot \int_0^\infty \Big(\frac{t^2}{b_0^2+t^2}-\frac{1}{\sqrt{s}}\Big(\big(C_2+b_0^2\,C_3\big)\,\frac{t^2}{(b_0^2+t^2)^2}-C_3\,\frac{t^2}{b_0^2+t^2}\Big)\Big)\\ & \cdot\ \Big(\frac{\exp(-b_0^2-t^2)}{1-\exp(-b_0^2-t^2)}+\frac{C_4}{\sqrt{s}}\,\frac{\exp(-b_0^2-t^2)}{(1-\exp(-b_0^2-t^2))^2}\Big)\mathrm{d}t\\ = {} & \frac{2\sigma}{\pi}\,\sqrt{\frac{s}{2\mu}}\,G_0(b_0)\\ & + \frac{2\sigma}{\pi}\,\sqrt{\frac{1}{2\mu}}\,\big((C_1+C_3)\,G_0(b_0)-(C_2+b_0^2\,C_3)\,G_3(b_0)+C_4\,G_4(b_0)\big),\end{aligned} \tag{2.70}$$

where G_0 is as in (2.55), and

$$G_3(b_0)=\int_0^\infty \frac{t^2}{(b_0^2+t^2)^2}\,\frac{\exp(-b_0^2-t^2)}{1-\exp(-b_0^2-t^2)}\,\mathrm{d}t, \tag{2.71}$$

$$G_4(b_0)=\int_0^\infty \frac{t^2}{b_0^2+t^2}\,\frac{\exp(-b_0^2-t^2)}{(1-\exp(-b_0^2-t^2))^2}\,\mathrm{d}t. \tag{2.72}$$

We shall express the integrals in (2.71) and (2.72) in terms of ζ-functions. By partial integration

$$\begin{aligned}G_3(b) = {} & \frac{1}{2}\int_0^\infty \frac{1}{b^2+t^2}\,\frac{\exp(-b^2-t^2)}{1-\exp(-b_0^2-t^2)}\,\mathrm{d}t\\ & -\int_0^\infty \frac{t^2}{b^2+t^2}\,\frac{\exp(-b^2-t^2)}{(1-\exp(-b^2-t^2))^2}\,\mathrm{d}t\\ = {} & \frac{1}{2b^2}\,G_2(b)-G_4(b),\end{aligned} \tag{2.73}$$

see (2.55) and (2.72). Since $G_2(b)$ is expressed in terms of ζ-functions in (2.60), it is sufficient to consider $G_4(b)$.

As to $G_4(b)$,

$$G_4(b)=G_5(b)-G_6(b),$$

where

$$G_5(b)=\int_0^\infty \frac{\exp(-b^2-t^2)}{(1-\exp(-b^2-t^2))^2}\,\mathrm{d}t,$$

$$G_6(b)=\int_0^\infty \frac{b^2}{b^2+t^2}\,\frac{\exp(-b^2-t^2)}{(1-\exp(-b^2-t^2))^2}\,\mathrm{d}t.$$

We have, compare (2.57),

$$
\begin{aligned}
G_5(b) &= \sum_{k=0}^{\infty} (k+1) \int_0^{\infty} \mathrm{e}^{-(k+1)(b^2+t^2)}\,\mathrm{d}t \\
&= \frac{\sqrt{\pi}}{2}\,\mathrm{e}^{-b^2}\,\Phi(\mathrm{e}^{-b^2},-\tfrac{1}{2},1) = \frac{\pi}{4b^3} + \frac{\sqrt{\pi}}{2}\sum_{r=0}^{\infty} \zeta(-\tfrac{1}{2}-r)\,\frac{(-1)^r\,b^{2r}}{r!}, \qquad (2.74)
\end{aligned}
$$

the last identity being valid when $0 < b < \sqrt{2\pi}$. Next we have, compare (2.58),

$$
\begin{aligned}
G_6(b) &= \sum_{k=0}^{\infty} (k+1)\,b^2 \int_0^{\infty} \frac{\exp(-(k+1)(b^2+t^2))}{b^2+t^2}\mathrm{d}t \\
&= \frac{\pi}{2}\,b \sum_{k=0}^{\infty} (k+1)\,\mathrm{erfc}(b\sqrt{k+1}).
\end{aligned}
$$

From [115, Eq. (5.4) & (5.21)] we have

$$
\sum_{n=1}^{\infty} \frac{n}{\sqrt{2\pi}} \int_{\beta\sqrt{n}}^{\infty} \mathrm{e}^{-x^2/2}\,dx = \frac{3}{4\beta^4} - \frac{1}{24} - \frac{1}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-3/2-r)(-1/2)^r}{r!\,(2r+1)}\,\beta^{2r+1} \qquad (2.75)
$$

when $0 < \beta < 2\sqrt{\pi}$. Taking $\beta = b\sqrt{2}$ in (2.75), we get

$$
G_6(b) = \frac{3\pi}{16b^2} - \frac{\pi b}{24} - \sqrt{\pi} \sum_{r=0}^{\infty} \frac{\zeta(-3/2-r)(-1)^r}{r!\,(2r+1)}\,b^{2r+2} \qquad (2.76)
$$

when $0 < b < \sqrt{2\pi}$. The two results (2.74) and (2.76) can be combined, as in [115, Sec. 5] and this yields

$$
G_4(b) = \frac{\pi}{16b^3} + \frac{\pi b}{24} + \tfrac{1}{2}\,\zeta(-1/2)\,\sqrt{\pi} + \sqrt{\pi} \sum_{r=0}^{\infty} \frac{\zeta(-3/2-r)(-1)^r\,b^{2r+2}}{r!\,(2r+1)(2r+2)} \qquad (2.77)
$$

when $0 < b < \sqrt{2\pi}$. Finally, we can rewrite

$$
\begin{aligned}
\frac{1}{2b^2}G_2(b) &= \frac{\pi}{8b^3} - \frac{\pi}{8b} - \frac{\sqrt{\pi}}{2}\sum_{r=0}^{\infty}\frac{\zeta(-1/2-r)(1-)^r b^{2r}}{r!(2r+1)} \\
&= \frac{\pi}{8b^3} - \frac{\pi}{8b} - \frac{\sqrt{\pi}}{2}\sum_{r=-1}^{\infty}\frac{\zeta(-3/2-r)(-1)^{r+1} b^{2r+2}}{(r+1)!(2r+3)} \\
&= \frac{\pi}{8b^3} - \frac{\pi}{8b} - \tfrac{1}{2}\zeta(-1/2)\sqrt{\pi} + \sqrt{\pi}\sum_{r=0}^{\infty}\frac{\zeta(-3/2-r)(-1)^r b^{2r+2}}{r!\,(2r+2)(2r+3)} \qquad (2.78)
\end{aligned}
$$

and use (2.77) and (2.78) in (2.73), by which we obtain for $0 < b < \sqrt{2\pi}$,

$$\begin{aligned} G_3(b) &= \frac{\pi}{16b^3} - \frac{\pi}{8b} - \frac{\pi b}{24} - \zeta(-1/2)\sqrt{\pi} \\ &\quad + \sqrt{\pi}\sum_{r=0}^{\infty} \frac{\zeta(-3/2-r)(-1)^r b^{2r+2}}{r!\,(2r+2)} \Big[\frac{1}{2r+3} - \frac{1}{2r+1}\Big] \\ &= \frac{\pi}{16b^3} - \frac{\pi}{8b} - \frac{\pi b}{24} - \zeta(-1/2)\sqrt{\pi} - 2\sqrt{\pi}\sum_{r=0}^{\infty} \frac{\zeta(-3/2-r)(-1)^r\, b^{2r+2}}{r!\,(2r+1)(2r+2)(2r+3)}. \end{aligned} \tag{2.79}$$

The right-hand side of (2.79) equals the right-hand side of [115, Eq. (2.3)] multiplied by $\pi/(2b)$ with $\beta = b\sqrt{2}$.

2.5.2 Variance of the congestion level

We have from (2.36) in Section 2.2, using the same approach and notation as in Section 2.4.1 for μ_Q, that σ_Q^2 is given with exponentially small error by

$$\frac{-s\,\alpha}{2\pi i}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{v\,z(v)}{(z(v)-1)^2}\, \frac{B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}{1-B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}\mathrm{d}v, \tag{2.80}$$

with B and α given in (2.40). From $z(-v) = z^*(v)$ for real v we now compute

$$\frac{z(v)}{(z(v)-1)^2} - \frac{z(-v)}{(z(-v)-1)^2} = -2i\,\frac{|z(v)|^2-1}{|z(v)-1|^4}\,\mathrm{Im}(z(v)),$$

and so (2.80) becomes

$$\frac{s\alpha}{\pi}\int_0^{\frac{1}{2}\delta} \frac{|z(v)|^2-1}{|z(v)-1|^4}\, v\,\mathrm{Im}(z(v))\, \frac{B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}{1-B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}\mathrm{d}v. \tag{2.81}$$

From

$$\mathrm{Im}(z(v)) = v + O(v^3), \qquad |z(v)|^2 - 1 = z_{\mathrm{sp}}^2 - 1 + O(v^2),$$

we get for the expression in (2.81)

$$\frac{s\alpha}{\pi}\int_0^{\frac{1}{2}\delta} \frac{v^2\,(z_{\mathrm{sp}}^2-1+O(v^2))(1+O(v^2))}{((z_{\mathrm{sp}}-1)^2+v^2+O((z_{\mathrm{sp}}-1)\,v^2)+O(v^4))^2}\, \frac{B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}{1-B\,\exp(-\frac{1}{2}\,s\,\alpha\,v^2)}\mathrm{d}v. \tag{2.82}$$

When $2\eta - 1 < 0$, we have as for the case of μ_Q in Section 2.4.1 that the whole expression in (2.82) is $O(\exp(-b^2\,s^{1-2\eta}))$ for any $b \in (0, b_0)$, as $s \to \infty$. When $2\eta - 1 \geq 0$, we get as in the case of μ_Q after substitution $v = t\sqrt{2/(s\,\alpha)}$ for the expression in (2.82)

$$\frac{2}{\pi}\Big(\frac{s\,\alpha}{2}\Big)^{3/2}\int_0^{\infty} \frac{t^2\,(z_{\mathrm{sp}}^2-1+O(t^2/s))(1+O(t^2/s))}{(d^2(s)+t^2)^2\,(1+O(1/s^{\eta})+O(t^2/s))}\, \frac{B\,\mathrm{e}^{-t^2}}{1-B\,\mathrm{e}^{-t^2}}\mathrm{d}t.$$

When $2\eta - 1 \geq 0$, the leading order behavior of σ_Q^2 depends crucially on the factor $z_{\rm sp}^2 - 1 + O(t^2/s)$, where

$$z_{\rm sp}^2 - 1 = \frac{2\,\gamma\,\mu}{\sigma^2\,s^\eta}\,\left(1 + O(s^{-\eta})\right)$$

is dominant when $\eta < 1$, while the $O(t^2/s)$ is dominant when $\eta > 1$. In the case that $\eta \in (1/2, 1)$, we get for the leading order behavior of σ_Q^2

$$\begin{aligned}
&\frac{2}{\pi}\left(\frac{s\,\alpha}{2}\right)^{3/2}\frac{2\,\gamma\,\mu}{\sigma^2\,s^\eta}\int_0^\infty \frac{t^2}{(d^2(s)+t^2)^2}\cdot\frac{\mathrm{e}^{-d^2(s)-t^2}}{1-\mathrm{e}^{-d^2(s)-t^2}}\mathrm{d}t\,\left(1+O(s^{\eta-1})\right)\\
&\qquad = \frac{\gamma\,\sigma}{\pi}\left(\frac{2}{\mu}\right)^{1/2} s^{3/2-\eta}\,G_3(d(s))\,\left(1+O(s^{\eta-1})\right),
\end{aligned}$$

where (2.25), (2.26) and (2.40) have been used for $\alpha = g''(z_{\rm sp})$ and where G_3 is given in (2.71).

When we insert the expansion (2.79) for $G_3(b)$, with the whole series on the second line being $O(b^2)$, we get the leading order behavior of σ_Q^2 as

$$\begin{aligned}
s^{2\eta}\Bigg(\frac{\sigma^4}{4\,\gamma^2\mu^2} - \frac{\sigma^2}{4\,\mu}\,\frac{1}{s^{2\eta-1}} - \left(\frac{2\,\sigma^2}{\pi\,\mu}\right)^{1/2}\frac{\gamma\,\zeta(-1/2)}{s^{3\eta-3/2}}&\\
- \frac{\gamma^2}{24\,s^{5\eta-5/2}} + O(s^{1-4\eta})\Bigg)&\,\left(1+O(s^{\eta-1})\right)\\
= s^{2\eta}\,\frac{\sigma^4}{4\,\gamma^2\,\mu^2}&\,\left(1+O(s^{\max(1-2\eta,\eta-1)})\right)
\end{aligned}\tag{2.83}$$

when $\eta \in (1/2, 1)$. For the case $\eta = 1/2$, we get the leading order behavior, assuming $0 < b_0 < \sqrt{2\pi}$,

$$\frac{\sigma^2 s}{\mu}\left[\frac{1}{8\,b_0^2} - \frac{1}{4} - \frac{1}{12}\,b_0^2 - \frac{2\,\zeta(-1/2)}{\sqrt{\pi}}\,b_0 - \frac{4}{\sqrt{\pi}}\sum_{r=0}^{\infty}\frac{\zeta(-3/2-r)\,(-1)^r\,b_0^{2r+3}}{r!\,(2r+1)\,(2r+2)\,(2r+3)}\right] \tag{2.84}$$

with relative error $O(s^{-1/2})$. The expression between brackets in (2.84) coincides with the right-hand side of [115], (2.3) with $\beta = b_0\,\sqrt{2}$.

This leads to the following two results.

Theorem 2.2. *For $\eta \in [1/2, 1)$,*

$$\sigma_Q^2 = \frac{\gamma\,\sigma_X}{\pi}\sqrt{\frac{2}{\mu}}\,s^{3/2-\eta}\,G_3(d(s))\,\left(1+O(s^{\eta-1})\right)$$

with G_3 given in (2.71).

Proposition 2.2. *For $\eta \in (0,1/2)$, and for all $b < b_0$,*

$$\sigma_Q^2 = O(\exp(-b^2\, s^{1-2\eta})).$$

For $\eta = 1/2$, σ_Q^2 equals expression (2.84) *with relative error $O(s^{-1/2})$. For $\eta \in (1/2,1)$ and $b_0 \in (0,\sqrt{2\pi})$, σ_Q^2 has the form in* (2.83).

As in Section 2.5.1 for the mean congestion level with $\eta = 1/2$, it is possible to give a correction term which involves now integrals and series with ζ-functions as considered in [116, Secs. 4-5].

2.5.3 The empty-system probability

We have from (2.9) by proceeding as in (2.17)–(2.19) that

$$\begin{aligned}
\ln\left[\mathbb{P}(Q=0)\right] &= \frac{s}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{z}{z-1}\Big)\,\frac{g'(z)\,\exp(s\,g(z))}{1-\exp(s\,g(z))}\,\mathrm{d}z \\
&= \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{1}{z(z-1)}\ln\left(1-\exp(s\,g(z))\right)\,\mathrm{d}z, \qquad (2.85)
\end{aligned}$$

where in the last step we used partial integration (noting that $\mathrm{Re}\,[g(z)] < 0$ on $|z| = 1+\varepsilon$). Then, as in Section 2.2 for μ_Q, the last integral in (2.85) is, with exponentially small error, given by

$$\frac{1}{2\pi i}\int_{-\frac{1}{2}\delta}^{\frac{1}{2}\delta} \frac{z'(v)}{z(v)(z(v)-1)}\ln\Big(1-B\,\exp\big(-\tfrac{1}{2}s\,\alpha v^2\big)\Big)\,\mathrm{d}v. \qquad (2.86)$$

Now for $v \geq 0$ from $z(-v) = z^*(v)$, $z'(-v) = -(z'(v))^*$,

$$\begin{aligned}
&\frac{z'(v)}{z(v)(z(v)-1)} + \frac{z'(-v)}{z(-v)(z(-v)-1)} = 2i\,\mathrm{Im}\Big[\frac{z'(v)}{z(v)(z(v)-1)}\Big] \\
&\quad = 2i\,\mathrm{Im}\Big[\frac{z'(v)\,z^*(v)(z^*(v)-1)}{|z(v)|^2\,|z(v)-1|^2}\Big] \\
&\quad = 2i\,\frac{z_{\mathrm{sp}}-1+O(v^2)}{(z_{\mathrm{sp}}+O(v^2))((z_{\mathrm{sp}}-1)^2+v^2-2c_2(z_{\mathrm{sp}}-1)\,v^2+O(v^4))},
\end{aligned}$$

where we used (2.30) and the fact that z_{sp} and c_2 are real with $z_{\mathrm{sp}} > 1$. Therefore, we get for the expression in (2.86)

$$\begin{aligned}
\frac{1}{\pi}\int_0^{\frac{1}{2}\delta} &\frac{1}{z_{\mathrm{sp}}+O(v^2)}\,\frac{z_{\mathrm{sp}}-1+O(v^2)}{(z_{\mathrm{sp}}-1)^2+v^2+O((z_{\mathrm{sp}}-1)v^2)+O(v^4)} \\
&\cdot\ln\Big(1-B\exp(-\tfrac{1}{2}s\,\alpha v^2)\Big)\,\mathrm{d}v. \qquad (2.87)
\end{aligned}$$

In the case that $2\eta - 1 < 0$, we have as earlier that the whole expression in (2.87) is $O(\exp(-b^2\, s^{1-2\eta}))$ for any $b \in (0, b_0)$, as $s \to \infty$. In the case that $2\eta - 1 \geq 0$, we substitute $v = t\sqrt{s/(2\,\alpha)}$, and we get as earlier for the expression (2.87), assuming also that $\eta < 1$,

$$\begin{aligned}
&\frac{1}{\pi}\sqrt{s\,\alpha/2}\int_0^\infty \frac{z_{\rm sp}-1+O(t^2/s)}{(d^2(s)+t^2)\,(1+O(s^{-\eta})+O(t^2/s))}\ln(1-B\,\mathrm{e}^{-t^2})\,\mathrm{d}t\\
&= \frac{1}{\pi}\int_0^\infty \frac{\sqrt{s\,\alpha/2}\;(z_{\rm sp}-1)}{d^2(s)+t^2}\ln(1-B\,\mathrm{e}^{-t^2})\,\mathrm{d}t\,\Big(1+O(s^{\eta-1})\Big)\\
&= \frac{1}{\pi}\int_0^\infty \frac{d(s)}{d^2(s)+t^2}\ln(1-\mathrm{e}^{-d^2(s)-t^2})\,\mathrm{d}t\,\Big(1+O(s^{\eta-1})\Big)\,.
\end{aligned}$$

Here we also used (2.44) and that $1/s^{3\eta-1} = O(d^2(s)/s^\eta)$, so that

$$(\tfrac{1}{2}\,s\,\alpha)^{1/2}\,(z_{\rm sp}-1) = d(s)\,\big(1+O(s^{-\eta})\big) = d(s)\,\Big(1+O(s^{\eta-1})\Big)\,,$$

since $\eta \geq 1/2$.

We have for $b > 0$

$$\frac{1}{\pi}\int_0^\infty \frac{b}{b^2+t^2}\ln(1-\exp(-b^2-t^2))\mathrm{d}t = -\frac{1}{2}\sum_{k=0}^\infty \frac{1}{k+1}\,\mathrm{erfc}(b\sqrt{k+1}) = -F(b\sqrt{2}), \tag{2.88}$$

where according to [115, Eq. (3.3) & (3.12)] for $\beta > 0$

$$\begin{aligned}
F(\beta) &= \sum_{n=1}^\infty \frac{1}{n}\frac{1}{\sqrt{2\pi}}\int_{\beta\sqrt{n}}^\infty \mathrm{e}^{-x^2/2}dx\\
&= -\ln\beta - \frac{1}{2}\ln 2 - \frac{1}{\sqrt{2\pi}}\sum_{r=0}^\infty \frac{\zeta(1/2-r)\,(-1/2)^r\,\beta^{2r+1}}{r!\,(2r+1)},
\end{aligned} \tag{2.89}$$

the last identity being valid for $0 < \beta < 2\sqrt{\pi}$.

Using (2.89) with $\beta^2 = d^2(s) = b_0^2/s^{2\eta-1}$, with the entire series on the second line being $O(\beta)$, we get the leading order behavior of $\ln[\mathbb{P}(Q=0)]$ as

$$\Big(-(\eta-1/2)\ln s + \ln(2\,b_0) + O(s^{1/2-\eta})\Big)\,\Big(1+O(s^{\eta-1})\Big) \tag{2.90}$$

when $\eta \in (1/2, 1)$. For $\eta = 1/2$, we get the leading order behavior, assuming $0 < b_0 < \sqrt{2\pi}$,

$$\ln(2\,b_0) + \frac{1}{\sqrt{\pi}}\sum_{r=0}^\infty \frac{\zeta(1/2-r)\,(-1)^r}{r!\,(2r+1)}\,b_0^{2r+1} \tag{2.91}$$

with relative error $O(s^{-1/2})$. The expression (2.91) coincides with $\ln[\mathbb{P}(M=0)]$ as given by [115, Eq. (2.1)] with $\beta = b_0\sqrt{2}$. The next two results summarize the above.

Theorem 2.3. *For $\eta \in (1/2, 1)$,*

$$\ln[\mathbb{P}(Q=0)] = -F\big(d(s)\sqrt{2}\big)\,\Big(1+O(s^{\eta-1})\Big)$$

with F given by (2.89).

Proposition 2.3. *For $\eta \in (0, 1/2)$, and for all $b < b_0$,*

$$\ln[\mathbb{P}(Q = 0)] = O(\exp(-b^2\, s^{1-2\eta})).$$

For $\eta = 1/2$, $\ln[\mathbb{P}(Q = 0)]$ equals $-F(b_0\sqrt{2})$ with a relative error $O(1/\sqrt{s})$. For $\eta \in (1/2, 1)$ and $0 < b_0 < \sqrt{2\pi}$, $\ln[\mathbb{P}(Q = 0)]$ has leading order behavior as in (2.90).

As in Section 2.5.1 for the mean congestion level case with $\eta = 1/2$, it is possible to give a correction term which involves now the integrals in (2.88) and (2.57).

2.6 Numerical examples

2.6.1 Accuracy of the approximations

In this subsection we present a numerical example that serves to illustrate the accuracy of the derived heavy-traffic approximations. Consider the Poisson case

$$\tilde{X}(z) = \mathrm{e}^{z-1}, \quad \mu = \sigma^2 = 1.$$

We fix μ and vary n with the value of s, according to

$$\vartheta = \frac{n}{s} = 1 - \frac{\gamma}{s^{\eta}}$$

for some $\gamma > 0$ and $\eta \geq 1/2$. To calculate the exact value of the mean congestion level we use the expression, see [45],

$$\mu_Q = \frac{\sigma_A^2}{2(s - \mu_A)} - \frac{s - 1 + \mu_A}{2} + \sum_{k=1}^{s-1} \frac{1}{1 - z_k}.$$

Here $z_1, \ldots, z_{s-1}$ are the zeros of $z^s - A(z)$ in $|z| < 1$. We apply the method of successive substitution described in [113] to obtain accurate numerical approximations for $z_1, ..., z_{s-1}$ and consequently μ_Q.

From Theorem 2.1, we find that the leading order behavior of μ_Q is given by

$$\frac{\sqrt{2s}}{\pi}\, G_0\Big(\frac{\gamma}{\sqrt{2}\, s^{\eta-\frac{1}{2}}}\Big). \tag{2.92}$$

In order to find the correction terms, we proceed by setting $\eta = 1/2$. Deriving constants C_1, C_2, C_3, and C_4 for our setting and substituting these into (2.70), we get for μ_Q, with an absolute error of $O(s^{-1/2})$, the approximation

$$\frac{\sqrt{2\,s}}{\pi}\Big(\Big(1 - \frac{\gamma}{3\sqrt{s}}\Big)\, G_0(b_0) - \frac{\gamma^3}{3\sqrt{s}}\,(\,G_3(b_0) + G_4(b_0))\Big),$$

which by (2.55) and (2.73) reduces to

$$\frac{\sqrt{2\,s}}{\pi}\, G_0(b_0) - \frac{\sqrt{2}\,\gamma}{3\,\pi}\, G_1(b_0). \tag{2.93}$$

s	ρ	μ_Q	(2.92)	(2.93)
10	0.683	0.244	0.399	0.247
20	0.776	0.410	0.565	0.412
50	0.858	0.739	0.893	0.741
100	0.900	1.110	1.263	1.111
200	0.929	1.633	1.787	1.634
500	0.955	2.672	2.825	2.673
1000	0.968	3.843	3.996	3.843

Table 2.1: Numerical results for $\gamma = 1$.

s	ρ	μ_Q	(2.92)	(2.93)
10	0.968	13.707	14.046	13.732
20	0.977	19.533	19.865	19.551
50	0.985	31.084	31.409	31.095
100	0.990	44.097	44.419	44.106
200	0.992	62.499	62.819	62.505
500	0.995	99.008	99.325	99.011
1000	0.996	140.152	140.468	140.154

Table 2.2: Numerical results for $\gamma = 0.1$.

	$\eta = 0.6$		$\eta = 0.75$		$\eta = 0.9$	
s	μ_Q	(2.92)	μ_Q	(2.92)	μ_Q	(2.92)
10	17.781	18.125	25.970	26.318	37.553	37.905
20	27.309	27.647	44.391	44.734	71.195	71.541
50	47.948	48.281	89.623	89.961	164.637	164.978
100	73.245	73.574	152.031	152.367	309.353	309.692
200	111.752	112.079	257.435	257.769	580.170	580.507
500	195.082	195.409	515.443	515.776	1329.581	1329.917
1000	297.122	297.448	870.524	870.857	2487.227	2487.562

Table 2.3: Numerical results for $\gamma = 0.1$ and several values of η.

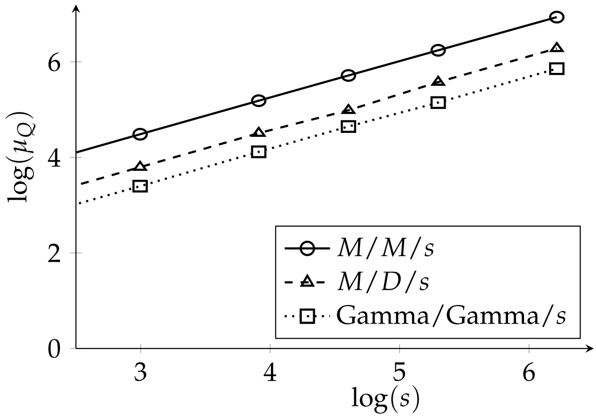


Figure 2.1: μ_Q plotted against s on log scale for 3 queues for $\eta = 0.75$.

Numerical results for $\eta = 1/2$ and various values of s are given in Table 2.1 and 2.2, for $\gamma = 1$ and $\gamma = 0.1$, respectively. We note that for small s the leading order approximation is still off by a significant amount, while the refinement only shows an error in the second decimal for $\gamma = 0.1$. This seems to justify the use of the correction term. In Table 2.3 we compare the approximation (2.92) against the exact value of μ_Q for three values of $\eta \geq 1/2$ to assess the influence of η. Clearly, the leading order approximation is relatively accurate for all three scenarios. As expected, the mean congestion increases along with η, since utilization approaches 1 more rapidly in this case.

2.6.2 Connection to other queueing models

As argued in the introduction, we believe that the heavy-traffic behavior for the discrete model in this chapter will up to leading order be universal for a wide range of other models (when subjected to the same heavy-traffic regime (2.2)). We shall now substantiate this for many-server systems, for which under (2.2), it turns out that the mean congestion is $O(s^\eta)$. We compare the mean congestion level in our discrete queue with that in the multi-server systems $M/M/s$, $M/D/s$ and Gamma/Gamma/s, all with unit mean service time and occupation rate $1 - \gamma/s^\eta$.

Figure 2.1 shows on logarithmic scale the mean congestion levels for $\gamma = 0.1$ and $\eta = 0.75$ under the specified scaling for three systems. We also display three lines with slope 0.75 for comparison, which confirms that mean congestion levels are of the order s^η, also in these multi-server system. Formally establishing this heavy-traffic behavior for these multi-server system is an important open problem and requires other mathematical approaches than the ones taken in this chapter (see the introduction for more details).

Figure 2.2 shows the mean queue length in the $M/M/s$ system for several values of η, again on logarithmic scale, together with lines with slope η. For $\eta \geq 1/2$, we

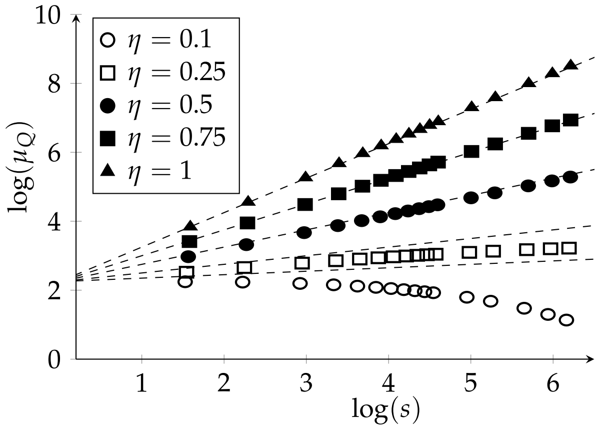


Figure 2.2: μ_Q of $M/M/s$ plotted against s on log scale for different values of η.

see the same $O(s^\eta)$ behavior, similar as for μ_Q in our discrete model. For $\eta < 1/2$ the mean queue length decays, again in agreement with our results for μ_Q. We note that this qualitative behavior of the $M/M/s$ system was also observed by [150, Thm. 4.1], by proving that the mean waiting time in the $M/M/s$ queue under (2.2) is of order $1/s^{1-\eta}$, which by Little's law implies the mean queue length to be of order s^η.

3
Overdispersion

Arrival processes to service systems often display fluctuations that are larger than anticipated under the Poisson assumption, a phenomenon that is referred to as *overdispersion*. Motivated by this, we analyze a class of discrete stochastic models for which we derive heavy-traffic approximations that are scalable in the system size. Subsequently, we show how this leads to novel capacity sizing rules that acknowledge the presence of overdispersion. This, in turn, leads to robust approximations for performance characteristics of systems that are of moderate size and/or may not operate in heavy traffic.

Based on
Robust heavy-traffic approximations for service systems facing overdispersed demand
Britt Mathijsen, Guido Janssen, Johan van Leeuwaarden & Bert Zwart
arxiv.org/abs/1512.05581

3.1 Introduction

In the previous chapter, we analyzed the scaling limit of a queueing model in which demand exhibits stochastic fluctuations that are asymptotically proportional to the square-root of the nominal load, while we deliberately chose to deviate from the square-root staffing principle by allocating a variability hedge that does not match the order of these fluctuations. This chapter in some ways does the opposite. We assume the demand faced by the queueing system is more volatile than anticipated by the independent many-sources paradigm that leads to Poisson traffic models. As will become clear in this chapter, this in fact *requires* an adaptation of the square-root staffing principle in order to maintain the desirable properties of the QED regime. We start by motivating our research through empirical evidence of the presence of so-called *overdispersion* in arrival processes faced by service systems reported by recent literature.

Motivation. The bulk of the queueing literature assumes perfect knowledge about the model primitives, including the mean demand per time period. For large-scale service systems, like health care facilities, communication systems or call centers, the dominant assumption is that demand arrives according to a (non)homogeneous Poisson process, which in practice translates into the assumption that arrival rates are known for each basic time period (second, hour or day). Although natural and convenient from a mathematical viewpoint, the Poisson assumption often fails to be confirmed in practice. A deterministic arrival rate implies that the demand over any given period is a Poisson random variable, whose variance equals its expectation. A growing number of empirical studies shows that the variance of demand typically deviates from the mean significantly. Recent work [137, 139] reports variance being strictly less than the mean in health care settings employing appointment booking systems. This reduced variability, known as underdispersion, can be accredited to the goal of the booking system to create a more predictable arrival pattern. On the other hand, in other scenarios with no control over the arrivals, the variance typically dominates the mean, see [26, 29, 30, 49, 58, 81, 99, 127, 138, 150, 165, 187, 200, 228]. The feature that variability is higher than one expects from the Poisson assumption is referred to as overdispersion. The latter concept will be the center of our attention in this chapter.

Stochastic models with the Poisson assumption have been widely applied to optimize capacity levels in service systems. The goal is to minimize operating costs while providing sufficiently high QoS in terms of performance measures such as mean delay or excess delay. When stochastic models, however, do not take into account overdispersion, resulting performance estimates are likely to be overoptimistic. The system then ends up being underprovisioned, which possibly causes severe performance problems, particularly under critical loading.

Causes of overdispersion. The literature discussed above proves that the presence of overdispersion is widespread across applications. It however does not specify what causes the increased variability in the arrival process. We name two possible explanations.

First, we revisit the many-sources characterization of demand inflow discussed in Chapter 2. Recall that in this setting, demand is generated by n stochastically identical and independent sources, with n large, so that workload arriving to the system in period j is given by $A_j^{(n)} = \sum_{i=1}^{n} A_{i,j}$, where $A_{i,j}$, $i = 1,2,\ldots,n$ are i.i.d. random variables. This resulted in nominal workload $\mu_n = n\mu$ and $\sigma_n^2 = n\sigma^2$, thus both of order n. If we now relax the assumption on the (pairwise) independence of the sources, but rather consider the scenario in which these are positively correlated, then the nominal load remains to be equal to $n\mu$, while the variance of demand becomes

$$\sigma_n^2 = \mathrm{Var}\, A_j^{(n)} = n\,\mathrm{Var}\, A_{1,j} + n(n-1)\,\mathrm{Cov}(A_{1,j}, A_{2,j}),$$

which is of higher order than n if $n\,\mathrm{Cov}(A_{1,j}, A_{2,j}) \to \infty$ as $n \to \infty$.

A second interpretation of overdispersion in arrival processes relates to *arrival rate uncertainty*. The canonical process for modeling the arrival process of a service system is the Poisson process with a given arrival rate λ. Since model primitives, in particular the arrival rate, are typically estimated through historical data, these are prone to be subject to forecasting errors. In the realm of Poisson processes, this inherent uncertainty can be acknowledged by viewing the arrival rate Λ_n itself as being stochastic. The resulting doubly stochastic Poisson process, also known as Cox process (first presented in [62]), implies that demand in a given interval A_j follows a mixed Poisson distribution. In this case, the expected demand per period equals $\mu_n = \mathbb{E}[\Lambda_n]$, while the variance is $\sigma_n^2 = \mathbb{E}[\Lambda_n] + \mathrm{Var}\,\Lambda_n$. By selecting the distribution of the mixing factor Λ_n, the magnitude of overdispersion can be made arbitrarily large, and only a deterministic Λ_n leads to a true Poisson process.

The mixed Poisson model presents a useful way to fit both the mean and variance to real data, particularly in case of overdispersion. The mixing distribution can be estimated parametrically or non-parametrically, see [127, 150]. A popular parametric family is the Gamma distribution, which gives rise to an effective data fitting procedure that uses the fact that a Gamma mixed Poisson random variable follows a negative binomial distribution. We will in this chapter adopt the assumption of a Gamma-Poisson mixture as the demand process.

Adapted QED scaling. To deal with overdispersion new models are needed, scaling rules must be adapted, and existing capacity sizing rules need to be modified in order to incorporate a correct hedge against (increased) variability. In this chapter, we consider an extension of the discrete queueing model of Chapter 2 that has a doubly stochastic Poisson process as input, $A_j \sim \mathrm{Pois}(\Lambda_n)$ and we identify the heavy-traffic regime in which it displays QED behavior. That is, it fits the three asymptotic characteristics in Section 1.2.3 of this thesis. As we argued in that particular section, a sensible candidate capacity allocation rule is $s_n = \mu_n + \beta\sigma_n$ for some $\beta > 0$, which is equivalent to the scaling

$$\frac{\mu_n}{\sigma_n}(1-\rho_n) \to \beta, \qquad \text{as } n \to \infty.$$

We will verify mathematically that this is asymptotically the appropriate choice. Studies that have adressed similar capacity allocation problems with stochastic arrival rates include [144, 150, 220, 223]. Of the aforementioned papers, our work best relates to [150], in the sense that we also assess the asymptotic performance of a queueing system having a stochastic arrival rate in heavy traffic. We therefore expand the paradigm of the QED regime, in order to have it accommodate for overdispersed demand that follows from a doubly stochastic Poisson process.

Structure of the chapter. The remainder of this chapter is structured as follows. Our model is introduced in Section 3.2 together with some preliminary results. In Section 3.3 we derive the classical heavy-traffic scaling limits for the queue length process in the presence of overdispersed arrivals both for the moments and the distribution itself. Section 3.4 presents our main theoretic result, which provides a robust refinement to the heavy-traffic characterization of the queue length measures in pre-limit systems. In Section 3.5, we describe the numerical results and demonstrate the heavy-traffic approximation for a real data set coming from a health care setting. Section 3.6 provides some concluding remarks.

3.2 Model description

We consider the same mathematical model as in Section 2.2, in which time is divided into periods of equal length. At the beginning of each period $j = 1, 2, 3, \ldots$ new demand $A_j^{(n)}$ arrives to the system. The demands per period $A_1^{(n)}, A_2^{(n)}, \ldots$ are assumed independent and equal in distribution to some non-negative integer-valued random variable $A^{(n)}$. The system has a service capacity $s_n \in \mathbb{N}$ per period, the steady-state queue length can be characterized as, see (1.27),

$$Q^{(n)} \stackrel{d}{=} \max_{k\geq 0}\Big\{\sum_{i=1}^{k}(A_i^{(n)} - s_n)\Big\}. \tag{3.1}$$

For brevity, we define $\mu_n := \mathbb{E}[A_1^{(n)}]$ and $\sigma_n^2 = \operatorname{Var} A_1^{(n)}$. The behavior of $Q^{(n)}$ predominantly depends on the characteristics of $A^{(n)}$ and s_n. As noted before, $\mu_n < s_n$ is a necessary condition for the maximum in (3.1) to be finite and consequently for the queue to be stable. Before continuing the analysis of $Q^{(n)}$, we impose a set of conditions on the asymptotic properties of s_n, μ_n and σ_n.

Assumption 3.1.

(a) (Asymptotic growth)

$$\mu_n, \sigma_n \to \infty, \quad \text{for } n \to \infty.$$

(b) (Persistence of overdispersion)

$$\sigma_n^2/\mu_n \to \infty \quad \text{for } n \to \infty.$$

(c) (Heavy-traffic condition) *The utilization* $\rho_n := \mu_n/s_n \to 1$ *as* $n \to \infty$, *while*

$$s_n = \mu_n + \beta\,\sigma_n, \tag{3.2}$$

for some $\beta > 0$. *This is equivalent to requiring*

$$(1-\rho_n)\frac{\mu_n}{\sigma_n} \to \beta, \qquad \text{for } n \to \infty. \tag{3.3}$$

Assumption 3.1 is assumed to hold throughout the remainder of this chapter. Since we are mainly interested in the system behavior in heavy traffic, it is appropriate to study the queue length process in a scaled form. Substituting s_n as in Assumption 3.1(c), and dividing both sides of (3.1) by σ_n, gives

$$\frac{Q^{(n)}}{\sigma_n} = \max_{k\geq 0}\Big\{\sum_{i=1}^{k}\Big(\frac{A_i^{(n)}-\mu_n}{\sigma_n} - \beta\Big)\Big\}. \tag{3.4}$$

By defining $\hat{Q}^{(n)} := Q^{(n)}/\sigma_n$ and $\hat{A}_i^{(n)} := (A_i^{(n)} - \mu_n)/\sigma_n$, we see that the scaled queue length process is in distribution equal to the maximum of a random walk with i.i.d. increments distributed as $\hat{A}^{(n)} - \beta$. Besides $\mathbb{E}[\hat{A}^{(n)}] = 0$ and $\operatorname{Var}\hat{A}^{(n)} = 1$, the scaled and centered arrival count $\hat{A}^{(n)}$ has a few other nice properties which we turn to later in this section.

The model in (3.1) is valid for any distribution of $A^{(n)}$, also for the original case where the number of arrivals follows a Poisson distribution with fixed parameter λ_n, but in that case Assumption 3.1(b) does not hold. Instead, we assume $A^{(n)}$ to be Poisson distributed with uncertain arrival rate rendered by the non-negative random variable Λ_n. This Λ_n is commonly referred to as the *prior* distribution, while $A^{(n)}$ is given the name of a Poisson mixture, see [87]. Given that the moment generation function of Λ_n, denoted by $M_n^{\Lambda}(\cdot)$, exists, we are able to express the probability generating function (pgf) of $A^{(n)}$ through the former. Namely,

$$\tilde{A}^{(n)}(z) = \mathbb{E}[\mathbb{E}[z^{A^{(n)}}|\Lambda_n]] = \mathbb{E}[\exp(\Lambda_n(z-1))] = M_n^{\Lambda}(z-1). \tag{3.5}$$

From (3.5), we get

$$\mu_n = \mathbb{E}[A^{(n)}] = \mathbb{E}[\Lambda_n], \qquad \sigma_n^2 = \operatorname{Var} A^{(n)} = \operatorname{Var}\Lambda_n + \mathbb{E}[\Lambda_n], \tag{3.6}$$

so that $\mu_n < \sigma_n^2$ if Λ_n is non-deterministic. Assumption 3.1(b) hence translates to

$$\operatorname{Var}\Lambda_n/\mathbb{E}[\Lambda_n] \to \infty, \qquad n \to \infty.$$

The next result relates the converging behavior of the centered and scaled Λ_n to that of $\hat{A}^{(n)}$.

Lemma 3.1. *Let $\mu_n, \sigma_n^2 \to \infty$ and $\sigma_n^2/\mu_n \to \infty$. If*

$$\hat{\Lambda}_n := \frac{\Lambda_n - \mu_n}{\sigma_n} \overset{d}{\Rightarrow} \mathcal{N}(0,1), \qquad \text{for } n \to \infty,$$

then $\hat{A}^{(n)}$ converges weakly to a standard normal variable as $n \to \infty$.

The proof can be found in Appendix 3.A. The prevalent choice for Λ_n is the Gamma distribution. The Gamma-Poisson mixture turns out to provide a very good fit to arrival counts experienced by service systems, as was observed by [127]. Assuming Λ_n to be of Gamma type with scale and rate parameters a_n and $1/b_n$, respectively, we get for the pgf of $A^{(n)}$:

$$\tilde{A}^{(n)}(z) = \Big(\frac{1}{1 + b_n(1-z)}\Big)^{a_n}, \tag{3.7}$$

in which we recognize the pgf of a negative binomial distribution with parameters a_n and $1/(b_n+1)$, so that

$$\mu_n = a_n b_n, \qquad \sigma_n^2 = a_n b_n (b_n + 1).$$

Note that in the context of a Gamma prior, the restrictions in Assumption 3.1 reduce to only two rules. For completeness, we include the revised list below.

Assumption 3.2.

1. (Asymptotic regime and persistence of overdispersion)

$$a_n, b_n \to \infty, \quad \text{for } n \to \infty.$$

2. (Heavy-traffic condition) *Let*

$$s_n = a_n b_n + \beta\sqrt{a_n b_n (b_n + 1)},$$

for some $\beta > 0$, or equivalently

$$(1-\rho_n)\sqrt{a_n} \to \beta, \quad \text{for } n \to \infty.$$

The next result follows from the fact that Λ_n is a Gamma random variable:

Corollary 3.1. *Let $\Lambda_n \sim \mathrm{Gamma}(a_n, 1/b_n)$, $A^{(n)} \sim \mathrm{Pois}(\Lambda_n)$ and $a_n, b_n \to \infty$. Then $\hat{A}^{(n)}$ converges weakly to a standard normal random variable as $n \to \infty$.*

Proof. By Lemma 3.1, it is sufficient to prove that $\hat{\Lambda}_n \overset{d}{\Rightarrow} \mathcal{N}(0,1)$ for this particular choice of Λ_n. We do this by proving the pointwise convergence of the characteristic function (cf) of $\hat{\Lambda}_n$ to $\exp(-t^2/2)$, the cf of the standard normal distribution. Let

$\phi_G(\cdot)$ denote the characteristic function of a random variable G. By basic properties of the cf,

$$\begin{aligned}
\phi_{\hat{\Lambda}_n}(t) &= \mathrm{e}^{-i\mu_n t/\sigma_n}\,\phi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n}\Big(1-\frac{ib_n t}{\sigma_n}\Big)^{-a_n} \\
&= \exp\Big[-\frac{i\mu_n t}{\sigma_n} - a_n \ln\Big(1-\frac{ib_n t}{\sigma_n}\Big)\Big] \\
&= \exp\Big[-\frac{i\mu_n t}{\sigma_n} - a_n\Big(-\frac{i\,b_n t}{\sigma_n}+\frac{b_n^2 t^2}{2\sigma_n^2}+O(b_n^3/\sigma_n^3)\Big)\Big] \\
&= \exp\Big[-\frac{b_n\,t^2}{2(b_n+1)}+O\left(1/\sqrt{a_n}\right)\Big] \to \exp\left(-t^2/2\right),
\end{aligned}$$

for $n\to\infty$. By Lévy's continuity theorem this implies $\hat{\Lambda}_n$ is indeed asymptotically standard normal. □

The characterization of the arrival process as a Gamma-Poisson mixture is of vital importance in later sections.

Expressions for the stationary distribution. Our main focus is on the stationary queue length distribution, denoted by

$$\mathbb{P}(Q^{(n)}=i) = \lim_{k\to\infty}\mathbb{P}(Q^{(n)}(k)=i).$$

Denote the pgf of $Q^{(n)}$ by

$$\tilde{Q}^{(n)}(w) := \sum_{i=0}^{\infty}\mathbb{P}(Q^{(n)}=i)w^i.$$

Furthermore, let $\mu_Q := \mathbb{E}[Q^{(n)}]$ and $\sigma_Q^2 := \mathrm{Var}\,Q^{(n)}$ denote the stationary mean and variance of the queue length, respectively. To avoid notational complexity, we omit the superscript (n) in these definitions. To continue our analysis of $Q^{(n)}$, we need one more condition on $A^{(n)}$.

Assumption 3.3. *The pgf of $A^{(n)}$, denoted by $\tilde{A}^{(n)}(w)$, exists for $|z|<r_0$, for some $r_0>1$, so that all moments of $A^{(n)}$ are finite.*

We next recall two characterizations of $\tilde{Q}^{(n)}(w)$ that play prominent roles in the remainder of our analysis. The first characterization of $\tilde{Q}^{(n)}(w)$ originates from a random walk perspective. As we see from (3.1), the (scaled) stationary queue length is equal in distribution to the all-time maximum of a random walk with i.i.d. increments distributed as $A^{(n)}-\beta$ (or $\hat{A}^{(n)}-\beta$ in the scaled setting). Spitzer's identity, see e.g. [20, Theorem VIII4.2] and Section 1.2.2 of this thesis, then gives

$$\tilde{Q}^{(n)}(w) = \exp\left\{\sum_{k=1}^{\infty}\frac{1}{k}\left(\mathbb{E}\Big[w^{\left(\sum_{i=1}^{k}\{A_i^{(n)}-s_n\}\right)^{+}}\Big]-1\right)\right\},$$

where $(x)^+ = \max\{x, 0\}$. Hence,

$$\mu_Q = \mathbb{E}[Q^{(n)}] = \tilde{Q}^{(n)\prime}(1) = \sum_{k=1}^{\infty} \frac{1}{k}\, \mathbb{E}\Big[\sum_{i=1}^{k}(A_i^{(n)} - s_n)\Big]^+,$$

$$\sigma_Q^2 = \operatorname{Var} Q^{(n)} = \tilde{Q}^{(n)\prime\prime}(1) + Q^{(n)\prime}(1) - \Big(\tilde{Q}^{(n)\prime}(1)\Big)^2 = \sum_{k=1}^{\infty} \frac{1}{k}\, \mathbb{E}\Big[\Big(\sum_{i=1}^{k}(A_i^{(n)} - s_n)\Big)^+\Big]^2,$$

$$\mathbb{P}(Q^{(n)} = 0) = \tilde{Q}_n(0) = \exp\Big\{-\sum_{k=1}^{\infty} \frac{1}{k}\, \mathbb{P}\Big(\textstyle\sum_{i=1}^{k}(A_i^{(n)} - s_n) > 0\Big)\Big\}.$$

A second characterization follows from Pollaczek's formula, see [2] and Section 2.2.2 of this thesis:

$$\tilde{Q}^{(n)}(w) = \exp\Big\{\frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{w-z}{1-z}\Big)\, \frac{(z^{s_n} - \tilde{A}^{(n)}(z))'}{z^{s_n} - \tilde{A}^{(n)}(z)} \mathrm{d}z\Big\}, \tag{3.8}$$

which is analytic for $|w| < r_0$, for some $r_0 > 1$. Therefore, $\varepsilon > 0$ has to be chosen such that $|w| < 1 + \varepsilon < r_0$. This gives

$$\mu_Q = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{1}{1-z}\, \frac{(z^{s_n} - \tilde{A}^{(n)}(z))'}{z^{s_n} - \tilde{A}^{(n)}(z)} \mathrm{d}z, \tag{3.9}$$

$$\sigma_Q^2 = \frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \frac{-z}{(1-z)^2}\, \frac{(z^{s_n} - \tilde{A}^{(n)}(z))'}{z^{s_n} - \tilde{A}^{(n)}(z)} \mathrm{d}z, \tag{3.10}$$

$$\mathbb{P}(Q^{(n)} = 0) = \exp\Big\{\frac{1}{2\pi i}\int_{|z|=1+\varepsilon} \ln\Big(\frac{z}{z-1}\Big)\, \frac{(z^{s_n} - \tilde{A}^{(n)}(z))'}{z^{s_n} - \tilde{A}^{(n)}(z)} \mathrm{d}z\Big\}. \tag{3.11}$$

Pollaczek-type integrals like (3.8)-(3.11) first occurred in the work of Pollaczek on the classical single-server queue (see [2, 61, 117] for historical accounts). These integrals are fairly straightforward to evaluate numerically and hence give rise to efficient algorithms for performance evaluation [2, 41]. The integrals also proved useful in establishing heavy-traffic results by asymptotic evaluation of the integrals in various heavy-traffic regimes [141, 61, 118, 40], and in this paper we follow that approach for a heavy-traffic regime that is suitable for overdispersion.

3.3 Heavy-traffic limits

In this section we present the result on the convergence of the discrete process $\hat{Q}^{(n)}$ to a non-degenerate limiting process and of the associated stationary moments. The latter requires an interchange of limits. Using this asymptotic result, we derive two sets of approximations for μ_Q, σ_Q^2 and $\mathbb{P}(Q^{(n)} = 0)$, that capture the limiting behavior of $Q^{(n)}$. The first set provides a rather crude estimation for the first cumulants of the queue length process for any arrival process $A^{(n)}$ satisfying Assumption 3.1.

The second set, which is the subject of the next section, is derived for the specific case of a Gamma prior and is therefore expected to provide more accurate, robust approximations for the performance metrics.

We start by indicating how the asymptotic properties of the scaled arrival process give rise to a proper limiting random variable describing the stationary queue length. The asymptotic normality of $\hat{A}^{(n)}$ provides a link with the Gaussian random walk and nearly deterministic queues [192, 193]. The main results in [192, 193] were obtained under the assumption that $\rho_n \sim 1 - \beta/\sqrt{n}$, in which case it follows from [193, Thm. 3] that the rescaled stationary waiting time process converges to a reflected Gaussian random walk.

We shall also identify the Gaussian random walk as the appropriate scaling limit for our stationary system. However, since the normalized natural fluctuations of our system are given by μ_n/σ_n instead of $\sqrt{n}$, we assume that the load grows like $\rho_n \sim 1 - \frac{\beta}{\mu_n/\sigma_n}$. Hence, in contrast to [192, 193], our systems' characteristics display larger natural fluctuations, due to the mixing factor that renders the arrivals. Yet, by matching this overdispersed demand with the appropriate hedge against variability, we again obtain Gaussian limiting behavior. This is not surprising, since we saw in Lemma 3.1 that the increments start resembling Gaussian behavior for $n \to \infty$. The following result summarizes this.

Theorem 3.1. *Let Λ_n be a non-negative random variable such that $(\Lambda_n - \mu_n)/\sigma_n$ is asymptotically standard normal, with μ_n and σ_n as defined in* (3.6), *and $\mathbb{E}[\Lambda_n^3] < \infty$ for all $n \in \mathbb{N}$. Then under Assumption 3.1, for $n \to \infty$,*

(i) $\hat{Q}^{(n)} \overset{d}{\Rightarrow} M_\beta$,

(ii) $\mathbb{P}(Q^{(n)} = 0) \to \mathbb{P}(M_\beta = 0)$,

(iii) $\mathbb{E}[\hat{Q}^{(n)}] \to \mathbb{E}[M_\beta]$,

(iv) $\mathrm{Var}\,\hat{Q}_n \to \mathrm{Var}\ M_\beta$,

where M_β is the all-time maximum of a random walk with i.i.d. normal increments with mean $-\beta$ and unit variance.

The proof of Theorem 3.1 is given in Appendix 3.A. The following result shows that Theorem 3.1 also applies to Gamma mixtures, which is a direct consequence of Corollary 3.1.

Corollary 3.2. *Let $\Lambda_n \sim$* $\mathrm{Gamma}(a_n, b_n)$. Then under Assumption 3.2 the four convergence results of Theorem 3.1 hold true.

It follows from Theorem 3.1 that the scaled stationary queueing process converges under (3.3) to a reflected Gaussian random walk. Hence, the performance measures of the original system should be well approximated by the performance measures of the reflected Gaussian random walk, yielding heavy-traffic approximations.

Like our original system, the Gaussian random walk falls in the classical setting of the reflected one-dimensional random walk, whose behavior is characterized by both Spitzer's identity and Pollaczek's formula. In particular, Pollaczek's formula gives rise to contour integral expressions for performance measures that are easy to evaluate numerically, also in heavy-traffic conditions. The numerical evaluation of such integrals is considered in [2]. For $\mathbb{E}[M_\beta]$ such an integral is as follows

$$\mathbb{E}[M_\beta] = -\frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big[\frac{1-\phi(-z)}{z^2}\Big]\mathrm{d}y, \tag{3.12}$$

where $z = x + iy$ with an appropriately chosen real part x, with $\phi(z) = \exp(-\beta z + \frac{1}{2}z^2)$, the Laplace transform of a normal random variable with mean $-\beta$ and unit variance. Note that this integral involves complex-valued functions with complex arguments. Similar Pollaczek-type integrals exist for $\mathbb{P}(M_\beta = 0)$ and $\mathrm{Var}\, M_\beta$, see [2]. The following result simply rewrites these integrals in terms of a real integral and uses the fact that the scaled queue length process mimics the maximum of the Gaussian random walk for large n.

Corollary 3.3. *Under Assumption 3.1, the leading order behavior of $\mathbb{P}(Q^{(n)} = 0)$, μ_Q and σ_Q^2 as $n \to \infty$ are given by, respectively,*

$$\exp\Big[\frac{1}{\pi}\int_0^\infty \frac{\beta/\sqrt{2}}{\frac{1}{2}\beta^2 + t^2}\ln\Big(1 - e^{-\frac{1}{2}\beta^2 - t^2}\Big)\mathrm{d}t\Big], \tag{3.13}$$

$$\frac{\sqrt{2}\sigma_n}{\pi}\int_0^\infty \frac{t^2}{\frac{1}{2}\beta^2 + t^2}\,\frac{\exp(-\frac{1}{2}\beta^2 - t^2)}{1-\exp(-\frac{1}{2}\beta^2 - t^2)}\mathrm{d}t, \tag{3.14}$$

$$\frac{\sqrt{2}\beta\sigma_n^2}{\pi}\int_0^\infty \frac{t^2}{(\frac{1}{2}\beta^2 + t^2)^2}\,\frac{\exp(-\frac{1}{2}\beta^2 - t^2)}{1-\exp(-\frac{1}{2}\beta^2 - t^2)}\mathrm{d}t. \tag{3.15}$$

Proof. According to [2, Eq. (15)],

$$-\ln[\mathbb{P}(M_\beta = 0)] = c_0, \quad \mathbb{E}[M_\beta] \;= c_1, \quad \mathrm{Var}\; M_\beta = c_2,$$

where

$$c_n = \frac{(-1)^n n!}{\pi}\mathrm{Re}\Big[\int_0^\infty \frac{\ln(1-\exp(\beta z + \frac{1}{2}z^2))}{z^{n+1}}\mathrm{d}y\Big],$$

in which $z = -x + i\,y$, $y \geq 0$, and x is any fixed number between 0 and 2β. Take $x = \beta$, so that

$$\beta z + \tfrac{1}{2}z^2 = -\tfrac{1}{2}\beta^2 - \tfrac{1}{2}y^2 \leq 0, \quad y \geq 0.$$

For $n = 0$, this gives

$$\begin{aligned}
c_0 &= \frac{1}{\pi}\mathrm{Re}\Big[\int_0^\infty \frac{\ln(1-\exp(-\frac{1}{2}\beta^2 - \frac{1}{2}y^2))}{-\beta + i\,y}\mathrm{d}y\Big] \\
&= -\frac{1}{\pi}\int_0^\infty \frac{\beta}{\beta^2 + y^2}\ln(1-\exp(-\tfrac{1}{2}\beta^2 - \tfrac{1}{2}y^2))\mathrm{d}y \\
&= -\frac{1}{\pi}\int_0^\infty \frac{\beta/\sqrt{2}}{\frac{1}{2}\beta^2 + t^2}\ln(1-\exp(-\tfrac{1}{2}\beta^2 - t^2))\mathrm{d}t,
\end{aligned}$$

where we used that

$$\mathrm{Re}\Big[\frac{1}{-\beta+iy}\Big]=\frac{-\beta}{\beta^2+y^2},$$

together with the substitution $y=t\sqrt{2}$. For $n=1,2,\dots$, partial integration gives

$$\begin{aligned}
c_n &= \frac{(-1)^n n!}{\pi}\,\mathrm{Re}\Big[\int_0^\infty \frac{\ln(1-\exp(-\frac12\beta^2-\frac12 y^2))}{(-\beta+iy)^{n+1}}\mathrm{d}y\\
&= \frac{(-1)^{n-1}(n-1)!}{\pi}\,\mathrm{Im}\Big[\int_0^\infty \ln(1-\exp(-\tfrac12\beta^2-\tfrac12 y^2))\mathrm{d}\Big(\frac{1}{(-\beta+iy)^n}\Big)\Big]\\
&= -\frac{(-1)^{n-1}(n-1)!}{\pi}\mathrm{Im}\Big[\int_0^\infty \frac{y}{(-\beta+iy)^n}\,\frac{\exp(-\frac12\beta^2-\frac12 y^2)}{1-\exp(-\frac12\beta^2-\frac12 y^2)}\mathrm{d}y\Big],
\end{aligned}$$

where we have used that

$$\mathrm{Im}\Big[\frac{\ln(1-\exp(-\frac12\beta^2-\frac12 y^2))}{(-\beta+iy)^n}\Big]\Big|_0^\infty=0.$$

Using

$$\frac{1}{(-\beta+iy)^n}=(-1)^n\,\frac{(\beta+iy)^n}{(\beta^2+y^2)^n},$$

we then get

$$c_n=\frac{(n-1)!}{\pi}\,\mathrm{Im}\,\Big[\int_0^\infty \frac{y(\beta+iy)^n}{(\beta^2+y^2)^n}\,\frac{\exp(-\frac12\beta^2-\frac12 y^2)}{1-\exp(-\frac12\beta^2-\frac12 y^2)}\mathrm{d}y\Big],$$

which after the substitution of $y=t\sqrt{2}$ gives

$$\begin{aligned}
c_1 &= \frac{1}{\pi}\int_0^\infty \frac{y^2}{\beta^2+y^2}\,\frac{\exp(-\frac12\beta^2-\frac12 y^2)}{1-\exp(-\frac12\beta^2-\frac12 y^2)}\mathrm{d}y\\
&= \frac{\sqrt{2}}{\pi}\int_0^\infty \frac{t^2}{\frac12\beta^2+t^2}\,\frac{\exp(-\frac12\beta^2-t^2)}{1-\exp(-\frac12\beta^2-t^2)}\mathrm{d}t, \qquad (3.16)
\end{aligned}$$

$$\begin{aligned}
c_2 &= \frac{2\beta}{\pi}\int_0^\infty \frac{y^2}{(\beta^2+y^2)^2}\,\frac{\exp(-\frac12\beta^2-\frac12 y^2)}{1-\exp(-\frac12\beta^2-\frac12 y^2)}\mathrm{d}y\\
&= \frac{\beta\sqrt{2}}{\pi}\int_0^\infty \frac{t^2}{(\frac12\beta^2+t^2)^2}\,\frac{\exp(-\frac12\beta^2-t^2)}{1-\exp(-\frac12\beta^2-t^2)}\mathrm{d}t.
\end{aligned}$$

□

3.4 Robust heavy-traffic approximations

We shall now establish robust heavy-traffic approximations for the canonical case of Gamma-POisson mixutres; see (3.7). As noted earlier, Gamma mixing yields an arrival process that has a negative binomial distribution, which allows us to establish the detailed asymptotic results in the next theorem.

Theorem 3.2. *Let a_n, b_n and s_n be as in Assumption 3.2. Then the leading order behavior of μ_Q is given by*

$$\frac{\sqrt{2}\,\beta_n}{\pi}\Big(\frac{b_n+\rho_n}{1-\rho_n}\Big)\int_0^\infty \frac{t^2}{\frac{1}{2}\beta_n^2+t^2}\,\frac{\exp(-\frac{1}{2}\beta_n^2-t^2)}{1-\exp(-\frac{1}{2}\beta_n^2-t^2)}\mathrm{d}t\,(1+o(1)), \tag{3.17}$$

where

$$\beta_n^2 = s_n\Big(\frac{1-\rho_n}{b_n+1}\Big)^2\Big(1+\frac{b_n}{\rho_n}\Big). \tag{3.18}$$

Furthermore, the leading order behavior of $\mathbb{P}(Q^{(n)}=0)$ and σ_Q^2 is given by

$$\exp\Big[\frac{1}{\pi}\,\frac{b_n+\rho_n}{b_n+1}\int_0^\infty \frac{\beta_n/\sqrt{2}}{\frac{1}{2}\beta_n^2+t^2}\ln\Big(1-\mathrm{e}^{-\frac{1}{2}\beta_n^2-t^2}\Big)\mathrm{d}t\Big],$$

and

$$\frac{\beta_n^3/\sqrt{2}}{\pi}\Big(\frac{b_n+\rho_n}{1-\rho_n}\Big)^2\Big(\frac{b_n+1}{b_n+\rho_n}+1\Big)\int_0^\infty \frac{t^2}{(\frac{1}{2}\beta_n+t^2)^2}\,\frac{\exp(-\frac{1}{2}\beta_n-t^2)}{1-\exp(-\frac{1}{2}\beta_n^2-t^2)}\mathrm{d}t, \tag{3.19}$$

respectively.

The proof of Theorem 3.2 requires asymptotic evaluation of the Pollaczek-type integrals (3.8)-(3.11), for which shall use the *non-standard* saddle point method—originally proposed by [67] and also applied in Chapter 2 of this thesis—to turn these contour integrals into practical approximations.

In contrast to the setting of Chapter 2, both the relevant saddle point and the analyticity radius tend to one as $n\to\infty$, which is a singular point of the integrand, in the setting with overdispersion. For the proof of Theorem 3.2, we therefore modify the special saddle point method developed in Chapter 2 to account for this circumstance.

Proof. Our starting point is the probability generating function of the number of arrivals per time slot, given in (3.7), which is analytic for $|z|<1+1/b_n =: r$. Under Assumption 3.2, we consider μ_Q as given in (3.9). We set

$$g(z) = -\ln z + \frac{1}{s_n}\ln\Big[\tilde{A}^{(n)}(z)\Big] = -\ln z - \frac{a_n}{s_n}\ln\left(1+(1-z)b_n\right), \tag{3.20}$$

to be considered in the entire complex plane with branch cuts $(-\infty,0]$ and $[r,\infty)$. The relevant saddle point $z_{\rm sp}$ is the unique zero z of $g'(z)$ with $z\in(1,r_0)$. Since

$$g'(z) = -\frac{1}{z}+\frac{\rho_n}{1+(1-z)b_n}, \tag{3.21}$$

this yields,

$$1+(1-z_{\rm sp})b_n = \rho_n z_{\rm sp}, \quad \text{i.e.,} \quad z_{\rm sp} = 1+\frac{1-\rho_n}{\rho_n+b_n}. \tag{3.22}$$

We then find

$$\mu_Q = \frac{s_n}{2\pi i}\int_{|z|=1+\varepsilon} \frac{g'(z)}{z-1}\,\frac{\exp(s_n\, g(z))}{1-\exp(s_n\, g(z))}\mathrm{d}z, \tag{3.23}$$

and we take here $1+\varepsilon = z_{\mathrm{sp}}$. There are no problems with the branch cuts since we consider $\exp(s_n g(z))$ with integer s_n.

We continue as in Chapter 2, Section 3 and thus we intend to substitute $z = z(v)$ in the integral in (3.23), where $z(v)$ satisfies

$$g(z(v)) = g(z_{\mathrm{sp}}) - \tfrac{1}{2}\, v^2\, g''(z_{\mathrm{sp}}) =: q(v)$$

on a range $-\frac{1}{2}\delta_n \le v \le \frac{1}{2}\delta_n$ with $\delta_n \to 0$ as $n \to \infty$. Note that, this range depends on n, whereas these bounds $\pm\frac{1}{2}\delta_n$ remained bounded away from zero in [118]. This severely complicates the present analysis. We consider the approximate representation

$$\frac{-s_n\, g''(z_{\mathrm{sp}})}{2\pi i}\int_{-\frac{1}{2}\delta_n}^{\frac{1}{2}\delta_n} \frac{v}{z(v)-1}\,\frac{\exp(s_n\, q(v))}{1-\exp(s_n\, q(v))}\mathrm{d}v \tag{3.24}$$

of μ_Q. We have to operate here with additional care, since both the analyticity radius $r = 1 + 1/b_n$ and the saddle point z_{sp} outside zero r_0 tend to 1 as $n \to \infty$. Specifically, proceeding under the assumptions that $(1-\rho_n)^2 a_n$ is bounded while $a_n \to \infty$ and $b_n \ge 1$, see Assumption 3.2, we have from (3.22) that

$$z_{\mathrm{sp}} - 1 = \frac{1-\rho_n}{b_n+\rho_n} = \frac{1-\rho_n}{b_n} + O\Big(\frac{1-\rho_n}{b_n^2}\Big), \tag{3.25}$$

where the O-term is small compared to $(1-\rho_n)/b_n$ when $b_n \to \infty$. Next, we approximate r_0, using that $r_0 > 1$ satisfies

$$-\ln r_0 - \frac{\rho_n}{b_n}\ln\left(1 + (1-r_0)b_n\right) = 0.$$

Write $r_0 = 1 + u/b_n$, so that we get the equation

$$\begin{aligned} 0 &= -\ln\left(1+\frac{u}{b_n}\right) - \frac{\rho_n}{b_n}\ln(1-u) \\ &= -\frac{u}{b_n}\Big(1-\rho_n - \tfrac{1}{2}\Big(\frac{1}{b_n}+\rho_n\Big)u - \tfrac{1}{3}\Big(\frac{-1}{b_n^2}+\rho_n\Big)u^2 + \cdots\Big), \end{aligned}$$

where we have used the Taylor expansion of $\ln(1+x)$ at $x = 0$. Thus we find

$$u = \frac{2(1-\rho_n)}{\rho_n + 1/b_n} + O(u^2) = 2(1-\rho_n) + O((1-\rho_n)^2) + O\Big(\frac{1-\rho_n}{b_n}\Big),$$

and so,

$$r_0 = 1 + 2\,\frac{1-\rho_n}{b_n} + O\Big(\frac{(1-\rho_n)^2}{b_n}\Big) + O\Big(\frac{1-\rho_n}{b_n^2}\Big).$$

In (3.24) we choose δ_n so large that the integral has converged within exponentially small error using $\pm\delta_n$ as integration limits, and, at the same time, so small that there is a convergent power series

$$z(v) = z_{\text{sp}} + iv + \sum_{k=2}^{\infty} c_k (iv)^k, \qquad \text{for } |v| \leq \tfrac{1}{2}\delta_n. \tag{3.26}$$

To achieve these goals, we supplement the information on $g(z)$, as given by (3.20) – (3.22), by

$$g''(z) = \frac{1}{z^2} + \frac{\rho_n b_n}{(1+(1-z)b_n)^2}, \quad g''(1) = 1 + \rho_n b_n, \quad g''(z_{\text{sp}}) = \frac{1}{z_{\text{sp}}^2}\Big(1 + \frac{b_n}{\rho_n}\Big). \tag{3.27}$$

Now

$$\exp(s_n\, q(v)) = \exp(s_n\, g(z_{\text{sp}}))\exp(-\tfrac{1}{2}\, s_n\, g''(z_{\text{sp}})\, v^2),$$

and

$$s_n\, g''(z_{\text{sp}})v^2 = s_n\, b_n v^2(1+o(1)) = a_n(b_n\, v)^2(1+o(1)).$$

Therefore, (3.24) approximates μ_Q with exponentially small error when we take $\frac{1}{2}\delta_n$ of the order $1/b_n$.

We next aim at showing that we have a power series for $z(v)$ as in (3.26) that converges for $|v| \leq \frac{1}{2}\delta_n$ with $\frac{1}{2}\delta_n$ of the order $1/b_n$.

Lemma 3.2. *Let*

$$r_n := \frac{1}{2\,b_n} - (z_{\text{sp}} - 1), \quad m_n := \tfrac{2}{3}\rho_n r_n \sqrt{\frac{b_n + \rho_n^{-1}}{b_n + \rho_n}},$$

where we assume $r_n > 0$. Then (3.26) *holds with real coefficients c_k satisfying*

$$|c_k| \leq \frac{r_n}{m_n^k}, \quad k = 2, 3, \ldots. \tag{3.28}$$

Proof. We let

$$G(z) := \frac{2(g(z) - g(z_{\text{sp}}))}{g''(z_{\text{sp}})(z - z_{\text{sp}})^2}. \tag{3.29}$$

Then $G(z_{\text{sp}}) = 1$ and so we can write (3.4) as

$$F(z) := (z - z_{\text{sp}})\sqrt{G(z)} = iv \tag{3.30}$$

when $|z - z_{\text{sp}}|$ is sufficiently small. Since $F(z_{\text{sp}}) = 0$, $F'(z_{\text{sp}}) = 1$, the Bürmann-Lagrange inversion theorem implies validity of a power series as in (3.26), with real c_k since $G(z)$ is positive and real for real z close to z_{sp}. We therefore just need to estimate the convergence radius of this series from below.

To this end, we start by showing that

$$\operatorname{Re}[g''(z)] > \frac{4}{9}\rho_n^2 \frac{b_n+\rho_n^{-1}}{b_n+\rho_n}, \quad |z-z_{\mathrm{sp}}| \le r_n. \tag{3.31}$$

For this, we consider the representation

$$G(z) = 2\int_0^1 \int_0^1 \frac{g''(z_{\mathrm{sp}} + s\,t(z-z_{\mathrm{sp}}))}{g''(z_{\mathrm{sp}})}\, t \mathrm{d}s\mathrm{d}t. \tag{3.32}$$

We have for $\zeta \in \mathbb{C}$ and $|\zeta - 1| \le 1/2b_n \le 1/2$ from (3.27) that

$$\operatorname{Re}[g''(\zeta)] = \operatorname{Re}(1/\zeta^2) + \rho_n b_n \operatorname{Re}\Big[\Big(\frac{1}{1+(1-\zeta)b_n}\Big)^2\Big] \ge \tfrac{4}{9}(1+\rho_n b_n). \tag{3.33}$$

To show the inequality in (3.33), it suffices to show that

$$\min_{|\xi-1|\le 1/2} \operatorname{Re}\Big(\frac{1}{\xi^2}\Big) = \frac{4}{9}. \tag{3.34}$$

The minimum in (3.34) is assumed at the boundary $|\xi - 1| = 1/2$, and for a boundary point ξ, we write

$$\xi = 1 + \tfrac{1}{2}\cos\theta + \tfrac{1}{2} i \sin\theta, \quad 0 \le \theta \le 2\pi,$$

so that

$$\operatorname{Re}\Big(\frac{1}{\xi^2}\Big) = \frac{1+\cos\theta + \frac{1}{4}\cos 2\theta}{(\frac{5}{4}+\cos\theta)^2}.$$

Now

$$\frac{\mathrm{d}}{d\theta}\Big[\frac{1+\cos\theta + \frac{1}{4}\cos 2\theta}{(\frac{5}{4}+\cos\theta)^2}\Big] = \frac{\sin\theta\,(1-\cos\theta)}{4(\frac{5}{4}+\cos\theta)^3}$$

vanishes for $\theta = 0, \pi, 2\pi$, where $\operatorname{Re}(1/\xi^2)$ assumes the values 4/9, 4, 4/9, respectively. This shows (3.34).

We use (3.34) with $\xi = \zeta$ and with $\xi = 1 + (1-\zeta)b_n$, with

$$\zeta = \zeta(s,t) = z_{\mathrm{sp}} + st\,(z - z_{\mathrm{sp}}), \quad 0 \le s,\, t \le 1, \tag{3.35}$$

where we take ζ such that $|\zeta - 1| \le 1/2b_n$. It is easy to see from $1 < z_{\mathrm{sp}} < 1 + 1/2b_n$ that $|\zeta - 1| \le 1/2b_n$ holds when $|z - z_{\mathrm{sp}}| \le r_n = 1/2b_n - (z_{\mathrm{sp}} - 1)$. We have, furthermore, from (3.22) that $0 < g''(z_{\mathrm{sp}}) \le 1 + b_n/\rho_n$. Using this, together with (3.33) where ζ is as in (3.35), yields

$$\operatorname{Re}[G(z)] \le \frac{4}{9}\,\frac{1+\rho_n b_n}{1+b_n/\rho_n}\, 2\int_0^1\int_0^1 t\,\mathrm{d}s\mathrm{d}t = \tfrac{4}{9}\rho_n^2\,\frac{b_n+\rho_n^{-1}}{b_n+\rho_n}$$

when $|z - z_{\mathrm{sp}}| \le r_n$, and this is (3.31). We therefore have from (3.30) that

$$|F(z)| > r_n \cdot \frac{2}{3}\rho_n \sqrt{\frac{b_n+\rho_n^{-1}}{b_n+\rho_n}} = m_n, \quad |z - z_{\mathrm{sp}}| = r_n.$$

Hence, for any v with $|v| \le m_n$, there is exactly one solution $z = z(v)$ of the equation $F(z) - iv = 0$ in $|z - z_{\mathrm{sp}}| \le r_n$ by Rouché's theorem. This $z(v)$ is given by

$$z(v) = \frac{1}{2\pi i} \int_{|z-z_{\mathrm{sp}}|=r_n} \frac{F'(z)\, z}{F(z) - iv} \mathrm{d}z,$$

and depends analytically on v, $|v| \le m_n$. From $|z(v) - z_{\mathrm{sp}}| \le r_n$, we can finally bound the power series coefficients c_k according to

$$|c_k| = \Big| \frac{1}{2\pi i} \int_{|iv|=m_n} \frac{z(v) - z_{\mathrm{sp}}}{(iv)^{k+1}} \mathrm{d}(iv) \Big| \le \frac{r_n}{m_n^k},$$

and this completes the proof of the lemma. □

Remark 3.1. We have $z_{\mathrm{sp}} - 1 = o(1/b_n)$, see (3.25), and so

$$r_n = \frac{1}{2b_n}(1 + o(1)), \quad m_n = \frac{1}{3b_n}(1 + o(1)),$$

implying that the radius of convergence of the series in (3.26) is indeed of order $1/b_n$ (since we have assumed $b_n \ge 1$).

We let $\delta_n = m_n$, and we write for $0 \le v \le \frac{1}{2}\delta_n$

$$\frac{v}{z(v) - 1} + \frac{-v}{z(-v) - 1} = \frac{-2iv\, \mathrm{Im}(z(v))}{|z(v) - 1|^2},$$

where we have used that all c_k are real, so that $z(-v) = z(v)^*$, where * denotes the complex conjugate. Now from (3.28) and realness of the c_k, we have

$$\mathrm{Im}(z(v)) = v + \sum_{l=1}^{\infty} c_{2l+1}(-1)^l\, v^{2l+1} = v + O(v^3), \tag{3.36}$$

and in similar fashion

$$|z(v) - 1|^2 = (z_{\mathrm{sp}} - 1)^2 + v^2 + O((z_{\mathrm{sp}} - 1)^2 v^2) + O(v^4) \tag{3.37}$$

when $0 \le v \le \frac{1}{2}\delta_n$. The order terms in (3.36)-(3.37) are negligible in leading order, and so we get for $\mu_{Q^{(n)}}$ via (3.24) the leading order expression

$$\frac{-s_n\, g''(z_{\mathrm{sp}})}{2\pi i} \int_0^{\frac{1}{2}\delta_n} \frac{-2iv^2}{(z_{\mathrm{sp}} - 1)^2 + v^2} \frac{\exp(s_n\, q(v))}{1 - \exp(s_n\, q(v))} \mathrm{d}v.$$

We finally approximate $q(v) = g(z_{\mathrm{sp}}) - \frac{1}{2} g''(z_{\mathrm{sp}}) v^2$. There is a z_1, $1 \le z_1 \le z_{\mathrm{sp}}$ such that

$$g(z_{\mathrm{sp}}) = -\tfrac{1}{2}(z_{\mathrm{sp}} - 1)^2\, g''(z_1),$$

and, see (3.25) and (3.27),

$$g''(z_1) = g''(z_{\mathrm{sp}}) + O((1 - \rho_n) b_n).$$

Hence

$$\begin{aligned} s_n\, q(v) &= -\tfrac{1}{2} s_n\, g''(z_{\rm sp})\, [(z_{\rm sp}-1)^2+v^2] + O((1-\rho_n) b_n s_n (z_{\rm sp}-1)^2) \\ &= -\tfrac{1}{2} s_n\, g''(z_{\rm sp}) [(z_{\rm sp}-1)^2+v^2] + O((1-\rho_n)^2 a_n), \end{aligned} \tag{3.38}$$

where (3.25) has been used and $a_n b_n = s_n(1+o(1))$ Therefore, the O-term in (3.38) tends to 0 by our assumption that $(1-\rho_n)^2 a_n$ is bounded. Thus, we get for $\mu_{Q^{(n)}}$ in leading order

$$\frac{s_n g''(z_{\rm sp})}{\pi} \int_0^{\frac{1}{2}\delta_n} \frac{v^2}{(z_{\rm sp}-1)^2+v^2} \frac{\exp(-\frac{1}{2} g''(z_{\rm sp}) s_n ((z_{\rm sp}-1)^2+v^2))}{1-\exp(-\frac{1}{2} g''(z_{\rm sp}) s_n ((z_{\rm sp}-1)^2+v^2))} \mathrm{d}v, \tag{3.39}$$

When we substitute $t = v\sqrt{s_n\, g''(z_{\rm sp})/2}$ and extend the integration in (3.39) to all $t \geq 0$ (at the expense of an exponentially small error), we get for $\mu_{Q^{(n)}}$ in leading order

$$\frac{1}{\pi} \sqrt{2\, s_n\, g''(z_{\rm sp})} \int_0^\infty \frac{t^2}{\frac{1}{2}\beta_n^2} \frac{\exp(-\frac{1}{2}\beta_n^2 - t^2)}{1-\exp(-\frac{1}{2}\beta_n^2 - t^2)} \mathrm{d}t,$$

where

$$\beta_n^2 = s_n\, g''(z_{\rm sp})(z_{\rm sp}-1)^2.$$

Now using (3.22) and (3.27), we get the result of Theorem 3.2. A separate analysis of β_n is provided in Subsection 3.5.1.

3.5 Numerical & empirical studies

A similar analysis, modeled after the one given in Chapter 2 gives under Assumption 3.1 the leading-order expression

$$\frac{1}{z_{\rm sp}\pi} \int_0^\infty \frac{\beta_n/\sqrt{2}}{\frac{1}{2}\beta_n^2 + t^2} \ln(1-e^{-\frac{1}{2}\beta_n^2 - t^2}) \mathrm{d}t \tag{3.40}$$

for $\ln \mathbb{P}(Q^{(n)} = 0)$. Observe that the quantity in (3.40) is negative, but bounded away from $-\infty$ when β_n is bounded away from 0. Furthermore, we find for the variance of $Q^{(n)}$ the approximation

$$\frac{\beta_n^3/\sqrt{2}}{\pi} \frac{z_{\rm sp}+1}{(z_{\rm sp}-1)^2} \int_0^\infty \frac{t^2}{(\frac{1}{2}\beta_n + t^2)^2} \frac{\exp(-\frac{1}{2}\beta_n - t^2)}{1-\exp(-\frac{1}{2}\beta_n^2 - t^2)} \mathrm{d}t.$$

□

Note that we can write (3.17) as

$$\mu_Q \approx \tilde{\sigma}_n\, \mathbb{E}[M_{\beta_n}] \quad \text{and} \quad \sigma_Q^2 \approx \tilde{\sigma}_n^2 \operatorname{Var} M_{\beta_n}$$

with

$$\tilde{\sigma}_n = \beta_n \Big(\frac{b_n + \rho_n}{1 - \rho_n} \Big). \tag{3.41}$$

This robust approximation for μ_Q is suggestive of the following two properties that extend beyond the mean system behavior, and hold at the level of approximating the queue by σ_n times the Gaussian random walk:

(i) At the process level, the space should be normalized with σ_n, as in (3.5). The approximation (3.17) suggests that it is better to normalize with $\tilde{\sigma}_n$. Although $\tilde{\sigma}_n \to \sigma_n$ for $n \to \infty$, the $\tilde{\sigma}_n$ is expected to lead to sharper approximations for finite n.

(ii) Again at the process level, it seems better to replace the original hedge β by the robust hedge β_n. This thus means that the original system for finite n is approximated by a Gaussian random walk with drift $-\beta_n$. Apart from this approximation being asymptotically correct for $n \to \infty$, it is also expected to approximate the behavior better for finite n.

3.5.1 Convergence of the robust hedge

We next examine the accuracy of the heavy-traffic approximations for μ_Q and σ_Q^2, following Corollary 3.3 and Theorem 3.2. We expect the robust approximation to be considerably better than the classical approximation when β_n and $\tilde{\sigma}_n$ differ substantially from their limiting counterparts. Before substantiating this claim numerically, we present a result on the convergence rates of β_n to β and $\tilde{\sigma}_n$ to σ_n.

Proposition 3.1. *Let a_n, b_n and s_n as in Assumption 3.2. Then*

$$\beta_n^2 = \beta^2 \Big(1 - \frac{1}{1 + b_n + \sigma_n/\beta} \Big). \tag{3.42}$$

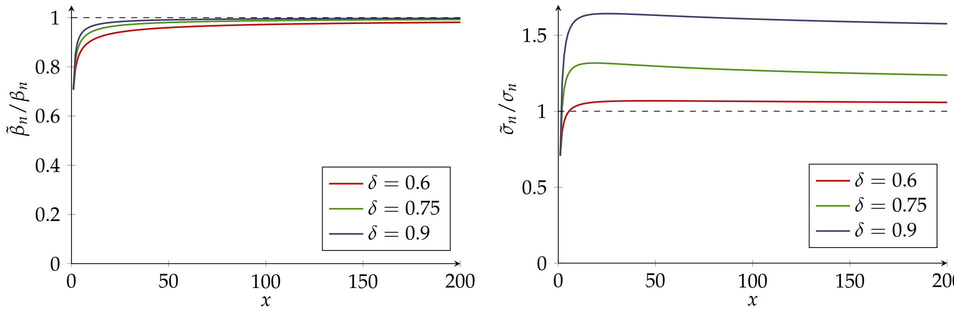
Proof. From (3.18), we have

$$\begin{aligned}
\beta_n^2 &= s_n \Big(\frac{1-\rho_n}{b_n+1} \Big)^2 \Big(1 + \frac{b_n}{\rho_n} \Big) = \frac{1}{s_n} \Big(\frac{s_n - a_n b_n}{b_n + 1} \Big)^2 \Big(1 + \frac{s_n}{a_n} \Big) \\
&= \frac{1}{s_n} \frac{\beta^2\, a_n b_n (b_n+1)}{(b_n+1)^2} \Big(1 + \frac{s_n}{a_n} \Big) = \beta^2 \, \frac{b_n}{b_n+1} \Big(1 + \frac{a_n}{s_n} \Big) =: \beta^2 \, \bar{F}_n.
\end{aligned}$$

Now,

$$\begin{aligned}
\bar{F}_n &= \frac{b_n}{b_n+1} \Big(1 + \frac{a_n}{s_n} \Big) = \frac{b_n}{b_n+1} + \frac{1}{b_n+1} \frac{a_n b_n}{s_n} \\
&= 1 - \frac{1}{b_n+1} \Big(1 - \frac{a_n b_n}{s_n} \Big) = 1 - \frac{1}{b_n+1} \frac{\beta\,\sigma_n}{s_n} \\
&= 1 - \frac{1}{b_n+1} \frac{1}{1 + \frac{\mu_n}{\beta\sigma_n}} = 1 - \frac{1}{b_n + 1 + \frac{1}{\beta}\sqrt{a_n b_n (b_n+1)}},
\end{aligned}$$

which together with $\sigma_n^2 = a_n b_n (b_n + 1)$ proves the proposition. $\square$



(a) Convergence of β_n.

(b) Convergence of $\tilde{\sigma}_n$.

Figure 3.1

Note that β_n always approaches β from below. Also, (3.42) shows that b_n is the dominant factor in determining the rate of convergence of β_n.

Proposition 3.2. *Let $\tilde{\sigma}_n$ as in* (3.41). *Then*

$$\tilde{\sigma}_n = \sigma_n + b_n\beta_n + O(1).$$

Proof. Straightforward calculations give

$$\begin{aligned}\tilde{\sigma}_n &= \beta_n\left(\frac{s_n b_n + a_n b_n}{s_n - a_n b_n}\right)\\ &= \frac{\beta_n}{\beta}\frac{b_n}{\sigma_n}(s_n + a_n) = \frac{\beta_n}{\beta}\sqrt{\frac{b_n}{a_n(b_n+1)}}\left(a_n(b_n+1) + \beta\sqrt{a_n b_n(b_n+1)}\right)\\ &= \frac{\beta_n}{\beta}\left(\sqrt{a_n b_n(b_n+1)} + \beta b_n\right) = \frac{\beta_n}{\beta}\sigma_n + \beta_n b_n.\end{aligned}$$

Applying Proposition 3.1 together with the observation

$$\sigma_n\sqrt{1 - \frac{1}{1 + b_n + \sigma_n/\beta}} = \sigma_n(1 + O(1/\sqrt{a_n}b_n)) = \sigma_n + O(1)$$

yields the result. $\square$

In Figure 3.1, we visualize the convergence speed of both parameters in case $\mu_n = n$, $\sigma_n = n^\delta$ with $\delta = 0.7$ and $\beta = 1$. This implies $a_n = n/(n^{2\delta} - 1)$ and $b_n = n^{2\delta} - 1$.

We observe that β_n starts resembling β fairly quickly, as predicted by Proposition 3.1; $\tilde{\sigma}_n$, on the other hand, converges extremely slowly to its limiting counterpart. Since μ_Q and σ_Q^2 are approximated by $\tilde{\beta}_n$ and $\tilde{\sigma}_n^2$, multiplied by a term that remains almost constant as n grows, the substitution of σ_n by $\tilde{\sigma}_n$, is essential for obtaining accurate approximations, as we illustrate further in the next subsection.

s_n	ρ_n	μ_Q	(3.14)	(3.17)	σ_Q	(3.15)	(3.19)
5	0.609	0.343	0.246	0.363	1.002	0.835	0.978
10	0.683	0.535	0.400	0.551	1.239	1.063	1.216
50	0.815	1.405	1.168	1.405	1.995	1.817	1.971
100	0.855	2.113	1.824	2.105	2.445	2.270	2.420
500	0.920	5.446	5.006	5.412	3.923	3.762	3.899

Table 3.1: Numerical results for the Gamma-Poisson case with $\beta = 1$ and $\delta = 0.6$.

s_n	ρ_n	μ_Q	(3.14)	(3.17)	σ_Q	(3.15)	(3.19)
5	0.550	0.462	0.284	0.479	1.162	0.896	1.130
10	0.587	0.852	0.521	0.855	1.570	1.213	1.528
50	0.668	3.197	2.093	3.106	3.025	2.433	2.947
100	0.700	5.561	3.784	5.377	3.983	3.270	3.887
500	0.766	19.887	14.741	19.202	7.514	6.455	7.361

Table 3.2: Numerical results for the Gamma-Poisson case with $\beta = 1$ and $\delta = 0.8$.

3.5.2 Comparison between heavy-traffic approximations

We set $\mu_n = n$ and $\sigma_n^2 = n^{2\delta}$ with $\delta > \frac{1}{2}$, so that $s_n = n + \beta n^\delta$, and $a_n = n/(n^{2\delta-1} - 1)$ and $b_n = n^{2\delta-1} - 1$.

Tables 3.1-3.4 present numerical results for various parameter values. In these tables, we fixed s_n to integer values, and use the associated value of n in our calculations. The exact values of the performance measures are calculated using the method in Appendix 3.B. Several conclusions are drawn from these tables. Observe that the heavy-traffic approximations based on the Gaussian random walk, (3.14) and (3.15), capture the right order of magnitude for both μ_Q and σ_Q. However, the values are off, in particular for small s_n and relatively low $\rho_n := \mathbb{E}[A^{(n)}]/s_n$. The inaccuracy also increases with the level of overdispersion. In contrast, the approximations that follow from Theorem 3.2, (3.17) and (3.19) are remarkably accurate. Even for small systems with $s_n = 5$ or 10, the approximations for μ_Q are within

s_n	ρ_n	μ_Q	(3.14)	(3.17)	σ_Q	(3.15)	(3.19)
5	0.949	11.532	11.306	11.495	3.634	3.559	3.602
10	0.961	17.565	17.268	17.548	4.474	4.398	4.444
50	0.979	46.368	45.869	46.418	7.241	7.168	7.218
100	0.984	70.340	69.735	70.430	8.910	8.839	8.888
500	0.991	184.900	183.989	185.108	14.422	14.357	14.404

Table 3.3: Numerical results for the Gamma-Poisson case with $\beta = 0.1$ and $\delta = 0.6$.

s_n	ρ_n	μ_Q	(3.14)	(3.17)	σ_Q	(3.15)	(3.19)
5	0.931	15.730	15.209	15.909	4.276	4.127	4.233
10	0.939	27.561	26.672	27.958	5.652	5.466	5.605
50	0.955	100.660	97.967	102.070	10.760	10.476	10.698
100	0.961	175.591	171.360	177.818	14.189	13.855	14.117
500	0.971	638.097	626.346	644.105	26.963	26.490	26.864

Table 3.4: Numerical results for the Gamma-Poisson case with $\beta = 0.1$ and $\delta = 0.8$.

6% of the exact value for small ρ_n and within 2% for ρ_n close to 1. For σ_Q^2, these percentages even reduce to 3% and 1%, respectively. For larger values of s_n these relative errors naturally reduce further. Overall, we observe that the approximations improve for heavily loaded systems, and the corrected approximations are particularly useful for systems with increased overdispersion.

3.5.3 Capacity allocation in health care

We next apply our model and robust approximations to real-life patient arrivals. We consider emergency patients who require diagnostic tests at the radiology department of a hospital. Green [89] points out that patients at the radiology department can be roughly categorized into three groups: Inpatients, outpatients and emergency patients. Inpatient and outpatient arrivals are relatively predictable as these are usually scheduled by appointment. Emergency patients, on the other hand, are inherently unpredictable: They typically require urgent care and therefore timely access to testing facilities, as well as a quick assessment of the test results. This leads to prioritization of emergency patients over the other two groups in such settings, so that they do not experience any delay caused by the groups of lower priority. However, patients from the same top-priority group can still cause considerable congestion. A careful evaluation of capacity allocation is hence required, bearing in mind that additional sophisticated pieces of medical equipment are very costly.

In the setting we study, capacity is defined by the number of time slots available to perform radiology tests on emergency patients in a given time period, which we set at 24 hours. As radiology tests are commonly performed in appointment slots of fixed length, the number of slots available per day is also indirectly fixed. In terms of our model parameters, see Section 3.2, we have s as the number of slots per day allocated to emergency patients, and $A(k)$ the number of test requests received by the department on day k. We omit the subscript n in this section due to the absence of limits. Then $\mathbb{E}[Q]$ can be interpreted as the expected number of patients whose test is carried over to the next day. A more natural performance measure in this setting is the expected waiting time, namely the time between the physician's request and the actual start of the test. However, Little's law implies that there is a linear relation between the two, hence we choose to only evaluate $\mathbb{E}[Q]$.

The data set on which our empirical study is based originates from the emergency department of SKHospital, monitored over a period of 76 days from September to November 2012. We extracted information of ED patients referred to the radiology department by the ED physicians, which includes the time the test request was made and the exact test type performed. The two test types, X-ray and CT scans, are performed on different equipment and hence it makes sense to analyze their queueing processes separately.

We refer to test requests as arrivals. The empirical cumulative distribution functions of the number of arrivals per day, for each type, are depicted by the black lines in Figure 3.2. The sample means equal 69.81 and 17.47, for the X-ray and CT scans respectively, whereas the sample variances are 121.8 and 26.12. These values suggest that fitting a Poisson distribution is inappropriate, which is visually backed up by the fitted Poisson cdf, depicted in Figure 3.2 by the red line. To strengthen this claim, we tested both samples for the Poisson assumption using the *dispersion test*, see Appendix 3.C, and obtained p-values equal $7.01 \cdot 10^{-3}$ and $3.57 \cdot 10^{-3}$ respectively, which allow us to safely reject the Poisson hypothesis in both cases.

In search for a better distributional fit with the arrivals count, we resort to Gamma-Poisson mixtures. Here we employ the procedure in [127], which is basically a maximum log-likelihood method, to obtain estimates for the parameters a and b. This yields

$$\hat{a}_{\text{X-ray}} = 95.68, \quad \hat{b}_{\text{X-ray}} = 0.7297, \quad \hat{a}_{\text{CT}} = 37.19, \quad \hat{b}_{\text{CT}} = 0.4698.$$

Applying the bootstrapping method to the data and the fitted model, also described in the appendix of [127], returns p-values that equal 0.7354 and 0.2120 for X-ray and CT scans, respectively. Therefore, the null hypothesis, stating that the data originated from a Gamma-Poisson mixture, cannot be rejected. The cdfs of the fitted Gamma-Poisson distributions, plotted in Figure 3.2, give visual confirmation of this claim as well. Naturally, we also compared the estimated densities to the empirical pdf of the data. However, these fail to give a convincing visual fit due to the relatively small sample size and are therefore omitted here.

We now have clear evidence that both the X-ray and CT scan facilities face an overdispersed arrival stream. In our final step of the empirical study we now evaluate the accuracy of our performance measure of interest $\mathbb{E}[Q]$, and the consequences of assessing system performance while ignoring the presence of overdispersion. We take the following approach: Trivially, we need to choose $s > \mathbb{E}[A]$ in order for the system to be stable. Hence, we vary s from 70 to 80 for X-rays and from 18 to 24 for CT scans and simulate the queue length process by sampling the number of requests per day from the actual arrival counts. The number of iterations performed is 10^8 for each configuration, excluding a warm-up interval of length 10^7 (days). Around the mean of Q obtained from this simulation, we create a 95% confidence interval. Next, we approximate the expected stationary queue length under two scaling rules. Assuming Poisson arrivals, the appropriate capacity allocation rule would be $s = \hat{\mu} + \beta\sqrt{\hat{\mu}}$, for some $\beta > 0$. Our novel capacity sizing rule prescribes $s = \hat{\mu} + \beta\hat{\sigma} = \hat{a}\hat{b} + \beta\sqrt{\hat{a}\hat{b}(\hat{b}+1)}$. We compute the first approximation based on

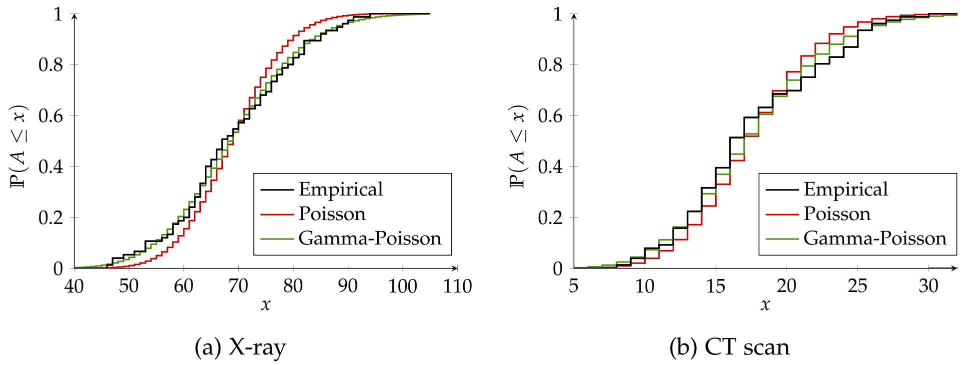


Figure 3.2: Empirical, fitted Poisson and fitted Gamma-Poisson cumulative distribution functions of the number of arrivals.

square-root safety capacity sizing by deducing β for each s, which we substitute in $\mathbb{E}[Q^{\mathrm{srs}}] \approx \sqrt{\hat{\mu}}\,\mathbb{E}[M_\beta]$. Similarly, we obtain β from the new rule, and plug this value, together with the fitted parameters $\hat{a}$ and $\hat{b}$, into (3.17). The results are given in Tables 3.5 and 3.6. The last column shows the 95% relative error bound of the second approximation.

s	ρ	$\mathbb{E}[Q]$ ($\pm$ conf. iv.)	$\mathbb{E}[Q^{\mathrm{srs}}]$	(3.14)	(3.17)	rel. error
70	0.997	$328.313 \pm 6.6 \cdot 10^{-2}$	186.664	324.231	325.221	$9.6 \cdot 10^{-3}$
71	0.983	$45.073 \pm 1.0 \cdot 10^{-2}$	24.946	45.290	45.308	$5.4 \cdot 10^{-3}$
72	0.970	$21.988 \pm 5.4 \cdot 10^{-3}$	11.650	21.982	22.129	$6.6 \cdot 10^{-3}$
73	0.956	$13.546 \pm 3.6 \cdot 10^{-3}$	6.847	13.455	13.649	$7.8 \cdot 10^{-3}$
74	0.943	$9.230 \pm 2.7 \cdot 10^{-3}$	4.438	9.106	9.319	$1.0 \cdot 10^{-2}$
75	0.931	$6.655 \pm 2.1 \cdot 10^{-3}$	3.031	6.513	6.731	$1.2 \cdot 10^{-2}$
76	0.919	$4.949 \pm 1.7 \cdot 10^{-3}$	2.136	4.821	5.037	$1.8 \cdot 10^{-2}$
77	0.907	$3.755 \pm 1.4 \cdot 10^{-3}$	1.534	3.650	3.861	$2.8 \cdot 10^{-2}$
78	0.895	$2.884 \pm 1.1 \cdot 10^{-3}$	1.115	2.807	3.009	$4.4 \cdot 10^{-2}$
79	0.884	$2.230 \pm 1.0 \cdot 10^{-3}$	0.816	2.183	2.374	$6.5 \cdot 10^{-2}$
80	0.873	$1.734 \pm 8.5 \cdot 10^{-4}$	0.600	1.710	1.890	$9.1 \cdot 10^{-2}$

Table 3.5: Computational results for X-ray.

Based on these figures, we make several remarks. First, assuming the conventional regime (neglecting overdispersion) the approximation severely overestimates system performance in both queues. Because the square-root safety margin underestimates the stochastic fluctuations in the arrival process, the safety parameter β is overestimated, which leads to a smaller magnitude of the approximated queue length process. This clearly illustrates the distorted view estimated performance

s	ρ	$\mathbb{E}[Q]$ ($\pm$ conf.iv.)	$\mathbb{E}[Q^{\mathrm{srs}}]$	(3.14)	(3.17)	rel. error
18	0.970	$22.116 \pm 4.9 \cdot 10^{-3}$	14.235	21.965	21.724	$1.8 \cdot 10^{-2}$
19	0.919	$6.289 \pm 1.7 \cdot 10^{-3}$	3.640	5.941	6.040	$4.0{\cdot}10^{-2}$
20	0.873	$3.101 \pm 1.0 \cdot 10^{-3}$	1.589	2.772	2.917	$6.0{\cdot}10^{-2}$
21	0.832	$1.767 \pm 6.6 \cdot 10^{-4}$	0.800	1.507	1.658	$6.1{\cdot}10^{-2}$
22	0.794	$1.066 \pm 4.6 \cdot 10^{-4}$	0.425	0.874	1.016	$4.7{\cdot}10^{-2}$
23	0.760	$0.653 \pm 3.3 \cdot 10^{-4}$	0.230	0.522	0.649	$7.1{\cdot}10^{-3}$
24	0.728	$0.377 \pm 2.3 \cdot 10^{-4}$	0.124	0.315	0.424	$1.2{\cdot}10^{-1}$

Table 3.6: Computational results for CT scan.

characteristics can give under the wrong scaling. Secondly, it is worth noticing the very good proximity of (3.17) to the values obtained through simulation. As we expected, the quality of the approximation deteriorates with increasing values of s. This makes sense, because we assumed the system to be in heavy traffic in the derivation of the formulas. What is surprising, on the other hand, is the fact that the approximation performs very well, even though the parameter b is very small for these particular data sets, while the analysis of Theorem 3.2 assumes that a and b are large. This demonstrates that the approximation scheme is remarkably robust and is able to capture the pre-limit behavior of these types of queues very well.

3.6 Conclusion & future research

In this chapter, we proposed an adaptation to the square-root staffing rule for service systems facing overdispersed demand, using the bulk service queue as a vehicle for our analysis. Subsequently, we derive two sets of asymptotic approximations for the scaled steady-state queue length moments for large arrival volumes. The first set being based on the resemblance with the maximum of a Gaussian random walk, the second set being derived through a non-standard saddle point method, assuming arrivals follow a Gamma-Poisson mixture. Numerical experiments indicated that our approximations capture the pre-limit behavior well under different order of overdispersion, and are robust against any parameter estimation errors.

Although our method provides a robust way to approximate and dimension queues with overdispersed arrival processes, we see some interesting directions for future research.

First, we accentuate that our model is a discrete time queueing model in which a deterministic amount of workload can we handled within each period. This approach allowed us to use Pollaczek's formula as a starting point to obtain more refined asymptotic approximations for the performance indicators of the system. In case we consider queueing models of birth-death-type, such as the $M/M/s$ queue, in the presence of overdispersion demand, different techniques are required to provide scaling limits and corresponding capacity allocation rules, see e.g. [150]. Although we expect that, just as in the novel scaling regimes of Chapter 2, the asymptotic behavior of the bulk service queue and the multi-server queueing models to be similar, this needs proper analysis and understanding.

Second, empirical work, see e.g. [26], shows that in real-life settings, demand in consecutive time periods is often positively correlated, rather than independently distributed as assumed in this chapter. This correlation structure obviously alters the queue's dynamics and presumably requires an adaptation of the square-root staffing rule as well, making it a worthwhile direction for further analysis.

Last, we have only considered the analysis of the queueing model in steady state. Typical service systems however do not face a constant expected arrival rate, nor do they run infinitely long. Henceforth, it would be interesting to study the influence of overdispersion on the transient dynamics of the queue and to investigate the capacity allocation problem in scenarios with time-varying demand. The theory developed in this chapter may serve as a building block to tackle these more profound questions.

Appendix

3.A Proofs of convergence results

This section presents the details of the proof of Lemma 3.1 and Theorem 3.1, using the random walk perspective of the process $\{Q^{(n)}(k)\}_{k=0}^{\infty}$. This section is structured as follows. The next two lemmata are necessary for proving the first assertion of Theorem 3.1, concerning the weak convergence of the scaled process to the maximum of the Gaussian random walk, which is summarized in Proposition 3.4. The two remaining propositions of this section show convergence of $\hat{Q}^{(n)}$ at the process level as well as in terms of the three characteristics.

Let us first fix some notation:

$$Y_k^{(n)} := \hat{A}_k^{(n)} - \beta, \quad S_k^{(n)} = \sum_{i=1}^{k} Y_i^{(n)}, \tag{3.43}$$

with $S_0^{(n)} = 0$ and $k = 1, 2, ...$ Then (3.4) can be rewritten as

$$\hat{Q}^{(n)} \stackrel{d}{=} \max_{k\geq 0}\left\{\textstyle\sum_{i=1}^{k} Y_i^{(n)}\right\} =: M_\beta^{(n)}, \tag{3.44}$$

Last, we introduce the sequence of independent normal random variables $Z_1, Z_2, \ldots$ with mean β and unit variance 1, and

$$M_\beta \stackrel{d}{=} \max_{k\geq 0}\{\textstyle\sum_{i=1}^{k} Z_i\}.$$

3.A.1 Proof of Lemma 3.1

Proof. We show weak convergence of the random variable $\hat{A}^{(n)}$, as defined in Section 3.2, to a standard normal random variable. Since $\hat{\Lambda}_n$ is asymptotically standard normal, its characteristic function converges pointwise to the corresponding limiting characteristic function, i.e.

$$\lim_{n\to\infty} \phi_{\hat{\Lambda}_n}(t) = \lim_{n\to\infty} \mathrm{e}^{-i\mu_n t/\sigma_n}\, \phi_{\Lambda_n}(t/\sigma_n) = \mathrm{e}^{-t^2/2}, \qquad \forall t \in \mathbb{R}. \tag{3.45}$$

Furthermore, by definition of $A^{(n)}$,

$$\phi_{A^{(n)}}(t) = \mathbb{E}\left[\exp(\Lambda_n(\mathrm{e}^{it}-1))\right] = \phi_{\Lambda_n}\left(-i(\mathrm{e}^{it}-1)\right),$$

so that

$$\phi_{\hat{A}_k^{(n)}}(t) = \mathrm{e}^{-i\mu_n t/\sigma_n}\, \phi_{A_k^{(n)}}(t/\sigma_n) = \mathrm{e}^{-i\mu_n t/\sigma_n} \phi_{\Lambda_n}\left(-i(\mathrm{e}^{it/\sigma_n}-1)\right). \tag{3.46}$$

Now fix $t \in \mathbb{R}$. By using

$$-i(\mathrm{e}^{it/\sigma_n}-1) = \frac{t}{\sigma_n} - \frac{it^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right),$$

we expand the last term in (3.46),

$$\phi_{\Lambda_n}(t/\sigma_n) + \Big(-\frac{i\,t^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right)\Big)\phi'_{\Lambda_n}(t/\sigma_n) + O\Big(\Big(-\frac{i\,t^2}{2\sigma_n^2} + O\left(\frac{t^3}{\sigma_n^3}\right)\Big)^2 \phi''_{\Lambda_n}\Big(\frac{t}{\sigma_n}\Big)\Big)$$

$$= \phi_{\Lambda_n}(t/\sigma_n) - \Big(\frac{i\,t^2}{2\sigma_n^2} + O\left(t^3/\sigma_n^3\right)\Big)\phi'_{\Lambda_n}(\zeta)$$

for some ζ such that $|\zeta - t/\sigma_n| < |i(1-\mathrm{e}^{it/\sigma_n}) - t/\sigma_n|$. Also,

$$\begin{aligned}|\phi'_{\Lambda_n}(u)| &= \left|\frac{\delta}{\mathrm{d}u}\int_{-\infty}^{\infty} \mathrm{e}^{iux}\mathrm{d}F_{\Lambda_n}(x)\right| = \left|\int_0^{\infty} ix\,\mathrm{e}^{iux}\mathrm{d}F_{\Lambda_n}(x)\right| \\ &\le \int_{-\infty}^{\infty} |ix\,\mathrm{e}^{iux}|\,\mathrm{d}F_{\Lambda_n}(x) = \int_0^{\infty} x\mathrm{d}F_{\Lambda_n}(x) = \mu_n \end{aligned} \tag{3.47}$$

for all $u \in \mathbb{R}$. Hence, by substituting (3.46),

$$\begin{aligned}\left|\phi_{\hat{A}_k^{(n)}}(t) - \mathrm{e}^{-i\mu_n t/\sigma_n}\phi_{\Lambda_n}(t/\sigma_n)\right| &= \left|\mathrm{e}^{-i\mu_n t/\sigma_n}\, \left(\frac{i\,t^2}{2\sigma_n^2} + O(t^3/\sigma_n^3)\right)\, \phi'_{\Lambda_n}(\zeta)\right| \\ &\le \left(\frac{t^2}{2\sigma_n^2} + O(t^3/\sigma_n^3)\right) |\phi'_{\Lambda_n}(\zeta)| \\ &= \frac{\mu_n t^2}{\sigma_n^2} + O\left(\frac{\mu_n t^3}{\sigma_n^3}\right), \end{aligned} \tag{3.48}$$

which tends to zero as $n \to \infty$ by our assumption that $\mu_n/\sigma_n^2 \to 0$. Finally,

$$\left|\phi_{\hat{A}_k^{(n)}}(t) - \mathrm{e}^{-\frac{1}{2}t^2}\right| \leq \left|\phi_{\hat{A}_k^{(n)}}(t) - \mathrm{e}^{-i\mu_n t/\sigma_n}\phi_{\Lambda_n}(t/\sigma_n)\right| + \left|\mathrm{e}^{-i\mu_n t/\sigma_n}\phi_{\Lambda_n}(t/\sigma_n) - \mathrm{e}^{-\frac{1}{2}t^2}\right|,$$

in which both terms go to zero for $n \to \infty$, by (3.45) and (3.48). Hence $\phi_{\hat{A}_k^{(n)}}(t)$ converges to $\mathrm{e}^{-t^2/2}$ for all $t \in \mathbb{R}$, so that we can conclude by Lévy's continuity theorem that $\hat{A}_k^{(n)} \stackrel{d}{\Rightarrow} \mathcal{N}(0,1)$. □

3.A.2 Proof of Theorem 3.1

To secure convergence in distribution of $\hat{Q}^{(n)}$ to M_β, i.e. the maximum of a Gaussian random walk with negative drift, the first assertion of Theorem 3.1, the following property of the sequence $\{Y_k^{(n)}\}_{n\in\mathbb{N}}$ needs to hold.

Lemma 3.3. *Let $Y_k^{(n)}$ be defined as in* (3.43) *with $\mu_n, \sigma_n^2 < \infty$ for all $n \in \mathbb{N}$. Then the sequence $\{(Y_k^{(n)})^+\}_{n\in\mathbb{N}}$ is uniform integrable, i.e.*

$$\lim_{K\to\infty} \sup_n \mathbb{E}\left[Y_k^{(n)+}|\mathbb{1}_{\{|Y_k^{(n)+}|\geq K\}}\right] = 0.$$

Proof. Because the sequence $\{Y_k^{(n)}\}_{k\in\mathbb{N}}$ is i.i.d. for all n, we omit the index k in this proof. First, fix $K > 0$ and note that

$$\mathbb{E}[|Y^{(n)+}|\mathbb{1}\{|Y^{(n)+}| \geq K\}] = \mathbb{E}[Y^{(n)+}\mathbb{1}\{Y^{(n)+} \geq K\}] = \mathbb{E}[Y^{(n)}\mathbb{1}_{\{Y^{(n)}\geq K\}}].$$

This last expression can be bounded from above using the Cauchy-Schwarz inequality, so that

$$\mathbb{E}[Y^{(n)}\mathbb{1}_{\{Y^{(n)}\geq K\}}] \leq \mathbb{E}[Y^{(n)\,2}]^{1/2}\,\mathbb{P}(Y^{(n)} \geq K)^{1/2}.$$

By the definition of $Y^{(n)}$, we know $\mathbb{E}[Y^{(n)}] = -\beta$ and $\operatorname{Var} Y^{(n)} = \operatorname{Var} A^{(n)}/\sigma_n^2 = 1$. Using this information, we find

$$\mathbb{E}[Y^{(n)\,2}] = \operatorname{Var} Y^{(n)} + (\mathbb{E}[Y^{(n)}])^2 = 1 + \beta^2$$

and

$$\begin{aligned}\mathbb{P}(Y^{(n)} \geq K) &= \mathbb{P}(Y^{(n)} + \beta \geq K + \beta) \leq \mathbb{P}(|Y^{(n)} + \beta| \geq K + \beta)\\ &\leq \frac{\operatorname{Var} Y^{(n)}}{(K+\beta)^2} = \frac{1}{(K+\beta)^2},\end{aligned}$$

where we used Chebyshev's inequality for the last upper bound. Therefore,

$$\begin{aligned}\lim_{K\to\infty} \sup_n \mathbb{E}[|Y^{(n)+}|\mathbb{1}_{\{|Y^{(n)+}|\geq K\}}] &= \lim_{K\to\infty} \sup_n \mathbb{E}[Y^{(n)}\mathbb{1}_{\{Y^{(n)}\geq K\}}]\\ &\leq \lim_{K\to\infty} \sup_n \mathbb{E}[Y^{(n)\,2}]^{1/2}\,\mathbb{P}(Y^{(n)} \geq K)^{1/2}\\ &\leq \lim_{K\to\infty} \frac{\sqrt{1+\beta^2}}{K+\beta} = 0.\end{aligned}$$

□

By combining the properties proved in Lemma 3.1 and 3.3 with Assumption 3.2, the next result follows directly by [20, Thm. X6.1].

Proposition 3.3. *Let $\hat{Q}^{(n)}$ as in (3.44). Then*

$$\hat{Q}^{(n)} \overset{d}{\Rightarrow} M_\beta, \qquad \text{as } n \to \infty.$$

Although Proposition 3.3 tells us that the properly scaled $Q^{(n)}$ converges to a non-degenerate limiting random variable, it does not cover the convergence of its mean, variance and the empty-queue probability. In order to secure convergence of these performance measures as well, we follow the approach similar to [193], using Assumptions 3.2 and 3.3.

Proposition 3.4. *Let $\hat{Q}^{(n)}$ as in (3.44), $\mu_n, \sigma_n^2 \to \infty$ such that both $\sigma_n^2/\mu_n \to \infty$ and $\mathbb{E}[\hat{A}^{(n)3}] < \infty$. Then*

$$\begin{aligned} \mathbb{P}(\hat{Q}^{(n)} = 0) &\to \mathbb{P}(M_\beta = 0), \\ \mathbb{E}[\hat{Q}^{(n)}] &\to \mathbb{E}[M_\beta], \\ \operatorname{Var} \hat{Q}^{(n)} &\to \operatorname{Var} M_\beta, \end{aligned}$$

as $n \to \infty$.

Proof. First, we recall that $\hat{Q}^{(n)} \overset{d}{=} M_\beta^{(n)}$ for all $n \in \mathbb{N}$, so that $\mathbb{P}(\hat{Q}^{(n)} = 0) = \mathbb{P}(M_\beta^{(n)} = 0)$, $\mathbb{E}[\hat{Q}^{(n)}] = \mathbb{E}[M_\beta^{(n)}]$ and $\operatorname{Var} \hat{Q}^{(n)} = \operatorname{Var} M_\beta^{(n)}$ as defined in (3.43). Our starting point is Spitzer's identity, see [20, p. 230],

$$\mathbb{E}[\mathrm{e}^{itM_\beta^{(n)}}] = \exp\Big(\sum_{k=1}^{\infty} \frac{1}{k}(\mathbb{E}[\mathrm{e}^{it(S_k^{(n)})^+}] - 1)\Big), \tag{3.49}$$

with $S_k^{(n)}$ as in (3.43), and $M_\beta^{(n)}$ the all-time maximum of the associated random walk. Simple manipulations of (3.49) give

$$\ln \mathbb{P}(M_\beta^{(n)} = 0) = -\sum_{k=1}^{\infty} \frac{1}{k}\,\mathbb{P}(S_k^{(n)} > 0), \tag{3.50}$$

$$\mathbb{E}[M_\beta^{(n)}] = \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{E}[S_k^{(n)\,+}] = \sum_{k=1}^{\infty} \frac{1}{k}\int_0^\infty \mathbb{P}(S_k^{(n)} > x)\mathrm{d}x, \tag{3.51}$$

$$\operatorname{Var} M_\beta^{(n)} = \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{E}[(S_k^{(n)\,+})^2] = \sum_{k=1}^{\infty} \frac{1}{k}\int_0^\infty \mathbb{P}(S_k^{(n)} > \sqrt{x})\mathrm{d}x. \tag{3.52}$$

By Lemma 3.1, we know

$$\mathbb{P}(S_k^{(n)} > y) = \mathbb{P}\left(\sum_{i=1}^{k} Y_i^{(n)} > y\right) \to \mathbb{P}\left(\textstyle\sum_{i=1}^{k} Z_i > y\right),$$

for $n \to \infty$, where the Z_i's are independent and identically normally distributed with mean $-\beta$ and variance 1. Because equivalent expressions to (3.50)-(3.52) apply to the limiting Gaussian random walk, it is sufficient to show that the sums converge uniformly in n, so that we can apply dominated convergence to prove the result.

We start with the empty-queue probability. To justify interchangeability of the infinite sum and limit, note

$$\mathbb{P}(S_k^{(n)} > 0) \leq \mathbb{P}(|S_k^{(n)} + k\beta| > k\beta) \leq \frac{k}{\beta^2 k^2} = \frac{1}{\beta^2 k},$$

where we used that $\mathbb{E}[S_k^{(n)}] = k\mathbb{E}[Y_1^{(n)}] = -k\beta$ and $\operatorname{Var} S_k^{(n)} = k$ and apply Chebychev's inequality, so that

$$\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(S_k^{(n)} > 0) \leq \sum_{k=1}^{\infty} \frac{1}{\beta^2 k^2} < \infty, \qquad \forall n \in \mathbb{N}.$$

Hence,

$$\begin{aligned}\lim_{n\to\infty} \ln \mathbb{P}(\hat{Q}^{(n)} = 0) &= \lim_{n\to\infty} -\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(S_k^{(n)} > 0) = -\sum_{k=1}^{\infty} \frac{1}{k} \lim_{n\to\infty} \mathbb{P}(S_k^{(n)} > 0) \\ &= -\sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(\textstyle\sum_{i=1}^{k} Z_i > 0) = \ln \mathbb{P}(M_\beta = 0).\end{aligned}$$

Finding a suitable upper bound on $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}^{(n)} > x)\mathrm{d}x$ and $\frac{1}{k}\int_0^\infty \mathbb{P}(\hat{Q}^{(n)} > \sqrt{x})\mathrm{d}x$ requires a bit more work. We initially focus on the former, the latter follows easily. The following inequality from [167] proves to be very useful:

$$\mathbb{P}(\bar{S}(k) > y) \leq C_r \left(\frac{k\,\sigma^2}{y^2}\right)^2 + k\,\mathbb{P}(X > y/r), \tag{3.53}$$

where $\bar{S}(k)$ is the sum of k i.i.d. random variables distributed as X, with $\mathbb{E}[X] = 0$ and $\operatorname{Var} X = \sigma^2$, $y > 0$, $r > 0$ and C_r a constant only depending on r. We take $r = 3$ for brevity in the remainder of the proof, although any $r > 2$ will suffice. We analyze the integral in two parts, one for the interval $(0,k)$ and one for $[k,\infty)$. For the first part, we have

$$\begin{aligned}\int_0^k \mathbb{P}(S_k^{(n)} > x)\mathrm{d}x &= \int_0^k \mathbb{P}(\textstyle\sum_{i=1}^{\infty} \hat{A}_i^{(n)} > x + k\beta)\mathrm{d}x \;\leq\; \displaystyle\int_0^k \mathbb{P}(\textstyle\sum_{i=1}^{\infty} \hat{A}_i^{(n)} > k\beta)\mathrm{d}x \\ &= k\,\mathbb{P}(\textstyle\sum_{i=1}^{k} \hat{A}_i^{(n)} > k\beta) \;\leq\; \displaystyle\frac{C_3}{k^2\beta^6} + k^2\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}k),\end{aligned} \tag{3.54}$$

where we used (3.53) in the last inequality. Hence,

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{1}{k} \int_0^k \mathbb{P}(S_k^{(n)} > x)\mathrm{d}x &\leq \frac{C_3}{\beta^6} \sum_{k=1}^{\infty} k^{-3} + \sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}k) \\ &\leq C_1^* + \sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}k).\end{aligned} \tag{3.55}$$

With the help of the inequality (see [193]),

$$(b-a)a\,\mathbb{P}(X>b) \le \int_a^b x\,\mathbb{P}(X>x)\mathrm{d}x \qquad \text{for } 0<a<b, \tag{3.56}$$

we get by taking $a=(k-1)/3$ and $b=k/3$,

$$\begin{aligned} k\,\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}k) &\le \frac{9\,k}{k-1}\int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}^{(n)}>x)\mathrm{d}x \\ &\le 18\int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}^{(n)}>x)\mathrm{d}x, \end{aligned} \tag{3.57}$$

for $k\ge 2$. Since the tail probability for $k=1$ is obviously bounded by 1, this yields

$$\begin{aligned} \sum_{k=1}^{\infty} k\,\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}k) &\le 1+18\sum_{k=2}^{\infty}\int_{(k-1)/3}^{k/3} x\,\mathbb{P}(\hat{A}^{(n)}>x)\mathrm{d}x \\ &\le 1+18\int_0^{\infty} x\,\mathbb{P}(\hat{A}^{(n)}>x)\mathrm{d}x \le 1+18\,\mathbb{E}[\hat{A}^{(n)2}]<\infty, \end{aligned} \tag{3.58}$$

since $\hat{A}^{(n)}$ has finite variance by assumption. This completes the integral over the first interval. For the second part, we use (3.53) again to find

$$\begin{aligned} \int_k^{\infty}\mathbb{P}(S_k^{(n)}>x)dx = \int_k^{\infty}\mathbb{P}(\textstyle\sum_{i=1}^{\infty}\hat{A}^{(n)}>x+k\beta)dx &\le \int_k^{\infty}\mathbb{P}(\textstyle\sum_{i=1}^{\infty}\hat{A}^{(n)}>x)\mathrm{d}x \\ &\le C_3\int_k^{\infty}\frac{k^2}{x^6}\mathrm{d}x + k\int_k^{\infty}\mathbb{P}(\hat{A}^{(n)}>\tfrac{1}{3}x)\mathrm{d}x \\ &= \frac{5C_3}{k^3}+k\int_k^{\infty}\mathbb{P}(\hat{A}^{(n)}>\tfrac{1}{3}x)\mathrm{d}x. \end{aligned} \tag{3.59}$$

So,

$$\sum_{k=1}^{\infty}\frac{1}{k}\int_k^{\infty}\mathbb{P}(S_k^{(n)}>x)\mathrm{d}x \le C_2^* + \sum_{k=1}^{\infty}\int_k^{\infty}\mathbb{P}(\hat{A}_i^{(n)}>\tfrac{1}{3}x)\mathrm{d}x, \tag{3.60}$$

for some constant C_2^*. Last, we are able to upper bound the second term in (3.60) by Tonelli's theorem:

$$\begin{aligned} \sum_{k=1}^{\infty}\int_k^{\infty}\mathbb{P}(\hat{A}_i^{(n)}>\tfrac{1}{3}x)dx &\le \int_1^{\infty} x\mathbb{P}(\hat{A}^{(n)}>\tfrac{1}{3}x)\mathrm{d}x \\ &\le 9\int_0^{\infty} y\mathbb{P}(\hat{A}^{(n)}>y)dy = 9\mathbb{E}[\hat{A}^{(n)2}]<\infty. \end{aligned} \tag{3.61}$$

Combining the results in (3.55), (3.58), (3.60) and (3.61), we find

$$\sum_{k=1}^{\infty}\frac{1}{k}\int_0^{\infty}\mathbb{P}(S_k^{(n)}>x)\mathrm{d}x<\infty,$$

and thus

$$\lim_{n\to\infty} \mathbb{E}[\hat{Q}^{(n)}] = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(S_k^{(n)} > x)\mathrm{d}x$$
$$= \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(\textstyle\sum_{i=1}^{k} Z_i > x)\mathrm{d}x = \mathbb{E}[M_\beta].$$

Finally, we show how the proof changes for the convergence of $\mathrm{Var}\,\hat{Q}^{(n)}$. The expressions for $\mathbb{E}[\hat{Q}^{(n)}]$ and $\mathrm{Var}\,\hat{Q}^{(n)}$ in (3.50) and (3.51) only differ in the term $\sqrt{x}$. Hence only minor modifications are needed to also prove convergence of the variance. Note that boundedness of the integral over the interval $(0, k)$ in (3.54)-(3.58) remains to hold when substituting $\sqrt{x}$ for x. (3.59) changes into

$$\int_k^{\infty} \mathbb{P}(S_k^{(n)} > \sqrt{x})dx = \int_k^{\infty} \mathbb{P}(\textstyle\sum_{i=1}^{\infty} \hat{A}_i^{(n)} > \sqrt{x} + k\beta)\mathrm{d}x$$
$$\leq C_3 \int_k^{\infty} \frac{k^2}{(\sqrt{x}+k\beta)^6} dx + k \int_k^{\infty} \mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}\sqrt{x})\mathrm{d}x$$
$$\leq \frac{C_4^*}{k} + k \int_k^{\infty} \mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}\sqrt{x})\mathrm{d}x,$$

for some constant C_4^*, so that

$$\sum_{k=1}^{\infty} \frac{1}{k} \int_k^{\infty} \mathbb{P}(S_k^{(n)} > \sqrt{x})\mathrm{d}x \leq C_4^* + \sum_{k=1}^{\infty} \int_k^{\infty} \mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}\sqrt{x})\mathrm{d}x.$$

Lastly, we have

$$\sum_{k=1}^{\infty} \int_k^{\infty} \mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}\sqrt{x})\mathrm{d}x \leq \int_1^{\infty} x\mathbb{P}(\hat{A}^{(n)} > \tfrac{1}{3}\sqrt{x})\mathrm{d}x$$
$$\leq 18 \int_0^{\infty} y^2 \mathbb{P}(\hat{A}^{(n)} > y)\mathrm{d}y = 18\,\mathbb{E}[\hat{A}^{(n)3}] < \infty.$$

Therefore the sum describing the variance is also uniformly convergent in n, so that interchanging of infinite sum and limit is permitted and

$$\lim_{n\to\infty} \mathrm{Var}\ \hat{Q}^{(n)} = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(S_k^{(n)} > \sqrt{x})\mathrm{d}x$$
$$= \sum_{k=1}^{\infty} \frac{1}{k} \int_0^{\infty} \mathbb{P}(\textstyle\sum_{i=1}^{k} Z_i > \sqrt{x})\mathrm{d}x = \mathrm{Var}\, M_\beta.$$

3.B Numerical procedures

An alternative characterization of the stationary distribution is based on the analysis in [45] and considers a factorization in terms of (complex) roots:

$$Q^{(n)}(w) = \frac{(s_n - \mathbb{E}[A^{(n)}])(w-1)}{w^{s_n} - \tilde{A}^{(n)}(w)} \prod_{k=1}^{s_n - 1} \frac{w - z_k^n}{1 - z_k^n},$$

where $z_1^n, z_2^n ..., z_{s_n-1}^n$ are the $s_n - 1$ zeros of $z^{s_n} - \tilde{A}^{(n)}(z)$, in $|z| < 1$, yielding

$$\mu_Q = \frac{\sigma_n^2}{2(s_n - \mu_n)} - \frac{s_n - 1 + \mu_n}{2} + \sum_{k=1}^{s_n-1} \frac{1}{1 - z_k^n},$$

$$\mathbb{P}(Q^{(n)} = 0) = \frac{s_n - \mu_A}{\tilde{A}^{(n)}(0)} \prod_{k=1}^{s_n-1} \frac{z_k^n}{z_k^n - 1},$$

which for our choice of $\tilde{A}^{(n)}(z)$ becomes

$$\mu_Q = \frac{a_n b_n (b_n + 1)}{2\beta\sqrt{a_n} b_n} - \frac{2a_n b_n + \beta\sqrt{a_n b_n (b_n + 1)} - 1}{2} + \sum_{k=1}^{s_n-1} \frac{1}{1 - z_k^n},$$

$$\mathbb{P}(Q^{(n)} = 0) = \beta\sqrt{a_n b_n (b_n + 1)}(1 + b_n)^{a_n} \prod_{k=1}^{s_n-1} \frac{z_k^n}{z_k^n - 1},$$

where $z_1, ..., z_{s_n-1}$ denote the zeros of $z^{s_n} - \tilde{A}^{(n)}(z)$ in $|z| < 1$. A robust numerical procedure to obtain these zeros is essential for a base of comparison. We discuss two methods that fit these requirements. The first follows directly from [113].

Lemma 3.4. *Define the iteration scheme*

$$z_k^{n,l+1} = w_k^n [\tilde{A}^{(n)}(z_k^{n,l})]^{1/s_n}, \tag{3.62}$$

with $w_k^n = e^{2\pi i k/s_n}$ *and* $z_k^{n,0} = 0$ *for* $k = 0, 1, \ldots, s_{n-1}$. *Then* $z_k^{n,l} \to z_k^n$ *for all* $k = 0, 1, ..., s_n - 1$ *for* $l \to \infty$.

Proof. The successive substitution scheme given in (3.62) is the fixed point iteration scheme described in [113], applied to the pgf of our arrival process. The authors show that, under the assumption of $\tilde{A}^{(n)}(z)$ being zero-free within $|z| \leq 1$, the zeros can be approximated arbitrarily closely, given that the function $[\tilde{A}^{(n)}(z)]^{1/s_n}$ is a contraction for $|z| \leq 1$, i.e.

$$\left| \frac{\mathrm{d}}{\mathrm{d}z} [\tilde{A}^{(n)}(z)]^{1/s_n} \right| < 1.$$

In our case,

$$\left| \frac{\mathrm{d}}{\mathrm{d}z} [\tilde{A}^{(n)}(z)]^{1/s_n} \right| = \left| \frac{\mathrm{d}}{\mathrm{d}z} (1 + (1 - z) b_n)^{-a_n/s_n} \right| = \frac{a_n b_n}{s_n} \Big| 1 + (1 - z) b_n \Big|^{-a_n/s_n - 1}, \tag{3.63}$$

where $a_n b_n / s_n = \rho_n$ is close to, but less than 1 and

$$|1 + (1 - z) b_n| \geq |1 + b_n| - |z| b_n = 1 + (1 - |z|) b_n \geq 1,$$

when $|z| \leq 1$. Hence the expression in (3.63) is less than 1 for all $|z| \leq 1$. Evidently, $\tilde{A}^{(n)}(z)$ is also zero-free in $|z| \leq 1$. Thus [113, Lemma 3.8] shows that $z_k^{n,l}$ as in (3.62) converges to the desired roots z_k^n for all k as l tends to infinity. $\square$

Remark 3.2. The asymptotic convergence rate of the iteration in (3.62) equals $\frac{\mathrm{d}}{\mathrm{d}z}[\tilde{A}^{(n)}(z)]^{1/s_n}$ evaluated at $z = z_k^n$. Hence, convergence is slow for zeros near 1 and fast for zeros away from 1.

A different approach is based on the Bürmann-Lagrange inversion formula.

Lemma 3.5. *Let $w_k^n = e^{2\pi ik/s_n}$ and $\alpha_n = a_n/s_n$. Then the zeros of $z^{s_n} - \tilde{A}^{(n)}(z)$ are given by*

$$z_k^n = \sum_{l=1}^{\infty} \frac{1}{l!} \frac{\beta[l\alpha_n + l - 1)}{\beta(l\alpha_n)} \frac{b_n + 1}{b_n} \Big(\frac{b_n}{(b_n+1)^{\alpha_n+1}} \Big)^l (w_k^n)^l,$$

for $k = 0, 1, ..., s_n - 1$.

Proof. Note that we are looking for z's that solve

$$z\,[\tilde{A}^{(n)}(z)]^{-1/s_n} = z\,(1 + (1-z)b_n)^{a_n/s_n} = w,$$

where $w = w_k = \mathrm{e}^{2\pi ik/s_n}$. The Bürmann-Lagrange formula for $z = z(w)$, as can be found in [67, Sec. 2.2] for $z = z(w)$ is given by

$$\begin{aligned} z(w) &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^{l-1} \left[\left(\frac{z}{z(1 + (1-z)b_n)^{a_n/s_n}} \right)^l \right]_{z=0} w^l \\ &= \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^{l-1} \left[\left(1 + (1-z)b_n \right)^{-l\,a_n/s_n} \right]_{z=0} w^l. \end{aligned}$$

Set $\alpha_n = a_n/s_n$. We compute

$$\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^{l-1} \left[(1 + (1-z)b_n)^{-l\alpha_n} \right]_{z=0} = \frac{\beta(l\alpha_n + l - 1)}{\beta(l\alpha_n)} \frac{1 + b_n}{b_n} \left(\frac{b_n}{(1+b_n)^{\alpha_n+1}} \right)^l.$$

With $c_n = b_n/(1+b_n)^{\alpha_n+1}$ and $d_n = (1+b_n)/b_n$, we thus have

$$z(w) = d_n \sum_{l=1}^{\infty} \frac{\beta(l\alpha_n + l - 1)}{\beta(l+1)\beta(l\alpha_n)} c_n^l\, w^l.$$

By Stirling's formula

$$\frac{\beta(l\alpha_n + l - 1)}{\beta(l+1)\beta(l\alpha_n)} = \frac{D}{l\sqrt{l}} \left(\frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} \right)^l,$$

where $D = \alpha_n^{1/2}(\alpha_n+1)^{-3/2}(2\pi)^{-1/2}$. Now,

$$\frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} c_n = \frac{(\alpha_n+1)^{\alpha_n+1}}{\alpha_n^{\alpha_n}} \cdot \frac{b_n}{(1+b_n)^{\alpha_n+1}} = \left(\frac{b_n + \rho_n}{b_n + 1} \right)^{\rho_n/b_n+1} \left(\frac{1}{\rho_n} \right)^{\rho_n/b_n}.$$

This determines the radius of convergence r_{BL} of the above series for $z(w)$:

$$\frac{1}{r_{\mathrm{BL}}} := \left(\frac{b_n + \rho_n}{b_n + 1} \right)^{\rho_n/b_n+1} \left(\frac{1}{\rho_n} \right)^{\rho_n/b_n}. \tag{3.64}$$

The derivative with respect to ρ_n of the quantity

$$\left(1+\frac{\rho_n}{b_n}\right)\ln\left(\frac{b_n+\rho_n}{b_n+1}\right)+\frac{\rho_n}{b_n}\ln\left(\frac{1}{\rho_n}\right) \tag{3.65}$$

is given by

$$\frac{1}{b_n}\ln\Big(\frac{b_n+\rho_n}{b_n\rho_n+\rho_n}\Big)>0$$

for $0<\rho_n<1$ and $b_n>0$. Furthermore, the quantity in (3.65) vanishes at $\rho_n=1$ and is therefore negative for $0<\rho_n<1$ and $b_n>0$.

Remark 3.3. The formula for the radius of convergence in (3.64) clearly shows the decremental effect of both having a large b_n and of having ρ_n close to 1. The quantities $\beta(l\alpha+l-1)/(\beta(l+1)\beta(l\alpha))$ in the power series for $z(w)$ are not very convenient for recursive computation, although normally $\alpha_n=a_n/s_n$ is a rational number.

□

3.C Statistical procedures

To calibrate our model to real data, we now discuss some statistical procedures to show the presence of overdispersion and to estimate the parameters of the mixed Gamma-Poisson distribution. Let $x_1,...,x_n$ denote the observed arrival counts in consecutive time slots. These observations can be interpreted as realizations of the random variables $A_1,...,A_N$, and

$$\bar{a}_N=\frac{1}{N}\sum_{i=1}^{N}x_i,\qquad \bar{s}_N^2=\frac{1}{N-1}\sum_{i=1}(x_i-\bar{x}_i)^2,$$

the sample mean and variance with equivalent random variables $\bar{A}_N$ and S_N^2, respectively. Several tests with null hypothesis that $x_1,...,x_N$ originate from a (constant rate) Poisson distribution were discussed by [50]. We mention two of them. The first is frequently referred to as the *dispersion test*, and is based on the test statistic

$$D_N=\frac{(N-1)S_N^2}{\bar{A}_N},$$

which is approximately chi-squared distributed with $N-1$ degrees of freedom. When using a significance level α, the critical value is equal to the $(1-\alpha)$-th quantile of chi-squared distribution $\chi^2_{N-1,1-\alpha}$. The second test, which is also used in [127], involves the test statistic

$$T_N=\sqrt{N/2}\,\Big(\frac{S_N^2}{\bar{A}_N}-1\Big),$$

which is known as the Neyman-Scott test statistic. Under the null hypothesis T_N tends to a standard normal random variable for large N. Hence both test statistics rely on the ratio of the sample variance and sample mean, which should be 1 if $A_1, ..., A_N$ are indeed i.i.d. Poisson distributed. Excessive values of D_N and T_N therefore raise the suspicion of overdispersed arrivals.

Once either (or both) of the Poisson tests rejects the hypothesis of arrivals originating from a unicomponent Poisson process, we fit the data to the Gamma-Poisson mixture. Note that if we assume A_i to be distributed as a Poisson random variable with random rate Λ_i, which is in turn Gamma distributed with parameters a and $1/b$, then A_i is in fact a negative binomial random variable with parameters $r = a$ and $p = b/(b+1)$. Finding estimators $\hat{a}$ and $\hat{b}$ therefore is equivalent to fitting a negative binomial distribution to the data to obtain $\hat{r}$ and $\hat{p}$, followed by retrieving $\hat{a} = \hat{r}$ and $\hat{b} = \hat{p}/(1-\hat{p})$. We proceed by applying the maximum likelihood estimation method described in [127] to find $\hat{r}$ and $\hat{p}$. This method prescribes to set $\hat{r}$ to be the value of r for which the *profile loglikelihood function* defined by

$$L(r) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{a_i} \ln(r+j+1) + r \ln r - (r + \bar{a}_N) \ln(r + \bar{a}_N),$$

is attained. Subsequently, $\hat{p} = \hat{r}/(\hat{r} + \bar{a}_N)$, so that $\hat{a} = \hat{r}$ and $\hat{b} = \hat{r}/\bar{a}_N$.

Finally, given the estimators $\hat{a}$ and $\hat{b}$, we need statistical evidence that the obtained Poisson mixture indeed fits the data reasonably well. Here we again cite [127], who give a method to retrieve the p-value for the goodness-of-fit based on bootstrap and Monte-Carlo simulation. In our experiments, we work with 10^6 replications of the Monte-Carlo simulation to obtain the approximated p-value. We refer to the appendix of [127] for further details on this method.

4

Retrial queues in the QED regime

Large-scale queueing systems with retrying customers are intrinsically hard to evaluate analytically. We in this chapter explore and extend the asymptotic approximation technique proposed by Avram et al. [25], that is able to characterize the impact of slow retrials in the QED regime, in three queueing models. The technique evolves around a fixed-point equation that quantifies the increased inflow due to retrials implicitly. We translate this fixed-point method into a powerful and elegant dimensioning procedure that is able to deal with both stationary and time-varying demand.

Based on

Delayed workload shifting in many-server systems

Johan van Leeuwaarden, Britt Mathijsen & Fiona Slothaak

In *SIGMETRICS Performance Evaluation Review, 43(2), 10–12 (2015)*

and

Cloud provisioning in the QED regime

Johan van Leeuwaarden, Britt Mathijsen & Fiona Slothaak

In *Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools, 180–187 (2016)*

4.1 Introduction

Retrial queues. In the previous chapters, we analyzed queueing systems in which all arriving customers join the queue and stay until eventually completing service with one of the servers. From a practical perspective though, these assumptions are questionable. For instance, in call centers, customer impatience is known to play a crucial role in the queueing dynamics, see e.g. [80, 49, 229]. Similar features are also seen in health care [65, 17]. However, impatience may not be the only cause of customers leaving the system without being seen by a server. Physical constraints may force system managers to apply some sort of admission policy. The simplest example of such admission control is the busy-signal in call centers, in which arriving customers finding all servers busy are simply discarded. But more elaborate strategies can be considered. A straightforward relaxation of the busy-signal policy is to allow a finite amount of waiting space, and block customers who find a full waiting room upon arrival. Many other options, such as probabilistic and dynamic admission control policies may be considered, see e.g. [119, 18] and references therein.

Since customers arrive to the system for the purpose of getting assistance from one of the servers, it is reasonable to assume that these refused customers retry getting access tot the system at a later point in time. In fact, retrials are widely observed in telecommunication systems, see e.g. [60, 73, 151, 10], and customers typically repeat their attempt until successful. Naturally, retrials have a detrimental effect on the performance of the queueing system in terms of QoS, compared to the setting in which blocked customers do not return. Hence, one needs to account for their impact in both performance analysis and the staffing decisions.

Unfortunately, the modeling of retrials is analytically challenging [60, 73], and numerical approaches become computationally infeasible as the number of servers increases, which is precisely the regime we are interested in. We therefore aim to tackle the performance analysis of such retrial systems in an asymptotic manner. We do so through a clever technique that was recently documented by Avram et al. [25].

Fixed-point equation. In this chapter, we will show how the asymptotic approximation technique of [25] can be extended to more complex large-scale retrial queueing systems in the QED regime. In [25], the authors study the $M/M/s/s$ queue with slow retrials. That is to say, customers retry only after a (stochastic) delay period that is relatively long compared to the service time. Under this assumption, the authors combine QED limits with a fixed-point equation, which characterizes the impact of retrials implicitly. We summarize and reformulate their main ideas here for completeness.

Consider the standard $M/M/s/s$ queue with arrival rate λ and service rate $\mu = 1$, so that the offered load is $R = \lambda$. Customers finding upon arrival all s servers busy retry after a stochastic delay with mean $1/\delta$. The first ingredient of the method is Cohen's equation. This result, first reported by Cohen [60], says that the stationary distribution of a $M/M/s/s$ queue with retrials converges as $1/\delta \to \infty$ to that of a $M/M/s/s$ queue with increased arrival rate $R+\Omega$, where Ω is the unique positive solution to

$$\Omega = (R+\Omega)\ B(R+\Omega, s), \tag{4.1}$$

where $B(R,s)$ denotes the blocking probability in the $M/M/s/s$ queue with offered load R, i.e. the Erlang-B formula:

$$B(R,s) := \frac{R^s/s!}{\sum_{k=0}^{s} R^k/k!} = \frac{\mathbb{P}(\mathrm{Pois}(R)=s)}{\mathbb{P}(\mathrm{Pois}(R)\le s)}. \tag{4.2}$$

Equation (4.1) essentially equates the arrival volume generated by retrials, given by definition on the left-hand side, to the right-hand side which quantifies this volume as a fraction of customers blocked times the increased arrival volume. Indeed, the retrial stream can for long retrial times be considered as independent from the primary arrival stream, yielding a thinned Poisson process [60].

The second crucial observation is that under QED scaling, i.e. $s = R + \beta\sqrt{R} + o(\sqrt{R})$, we have

$$\sqrt{R}\cdot B(R, R+\beta\sqrt{R}) \to \frac{\varphi(\beta)}{\Phi(\beta)} =: f_0(\beta), \tag{4.3}$$

as $R \to \infty$ for all $\beta \in \mathbb{R}$, see e.g. Lemma 1 of [25] or the proof of Proposition 1.1 in this thesis. Hence, we heuristically deduce that $\Omega = \alpha\sqrt{R}$ for some $\alpha > 0$ and denote $R_{\mathrm{tot}} = R + \Omega$ – a rigorous argument can be found in [25]. Rewrite $R = R_{\mathrm{tot}} - \alpha\sqrt{R_{\mathrm{tot}}} + o(\sqrt{R_{\mathrm{tot}}})$ and note that $R_{\mathrm{tot}} = O(R)$. Then (4.1) becomes

$$\alpha\sqrt{R_{\mathrm{tot}}} = R_{\mathrm{tot}}\cdot B\big(R_{\mathrm{tot}}, R_{\mathrm{tot}} + (\beta-\alpha)\sqrt{R_{\mathrm{tot}}}\big) + o(\sqrt{R_{\mathrm{tot}}}). \tag{4.4}$$

Dividing both sides of (4.4) by $\sqrt{R_{\mathrm{tot}}}$ and letting $R_{\mathrm{tot}} \to \infty$ then together with (4.3) yields the fixed-point equation

$$\alpha = f_0(\beta - \alpha). \tag{4.5}$$

It can be shown that this fixed-point equation has a unique positive root for all $\beta > 0$, which can be computed numerically. Recalling Cohen's retrial queue characterization, Avram et al. [25] conclude that the loss system with slow retrials can in the QED regime be characterized in terms of the original loss system without retrials, but with a corrected QoS-parameter $\beta - \alpha$.

Structure of the chapter. The fixed-point method of Avram et al. provides a quick and elegant way to approximate the behavior of large-scale loss systems that experience retrials. In the remainder of this chapter we will explore if and how this technique extends to three more complex queueing settings. These three models have in common that (i) they exhibit QED limiting behavior, which can be quantified explicitly, and (ii) the blocking probability is $O(1/\sqrt{R})$, so that the retrial volume is $O(\sqrt{R})$. Note that these were the two essential features for the fixed-point method to work.

In Section 4.2 we describe a direct extension of the $M/M/s/s$ queue, in which some amount of waiting room is present. That is, we analyze the $M/M/s/n$ queue with retrials, where $n > s$. Naturally, this requires a scaling for both s and n as $R \to \infty$, which will become clear in this section. Motivated by a process related to cloud computing, we in Section 4.3 study a tandem queueing network, in which total number of concurrent admissions is limited. Section 4.4 analyzes a queueing model in which all customers are admitted upon arrival, but make the deliberate decision to abandon the queue and retry later in case their patience runs out. In Section 4.5 we show how the fixed-point method together with QED scaling can be used for dimensioning purposes in both stationary and time-varying environments. We end the chapter in Section 4.6 with some final remarks and suggestions for future research.

4.2 The $M/M/s/n$ queue

In this section, we discuss a simple extension of the loss model of [25], namely the $M/M/s/n$ queue with retrials with $n > s$, to expose typical behavior of retrial queues and the influence of the retrial rate δ. Second, we illustrate the fixed-point method for this model and perform numerical experiments to verify its accuracy.

4.2.1 Markov process

We consider the standard $M/M/s/n$ queue with arrival rate λ and service rate μ. Without loss of generality, we set $\mu = 1$ throughout this chapter, so that offered load R equals λ. A customer that finds upon arrival a free server occupies this server immediately, while customers that meet more than s but fewer than $n > s$ customers in the system are admitted and wait in a queue for a server to become available. Customers who meet upon arrival n customers are not admitted directly, but will retry after an exponentially distributed time with mean $1/\delta$. Each initially blocked customer performs retrials until admitted eventually; see Figure 4.1.

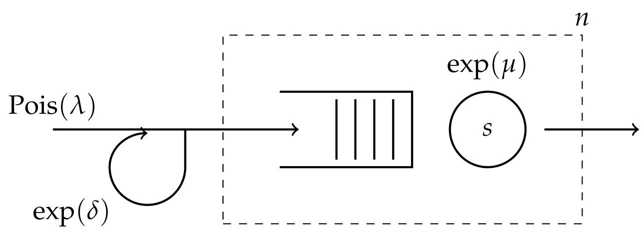


Figure 4.1: An $M/M/s$ queue with space constraints and retrials.

Quasi-birth-death process. The system state can be described by a two-dimensional process $\{(Q(t), N(t))\}_{t\geq 0}$ with $Q(t)$ the number of customers inside the system (either being served or waiting), and $N(t)$ the number of customers in the retrial

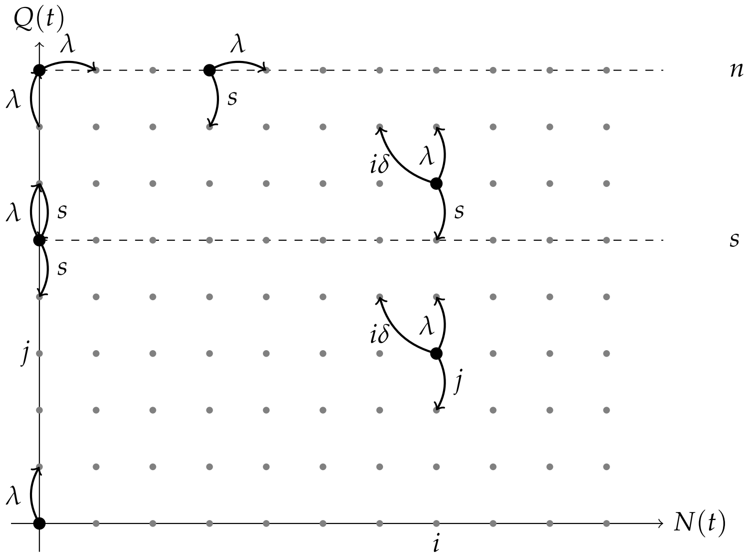


Figure 4.2: Transition diagram of the Markov process $(Q(t), N(t))$.

orbit. Under the above assumptions, this process is a continuous-time Markov chain on the semi-infinite strip $\{0, 1, \ldots, n\} \times \{0, 1, \ldots\}$. Its transition diagram is presented in Figure 4.2. From this diagram it is evident that the process is a quasi-birth-death (QBD) process. Under stability condition $R > s$, the QBD structure of the process allows for numerical computation of the stationary distribution $\pi(i, j)$, where

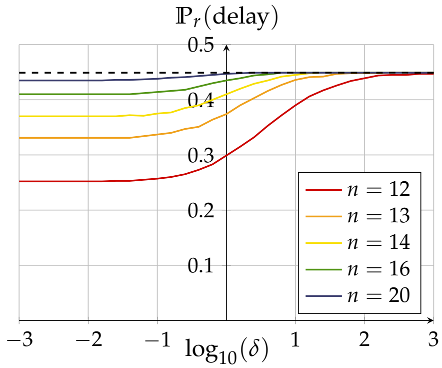
$$\pi(i,j) = \lim_{t\to\infty} \mathbb{P}\left(Q(t) = i, N(t) = j\right).$$

The stationary probability that an arriving customer has to wait or is blocked is given by, respectively,

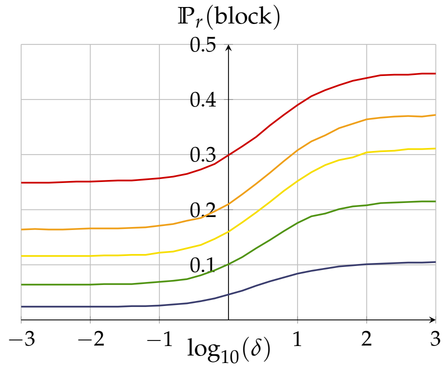
$$\mathbb{P}_r(\text{delay}) = \sum_{i=s}^{n}\sum_{j=0}^{\infty} \pi(i,j), \qquad \mathbb{P}_r(\text{block}) = \sum_{j=0}^{\infty} \pi(n,j). \tag{4.6}$$

Here, the subscript r is meant to indicate that we consider the system with retrials.

Influence of retrial rate δ. We first compute the stationary distribution of the Markov process numerically, in order to understand the influence of retrials on the queue performance. In particular, we investigate the effect of varying δ on the delay and blocking probability as defined in (4.6). In Figure 4.3 we fix $R = 10$ and $s = 12$, and plot the delay and blocking probability as a function of $\log_{10}(\delta)$ for several values of n. We see that the value of δ indeed does influence the performance of the queue, and its effect is particularly pronounced in systems with n close to s. Both the delay probability and the blocking probability increase with δ. This can



(a) Delay probability



(b) Blocking probability

Figure 4.3: Performance metrics of the basic model with $R = 10$ and $s = 12$ as a function of $\log_{10}(\delta)$ for several n.

be explained as follows. If a customer finds n customers on arrival (or retrial) and hence gets blocked, she is more likely to find a less congested system in case she retries after a relatively long amount of time than a short retrial time, because the system might not yet have had enough time to recover from the congested period. Slow retrials hence create an opportunity to smooth out workload over time, resulting in better quality-of-service. Figure 4.3 also suggests that performance no longer changes if δ is decreased below 10^{-1} or increased beyond 10^2.

Also, we note that the delay probability increases with n, and the blocking probability decreases with n, regardless of the value of δ. Fewer customers get blocked if the waiting room $(n - s)$ increases. On the other hand, this allows more customers to enter the system, creating higher congestion levels.

Finally, notice that the delay probability approaches a constant as $\delta \to \infty$. In fact, this constant equals the delay probability in the standard $M/M/s$ queue, see Equation (1.2), which under these parameter settings equals 0.449 and is represented by the dashed horizontal line. Indeed, when $\delta \to \infty$ blocked customers retry getting access to the system instantaneously and effectively create a queue (in random order) outside the system, which immediately fills up vacant spaces after service completions. Therefore, the $M/M/s/n$ queue with instant retrials essentially resembles the behavior of the $M/M/s$ queue. By similar reasoning, the blocking probability in the $M/M/s/n$ queue approaches as $\delta \to \infty$ the probability that the number of customers in the $M/M/s$ queue exceeds n.

Figure 4.3 shows that the influence of retrials on congestion can be significant. For fast retrials, we are able to characterize the performance metrics through the standard multi-server queue. However, for slow retrials, say $\delta < 10^{-1}$, the system behavior is not comparable to that of the open $M/M/s$ queue.

4.2.2 QED regime

Following the approach of Avram et al. [25], we choose to take a step back and consider the model in Figure 4.1 without the retrials first. When blocked customers are simply discarded, the process $\{(Q(t), N(t))\}_{t\geq 0}$ reduces to that of the $M/M/s/n$ queue. In this case $N(t) = 0$ and $Q(t)$ is a birth-death process with stationary distribution

$$\pi(i) = \lim_{t\to\infty} \mathbb{P}(Q(t) = i) = \begin{cases} \pi(0)\,\frac{R^i}{i!}, & \text{if } i < s, \\ \pi(0)\,\frac{R^i}{s!s^{i-s}}, & \text{if } s \leq i \leq n, \end{cases}$$

where

$$\pi(0) = \Big(\sum_{i=0}^{s-1} \frac{R^i}{i!} + \sum_{i=s}^{n} \frac{R^i}{s!s^{i-s}}\Big)^{-1}.$$

Hence,

$$\mathbb{P}(\text{delay}) = \pi(0)\sum_{i=s}^{n} \frac{R^i}{s!s^{i-s}}, \qquad \mathbb{P}(\text{block}) = \pi(0)\frac{R^n}{s!s^{n-s}}. \tag{4.7}$$

The $M/M/s/n$ queue is well understood. In particular, Massey & Wallace [160] identified the asymptotic scaling regime for s and n under which QED-type behavior prevails. Namely, under the two-fold scaling rule

$$\begin{aligned} s &= R + \beta\sqrt{R} + o(\sqrt{R}), \\ n &= s + \gamma\sqrt{R} + o(\sqrt{R}), \end{aligned} \tag{4.8}$$

for $\beta \in \mathbb{R}$ and $\gamma > 0$, they show that the delay probability converges to a value strictly between 0 and 1, while the blocking probability vanishes as $R \to \infty$. Note that this is in line with our reasoning in Section 1.4. In the next proposition, we cite the asymptotic results of [160] for completeness.

Proposition 4.1 ([160]). *If s and n scale according to* (4.8)*, then in the $M/M/s/n$ queue,*

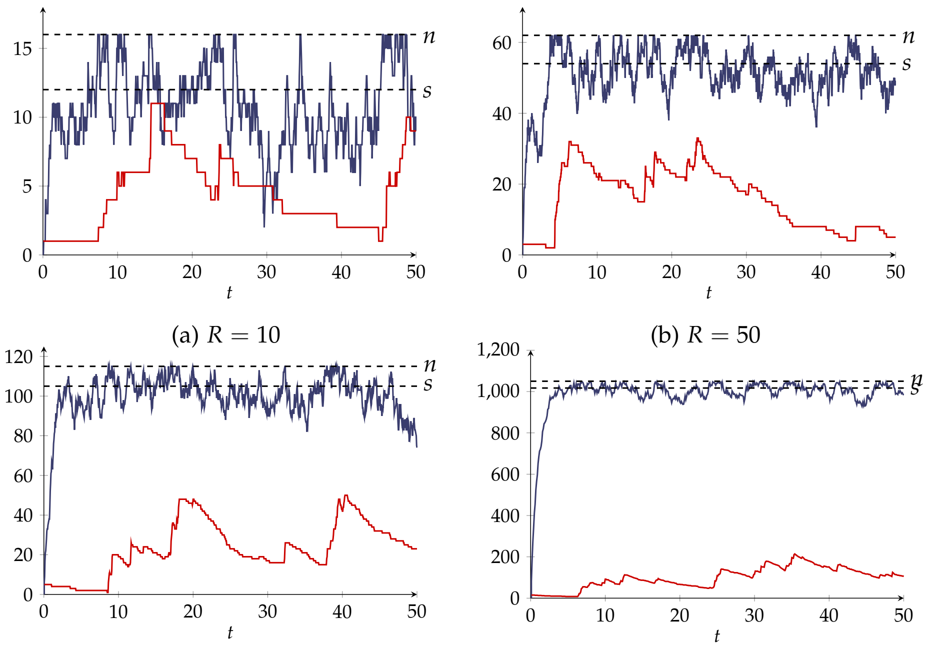
$$\mathbb{P}(\text{delay}) \to \frac{1 - \mathrm{e}^{-\beta\gamma}}{1 - \mathrm{e}^{-\beta\gamma} + \beta\,\Phi(\beta)/\varphi(\beta)} =: g(\beta,\gamma), \tag{4.9}$$

$$\sqrt{R}\,\mathbb{P}(\text{block}) \to \frac{\beta\,\mathrm{e}^{-\beta\gamma}}{1 - \mathrm{e}^{-\beta\gamma} + \beta\,\Phi(\beta)/\varphi(\beta)} =: f(\beta,\gamma), \tag{4.10}$$

as $R \to \infty$.

Before turning to the asymptotic analysis of the model with retrials, we check empirically whether the scaling in (4.8) also achieves the desirable limiting behavior in case blocked customers are not discarded. In Figure 4.4 we plot sample paths $Q(t)$ and $N(t)$ in the system with retrials with $\beta = 0.5$ and $\gamma = 1$ and slow retrials ($\delta = 0.1$) for increasing values of R.

From these sample paths, we observe that indeed the server utilization approaches unity as R tends to infinity, indicating efficient usage of resources. This



(a) $R = 10$

(b) $R = 50$

(c) $R = 100$

(d) $R = 1000$

Figure 4.4: Sample paths of $Q(t)$ (blue) and $N(t)$ (red) for increasing R while s and n are scaled as in (4.8) with $\beta = 0.5$ and $\gamma = 1$ and retrial rate $\delta = 0.1$.

should not be surprising, since although retrials occur, all customers eventually receive service, so that the server utilization equals $R/s = R/(R+\beta\sqrt{R}) \to 1$ as $R \to \infty$. Furthermore, we see that the number of customers in the system concentrates around the level s, implying a delay probability away from both 0 and 1. Observe that the order of magnitude of $N(t)$, the number of customers in the retrial orbit, is smaller than $Q(t)$ or R. This implies that as R grows large, only a small fraction of customers ends up retrying. Naturally, the order of $N(t)$ also depends on the mean retrial time $1/\delta$. It can be numerically verified that the expected retrial population grows linearly in $1/\delta$. Last, observe that $N(t)$ is increasing only if $Q(t) = n$, which is visible through the surges in the sample paths of $N(t)$ in Figure 4.4. This is illustrative for the dependency between the two coordinates of the process $\{(Q(t), N(t))\}_{t\geq 0}$ and therefore, we cannot expect to find a simple decoupling in the limit either. Instead, we propose to evaluate the model with retrials through a heuristic approach which builds upon the asymptotic behavior of the model without retrials.

4.2.3 Fixed-point method

We continue to translate the ideas behind the fixed-point method by noting that due to (4.10), the fraction of blocked customers is of order $1/\sqrt{R}$, which implies that the mean additional load due to retrials must be of order $\sqrt{R}$. We can thus assume that the total arrival rate R_{tot} takes the form $R_{\text{tot}} = R + \alpha\sqrt{R}$ for some $\alpha > 0$. Then, using that $R = O(R_{\text{tot}})$, the first scaling rule in (4.8) is asymptotically equivalent with

$$s = R_{\text{tot}} + (\beta - \alpha)\sqrt{R_{\text{tot}}} + o(\sqrt{R_{\text{tot}}}), \tag{4.11}$$

while the scaling for n remains unchanged. We thus argue that the retrial system in the QED regime mimics an $M/M/s/n$ queue with parameters $\beta_\alpha = \beta - \alpha$ and γ. Note that the volume of blocked users in this setting is $f(\beta-\alpha,\gamma)\sqrt{R_{\text{tot}}}$. This quantity must equal the mean additional load $\alpha\sqrt{R} \sim \alpha\sqrt{R_{\text{tot}}}$ and therefore we obtain the *fixed-point equation*

$$\alpha = f(\beta - \alpha, \gamma). \tag{4.12}$$

Numerically determining α is straightforward, particularly because it is uniquely defined.

Lemma 4.1. *Equation* (4.12) *has a unique solution for all* $\beta, \gamma > 0$.

Proof. Let $h(\beta) := \varphi(\beta)/\Phi(\beta)$ and $w(\beta) := (1 - \mathrm{e}^{-\beta\gamma})/\beta$. Write

$$f(\beta) := f(\beta,\gamma) = \frac{(1-\beta w(\beta))h(\beta)}{1 + w(\beta)h(\beta)}, \tag{4.13}$$

so that

$$\beta + f(\beta) = \frac{\beta + h(\beta)}{1 + w(\beta)h(\beta)} \tag{4.14}$$

For $h(\beta)$ it is known that, see [189], for $\beta \in \mathbb{R}$,

$$h(\beta) > -\beta, \quad -1 < h'(\beta) < 0, \quad h''(\beta) > 0, \tag{4.15}$$

so that $h(\beta)$ is non-negative and non-increasing in $\beta \in \mathbb{R}$, while $\beta + h(\beta)$ is positive and strictly increasing in $\beta \in \mathbb{R}$. Because $\mathrm{e}^x \geq 1 + x$,

$$w'(\beta) = \frac{\mathrm{e}^{-\beta\gamma}}{\beta^2}(1 + \beta\gamma - \mathrm{e}^{\beta\gamma}) \leq 0 \tag{4.16}$$

so $w(\beta)$ is also non-negative and non-increasing in $\beta \in \mathbb{R}$. It thus follows that $\beta + f(\beta)$ is strictly increasing in $\beta \in \mathbb{R}$. Moreover, $\beta + f(\beta) \to 0$ as $\beta \to -\infty$ and $\beta + f(\beta) \to \infty$ as $\beta \to \infty$. Let $\Delta = \beta - \alpha$, and rewrite (4.12) as

$$\beta = \Delta + f(\Delta). \tag{4.17}$$

Hence, for each fixed $\beta > 0$ there is a unique solution $\Delta \in \mathbb{R}$ from which $\alpha = \beta - \Delta$ follows. $\square$

As a result, the delay probability $\mathbb{P}_r(\text{delay})$ and the blocking probability $P_r(\text{block})$ in the model with retrials can be approximated in the QED regime by

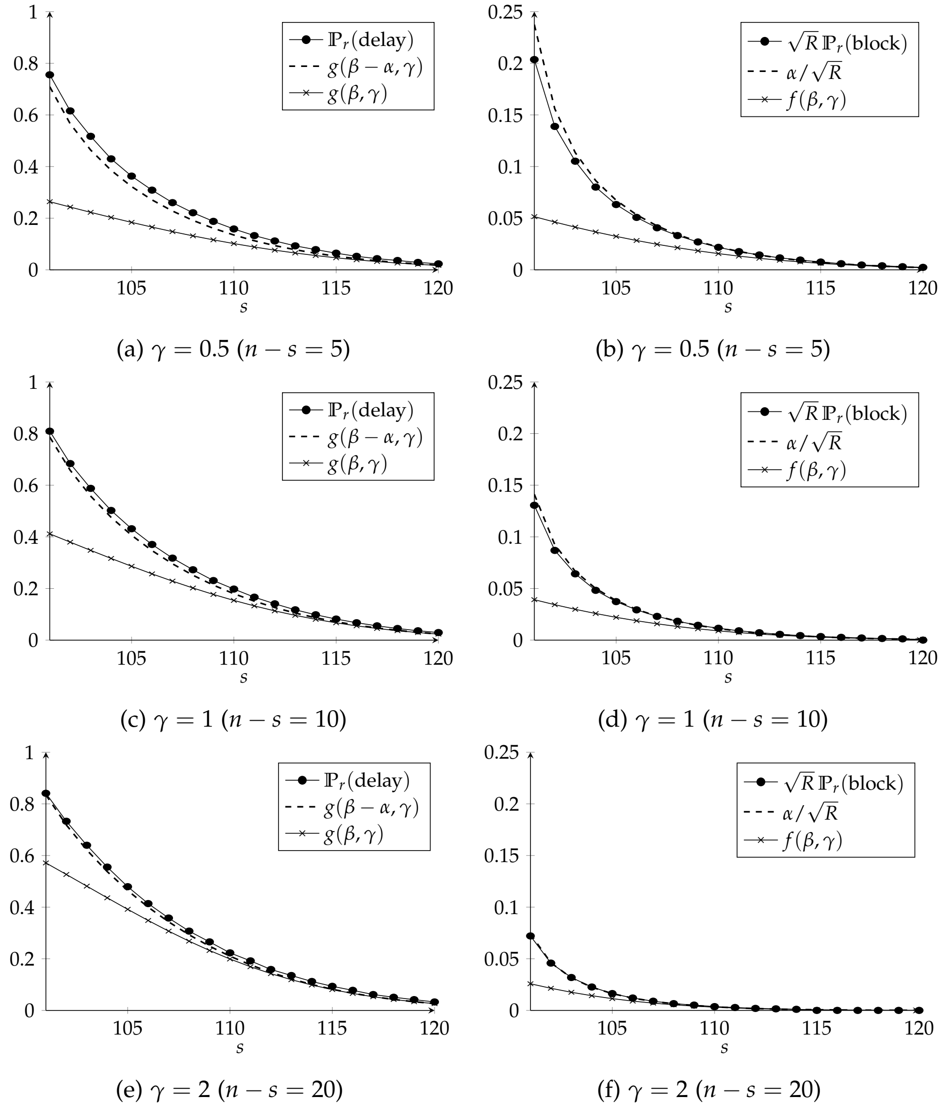
$$\mathbb{P}_r(\text{delay}) \approx g(\beta - \alpha, \gamma), \qquad \mathbb{P}_r(\text{block}) \approx \alpha/\sqrt{R}, \tag{4.18}$$

which should become more accurate as R grows large.

We next test the accuracy of the approximated delay probability in (4.18) in the basic model with slow retrials against the true values obtained through simulation. Given R, s, and n, we compute $\beta = (s - R)/\sqrt{R}$ and $\gamma = (n - s)/\sqrt{R}$ in order to approximate the delay and blocking probability as in (4.18) with α as in (4.12). First, we assess the quality of the fixed-point approximation for a large but finite system with $R = 100$. In Figure 4.5, we plot the simulated delay probability against the approximation as a function of s (or equivalently β). We consider different values of γ, namely $\gamma = 0.5$, 1 and 2, which corresponds to waiting room size $\gamma\sqrt{100} = 5$, 10 and 20, respectively. For comparison, we also include $g(\beta, \gamma)$, the asymptotic delay probability in the system with no retrials, in these plots.

We observe that the heuristic is remarkably accurate in describing both the delay and blocking probability over all values of $\beta, \gamma > 0$ considered here. The approximation improves as γ increases. Figure 4.5 also clearly illustrates the impact of retrials on the performance measures, which decreases with both β and γ.

Table 4.1 furthermore shows how the accuracy of the approximations increases as R increases. In this table, we used the simulated delay and blocking probability for systems of increasing size while adhering to the two-fold scaling rule of (4.8). The values of s and n are rounded to the nearest integer.



(a) $\gamma = 0.5$ $(n - s = 5)$

(b) $\gamma = 0.5$ $(n - s = 5)$

(c) $\gamma = 1$ $(n - s = 10)$

(d) $\gamma = 1$ $(n - s = 10)$

(e) $\gamma = 2$ $(n - s = 20)$

(f) $\gamma = 2$ $(n - s = 20)$

Figure 4.5: Accuracy of the delay probability approximation in basic model with $R = 100$ and $\delta = 0.01$.

	$(\beta,\gamma) = (0.5, 0.5)$				$(\beta,\gamma) = (1, 0.5)$			
R	s	n	$\mathbb{P}_r(\text{delay})$	$\sqrt{R}\mathbb{P}_r(\text{bl.})$	s	n	$\mathbb{P}_r(\text{delay})$	$\sqrt{R}\mathbb{P}_r(\text{bl.})$
5	6	7	0.5019	0.5982	7	8	0.2607	0.2610
10	12	14	0.3697	0.3679	13	15	0.2298	0.1966
50	54	58	0.3509	0.4931	57	61	0.1765	0.2019
100	105	110	0.3640	0.6336	110	115	0.1579	0.2178
500	511	522	0.3460	0.6780	522	533	0.1482	0.2297
1000	1016	1032	0.3333	0.6481	1032	1048	0.1412	0.2141
		Approx	0.3225	0.6734		Approx	0.1349	0.2206

	$(\beta,\gamma) = (0.5, 1)$				$(\beta,\gamma) = (1, 1)$			
R	s	n	$\mathbb{P}_r(\text{delay})$	$\sqrt{R}\mathbb{P}_r(\text{bl.})$	s	n	$\mathbb{P}_r(\text{delay})$	$\sqrt{R}\mathbb{P}_r(\text{bl.})$
5	6	8	0.5337	0.4065	7	9	0.2866	0.1612
10	12	15	0.3932	0.2701	13	16	0.2472	0.1374
50	54	61	0.3993	0.3171	57	64	0.2063	0.1183
100	105	115	0.4333	0.3754	110	120	0.1971	0.1143
500	511	533	0.4247	0.3986	522	544	0.1928	0.1202
1000	1016	1048	0.4115	0.3689	1032	1064	0.1831	0.1088
		Approx	0.4062	0.3828		Approx	0.1798	0.1106

Table 4.1: Numerical results of the fixed-point method for the basic model as $R \to \infty$.

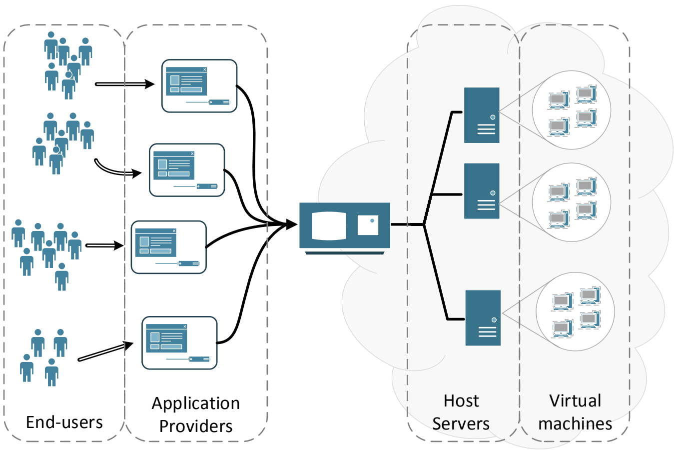


Figure 4.6: Cloud provisioning process

4.3 Cloud model

The second model we consider in this chapter is inspired by cloud computing services. We shall see how our fixed-point heuristic helps cloud providers in their provisioning process.

4.3.1 Practical context

Cloud computing enables network access to a shared pool of configurable computing resources, allowing users (e.g. companies, service providers) to store and process their data in third-party data centers, without investing in the operating equipment themselves. At the foundation of cloud computing lies the idea of sharing resources to achieve economies-of-scale in terms of maximizing computing power usage and reducing the overall cost of resources such as energy and infrastructure. Cloud providers, such as Amazon EC2, Windows Azure and Rackspace [15], offer virtual machine (VM) provisioning, which allows users to request VM instances configured to their preference. In a service system context, the provider thus serves users by supplying them with a VM that matches their requirements, running on one of the cloud's physical machines.

Let us describe the cloud provisioning process in more detail; see Figure 4.6. At the highest granularity level there are the *end-users*, devices typically directly operated by humans, using an *application provider* (AP), usually a company that provides software usage over the internet (e.g. SaaS [166]). To some extent, the AP will rely on a static set of computing resources, but certainly in case of sudden surges in workload, these might not be sufficient. When the AP recognizes the need for additional capacity, for instance by *auto-scaling* procedures [1], a VM request is submitted to the cloud provider. The request is handled by a *host server* that starts the set-up of the VM with requested specifications. This includes elementary operations such as copying the VM image and assigning an IP address. Each server is able to host multiple VM instances in parallel, although the VMs in set-up need their dedicated attention, due to concurrency level constraints incurred by large I/O activities. Once the set-up is completed, the VM is ready for use, and the AP may start using the additional computing resources.

Our focus lies on the capacity allocation within the cloud environment, so the right-hand side of Figure 4.6. Successful management of cloud systems requires the right scaling of both the number of host servers (denoted by s) at the first I/O queue and the maximum number of VMs (denoted by n) that can be hosted simultaneously. Moreover, this needs to be done in a dynamic way in order to respond effectively to the time-varying demand. The capacity n defines a hard constraint on whether a new VM request will be accepted immediately or not. Therefore, new requests will be delayed or even dropped if the available host capacity is insufficient, which is more likely to occur during periods in which the s host servers are overloaded.

4.3.2 Queueing model

To describe the cloud system in mathematical terms, we extend the model proposed by Tan et al. [205]. Each host server may host a number of VM instances at the same time, yielding a total number of n parallel VM instances. Requests, arriving to the system according to a Poisson process with rate λ, are granted only if one of these n positions is available. Otherwise, the user retries getting access after an exponentially distributed time with mean $1/\delta$. If granted, the request is assigned to a host server not busy initializing another VM instance, if available, or waits for one to become available. This start-up time is assumed to be exponentially distributed with mean $1/\mu$. On completion of the initialization phase, VM usage is initiated by the client. The VM continues to be occupied for a random amount of time, with mean $1/\kappa$, until release by the user. We note that the model of Tan et al. [205] has three queues in tandem, one $M/M/s$ queue, followed by two $M/M/\infty$ queues that separately model a second initialization phase and the actual VM usage by the cloud user. We thus replace the two $M/M/\infty$ queues by one $M/G/\infty$ queue with an aggregated service time, which does not alter the system performance analysis. This yields the queueing model in Figure 4.7.

Remark 4.1. We mention that a queueing model similar to the one in Figure 4.7 without retrials is analyzed by Khudyakov et al. [135] in a telecommunication environment. In their work, s and n represent the number of agents and trunk lines in a call center. Although the order of the two queues is switched, the stationary analysis of their model and the cloud model is the same, due to the product-form

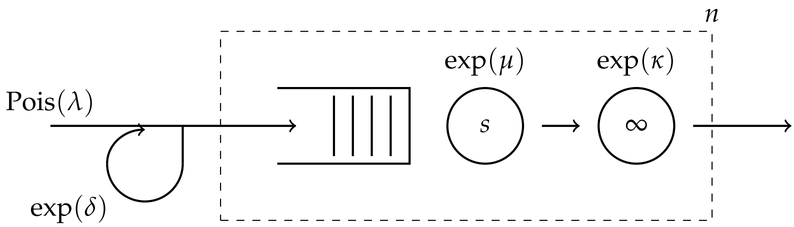


Figure 4.7: Abstracted model of VM provisioning process.

structure of the stationary distribution. In fact, Tan et al. [205] use the results of [135] in their asymptotic analysis.

An exact analysis of the cloud model is again obstructed by the absence of a product-form solution in case of retrials. We therefore turn to the QED paradigm to approximate the system behavior as $R \to \infty$.

Following the approach in [135, 205], we argue that the appropriate QED scaling for s and n should be

$$\begin{aligned} s &= R + \beta\sqrt{R} + o(\sqrt{R}), && \beta > 0, \\ n &= s + R/\kappa + \gamma\sqrt{R/\kappa} + o(\sqrt{R}), && \gamma > 0, \end{aligned} \tag{4.19}$$

where $R = \lambda/\mu = \lambda$. To understand why this is indeed the correct scaling regime to obtain non-degenerate limiting behavior, we recall the arguments we presented in Section 1.3. Namely, in order to achieve QED performance, one allocates the nominal workload brought towards the queue plus a variability hedge that is proportional to the square-root of this amount. For s, this results in the standard square-root staffing rule. For n, this is the sum of the capacity needed at the multi-server queue, i.e. s, and the consecutive infinite-server queue. Since the expected workload at the second queue equals R/κ, the capacity required at this stage equals $R/\kappa + \gamma\sqrt{R/\kappa}$ for some γ. In total, this yields the scaling for the number of VMs n as in (4.19). The limiting behavior of this queueing model without retrials is documented in [135, 205].

Proposition 4.2. *Let s and n in the cloud model of Figure 4.7 without retrials scale as in* (4.19)*. Then, as $R \to \infty$,*

$$\mathbb{P}^c(\text{delay}) \to \frac{\xi_1 - \xi_2}{\eta + \xi_1 - \xi_2} =: g_c(\beta,\gamma), \tag{4.20}$$

$$\sqrt{R}\cdot\mathbb{P}^c(\text{block}) \to \frac{\nu}{\eta + \xi_1 - \xi_2} =: f_c(\beta,\gamma), \tag{4.21}$$

where

$$\eta = \int_{-\infty}^{\beta} \Phi\Big(\gamma + (\beta - t)\sqrt{\kappa}\Big)\,\varphi(t)\,\mathrm{d}t, \qquad \xi_1 = \frac{\varphi(\beta)\Phi(\gamma)}{\beta},$$

$$\xi_2 = \frac{1}{\beta}\,\varphi\left(\sqrt{\beta^2+\gamma^2}\right)\mathrm{e}^{\frac{1}{2}(\gamma-\beta/\sqrt{\kappa})^2}\Phi(\gamma - \beta/\sqrt{\kappa}),$$

$$\nu = \sqrt{\frac{\kappa}{1+\kappa}}\,\varphi\Big(\frac{\gamma+\beta\sqrt{\kappa}}{\sqrt{1+\kappa}}\Big)\Phi\Big(\frac{\beta-\gamma\sqrt{\kappa}}{\sqrt{1+\kappa}}\Big) + \beta\,\xi_2.$$

4.3.3 Fixed-point method

Proposition 4.2 shows that also in this model, the blocking probability vanishes at rate $1/\sqrt{R}$, making it amenable to our fixed-point method for retrials. Let $\alpha\sqrt{R}$ the volume of retrials, so that a total arrival rate is $R_{\text{tot}} = R + \alpha\sqrt{R}$, or equivalently $R = R_{\text{tot}} - \alpha\sqrt{R_{\text{tot}}} + o(\sqrt{R_{\text{tot}}})$. Substituting this into the two-fold scaling rule in (4.19) gives

$$\begin{aligned} s &= R_{\text{tot}} + (\beta-\alpha)\sqrt{R_{\text{tot}}} + o(\sqrt{R_{\text{tot}}}),\\ n &= s + \frac{R_{\text{tot}}}{\kappa} + \Big(\gamma - \frac{\alpha}{\sqrt{\kappa}}\Big)\sqrt{\frac{R_{\text{tot}}}{\kappa}} + o(\sqrt{R_{\text{tot}}}). \end{aligned}$$

Accordingly, the constant α is defined as the solution of the fixed-point equation

$$f_c\left(\beta-\alpha, \gamma-\alpha/\sqrt{\kappa}\right) = \alpha. \tag{4.22}$$

Approximations for the delay and blocking probability in the cloud model with retrials are hence given by

$$\mathbb{P}^c_r(\text{delay}) \approx g_c\left(\beta-\alpha, \gamma-\alpha/\sqrt{\kappa}\right), \qquad \mathbb{P}^c_r(\text{block}) \approx \alpha/\sqrt{R}. \tag{4.23}$$

Note that in contrast to the fixed-point equation (4.12) for the basic model, the second argument $\gamma - \alpha/\sqrt{R}$ is also corrected.

We next test the accuracy of the fixed-point equation for several instances. In Table 4.2, we present the simulation results for $\kappa = 0.02$, 0.2 and 1, and two pairs of (β, γ) for increasing R.

First, observe from Table 4.2 that n now lives on a different scale than s. This is required to facilitate the long sojourn time of customers in the second stage, which is proportional to $1/\kappa$, creating the need for larger system size. Besides that, the numerical results show that the fixed-point approximation is again remarkably accurate over a wide range of parameter settings. Even for cloud systems as small as 50 servers, the fixed-point method gives accurate approximations.

4.4 Abandonments

Whereas in the basic model of Section 4.2, retrials were governed by the system architecture (arriving customers are requested to reattempt if n customers are present in the system), we now consider a setting in which departures from the queue are customer-initiated. That is, customers deliberately decide to leave the queue to return for service at a later time. Hence we consider a queueing system with abandonments and retrials.

	$(\beta,\gamma) = (0.5,1)$				$(\beta,\gamma) = (1,1)$			
R	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$
5	6	13	0.5309	0.5540	7	14	0.2800	0.2426
10	12	25	0.3864	0.3810	13	26	0.2393	0.2164
50	54	111	0.3904	0.4525	57	114	0.1965	0.2010
100	105	215	0.4300	0.5474	110	220	0.1859	0.1952
500	511	1033	0.4139	0.5586	522	1044	0.1787	0.2052
1000	1016	2048	0.4003	0.5479	1032	2064	0.1660	0.1803
		Approx	0.4029	0.5638		Approx	0.1709	0.1992

(a) $\kappa = 1$

	$(\beta,\gamma) = (0.5,1)$				$(\beta,\gamma) = (1,1)$			
R	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$
5	6	36	0.5664	0.3049	7	37	0.3079	0.1457
10	12	69	0.4263	0.2227	13	70	0.2683	0.1410
50	54	320	0.4444	0.2555	57	323	0.2293	0.1334
100	105	627	0.4826	0.3085	110	632	0.2187	0.1379
500	511	3061	0.4842	0.3235	522	3072	0.2182	0.1358
1000	1016	6087	0.4630	0.2906	1032	6103	0.2039	0.1332
		Approx	0.4687	0.3029		Approx	0.2042	0.1326

(b) $\kappa = 0.2$

	$(\beta,\gamma) = (0.5,1)$				$(\beta,\gamma) = (1,1)$			
R	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$	s	n	$\mathbb{P}_r^c(\text{del})$	$\sqrt{R}\,\mathbb{P}_r^c(\text{bl})$
5	6	272	0.5836	0.1085	7	273	0.3217	0.0738
10	12	534	0.4456	0.0945	13	535	0.2822	0.0781
50	54	2604	0.4683	0.1031	57	2607	0.2429	0.0764
100	105	5176	0.5106	0.1064	110	5181	0.2345	0.0759
500	511	25669	0.5130	0.1158	522	25680	0.2353	0.0795
1000	1016	51240	0.4946	0.0975	1032	51256	0.2223	0.0744
		Approx	0.4999	0.0862		Approx	0.2207	0.0595

(c) $\kappa = 0.02$

Table 4.2: Numerical results of the fixed-point method for the cloud model with slow retrials as $R \to \infty$.

4.4.1 The Erlang-A model

The canonical model for abandonments is the $M/M/s+M$ or Erlang-A model [175, 82]. The queueing dynamics of the Erlang-A model are similar to those in the $M/M/s$ queue, with the additional feature that each customer is assigned an i.i.d. patience time, which is exponentially distributed with mean $1/\theta$. If a customer's patience time expires before reaching an available server, she leaves (abandons) the system. As the number of customers in the Erlang-A queue is a birth-death process, its stationary distribution and associated performance measures are fairly well-understood, also in the QED regime [82, 229, 230]. Most importantly to us, Garnett et al. [82] and Zeltyn & Mandelbaum [229] identified the asymptotic delay and abandonment probability in the Erlang-A model under QED scaling.

Proposition 4.3. *[229, Thm. 4.1] Let $s = R + \beta\sqrt{R} + o(\sqrt{R})$ for some $\beta \in \mathbb{R}$. Then in the $M/M/s+M$ queue*

$$\mathbb{P}^a(\text{delay}) \to \left(1 + \sqrt{\theta}\,\frac{h(\beta/\sqrt{\theta})}{h(-\beta)}\right)^{-1} =: g_a(\beta) \tag{4.24}$$

$$\sqrt{R}\,\mathbb{P}^a(\text{abandon}) \to \frac{\sqrt{\theta}\,h(\beta/\sqrt{\theta}) - \beta}{1 + \sqrt{\theta}\,h(\beta/\sqrt{\theta})/h(-\beta)} =: f_a(\beta), \tag{4.25}$$

as $R \to \infty$ where $h(\beta) = \varphi(\beta)/\Phi(-\beta)$.

We remark that in [229], the QED limits for generally distributed patience time were derived. Although our heuristic also works for this more general setting, we focus on the exponentially distributed patience here to convey our main ideas.

Remark 4.2. Large-scale Markovian multi-server queues with abandonments and retrials have been thoroughly studied in a series of papers by Mandelbaum et al. [152, 154, 153]. In these works, the authors consider a system with time-varying arrivals and a retrial rate that remains bounded away from zero, for which they deduce fluid and diffusion limits as the system grows large. These limits provide approximations for the time-dependent queue length and virtual waiting time processes, including their means and variances. We in this section take a different approach by assuming $\delta \to 0$, which enables us to characterize the steady-state behavior of queues with abandonments and retrials.

4.4.2 Fixed-point method

Next, we include (slow) retrials. More specifically, we assume that customers who abandon the queue rejoin the queue after an exponentially distributed time with mean $1/\delta \gg 1$. Just as in the $M/M/s/n$ queue, the $M/M/s+M$ queue with retrials is analytically intractable, and therefore we apply our fixed-point method to approximate its performance in the QED regime.

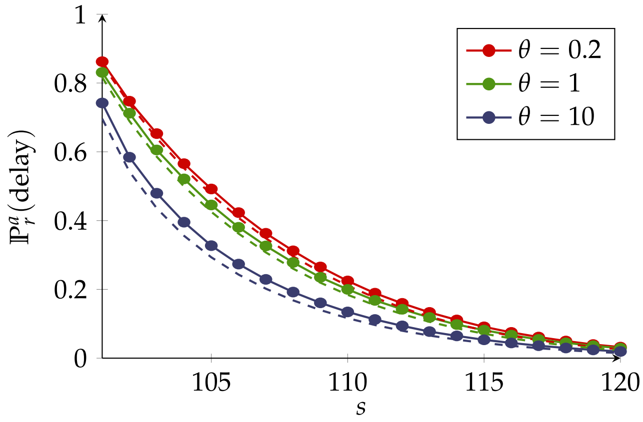


Figure 4.8: Simulated (solid) and approximated (dashed) delay probability in the $M/M/s+M$ queue with retrials and $R = 100$, $\delta = 0.01$.

Observe through Proposition 4.3 that the fraction of customers leaving before receiving service is roughly $\alpha/\sqrt{R}$. Following the reasoning of Section 4.2.3, the total arrival volume, consisting of new arrivals and reattempting customers, is $R_{\text{tot}} = R + \alpha\sqrt{R}$, with

$$\alpha = f_a(\beta - \alpha). \tag{4.26}$$

Accordingly, this yields the following approximations for the delay and abandonment probability in the system with retrials

$$\mathbb{P}_r^a(\text{delay}) \approx g_a(\beta - \alpha), \qquad \mathbb{P}_r^a(\text{abandon}) \approx \alpha/\sqrt{R}. \tag{4.27}$$

We test our heuristic in the model with abandonment with parameters $R = 100$, $\delta = 0.01$. For $\theta = 0.2$, customers are quite patient, as they are willing to wait on average 5 times their expected service time. Customer abandonment becomes more dominant for the cases in which customers are reasonably patient $\theta = 1$ and very impatient $\theta = 10$.

In Figures 4.8 and 4.9, we plot the simulated delay and (scaled) abandonment probability against approximations obtained through the fixed-point method.
Again, we see a very good match between approximated and actual values. As θ decreases, that is, customers become more patient, accuracy of the approximations improves. This makes sense, since the volume of retrials decreases, and the system behaves more and more like a standard $M/M/s$ queue.

In Table 4.3 we also check the asymptotic accuracy of the model with abandonments and retrials and see that the approximation indeed improves as R increases.

Remark 4.3. Note that even though the fixed-point approximations come close to the simulated values as R increases, a small gap remains, especially notable in $\beta = 0.5$. This can be attributed to both rounding errors and the heuristic assumption that the retrial stream is independent Poisson. The latter is obviously false, as the

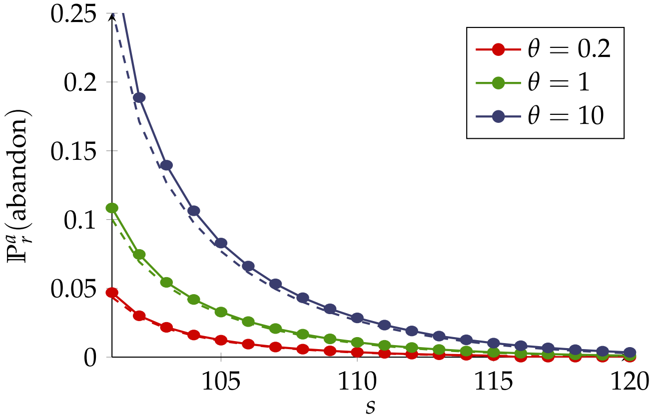


Figure 4.9: Simulated (solid) and approximated (dashed) abandonment probability in the $M/M/s+M$ queue with retrials and $R = 100$, $\delta = 0.01$.

		$\theta = 0.2$		$\theta = 1$		$\theta = 10$	
R	s	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$
5	6	0.5703	0.1522	0.5423	0.4158	0.4766	1.0980
10	12	0.6612	0.2152	0.6277	0.5619	0.5429	1.5099
50	54	0.5521	0.1517	0.5089	0.4009	0.3920	1.0251
100	105	0.4896	0.1218	0.4456	0.3276	0.3282	0.8321
500	511	0.4877	0.1228	0.4442	0.3302	0.3132	0.8135
1000	1016	0.4992	0.1274	0.4472	0.3359	0.3148	0.8244
Approx		0.4757	0.1182	0.4254	0.3120	0.2933	0.7695

(a) $\beta = 0.5$

		$\theta = 0.2$		$\theta = 1$		$\theta = 10$	
R	s	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$	$\mathbb{P}^a_r(\text{del})$	$\sqrt{R}\,\mathbb{P}^a_r(\text{ab})$
5	7	0.3130	0.0527	0.2908	0.1563	0.2382	0.4033
10	13	0.2732	0.0444	0.2503	0.1331	0.1917	0.3504
50	57	0.2344	0.0373	0.2090	0.1120	0.1442	0.2974
100	110	0.2244	0.0357	0.1999	0.1077	0.1355	0.2877
500	522	0.2232	0.0355	0.1973	0.1079	0.1282	0.2842
1000	1032	0.2210	0.0359	0.1979	0.1092	0.1287	0.2890
Approx		0.2105	0.0335	0.1842	0.1005	0.1162	0.2642

(b) $\beta = 1$

Table 4.3: Numerical results of the fixed-point method for the $M/M/s+M$ queue with slow retrials.

retrial process naturally depends on the history of the external arrival and service processes. Therefore, the fixed-point heuristic slightly underestimates congestion levels in the actual system. However, this error is relatively small and moreover in small to moderate-size systems negligible compared to the effects of rounding.

Remark 4.4. Our fixed-point heuristic easily extends to the case in which only a fraction of abandoning customers returns to the system later. If each customer who abandons decides (independent from others and his own retrial history) to return with probability $q \in [0,1]$, then the arrival stream due to retrial becomes $q \cdot \alpha\sqrt{R}$, so that the fixed-point becomes $f_a(\beta - q\alpha) = \alpha$. Approximations of the performance measures follow accordingly.

4.5 Dimensioning

The asymptotic QED expressions for the systems we considered in Sections 4.2-4.4 without retrials together with the corrections obtained through the fixed-point equation provide a method for dimensioning large-scale systems with retrials. For sufficiently large arrival volumes, we can tune the QoS-levels offered by the systems through the QoS-parameters β and γ. In this section we demonstrate how to do so in the cloud model, using the delay probability as a vehicle. First we explore the procedure under stationary conditions, then in a time-varying environment. The methods we propose easily translate to the two other model settings considered in this chapter, and the blocking probability.

4.5.1 Stationary dimensioning

We consider the dimensioning problem in the cloud model from a constraint satisfaction perspective. That is, given the offered load R, we search for the pair (s,n) that realizes a target delay probability $\varepsilon \in (0,1)$. Relying on the two-fold scaling in (4.19), this under large offered loads R is tantamount to finding the pair (β,γ) that achieves asymptotic delay probability ε. In a system without retrials, attaining this target performance boils down to finding a pair $(\beta_\varepsilon,\gamma_\varepsilon)$ such that $g_c(\beta_\varepsilon,\gamma_\varepsilon) = \varepsilon$. The fixed-point heuristic however tells us that the model with retrials performs slightly worse, namely as if the QoS-parameters were $(\beta_\varepsilon - \alpha, \gamma - \alpha/\sqrt{\kappa})$ for $\alpha > 0$. Henceforth, to attain the target delay probability ε in the limit with retrials, larger QoS parameters are required. To be precise, $\beta_\varepsilon^* = \beta_\varepsilon + \alpha$ and $\gamma_\varepsilon^* = \gamma_\varepsilon + \alpha/\sqrt{\kappa}$, with α satisfying $\alpha = f_c(\beta^* - \alpha, \gamma_\varepsilon^* - \alpha/\sqrt{\kappa}) = f_c(\beta_\varepsilon,\gamma_\varepsilon)$. Finally, we substitute β_ε^* and γ_ε^* in the scaling (4.19) to obtain capacity levels s and n. Altogether, this yields the QED dimensioning procedure in Algorithm 1, in which $[\cdot]$ denotes the integer rounding operator.

In Table 4.4 we performed this stationary dimensioning procedure for $\kappa = 1$, 0.2 and 0.02, and $\varepsilon = 0.1$, 0.25 and 0.4, and increasing offered loads R, and used simulation to obtain the actual delay probabilities. We immediately see that the

Input: Offered load R
Expected time spent in seconds stage $1/\kappa$
Target delay probability $\varepsilon \in (0, 1)$
Output: Capacity levels s and n.

1. Compute $(\beta_\varepsilon, \gamma_\varepsilon)$ such that $g_c(\beta_\varepsilon, \gamma_\varepsilon) = \varepsilon$.
2. Set $\beta^*_\varepsilon = \beta_\varepsilon + f_c(\beta_\varepsilon, \gamma_\varepsilon)$ and $\gamma^*_\varepsilon = \gamma_\varepsilon + f_c(\beta_\varepsilon, \gamma_\varepsilon)/\sqrt{\kappa}$.
3. Return $s = \lceil R + \beta^*_\varepsilon \sqrt{R} \rceil$ and $n = [s + R/\kappa + \gamma^*_\varepsilon \sqrt{R/\kappa}]$.

Algorithm 1: Stationary dimensioning for cloud model with retrials.

procedure yields remarkably good results, that are very close to the target delay probabilities.

4.5.2 Time-varying dimensioning

We next discuss how the parameters s and n can be adjusted in time-varying environments where the offered load $R(t)$ is a function of time. For this we use the mean-offered-load (MOL) method, which was developed in [125] to approximate and dimension the $M_t/G/s$ system by establishing a relation with the analytically tractable $M_t/G/\infty$ system. An underlying assumption of the MOL method is that a well-capacitated multi-server queue delays only a small portion of users and only for short periods. Therefore, the system can be approximated by an infinite-server system. The MOL approximation [125] combines the desirable QoS properties rendered by the QED regime with the analytic tractability of the $M/G/\infty$ queue, see [71], to establish a dynamic algorithm for choosing $s(t)$ that stabilizes the system behavior at some QoS-target.

To understand why the MOL approximation is likely to be accurate for the systems in this chapter, observe that under the QED scalings, the blocking probability vanishes asymptotically and hence the main assertion on which the MOL approximation is built continues to hold. Following the line of thought in [125], we consider the number of users in a system with $s = n = \infty$ to obtain $R_1(t) = \mathbb{E}[R(t - S)]\mathbb{E}[S]$, where S is the service requirement per customer taken to be unit exponentially distributed. Then,

$$R_1(t) = \int_0^\infty e^{-u} R(t-u) \, \mathrm{d}u. \tag{4.28}$$

Note that this transformation typically shifts and levels peaks in workload ahead in time with respect to those in $R(t)$. As the time-varying counterparts of s in (4.19), we then get

$$s(t) = R_1(t) + \beta\sqrt{R_1(t)}. \tag{4.29}$$

	$\varepsilon = 0.10$			$\varepsilon = 0.25$			$\varepsilon = 0.40$		
	$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (1.17, 0.35)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.73, 0.57)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.48, 0.78)$		
R	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$
10	14	25	0.1231	13	25	0.2268	12	24	0.3717
50	59	111	0.0969	56	110	0.2289	54	110	0.3844
100	112	215	0.1069	108	214	0.2426	105	213	0.4148
500	527	1035	0.0994	517	1030	0.2486	511	1028	0.3996
1000	1038	2049	0.0977	1024	2042	0.2442	1016	2041	0.3925

(a) $\kappa = 1$

	$\varepsilon = 0.10$			$\varepsilon = 0.25$			$\varepsilon = 0.40$		
	$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (1.34, 0.50)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.87, 0.65)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.59, 0.79)$		
R	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$
10	15	69	0.0893	13	68	0.2628	12	68	0.4238
50	60	318	0.1007	57	317	0.2210	55	318	0.3560
100	114	625	0.0989	109	624	0.2504	106	624	0.4099
500	530	3055	0.1051	520	3053	0.2449	514	3054	0.3857
1000	1043	6078	0.0986	1028	6074	0.2459	1019	6075	0.3981

(b) $\kappa = 0.2$

	$\varepsilon = 0.10$			$\varepsilon = 0.25$			$\varepsilon = 0.40$		
	$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (1.41, 0.68)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.93, 0.74)$			$(\beta^*_\varepsilon, \gamma^*_\varepsilon) = (0.59, 0.79)$		
R	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$	s	n	$\mathbb{P}^c_r(\text{del})$
10	15	530	0.0996	13	530	0.2814	13	531	0.2816
50	60	2594	0.1146	57	2594	0.2416	55	2595	0.3805
100	115	5163	0.0933	110	5163	0.2323	107	5163	0.3794
500	532	25640	0.1001	521	25638	0.2516	515	25641	0.3873
1000	1045	51198	0.1018	1030	51196	0.2437	1021	51199	0.3900

(c) $\kappa = 0.02$

Table 4.4: Results of the stationary dimensioning algorithm for $\varepsilon = 0.1$, 0.25 and 0.4.

Secondly, the number of customers present in the system strongly depends on the number of customers in the second phase of the system, especially since $\kappa \ll \mu$. Therefore, we moreover need an approximation for the workload offered to the second queue as a function of t, which we denote by $R_2(t)$. Continuing the reasoning of MOL, we argue that $R_2(t)$ is equal to the output process of the first queue $R_1(t)$. Then,

$$R_2(t) = \mathbb{E}[R_1(t - S_2)]\mathbb{E}[S_2] = \int_0^\infty \int_0^\infty e^{-u-\kappa v} R(t - u - v)\mathrm{d}u\,\mathrm{d}v \tag{4.30}$$

and the natural shape of $n(t)$ becomes

$$n(t) = s(t) + R_2(t) + \gamma\sqrt{R_2(t)}. \tag{4.31}$$

Combining the above ingredients then leads to Algorithm 2. Observe that if the

Input: Offered load function $R(t)$
Expected time spent in second stage $1/\kappa$
Target delay probability $\varepsilon \in (0, 1)$
Output: Capacity levels $s(t)$ and $n(t)$ achieving $P_r(\text{delay}) = \varepsilon$.

1. Compute β_ε^* and γ_ε^* according to Algorithm 1.
2. Compute $R_1(t)$ and $R_2(t)$ as in (4.28) and (4.30).
3. Return

$$s(t) = \left\lceil R_1(t) + \beta_\varepsilon^* \sqrt{R_1(t)} \right\rceil,$$
$$n(t) = \left\lceil s(t) + R_2(t) + \gamma_\varepsilon^* \sqrt{R_2(t)} \right\rceil.$$

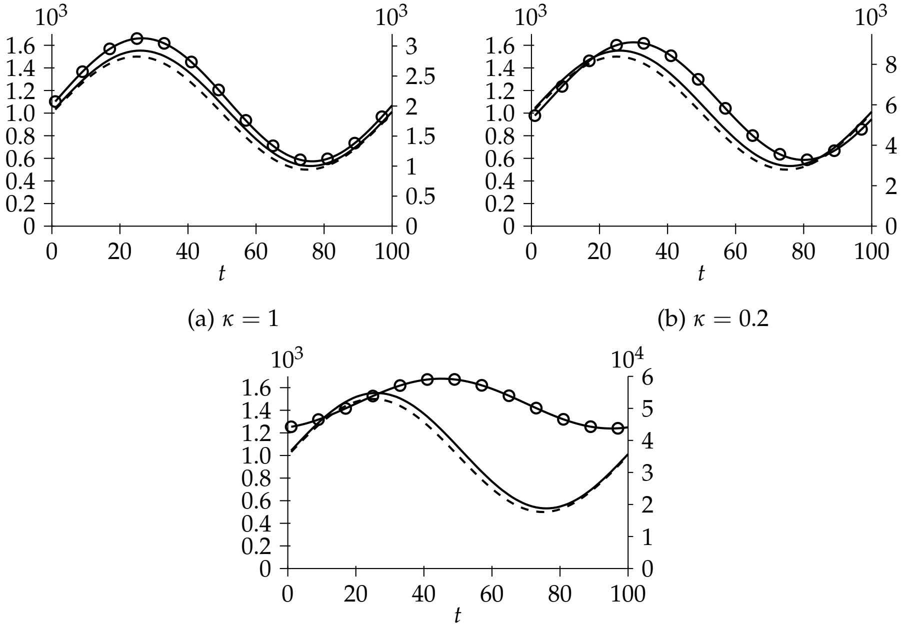
Algorithm 2: Time-varying dimensioning algorithm for cloud model with retrials.

service times at the infinite-server queue are relatively short compared with the rate of change of the load function, we have $R_2(t) \approx R_1(t)/\kappa$, so that $n(t)$ as in Algorithm 2 shows resemblance with 4.19.

To illustrate Algorithm 2 we consider the time-varying load

$$R(t) = a + b\,\sin\left(2\pi t/T\right), \tag{4.32}$$

where we set $a = 1000$ as the mode, $b = 500$ as the amplitude and $T = 100$ as the cycle length. This system experiences large fluctuations in load volume over the course of one cycle. Since $\mu = 1$, this implies that one cycle on average consists of 100 service times at the host server queue. Due to relatively short service times with respect to the cycle length, the MOL approximation for the number of customers at the first queue is roughly equal to the original load, i.e. $R_1(t) \approx R(t)$. These short services at the first queue compared to the cycle length are typical for cloud systems, in which case the cycle is usually one day.

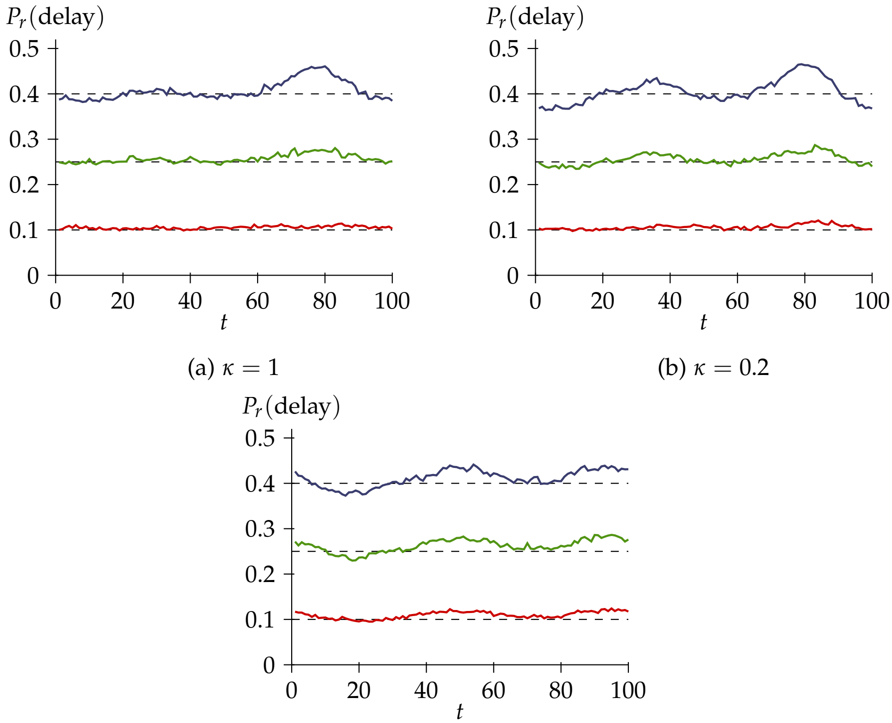


(a) $\kappa = 1$

(b) $\kappa = 0.2$

(c) $\kappa = 0.02$

Figure 4.10: Arrival rate function $R(t)$ (dashed) and staffing functions $s(t)$ (solid) and $n(t)$ (o) for different values of κ. The left vertical axis refers to $R(t)$ and $s(t)$, where the right axis refers to $n(t)$.



(a) $\kappa = 1$

(b) $\kappa = 0.2$

(c) $\kappa = 0.02$

Figure 4.11: Simulated time-dependent delay probabilities in the cloud model with $\delta = 10^{-2}$, targets $\varepsilon = 0.1, \varepsilon = 0.25$ and $\varepsilon = 0.4$, and capacity levels determined by Algorithm 2.

First, we examine the functions $s(t)$ and $n(t)$ as prescribed by Algorithm 2 for $\kappa = 1, 0.2, 0.02$ and $\varepsilon = 0.25$. The resulting values are depicted in Figure 4.10 together with the arrival rate function. Note that $n(t)$ lives on a different scale than $s(t)$, and has its own vertical axis at the right side of the plots. For small and hence realistic values of κ, the function $n(t)$ displays a shifted phase compared to the real-time offered load, due to the relatively long service time at the second station. The lag can be observed in (4.30). Hence, while the number of servers $s(t)$ allocated at time t is almost in phase with the arrival rate $R(t)$, $n(t)$ undergoes a shift of its peak capacity somewhat ahead in time. Observe also that $n(t)$ shows milder fluctuations when κ decreases. This can be attributed to the added hedge which is of order $\sqrt{R/\kappa}$. Remark that the overcapacity is relatively small. This illustrates the economies-of-scale that can be achieved in these large-scale systems. Next, we simulate the time-dependent process, given the staffing functions depicted in Figure 4.10, as well as the staffing functions designed for the target delay probabilities $\varepsilon = 0.1$ and $\varepsilon = 0.4$ for the three values of κ. The results of the simulations are depicted in Figure 4.11. In all cases, the time-dependent delay probability only mildly fluctuates around the target. As we increase ε, the stabilizing effect of the method weakens somewhat, which for other systems was also observed in [125].

4.6 Conclusion

In this chapter, we studied the impact of retrying customers in large-scale systems in the QED regime. The presence of retrials has a detrimental effect on congestion-related performance, compared to systems in which customers are simply discarded upon blockage/abandonment. On the other hand, compared to similar systems without physical size restrictions or customer impatience, the performance gain can be substantial, if retrial times are relatively long compared to the service times. Namely, retrials prompt temporary release of pressure from the system by shifting workload ahead in time.

Through our analysis, we have shown how the performance of large-scale queueing systems facing slow retrials can be approximated by appropriately combining a fixed-point technique with QED scaling. We showed the remarkable accuracy of this approximation scheme in various retrial settings, that are otherwise intractable to analyze. As we discussed in Section 4.5, our novel asymptotic analysis technique is furthermore a powerful and elegant tool for dimensioning large-scale systems with slow retrials, which is moreover amenable to deal with time-varying demand.

We illustrate a few directions for future research. As we explained before, the fixed-point method relies heavily on the premise that the blocking (or abandonment) probability vanishes at rate $1/\sqrt{R}$ in the QED regime, and on the availability of expressions for its limiting behavior. Since this description likely fits a wide range of queueing models, we henceforth believe that our fixed-point method and the related dimensioning scheme find application beyond the three models we discussed here.

Secondly, in the dimensioning procedure of Section 4.5 we took a constraint satisfaction perspective in which we aimed to achieve a preset target QoS-level. As an alternative approach, one could define a cost function to quantify the trade-off between capacity costs and customer dissatisfaction. Specifically, suppose a cost c_1 is associated with each server per unit of time, cost c_2 is charged for every waiting customer per time unit, and cost c_3 is the penalty for each blocked customer. Then in the $M/M/s/n$ queue with retrials in the QED regime, we use that $s = R + \beta\sqrt{R}$, the blocking probability is roughly $f(\beta-\alpha,\gamma)/\sqrt{R}$ and the expected waiting time is approximately $h(\beta-\alpha,\gamma)/\sqrt{R}$ for some function h, see [160], yielding total operational cost

$$c_1\left(R+\beta\sqrt{R}\right) + c_2\, R\,\frac{f(\beta-\alpha,\gamma)}{\sqrt{R}} + c_3\, R\,\frac{h(\beta-\alpha,\gamma)}{\sqrt{R}},$$

where α satisfies the fixed-point equation. Hence, asymptotic dimensioning of the system boils down to finding the parameters β^* and γ^* that minimize

$$c_1\beta^* + c_2\, f(\beta^*-\alpha^*,\gamma^*) + c_2\, h(\beta^*-\alpha^*,\gamma^*),$$

with corresponding fixed point α^*. Solving this optimization problem is not straightforward and a detailed study of this and related asymptotic dimensioning problems is an interesting avenue for future research.

Last, we remark that even though the fixed-point method works very well for systems with slow retrials, i.e. $\delta \to 0$, it may also serve as an approximation to systems with short to moderate retrial times. In these scenarios, the method is likely to underestimate congestion levels as it ignores dependencies between the primary and retrial stream of arrivals. In the extreme case that $\delta \to \infty$, that is, blocked customers retry immediately, the customers in the retrial orbit basically form a (random order) queue outside the service facility. When the inside of the facility consists of more than one queue, our fixed-point may be used as an heuristic approach to account for the increased workload that builds up outside the facility. We explore this heuristic idea in a health care context in the next chapter.

5

Finite-size effects in critically dimensioned emergency departments

Motivated by health care systems with repeated services that have both personnel (nurse/physician) and space (beds) constraints, we study a restricted version of the Erlang-R model. The space restriction policies we account for are blocking or holding in a pre-entrant queue. We develop many-server approximations for the system performance measures when either policy applies, and explore the connection between them. We show that capacity allocation of both resources should be determined simultaneously, and derive the methodology to determine it explicitly. We show that the system dynamics is captured by the fraction of needy time in the network, and that returning patients should be accounted for both in steady-state and time-varying conditions. We demonstrate the application of our policies in two case-studies of resource allocation in hospitals.

Based on
Finite-size effects in critically dimensioned emergency departments
Johan van Leeuwaarden, Britt Mathijsen, Fiona Sloothaak & Galit Yom-Tov
Submitted to *Operations Research*

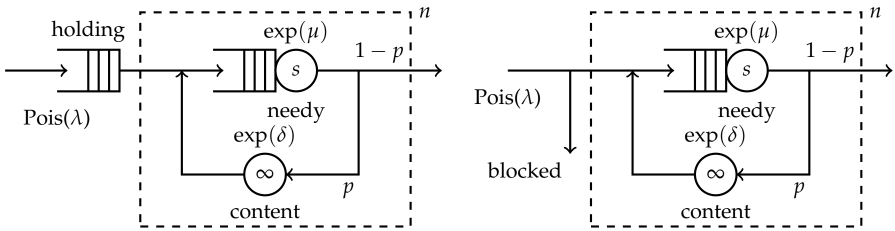
5.1 Introduction

In recent years, operations research techniques have received increased interest from the health care community, as they are able to design and improve workflow processes in health care facilities [17, 93, 68, 102, 103]. Because these processes are typically stochastic in nature, it is common practice to use queueing theory for performance analyses and workforce planning. As a first step towards understanding the processes going on in health care environments, systems are commonly modeled after a single station queue, such as the $M/M/s$ (Erlang-C), $M/M/s/s$ (Erlang-B) or $M/M/s+M$ (Erlang-A) models, and fluid and diffusion approximations are used to provide insights into the process dynamics. However, simple single station models often fail to capture the more intricate dynamics of the settings specific to health care contexts. Prime examples include the flows of patients in a hospital from one medical ward to another [17], within the Emergency Department (ED) between different stages of treatment [106], or between medical facilities [232]. Queueing networks can capture the dependency between several service stages and several types of resources. More specifically, we are interested in the ubiquitous feature, particularly present in health care environments, that patients during their stay in the system might require a specific resource multiple times, e.g. physicians and nurses who treat patients several times during their stay in the medical wards [124] or the ED [226], while multiple resources types are limited (e.g. medical staff and beds). In this chapter, we concentrate on the dynamics within EDs.

An often ignored yet essential feature of medical facilities concerns the restriction of the number of patients that can reside in the facility simultaneously. In Chapter 4, we already observed that finite-size restrictions can have a significant effect on the performance of queueing systems. In this chapter, we investigate the influence of such multiple restrictions on the network dynamics and the required staffing policies in the context of an ED.

The restricted Erlang-R model. The canonical model for service networks with returns is the Erlang-R model, introduced by Yom-Tov & Mandelbaum [226]. In this open two-station model, customers arrive according to a Poisson process to an $M/M/s$. After service completion, the customer with probability $1-p$ leaves the system and with probability p returns to the queue after a random delay. This delay is modeled as an infinite-server queue. A schematic visualization of the Erlang-R model is depicted in Figure 5.2b

in which customers, during their stay in the system receive a random number of services from the same pool of servers. Yom-Tov & Mandelbaum [226] showed that such a simple network model can be used to determine staffing in an ED both in stable and time-varying conditions. Nevertheless, empirical studies report that some countries, such as the US, use a different operational mode that applies strict restrictions on entering the ED [196]. In typical US EDs, a patient will not enter the ED until both a bed and a physician are available to treat her. Those restrictions can be either physical (beds) restrictions or managerial ones — for instance by imposing



(a) Erlang-R model with holding. (b) Erlang-R model with blocking.

Figure 5.1: Restricted Erlang-R models with maximally n customers in system.

a patient-to-physician ratio. In this work, we extend the Erlang-R model by enforcing a constraint on the maximum number of available places inside the facility. Our model hence incorporates two kinds of resource constraints: servers that provide the actual service and the maximum available places inside the service system. Both affect the system in a highly interdependent way. The model, presented in Figure 5.1, assumes s servers and a maximum capacity of n concurrent places. We assume that patients arrive according to a Poisson process with rate λ. In case a new arrival finds n or more patients already present, we consider two options: either she waits outside the service facility in a holding queue until a vacant space becomes available (Figure 5.1a) or she is blocked (Figure 5.1b), such as is the case when patients are sent to an alternative facility. Once a patient is admitted, she requires assistance from one of the s servers for an exponentially distributed duration with mean $1/\mu$. Then, with probability $1-p$, the patient leaves the system or, with probability p, returns to service again after an exponentially distributed time with mean $1/\delta$. Following Jennings & de Véricourt [124] and Yom-Tov & Mandelbaum [226], we call patients *needy* when they require attention from one of the servers and *content* when they are in the delayed return phase. In addition, we call patients *holding* when they are waiting outside the facility for an available space. We assume that the arrival process, the needy times and content times are mutually independent. In the holding queue and the needy queue, we apply the First-Come-First-Served (FCFS) discipline.

As mentioned, we consider two versions of the finite-capacity constraint. The first version is called *Erlang-R with holding*, in which patients wait for an available space in the system. The second version is called *Erlang-R with blocking*, in which patients meeting a full system are blocked. Naturally, intermediate scenarios can be constructed in which a proportion of the total arrival volume of patients indeed leaves upon finding a full system, while the rest joins the holding queue. While this chapter focuses on the two extreme cases, straightforward adaptations can fit these intermediate scenarios.

Examples of restricted Erlang-R. As noted before, an ED operated in the US can be modeled using a restricted Erlang-R model. Another health care example is medical units (MUs) in a hospital. Such units specialize in specific types of illnesses (cardiology, oncology, etc.) and have limited resources such as nurses and beds. If the unit is full, new patients are either allocated to an alternative medical unit, i.e. blocked, or wait for an available bed. Both policies are problematic in terms of quality-of-care, because the personnel in the alternative unit (or the ED) may be less knowledgeable about the patient's medical condition and waiting in the ED was shown to increase mortality. Moreover, ED waiting may reduce available capacity for treating ED patients [55, 35], hence endangering both the delayed patient as well as others. Both the number of personnel (nurses and physicians) and the number of beds impact service dynamics and quality-of-care. Research so far looked at the capacity allocation of those resources separately. Green & Yankovic [88] and Jennings & de Véricourt [123] looked at nurse staffing in medical units, while de Bruin et al. [68] looked into bed allocation. The unified model we suggest enables us to capture the dependency between those two decisions, and its impact on other medical units in the hospital. At the same time, we capture the two most commonly used modes of operation — blocking and holding of new patients.

Two-fold square-root staffing rule. Our main goal is to provide staffing policies for the ED that high resource utilization, while at the same time maintain good quality-of-care. This goal relates to the philosophy of the Quality-and-Efficiency-Driven (QED) regime that is the recurring theme of this thesis. In this chapter, we obtain asymptotic results for the Erlang-R model with blocking in the QED regime (Section 5.4.2). Following [123], we employ a two-fold QED staffing policy: $s = R_1 + \beta\sqrt{R_1}$ for the number of nurses and $n = R_1/r + \gamma\sqrt{R_1/r}$ for the number of patients in the system (beds), where β and γ are constants, R_1 is the offered load of the servers (nurses) and r is the fraction of time a patient spends in the needy state. We establish limiting expressions for performance measures, such as the probability of delay and blocking, in the form of explicit functions that depend solely on β and γ. In deriving these limit results, we use the available product-form solution for the stationary distribution.

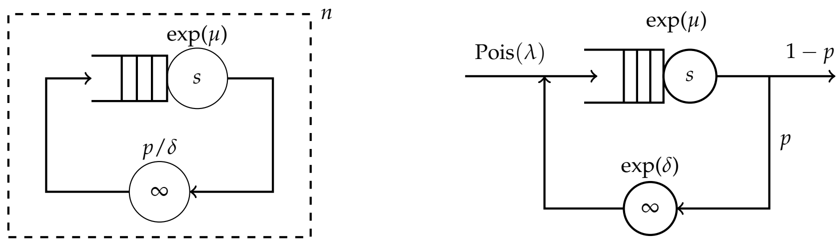
Likewise, we pursue QED performance for the Erlang-R model with holding. However, a direct analytic approach is obstructed by the absence of product-form solutions. We provide two solutions for establishing QED behavior. First, we provide stochastic performance bounds that stay meaningful in the QED regime, which demonstrate the non-degenerate behavior of the two-fold scaling in the large-system limit. Second, we develop a heuristic method that quantifies the difference between the holding model and the blocking model. This method is to a large extent related to the asymptotic approximation method for retrial queues discussed in Chapter 4, in the sense that we approximate the model with holding through the model with blocking, yet with an increased arrival rate. The increase in arrival rate turns out to be the solution of a fixed-point equation. Using our results on the asymptotic behavior of the model with blocking in the QED regime, we then obtain approximative QED performance measures for the model with holding. These theoretical findings ultimately yield algorithms for dimensioning and time-varying staffing.

Structure of the chapter. We first review related literature on the subject of staffing in health care environments in Section 5.2. In Section 5.3, we introduce the mathematical models more formally, and deduce preliminary results on their stability conditions and relative performance. Section 5.4 describes the scaling regime we use for our asymptotic study of the restricted Erlang-R models, and Sections 5.4.2 and 5.4.3 present our main theoretical findings. We turn to dimensioning problems in Section 5.5, and show how our asymptotic QED results can be used to make resource allocation decisions in realistic settings. Section 5.6 is devoted to the numerical and comparative analysis of the restricted Erlang-R models, and also shows how our method can be applied in time-varying environments through a case study. We summarize our findings and give directions for future research in Section 5.7.

5.2 Literature review

Due to increasing demand and tightening budgets in health care, there is a growing need for efficient workforce management [93]. Personnel (nurse and physician) expenditure is one of the biggest factors in hospital costs [131], and inadequate nursing levels have been mentioned as a significant factor in medical errors and ED overcrowding. In order to establish appropriate nursing levels, a staffing policy requires assessment of a wide range of variables, such as differing nurse expertise and patient acuity during the day. Current methods, such as the minimum nurse-to-patient ratios, are often too inflexible to capture those varying conditions. The American Hospital Association (AHA) and others call for dynamic staffing policies that can deal with the complex and evolving nature of health care [12]. Workforce management in health care systems has been studied extensively; see [69, 102, 103] for overviews. In recent years it has become apparent that queueing models can be helpful in developing staffing and routing recommendations, not just for large-scale service systems, but also for the small and complicated health care systems.

The first to try such an approach through queueing models were Green et al. [90, 93], who used the single station stationary Erlang-C model to set staffing levels in EDs and panel sizes for clinics. Using a similar approach, Bekker & de Bruin [32] used the Erlang-B model to determine bed allocation for medical wards. The first to observe the significant impact of interrupted services in a health care setting were Jennings & de Véricourt [123, 124]. Motivated by the need to set nurse-to-patient ratios for internal wards, they considered a closed queueing system with s nurses and n beds. This is essentially the Erlang-C model with the additional restriction that a finite population of the n patients requires care. In their model, all beds are always occupied, and patients alternate between two phases: the needy phase where patients require service of a nurse and the content phase where they do not; see Figure 5.2a. The system dynamics of the restricted Erlang-R model are equivalent to those of the closed ward model of [123] if the holding queue would never be empty.



(a) The closed ward model.

(b) The Erlang-R model.

Figure 5.2: Related queueing models.

Campello et al. [54] analyzed a similar operational decision, referred to as ED case management, which determines the maximal number of patients a physician should handle in parallel. They also used queueing networks and analyzed the stationary distribution. Note that in practice such a decision is not only affected by operational measurements such as waiting times, but also by psychological constraints that limit physician capability to manage multiple tasks (patients) in parallel. KC [70] provided empirical evidence that physicians should not treat more than 6-7 patients at the same time. Therefore, many hospitals in the US restrict entrance to EDs even if beds are available if physicians are overloaded. We too consider such constraints, and analyze their impact on performance. We take a different approach than [54]; instead of analyzing numerically steady-state distributions, we develop many-server approximations that can produce insight into the system dynamics, and can be incorporated into time-varying staffing procedures; see Section 5.6.4.

The model in [123, 124] was developed for modeling internal dynamics within an internal ward. However, in the ED, beds are not constantly occupied and the utilization level depends on the flow of patients that arrive from outside the system. Yom-Tov & Mandelbaum [226] highlight the interrupted services while accounting for the transient nature of patient's arrival process, and introduced the Erlang-R model as a model for an ED. The Erlang-R model is an open two-station queueing network that has the same layout as the restricted Erlang-R model, except that all patients find a bed available upon arrival, see Figure 5.2b. In both models patients experience the interrupted services, but the Erlang-R model has no further restrictions on the bed capacity, hence neglecting the finite-size effects. Yom-Tov & Mandelbaum [226] showed, using a simulator tailored to an Israeli ED, that the complicated small ED dynamics can be captured using the relatively simple Erlang-R model, and hence, its recommendations can be implemented in ED workforce management. Although the feature of interrupted services is present in many systems, it is particularly important for modeling EDs, because the duration of the interruption is typically much longer than the time patients require care from a nurse. This explains why the Erlang-R model is considered to be the canonical model for EDs. The restricted Erlang-R model with holding/blocking thus extends the Erlang-R model with finite-size constraints which, like interrupted services, are

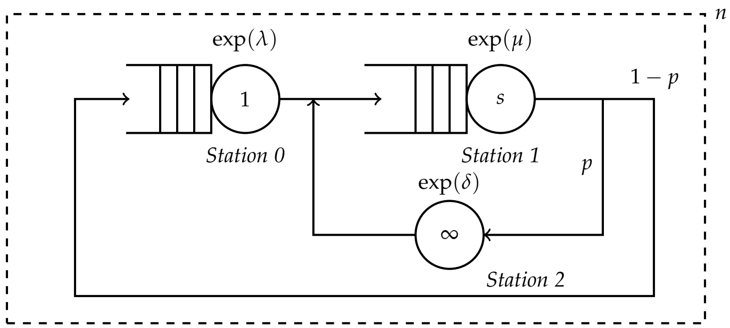


Figure 5.3: The Erlang-R model with blocking viewed as a closed Jackson network.

expected to have a decisive impact on performance.

5.3 Models and performance measures

5.3.1 Three-dimensional Markov process

Since in the restricted Erlang-R model described above the arrival process is taken Poisson, and all service and content times are assumed independent and exponential, the system can be characterized in terms of a Markov process. Let $Q(t) = (H(t), Q_1(t), Q_2(t))$ represent the number of patients in the *holding*, *needy* and *content* state at time t, respectively. In both variants, n is the maximum number of patients admitted to system, we have $Q_1(t) + Q_2(t) \leq n$ for all $t \geq 0$. Due to the absence of holding patients in the Erlang-R model with blocking, $H(t) = 0$ is enforced in this case, whereas $H(t)$ has unbounded support in the model with holding. This distinction requires us to explore the stationary distribution of the two variants separately. Before doing so, we introduce some additional notation. We define

$$R_1 := \frac{\lambda}{(1-p)\mu}, \qquad R_2 := \frac{p\lambda}{(1-p)\delta}, \tag{5.1}$$

where R_1 and R_2 can be interpreted as the offered workload brought towards the needy queue and the content (infinite-server) queue, respectively. Furthermore, we define

$$r := \frac{\delta}{\delta + p\mu}, \tag{5.2}$$

which is the fraction of time a patient spends in the needy state (in case she experienced no wait during her sojourn).

Erlang-R model with blocking. In case of the blocking model, $Q(t)$ reduces to a finite-state Markov process $Q(t) = (Q_1(t), Q_2(t))$, where $Q_1(t) + Q_2(t) \leq n$ for all

$t \geq 0$. In fact, this is equivalent to the closed Jackson network depicted in Figure 5.3 with finite population n. Station 1 in Figure 5.3 is an $M/M/s$ queue with service rate μ, modeling the number of needy patients $Q_1(t)$. Station 2 models the number of content patients $Q_2(t)$, and can therefore be represented as an infinite-server queue with service rate δ. A patient can enter the unit only if $Q_1(t) + Q_2(t) < n$. Station 0—a single-server queue—moderates this as it only produces output at rate λ in case its queue length is positive, i.e. if $n - Q_1(t) - Q_2(t) > 0$.

Observe that because patients finding a full network are blocked, the number of patients in the system cannot grow beyond n. Hence, the system is stable for all parameter settings, and hence a steady-state distribution exists. Moreover, the simplification of the model with blocking allows us to express the steady-state distribution of the system in explicit product-form. Let $\pi_b(j,k)$ denote the steady-state probabilities of having j needy and k content patients in the system. Then,

$$\pi_b(j,k) = \begin{cases} \pi_0 \, \frac{1}{\kappa(j)} \, \frac{1}{k!} \cdot R_1^j \cdot R_2^k, & \text{if } j+k \leq n, \\ 0, & \text{else,} \end{cases} \tag{5.3}$$

where

$$\kappa(j) := \begin{cases} j!, & \text{if } j \leq s, \\ s! \, s^{j-s}, & \text{else,} \end{cases}$$

and $\pi_0^{-1} = \sum_{j+k \leq n} \frac{1}{\kappa(j)} \frac{1}{k!} \cdot R_1^j \cdot R_2^k$.

Erlang-R model with holding. The Erlang-R model with holding does not lead to a Jackson network with an elegant product-form solution for the steady-state distribution, because the holding queue cannot be modeled as a station that is independent from the other queues in the system. However, we are able to describe the system as a two-dimensional Markov process without loss of information. To see this, define $N := \{N(t)\}_{t \geq 0}$ with $N(t) := H(t) + Q_1(t) + Q_2(t)$, the total number of patients in the system (including the holding queue). Using the restriction $Q_1(t) + Q_2(t) \leq n$ together with the fact that no bed is left vacant if a patient is waiting in the holding queue, this yields

$$H(t) = (N(t) - n)^+, \quad t \geq 0,$$

where $(\cdot)^+ := \max\{0, \cdot\}$. For the same reason, $Q_2(t) = N(t) - Q_1(t)$ if $H(t) = 0$, and $Q_2(t) = n - Q_1(t)$ otherwise. In other words,

$$Q_2(t) = \min\{N(t), n\} - Q_1(t), \quad t \geq 0.$$

Therefore, we can express the state of all three queues in the Erlang-R model with holding using a two-dimensional Markov process $X := \{X(t)\}_{t \geq 0}$, where

$$X(t) := (N(t), Q_1(t)).$$

The process X lives on the semi-infinite strip

$$X(t) \in \{\, (i,j) \mid j \leq \min\{i,n\}, i \in \mathbb{N}_0, j \in \{0,1,\ldots,n\} \,\},$$

and belongs to the class of Quasi-Birth-Death (QBD) processes. The reader is referred to Appendix 5.A for a detailed description of this process, in terms of its transition diagram and generator matrix.

Contrary to the model with blocking, the system with holding *can* become unstable in case capacity is insufficient to satisfy patient demand.

Proposition 5.1. *The Erlang-R model with holding is stable if and only if*

$$\frac{\lambda}{(1-p)\mu s} < \frac{\sum_{i=0}^{s} \frac{i}{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^{i} + \sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^{i}}{\sum_{i=0}^{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^{i} + \sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^{i}} =: \rho_{\max}(s,n). \tag{5.4}$$

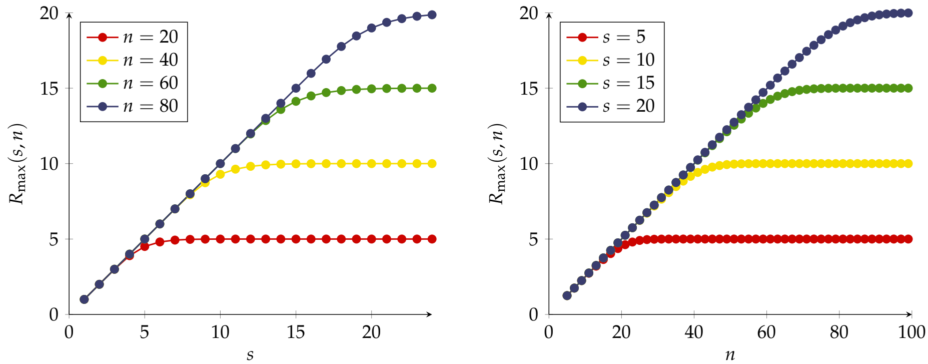
The proof is given in Appendix 5.A.2 and follows from the general theory for QBD processes.

Observe that $\rho_{\max}(s,n)$ poses an upper bound on the occupancy level of the servers in the holding model, which is clearly smaller than 1 for all s and n. In addition, this implies that the maximum workload $R_{\max}(s,n) := s \cdot \rho_{\max}(s,n)$ the system is able to handle is strictly less than s. If we compare this to the open Erlang-R model, in which the maximal attainable workload equals s, we observe the effect of finite-size constraints on operational performance. Figure 5.4 shows the influence of both s and n on the maximum feasible workload in case $r = 0.25$. From these graphs, note that if $s \ll rn$, $R_{\max}$ grows almost linearly with s. Furthermore, $R_{\max}(s,n)$ is increasing in n for s fixed. A logical practical consequence is that a larger number of beds allows for a larger patient volume to enter the ED with the same number of nurses. Moreover, $R_{\max}(s,n)$ is increasing in s, but as in Figure 5.4a, adding an extra nurse does not increase the stability region in case n is too tight. Conversely, adding extra beds does not increase $R_{\max}(s,n)$ if the number of nurses does not allow for an increase in offered load, see Figure 5.4b. Additionally, it is easily verified that $R_{\max}(s,n)$ is upper bounded by both s and $R_{\max}(n,n) = rn$. Therefore, a careful balance is called for between servers (nurses) and beds, so that resources will be efficiently utilized. We observe that when the ratio $s/n \approx r$, the system is better balanced. We will propose an appropriate balance between resources by defining a synchronized QED capacity recommendation for both servers and beds in Section 5.4.

Provided that the system is stable, the stationary distribution of the QBD process X can be obtained numerically by the matrix geometric method [169]. Subsequently, we can derive the stationary distribution of the original $Q(t)$, denoted by $\pi_h(\cdot,\cdot,\cdot)$.

5.3.2 Performance measures

In this work, we concentrate on five performance measures that are central to our analysis. In the definitions that follow, we present expressions for these measures in terms of a general three-dimensional measure π, which one can replace by either π_b or π_h, depending on the scenario considered. In the remainder of this work,



(a) $R_{\max}$ as a function of s.

(b) $R_{\max}$ as a function of n.

Figure 5.4: The maximum achievable workload in the restricted Erlang-R model with holding for $r = 0.25$.

we will augment the measures related to the Erlang-R model with blocking and holding by the superscript b and h, respectively[1].

As relevant performance measures, we consider the probability of holding (cq. blocking) at entering the system, the probability of delay at the needy queue, expected waiting time for a nurse, utilization of nurses and utilization of beds:

$$\mathbb{P}(\text{hold}) = \sum_{i=0}^{\infty}\sum_{j=0}^{n}\pi(i,j,n-j), \qquad \mathbb{P}(\text{delay}) \approx \sum_{i=0}^{\infty}\sum_{j=s}^{n}\sum_{k=0}^{n-j}\pi(i,j,k), \tag{5.5}$$

$$\mathbb{E}[W] \approx \sum_{i=0}^{\infty}\sum_{j=s}^{n}\sum_{k=0}^{n-j}\frac{\max\{0, j-s+1\}}{\mu}\pi(i,j,k), \tag{5.6}$$

$$\rho_s = \frac{1}{s}\sum_{i=0}^{\infty}\sum_{j=0}^{n}\sum_{k=0}^{n-j}\min\{j,s\}\pi(i,j,k), \qquad \rho_n = \frac{1}{n}\sum_{i=0}^{\infty}\sum_{j=0}^{n}\sum_{k=0}^{n-j}\min\{i,n\}\pi(i,j,k). \tag{5.7}$$

It should be stressed that the above expression for the delay probability and the expected waiting time for a nurse are not exact. For the blocking model one can use the Arrival Theorem, see e.g. [59], whereby the exact expression sums up to $n-1$ instead of n. Since we consider the system as $n \to \infty$, this discrepancy becomes negligible. For the holding model, a similar argument holds. We will therefore use the expressions in (5.5)-(5.7) as definitions for the performance measures.

5.3.3 Stochastic bounds

Although the two variants of the Erlang-R model differ with respect to the admission policy, and require different mathematical treatment, we would like to be able

[1] In line with $H(t) = 0$, we use $\pi_b(i,j,k) = \pi_b(j,k)$ if $i = 0$, with $\pi_b(j,k)$ as in (5.3), and $\pi_b(i,j,k) = 0$ otherwise, when considering the model with blocking.

to capture their relative performance. We substantiate the intuition that the holding room leads to more patients in the ED, in the following result.

Proposition 5.2. *Let Q_1^b, Q_2^b, Q_1^h, Q_2^h denote the nurse and content queue length processes in the Erlang-R model with blocking and holding, respectively. Let $H(0) = 0$, $Q_1^b(0) = Q_1^h(0)$ and $Q_2^b(0) = Q_2^h(0)$. For all $t \geq 0$,*

$$Q_1^b(t) + Q_2^b(t) \preceq_{\text{st}} Q_1^h(t) + Q_2^h(t) \preceq_{\text{st}} n, \tag{5.8}$$

$$Q_2^b(t) \preceq_{\text{st}} Q_2^h(t), \tag{5.9}$$

$$Q_1^b(t) \preceq_{\text{st}} Q_1^h(t) + H(t), \tag{5.10}$$

where $X \preceq_{\text{st}} Y$ implies $\mathbb{P}(X \geq k) \leq \mathbb{P}(Y \geq k)$ for all $k \geq 0$.

The proof of Proposition 5.2 uses sample path coupling and can be found in Appendix 5.B. Note that as an immediate consequence, we have

$$\mathbb{P}^b(\text{block}) = \lim_{t\to\infty} \mathbb{P}\big(Q_1^b(t) + Q_2^b(t) \geq n\big) \leq \lim_{t\to\infty} \mathbb{P}\big(Q_1^h(t) + Q_2^h(t) \geq n\big) = \mathbb{P}^h(\text{hold})$$

and by similar reasoning $\rho_n^b \leq \rho_n^h$. In other words, under similar offered load and capacity constraints, utilization levels for the nurses in the Erlang-R model with blocking are lower than in the Erlang-R model with holding. Moreover, the total number of waiting patients in the setting with holding is stochastically larger than in the setting with blocking, and in the open Erlang-R model. We further discuss the differences between both models in Section 5.5 and Section 5.6.

5.4 Two-fold QED regime

We do not want to waste capacity of either servers or beds without getting significant advantage in terms of performance. We therefore take an asymptotic approach that lets the external arrival rate λ grow to infinity, while scaling s and n accordingly. In doing so, we intend to establish QED-type system behavior, i.e. high occupancy levels of both nurses and beds and good quality-of-service.

5.4.1 Two-fold scaling rule

In order to identify the scaling of s and n as $\lambda \to \infty$, we draw inspiration from the two-fold scaling rule used by Jennings & de Véricourt [123] and Khudyakov et al. [136], which follows the celebrated square-root staffing principle. This principle suggests that, in the most general setting, capacity should be equal to the expected offered load entering the system, let us say R, plus an additional variability hedge that is proportional to $\sqrt{R}$. In the restricted Erlang-R model, we have two capacity sources, namely s and n, which experience different relevant amounts of work.

The offered load the servers in the needy queue experience is given by $R_{\text{nurse}} = R_1$, as in the regular Erlang-R model; it does not change due to the finite-size effects, since all patients are served eventually. Hence, we only need to account for the interrupted services. It follows that the appropriate staffing rule for the nurses in the QED regime remains $s = R_1 + \beta\sqrt{R_1}$ for some constant $\beta > 0$.

To establish the bed capacity level, we need to reflect on the load offered to the beds. Observe that beds remain occupied both in needy and content states. This suggests that $R_{\text{bed}} := R_1 + R_2 = R_1/r$, with R_1 and R_2 as in (5.1) and r is the expected fraction of time a patient spends at the nurse station defined in (5.2). As a result, the appropriate staffing rule is $n = R_{\text{bed}} + \gamma\sqrt{R_{\text{bed}}}$ for some constant $\gamma > 0$. In conclusion, the two-fold QED scaling rule is given by

$$\begin{aligned} s &= R_1 + \beta\sqrt{R_1} + o(\sqrt{R_1}) \\ n &= \tfrac{R_1}{r} + \gamma\sqrt{\tfrac{R_1}{r}} + o(\sqrt{R_1}) \end{aligned} \tag{5.11}$$

with $\beta, \gamma > 0$ constants and $R_1 := \lambda/((1-p)\mu)$.

Recall that we saw in Figure 5.4 that resources seem efficiently utilized if $s/n \approx r$. Scaling (5.11) is in line with this reasoning since

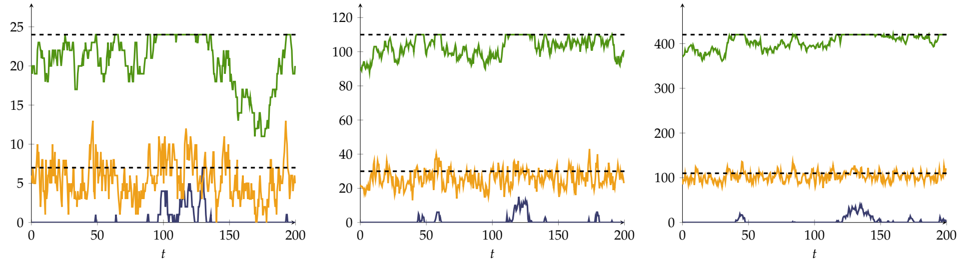
$$\frac{s}{n} = r\left(1 + \frac{\beta - \gamma\sqrt{r}}{\sqrt{R_1}} + O(1/R_1)\right).$$

Remark 5.1. In [123], a similar scaling regime is considered, which only relates s and n through a square-root scaling, namely the regime $s = rn + \hat{\gamma}\sqrt{n}$, which is equivalent to the second relation in (5.11) if $\hat{\gamma} = \beta\sqrt{r} - \gamma r$. Due to the absence of external arrivals in this closed system, they let the number of beds n approach infinity as opposed to λ in our settings. Nevertheless, this results in the same asymptotic regime.

Before turning to asymptotic expressions for the performance measures concerning the Erlang-R model with blocking or holding, we conduct a few numerical experiments to confirm that the scaling in (5.11) indeed leads to desired QED behavior.

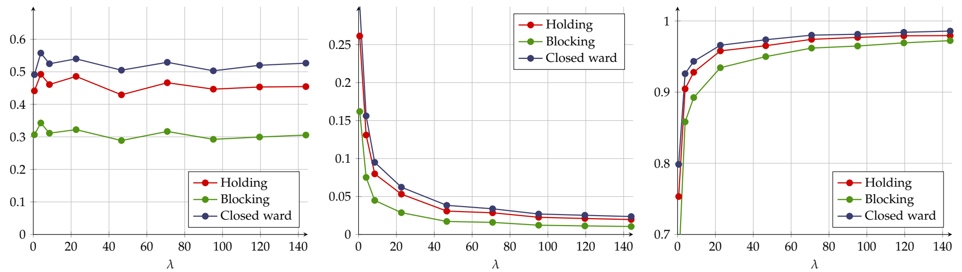
In Figure 5.5, we plotted the sample paths of the three-dimensional queue length process of the holding model in which β and γ are fixed, and R_1 is increased. Observe that the needy queue length $Q_1(t)$, plotted in orange in Figure 5.5, fluctuates around the values s, and stabilizes for larger values of R_1. This naturally implies that the server (nurses) utilization approaches 100%, while the number of patients waiting is $O(\sqrt{R_1})$. Furthermore, we see that the percentage of occupied beds also tends to 100%, while the holding queue length remains small. The holding queue is of much smaller order than R_1, which implies that the holding time of a patient becomes negligible as $R_1 \to \infty$. From these empirical findings we deduce that under scaling (5.11) the restricted Erlang-R model exhibits QED behavior on two levels: Outside the facility while waiting for an available bed, and inside the facility while waiting for attention of a nurse.

We also check how the Erlang-R model with blocking or holding and the closed ward model of [123] relate under scaling (5.11). In Figure 5.6, we plot the performance measures, obtained through simulation, for the three models in which we



(a) $R_1 = 5$ (b) $R_1 = 25$ (c) $R_1 = 100$

Figure 5.5: Sample paths of $H(t)$ (blue), $Q_1(t)$ (orange) and $Q_1(t) + Q_2(t)$ (green) of the Erlang-R model with holding with parameters $\mu = 1$, $\delta = 0.25$, $p = 0.75$ and $\beta = \gamma = 1$. The staffing levels s and n are depicted by the dashed lines.



(a) Delay probability nurse (b) Expected wait (c) Nurse utilization

Figure 5.6: Asymptotic behavior of the restricted Erlang-R model with holding and blocking, and the closed ward model for $\mu = 1$, $\delta = 0.2$, $p = 0.8$ and $\beta = \gamma = 0.5$.

fix $\beta = \gamma = 0.5$ and vary the arrival rate λ. First, we see that $\mathbb{P}(\text{delay})$ stabilizes as $\lambda \to \infty$ in all three models under scaling (5.11), and the delay probability of the model with holding lies in between the other two. Second, note that the expected waiting time for a nurse in all models converges to 0 as λ increases. In fact, the rate of decay is similar in all three models. We observe that ρ_s approaches unity in all models, and the rate of convergence seems again comparable. Finally, and most importantly, we notice an ordering between the three models. Namely, in all performance measures considered in Figure 5.6, Erlang-R with holding appears to be upper bounded by the closed ward and lower bounded by the Erlang-R with blocking. In a multitude of parameter settings of (β, γ), we have seen the same ordering, leading to the following conjecture:

Conjecture 5.1. *Let $Q_1^b(\infty)$, $Q_1^h(\infty)$ and $Q_1^J(\infty)$ denote the stationary number of needy patients in the Erlang-R model with blocking, holding and the closed ward, respectively. Then,*

$$Q_1^b(\infty) \preceq_{\text{st}} Q_1^h(\infty) \preceq_{\text{st}} Q_1^J(\infty). \tag{5.12}$$

Observe that Conjecture 5.1 poses a stronger statement than the third assertion in Proposition 5.2. The latter does give an upper bound to $Q_1^h(\infty)$ in terms of $Q_1^b(\infty)$, albeit supplemented with the stationary holding queue length.

5.4.2 QED limits for Erlang-R with blocking

We now continue our analysis by examining its limiting behavior under scaling (5.11), and obtain QED limits for some performance measures of the Erlang-R model with blocking. Using the explicit expressions for the blocking model in (5.3), we derive the limiting values of the relevant performance measures defined in Section 5.3.2 in terms of β and γ.

Theorem 5.1. *Let s and n scale as in* (5.11) *with $-\infty < \beta < \infty$, $\gamma > 0$ as $\lambda \to \infty$. Then,* *if* $\beta \neq 0$,

$$g^b(\beta,\gamma) := \lim_{\lambda\to\infty} \mathbb{P}^b(\text{delay})$$

$$= \left(1 + \frac{\beta \int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t)}{\varphi(\beta)\Phi(\eta) - \varphi(\sqrt{\beta^2+\eta^2})\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}\right)^{-1}, \tag{5.13}$$

$$f^b(\beta,\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block})$$

$$= \frac{\sqrt{r}\varphi(\gamma)\Phi(-\omega\sqrt{r}) + \varphi(\sqrt{\beta^2+\eta^2})\,\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t) + \frac{\varphi(\beta)\Phi(\eta)}{\beta} - \frac{\varphi(\sqrt{\beta^2+\eta^2})}{\beta}\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}, \tag{5.14}$$

$$h^b(\beta,\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1} \cdot \mathbb{E}[W]$$

$$= \frac{\frac{\varphi(\beta)\Phi(\eta)}{\beta^2} + \left(\frac{\beta}{r} - \frac{\gamma}{\sqrt{r}} - \frac{1}{\beta}\right) \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega) - \sqrt{\frac{1-r}{r}}\,\frac{\varphi(\beta)\varphi(\eta)}{\beta}}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t) + \frac{\varphi(\beta)\Phi(\eta)}{\beta} - \frac{\varphi(\sqrt{\beta^2+\eta^2})}{\beta}\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}, \tag{5.15}$$

and if $\beta = 0$,

$$g_0^b(\gamma) := \lim_{\lambda\to\infty} \mathbb{P}^b(\text{delay})$$

$$= \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t)}{\sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\,(\eta\,\Phi(\eta) + \varphi(\eta))}\right)^{-1}, \tag{5.16}$$

$$f_0^b(\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block})$$

$$= \frac{\sqrt{r}\,\varphi(\gamma)\Phi(-\omega\sqrt{r}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t) + \sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\,(\eta\,\Phi(\eta) + \varphi(\eta))}, \tag{5.17}$$

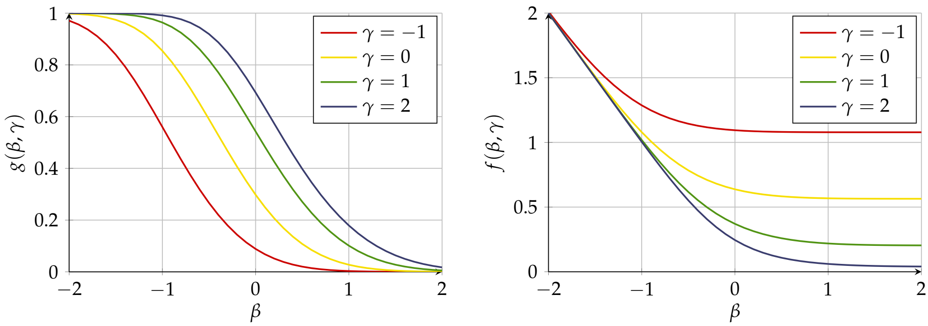
$$h_0^b(\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1} \cdot \mathbb{E}[W]$$

$$= \frac{1}{2\mu}\,\frac{\left(\gamma^2/r + 1\right)\Phi(\eta) + \eta\varphi(\eta)}{\frac{r}{1-r}\sqrt{2\pi}\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) \mathrm{d}\Phi(t) + \sqrt{\frac{r}{1-r}}\,(\eta\Phi(\eta) + \varphi(\eta))}, \tag{5.18}$$

where $\eta = \frac{\gamma - \beta\sqrt{r}}{\sqrt{1-r}}$ *and* $\omega := \frac{\gamma - \beta/\sqrt{r}}{\sqrt{1-r}}$.

The proof of Theorem 5.1 is given in Appendix C of [225] under a parameter transformation.

Theorem 5.1 proves that the scaling (5.11) results in QED behavior: the probability of waiting in Equations (5.13) and (5.16) converges to a limit that is strictly between 0 and 1. Notice that all limits in Theorem 5.1 are functions of three parameters: β and γ, which are decision variables, and the fraction of needy time r, which



(a) Delay probability (b) Scaled blocking probability

Figure 5.7: Asymptotic delay and scaled blocking probability for $r = 0.5$ based on Theorem 5.1.

is dictated by the physics of the system. Furthermore, the theorem also shows that the probability of blocking (Equations (5.14) and (5.17)) is of order $1/\sqrt{R_1}$. For example, assume that the fraction of needy time r is 0.5 and the system is large (100 servers). Using Figure 5.7, we observe that, by choosing the pair $\gamma = 1$ and $\beta = 0.245$, we actually aim at a probability of getting served immediately to be 40%. At the same time, the probability of getting immediately a bed is 97%. Thus, waiting inside the ED occurs at a reasonable level, while wait outside the facility becomes negligible.

Theorem 5.1 further shows that the expected waiting (Equations (5.15) and (5.18)) is of order $1/\sqrt{R_1}$ too and hence vanishes in the large-system limit.

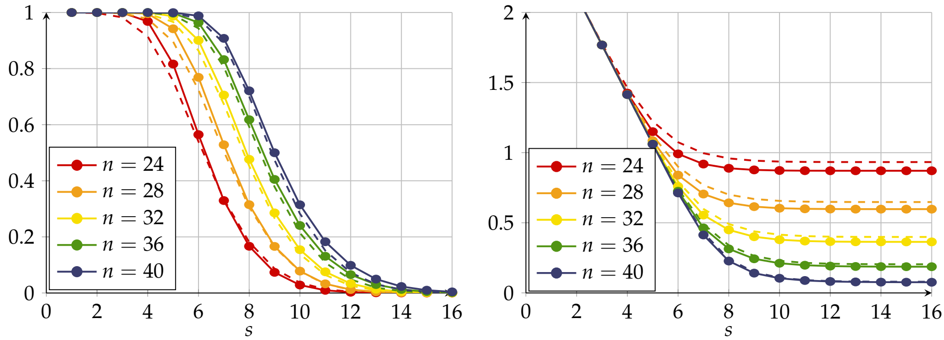
We see from Theorem 5.1 that achieving target service levels is always an interplay between β and γ. Figure 5.7a shows for instance that in order to keep $\mathbb{P}(\text{delay}) \in (0.25, 0.75)$, choosing $\gamma = -1$ requires β to stay within the range $[-1.4, -0.5]$, while $\gamma = 1$ corresponds to values of β in $[-0.4, 0.5]$.

While the two-fold scaling rule in (5.11) automatically captures the right dimensioning ratio as the system scales up, Theorem 5.1 shows that the parameters β and γ provide a means to fine-tune the performance. Figure 5.7b confirms how adding nurses, i.e. increasing β, does not improve the blocking probability if the number of beds, i.e. γ, is too tight. This is in accordance with our previous observations in Figure 5.4 for the exact steady-state distribution.

To test the accuracy of the asymptotic results in Theorem 5.1 as approximations in a realistic setting, we plot in Figure 5.8 the exact probability of delay and blocking for an Erlang-R model with $R = 8$ and $r = 0.25$, as a function of s. The exact probabilities are given by Equation (5.5), and their respective asymptotic approximations are based on Theorem 5.1. Despite the realistic moderate size of the system ($R = 8$), we see that the QED approximations are remarkably accurate for many settings (s, n). This fast relaxation is in line with observations made earlier in the QED literature [43, 120].

	μ	δ	p	r
Case 1	1	0.10	0.90	0.10
Case 2	1	0.25	0.75	0.25
Case 3	1	0.50	0.50	0.50

Table 5.1: Parameter settings for numerical experiments.



(a) Delay probability (b) Scaled blocking probability

Figure 5.8: Comparison of exact performance measures (solid) against asymptotic approximations (dashed) with $\beta = (s - R_1)/\sqrt{R_1}$ and $\gamma = (n - R_1/r)/\sqrt{R_1/r}$ for $\lambda = 2$, $\mu = 1$, $\delta = 0.25$ and $p = 0.75$.

We furthermore compare the asymptotic delay and blocking probability in the three scenarios given in Table 5.1. In Tables 5.2–5.4 we compute the exact probabilities of delay and blocking through the explicit forms in (5.5) for increasing values of the offered load, R_1.

The numerical results show that $g^b(\beta,\gamma)$, $f^b(\beta,\gamma)$ and $h^b(\beta,\gamma)$ provide accurate approximations to $\mathbb{P}(\text{delay})$, $\sqrt{R_1}\mathbb{P}(\text{block})$ and $\sqrt{R_1}\,\mathbb{E}[W]$ in pre-limit systems. The quality of the approximations increases with R_1. Naturally, fluctuations occur for relatively small values of R_1, because s and n need to be rounded to an integer.

5.4.3 QED limits for Erlang-R with holding

As explained in Section 5.4, the model with holding has no product-form steady-state distribution, which makes it hard (if not impossible) to obtain QED limits. Instead, we derive QED approximations by exploiting a connection with the blocking model.

We first prove that under scaling (5.11), the upper bound on the utilization level of the nurses needed to achieve stability in the model with holding, as given in Proposition 5.1, converges to unity as $R \to \infty$. This facilitates high utilization levels of both nurses and beds, a key characteristic of the QED regime.

	$\beta=1,\ \gamma=1$			$\beta=1,\ \gamma=2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1270	0.0900	0.2283	0.1553	0.0212	0.1085
10	0.1340	0.0910	0.1919	0.1628	0.0206	0.1205
25	0.1981	0.0945	0.1614	0.2356	0.0216	0.2145
50	0.1513	0.0963	0.1588	0.1830	0.0205	0.1496
100	0.1880	0.0956	0.1532	0.2231	0.0224	0.2055
250	0.1797	0.0971	0.1399	0.2143	0.0219	0.2057
	0.1767	*0.0981*	*0.1437*	*0.2108*	*0.0217*	*0.1947*

	$\beta=2,\ \gamma=1$			$\beta=2,\ \gamma=2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0237	0.0868	0.0282	0.0322	0.0192	0.0391
10	0.0206	0.0872	0.0188	0.0278	0.0183	0.0264
25	0.0277	0.0876	0.0123	0.0363	0.0174	0.0174
50	0.0185	0.0913	0.0116	0.0249	0.0175	0.0166
100	0.0232	0.0888	0.0103	0.0303	0.0183	0.0145
250	0.0203	0.0905	0.0079	0.0267	0.0179	0.0109
	0.0188	*0.0914*	*0.0084*	*0.0247*	*0.0177*	*0.0118*

Table 5.2: Exact numerical results for Erlang-R model with blocking for Case 1. The last row presents the asymptotic approximations.

Proposition 5.3. *Let s and n scale with $R_1 \to \infty$ as in* (5.11)*. Then for $\lambda \to \infty$,*

$$\rho_{\max}(s,n) \to 1.$$

The proof can be found in Appendix 5.C. Combining Proposition 5.3 with Proposition 5.1 shows that indeed the scaling we use results in a highly utilized system.

As observed before, the nature of the two variants of the model is similar up to the fact that a fraction of the patients is deferred on arrival in the setting with blocking, whereas all the arriving patients are eventually admitted into the system in the holding model. This implies that, given s and n, the nurses face an increased workload in case of a holding room. In fact, Theorem 5.1 shows that the blocking probability is of order $1/\sqrt{R_1}$, yielding a volume of blocked patients of order $\sqrt{R_1}$ in setting with blocking. Accordingly, if $R^b = R_1$ and R^h denote the nominal load arriving to the nurses in the model with blocking and holding, respectively, we can argue that

$$R^h = R^b + \alpha\sqrt{R^b} + o(\sqrt{R^b}),$$

for some $\alpha > 0$. Notice that this additional load is of the same order as the safety

	$\beta = 1,\ \gamma = 1$			$\beta = 1,\ \gamma = 2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0911	0.1538	0.0479	0.1431	0.0345	0.0909
10	0.1010	0.1498	0.0560	0.1520	0.0326	0.1025
25	0.1594	0.1509	0.1058	0.2192	0.0405	0.1785
50	0.1201	0.1506	0.0726	0.1697	0.0381	0.1248
100	0.1514	0.1539	0.1001	0.2088	0.0398	0.1704
250	0.1459	0.1524	0.0957	0.2003	0.0397	0.1618
	0.1429	*0.1569*	*0.0940*	*0.1976*	*0.0391*	*0.1617*

	$\beta = 2,\ \gamma = 1$			$\beta = 2,\ \gamma = 2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0130	0.1484	0.0044	0.0277	0.0294	0.0109
10	0.0121	0.1432	0.0042	0.0244	0.0267	0.0098
25	0.0182	0.1383	0.0070	0.0319	0.0295	0.0141
50	0.0119	0.1415	0.0043	0.0216	0.0301	0.0090
100	0.0154	0.1413	0.0059	0.0270	0.0290	0.0119
250	0.0136	0.1403	0.0051	0.0236	0.0291	0.0103
	0.0126	*0.1445*	*0.0048*	*0.0220*	*0.0284*	*0.0097*

Table 5.3: Exact numerical results for Erlang-R model with blocking for Case 2. The last row presents the asymptotic approximations.

staffing in the blocking model staffing rule (5.11). As s and n remain unchanged, we rewrite (5.11) in terms of R^h,

$$\begin{aligned} s &= R^h + (\beta - \alpha)\sqrt{R^h} + o(\sqrt{R^h}), \\ n &= \frac{R^h}{r} + \left(\gamma - \alpha/\sqrt{r}\right)\sqrt{\frac{R^h}{r}} + o(\sqrt{R^h}), \end{aligned} \tag{5.19}$$

where we have used $R^b = O(R^h)$. Observe that the square-root principle prevails also after this substitution, albeit with different hedging parameters. We therefore heuristically argue that the holding model under scaling (5.11) with parameters β and γ mimics the blocking model with parameters $\beta - \alpha$ and $\gamma - \alpha/\sqrt{r}$, respectively.

Observe, however, that we have not yet specified the value of α. By definition, $\alpha\sqrt{R^b}$ is the expected volume of patients that would be rejected in the model with blocking, that is, R^h times the probability of not being admitted to the ED directly. By the construction in (5.19), this volume asymptotically equals $R^h \cdot \mathbb{P}^b(\text{block})$, with parameters $\beta - \alpha$ and $\gamma - \alpha/\sqrt{r}$, which by Theorem 5.1 is approximated by

$$f^b\left(\beta - \alpha, \gamma - \alpha/\sqrt{r}\right)/\sqrt{R^h}$$

	$\beta=1,\ \gamma=1$			$\beta=1,\ \gamma=2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0547	0.1945	0.0221	0.1181	0.0604	0.0617
10	0.0579	0.2158	0.0237	0.1325	0.0526	0.0746
25	0.1113	0.2086	0.0544	0.1959	0.0641	0.1311
50	0.0813	0.2050	0.0363	0.1523	0.0562	0.0933
100	0.1060	0.2146	0.0509	0.1873	0.0632	0.1250
250	0.1006	0.2179	0.0475	0.1820	0.0596	0.1214
	0.1011	*0.2185*	*0.0478*	*0.1792*	*0.0605*	*0.1199*

	$\beta=2,\ \gamma=1$			$\beta=2,\ \gamma=2$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0034	0.1888	0.0009	0.0175	0.0510	0.0057
10	0.0030	0.2093	0.0008	0.0172	0.0416	0.0058
25	0.0070	0.1937	0.0020	0.0243	0.0440	0.0089
50	0.0043	0.1946	0.0011	0.0163	0.0414	0.0056
100	0.0061	0.1999	0.0017	0.0207	0.0431	0.0076
250	0.0052	0.2037	0.0014	0.0185	0.0401	0.0067
	0.0052	*0.2039*	*0.0014*	*0.0173*	*0.0404*	*0.0063*

Table 5.4: Exact numerical results for Erlang-R model with blocking for Case 3. . The last row presents the asymptotic approximations.

as R^h grows large. In conclusion, α is characterized as the solution of the fixed-point equation

$$\alpha = f^h\left(\beta-\alpha, \gamma-\alpha/\sqrt{r}\right), \tag{5.20}$$

and as a result, we are able to approximate the nurse delay probability in the Erlang-R model with holding as

$$\mathbb{P}^h(\text{delay}) \approx g^b(\beta-\alpha, \gamma-\alpha/\sqrt{r}) =: g^h(\beta,\gamma). \tag{5.21}$$

Likewise, the scaled mean waiting time for a nurse can be approximated by

$$\sqrt{R_1}\cdot\mathbb{E}[W] \approx h^b(\beta-\alpha, \gamma-\alpha/\sqrt{r}) =: h^h(\beta,\gamma). \tag{5.22}$$

This also implies that the holding queue is $O(\sqrt{R_1})$. Subsequently, we argue that the expected holding time (pre-entering wait) under the QED policy is $O(1/\sqrt{R_1})$ and hence asymptotically negligible. We justify this claim numerically in Section 5.6.

Remark 5.2. Notice that in the reasoning leading to (5.20), we implicitly assumed that the additional volume $\alpha\sqrt{R^b}$ is an independent Poisson process, which is obviously not the case. Therefore, (5.21)-(5.22) are approximations for pre-limit systems that are not asymptotically correct as $R_1 \to \infty$. Nevertheless, the heuristic approach seems to performs well as we confirm numerically next.

In Figure 5.9, we repeat the numerical experiments of Figure 5.8 for the model with holding. Since the heuristic does not provide an approximation for the holding probability, Figure 5.9b only plots the simulated holding probabilities. Those are provided to better understand the implication of operational decisions. Recall that the holding system is only stable (i.e. $\mathbb{P}(\text{hold}) < 1$) if both $s > R_1 = 8$ and $n > R_1/r = 32$. We nevertheless included the boundary case $n = 32$ for illustrative purposes. The graphs in Figure 5.9 show that the heuristic captures the congestion levels well, even for this moderate-size system.

To see how this heuristic approach performs under different settings, and particularly if $R_1 \to \infty$, we again compare the approximated delay probability in the Erlang-R model with holding as solution of the fixed-point procedure to the outcomes of simulation experiments for the three scenarios in Table 5.1. We performed 100 runs of length 10^4 for each parameter setting and all values of R, yielding the results presented in Tables 5.5–5.7, which are accurate up to a 95% confidence interval of width 10^{-3}.

We conclude from these tables that the approximation is good. As R increases, the simulated values move closer to the heuristic approximation. Extensive numerical experiments suggest that load is slightly underestimated in the limit. The best results in terms of accuracy are attained for small r. This suggests that the quality of the heuristic method improves as r gets smaller. These are exactly the parameter settings for which this model is relevant.

Remark 5.3. The approximation technique that evolves around the fixed-point method can be adapted to accommodate balking behavior of external arrivals. If we assume that an arriving patient finding all beds occupied leaves the system instantly with probability $1 - q$, for some $q \in (0, 1)$, independently of the rest of the arrivals, with the same argumentation, the volume of arrivals blocked is still $\alpha\sqrt{R_1}$, while the volume that will enter the ED eventually is $q \cdot \alpha\sqrt{R_1}$. Therefore, we may argue that in the QED regime, the system with holding and balking behaves as the system with blocking but with corrected parameters $(\beta - q\alpha, \gamma - q\alpha/\sqrt{r})$, where α satisfies

$$\alpha = f^b(\beta - q\alpha, \gamma - q\alpha/\sqrt{r}). \tag{5.23}$$

Note that the choice of q interpolates between the two system variants with holding ($q = 0$) and blocking ($q = 1$).

5.5 Dimensioning

We will now use the accurate asymptotic approximations of the previous section to define a procedure that determines resource capacity in the restricted Erlang-R models. That is, we aim to set the number of nurses s and the number of beds n,

	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1532	0.1031	0.1628	0.1216
10	0.1622	0.1272	0.1697	0.1331
25	0.2340	0.2116	0.2413	0.2342
50	0.1817	0.1468	0.1890	0.1678
100	0.2199	0.1931	0.2304	0.2269
250	0.2110	0.1852	0.2176	0.2230
	0.2076	*0.1777*	*0.2187*	*0.2050*

	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 1$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0310	0.0121	0.0344	0.0148
10	0.0267	0.0123	0.0292	0.0128
25	0.0348	0.0171	0.0373	0.0184
50	0.0240	0.0108	0.0258	0.0125
100	0.0293	0.0143	0.0317	0.0163
250	0.0256	0.0120	0.0276	0.0145
	0.0229	*0.0104*	*0.0257*	*0.0124*

Table 5.5: Simulated probability of delay and scaled expected waiting time in Erlang-R model with holding for Case 1. The last row gives the asymptotic approximations.

such that a preset performance level is achieved. We take the probability of delay at the needy queue and the probability of blocking/holding at the pre-entrant queue as the target performance objectives.

5.5.1 Capacity setting for Erlang-R with blocking

In the setting with blocking, we can readily use the asymptotic results of Theorem 5.1 to (numerically) find a pair of parameters (β^*, γ^*) to meet the performance requirements. For instance, given that we want the delay probability to be at most ε, we first solve the equation $g^b(\beta^*, \gamma^*) = \varepsilon$ and then assign $s = \lceil R_1 + \beta^* \sqrt{R_1} \rceil$ and $n = \lceil R_1/r + \gamma^* \sqrt{R_1/r} \rceil$. Note that there could be multiple solutions to that problem, i.e. there could be multiple combinations of number of beds and number of nurses that can result in the same value of a single performance level. The ED manager can ultimately decide which of these feasible solutions fits the environment best, for instance taking into account space and cost constraints.

We illustrate the resource allocation decisions in an MU setting, using data originated from two articles: [149] and [88]. Green & Yankovic describe an MU that has

	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1327	0.0740	0.1620	0.1096
10	0.1446	0.0894	0.1683	0.1207
25	0.2204	0.1631	0.2442	0.2203
50	0.1694	0.1122	0.1888	0.1507
100	0.2098	0.1524	0.2322	0.2111
250	0.2033	0.1534	0.2190	0.1979
	0.1840	*0.1277*	*0.2109*	*0.1759*

	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 1$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0219	0.0079	0.0322	0.0137
10	0.0199	0.0073	0.0284	0.0115
25	0.0283	0.0128	0.0375	0.0163
50	0.0190	0.0078	0.0255	0.0107
100	0.0244	0.0097	0.0314	0.0151
250	0.0214	0.0083	0.0272	0.0134
	0.0169	*0.0066*	*0.0234*	*0.0104*

Table 5.6: Simulated probability of delay and scaled expected waiting time in Erlang-R model with holding for Case 2. The last row gives the asymptotic approximations.

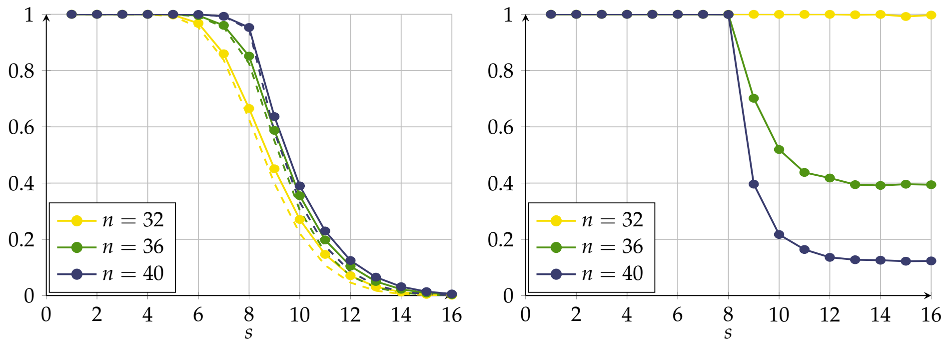
42 beds, with average occupancy level 78%, and Average Length of Stay (ALOS) 4.3 days. Lundgren & Segesten studied nurses' service times in a medical-surgical ward. They found that the average service time in their unit was 15.3 minutes per service, and that the average demand rate for each patient is 0.38 requests per hour. Therefore, we take an average service time of 15 minutes and assume that there are 0.4 requests per hour from each patient. Fitting this data to our model results in the following parameters values: $\lambda = 0.32, \mu = 4, \delta = 0.4$, $p = 0.975$ and the fraction of needy time is then approximately $r = 0.09$. This yields nominal offered load $R_1 = 3.2$ and $R_1/r = 34.4$.

Figure 5.10 visualizes the dimensioning procedure for this particular MU. The hospital management can find a pair of n and s to meet certain criteria, for example to achieve target delay probability $\varepsilon = 0.5$ with reasonable blocking probability. Figure 5.10a indicates that this target can be achieved by a variety of pairs, for instance $(\beta_1, \gamma_1) = (-0.06, -1)$, $(\beta_2, \gamma_2) = (0.16, 0)$, $(\beta_3, \gamma_3) = (0.36, 1)$ or $(\beta_4, \gamma_4) = (0.46, 2)$, among infinitely many others. According to Figure 5.10b, the pairs named above lead to blocking probabilities 0.293, 0.165, 0.071 and 0.021, respectively. If the

	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0977	0.0413	0.1521	0.0851
10	0.1070	0.0469	0.1648	0.1028
25	0.1926	0.1076	0.2421	0.1874
50	0.1431	0.0727	0.1876	0.1342
100	0.1855	0.1012	0.2282	0.1714
250	0.1775	0.0963	0.2217	0.1765
	0.1442	*0.0711*	*0.1981*	*0.1354*

	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 2$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0072	0.0019	0.0250	0.0081
10	0.0067	0.0018	0.0235	0.0082
25	0.0148	0.0043	0.0325	0.0133
50	0.0092	0.0025	0.0217	0.0081
100	0.0132	0.0038	0.0277	0.0105
250	0.0114	0.0033	0.0246	0.0099
	0.0078	*0.0022*	*0.0188*	*0.0069*

Table 5.7: Simulated probability of delay and scaled expected waiting time in Erlang-R model with holding for Case 3. The last row gives the asymptotic approximations.



(a) Delay probability

(b) Holding probability

Figure 5.9: Comparison of simulated delay probability (solid) against asymptotic approximations (dashed) with $\beta = (s - R_1)/\sqrt{R_1}$ and $\gamma = (n - R_1/r)/\sqrt{R_1/r}$ for $\lambda = 2$, $\mu = 1$, $\delta = 0.25$ and $p = 0.75$.

manager decides that probability of blocking of more than 10 percent is not acceptable, this leaves the choices $(\beta_3, \gamma_3) = (0.36, 1)$ or $(\beta_4, \gamma_4) = (0.46, 2)$ as candidate parameter pairs. Using the two-fold square-root staffing rule $s_i = \lceil R_1 + \beta_i \sqrt{R_1} \rceil$ and $n_i = \lceil R_1/r + \gamma_i \sqrt{R_1/r} \rceil$, this yields feasible staffing levels $(s_3, n_3) = (4, 40)$ and $(s_4, n_4) = (5, 46)$. The ultimate decision to apply any of these solutions can be based on external factors, such as operational costs or space limitations on the number of beds.

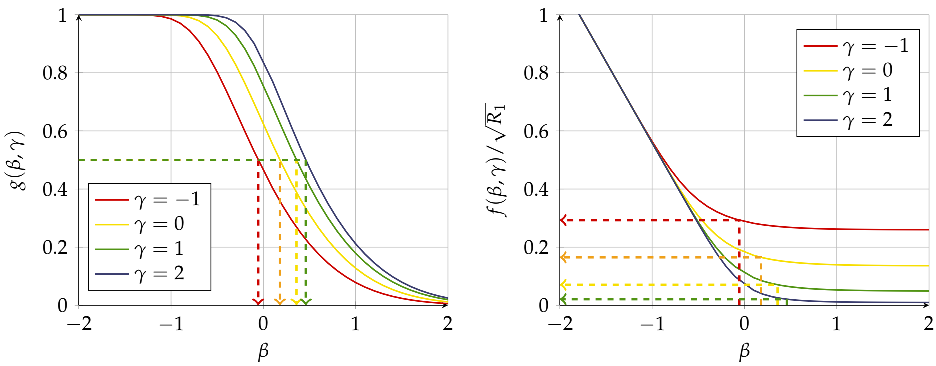
5.5.2 Capacity setting for Erlang-R with holding

For the holding model, we need a more sophisticated approach, exploiting the asymptotic approximation with the fixed-point equation in (5.20). We propose the following dimensioning procedure to achieve a preset target delay probability at the needy queue.

Remark 5.4. In Step 2 of Algorithm 3 infinitely many pairs (β^*, γ^*) satisfy the delay probability equation. For practical purposes, it is convenient to fix either β^* or γ^* beforehand, and then solve $g^b(\beta^*, \gamma^*) = \varepsilon$ for the remaining unknown. The preset value should however be chosen with care, since $g^b(\beta^*, \gamma^*)$ is upper bounded by the Halfin-Whitt delay probability formula

$$g_{\mathrm{HW}}(\beta^*) = \left(1 + \frac{\beta^* \Phi(\beta^*)}{\varphi(\beta^*)}\right)^{-1}.$$

Hence, if $\varepsilon > g_{\mathrm{HW}}(\beta^*)$, then no feasible solution to $g^b(\beta^*, \gamma^*) = \varepsilon$ exists. This should be considered when choosing β^*. Furthermore, it is evident from Step 3 that the final values (β, γ) are always larger than (β^*, γ^*).



(a) Delay probability (b) Blocking probability

Figure 5.10: Approximate performance of restricted Erlang-R with blocking for $r \approx 0.09$ and $R_1 = 3.2$, as functions of β.

Input: Target delay probability ε. Parameters λ, μ, δ and p.
Output: Staffing levels s and n.

---

1. Set $R_1 := \frac{\lambda}{(1-p)\mu}$ and $r = \frac{\delta}{\delta+p\mu}$.
2. Determine parameters (β^*, γ^*) such that $g^b(\beta^*, \gamma^*) = \varepsilon$.
3. Set $\beta = \beta^* + f^b(\beta^*, \gamma^*)$ and $\gamma = \gamma^* + f^b(\beta^*, \gamma^*)/\sqrt{r}$.
4. Return $s = \lceil R_1 + \beta\sqrt{R_1} \rceil$ and $n = \lfloor R_1/r + \gamma\sqrt{R_1/r} \rfloor$.

Algorithm 3: Stationary dimensioning algorithm for ED with holding.

We now use the same example as in Section 5.5.1 to demonstrate capacity allocation decisions for the model with holding. This can be viewed as the additional capacity the medical unit needs in terms of nurses and beds, in order to account for the fact that patients are waiting in the ED to be admitted instead of being blocked and transferred to a less preferred unit. Observe that the holding model leaves less flexibility for management in choosing system parameters due to stability constraints. For example, the policy with $n = 30$ ($\gamma = -0.75$) is infeasible in the holding model. For similar reasons, only nurse staffing levels with $\beta > 0$, or $s > R_1 = 3.2$ are feasible.

Targeting a delay probability of 0.5 with $n = 40$, Figure 5.11 shows that operating a MU with holding room requires $\beta = 0.475$ or $s = 5$. Recall that under the blocking policy, only $s = 4$ nurses were needed to achieve a delay probability of 0.5. This example hence shows how the managerial decision to have a holding room, rather than deferring patients to less preferred medical units, requires additional workforce in that unit (as well as the ED). This example also shows that the facility with holding room is able to treat fewer patients simultaneously than under blocking constraints, in line with the bounds in Section 5.3.3 and Conjecture 5.1.

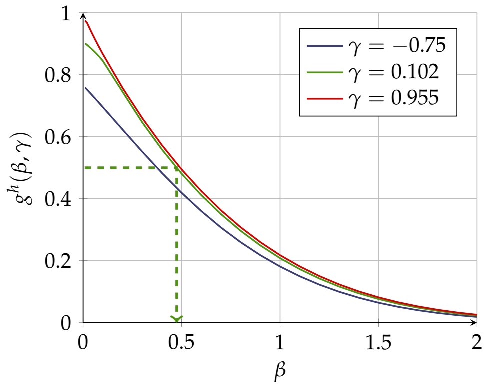
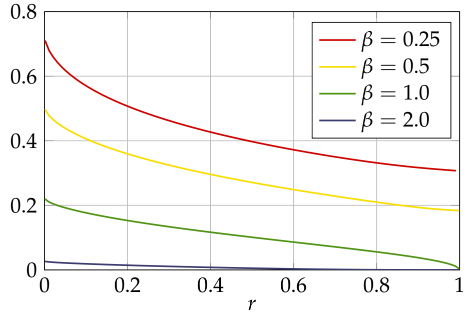


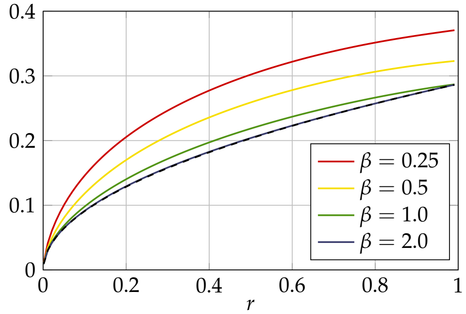
Figure 5.11: Approximate delay probability of restricted Erlang-R system with holding for $r \approx 0.09$ and $R_1 = 3.2$

5.6 Model analysis and managerial implications

In this section, we use the analysis and algorithms developed in earlier sections to gain insights into the importance of the capacity restrictions and patient returns in a restricted Erlang-R system by drawing a comparison to related models studied in the literature.



(a) Delay probability $g^b(\beta, \gamma)$



(b) Scaled blocking probability $f^b(\beta, \gamma)$.

Figure 5.12: Asymptotic performance measures as a function of r in the restricted Erlang-R model with blocking for $\gamma = 1$.

5.6.1 The influence of patient returns or the role of r

Here we study how the parameter r affects the service level in the restricted Erlang-R model with blocking, on the basis of the asymptotic expressions in Theorem 5.1.

To better understand the connection with the single-station model and the importance of returns we examine the role of r. Recall the interpretation of r as the fraction of time a patient is needy during his stay within the system in the idealized scenario with infinite capacity, i.e. for $r \in (0, 1)$. The case $r = 1$ corresponds to the setting in which patients are needy all the time, in this case patients get service in one time. When $r = 1$ the infinite-server queue, describing the number of content patients, disappears from the queueing system and we end up with a standard loss model—$M/M/s/n$ queue—in which capacity is scaled as

$$s = R_1 + \beta\sqrt{R_1}, \qquad n = R_1 + \gamma\sqrt{R_1}.$$

This staffing rule only makes sense in case $\beta < \gamma$, since no delay is experienced if $n \leq s$. If indeed $\gamma > \beta$, then the asymptotic delay probability and scaled blocking probability are given by [160],

$$g_B(\beta, \gamma) = \frac{1 - e^{-\beta(\gamma-\beta)}}{1 - e^{-\beta(\gamma-\beta)} + \beta\Phi(\beta)/\varphi(\beta)},$$
$$f_B(\beta, \gamma) = \frac{\beta e^{-\beta(\gamma-\beta)}}{1 - e^{-\beta(\gamma-\beta)} + \beta\Phi(\beta)/\varphi(\beta)}.$$

We can see that $f^b(\beta, \gamma)$ for increasing β approaches a lower bound that is a function of r. To understand this, observe that as β grows, delays at the nurse queue vanish. Then the sojourn time of an admitted patient only consists of a geometric number of needy and content periods with mean $(1/\mu + p/\delta)/(1-p) = 1/(r\mu(1-p))$. The blocking model can in this case be modeled as an $M/G/n/n$ queue, with offered load $\lambda/(r\mu(1-p)) = R_1/r$, in which the scaled blocking probability is known to be, see [25],

$$\sqrt{R_1}\,\mathbb{P}(\text{block}) = \sqrt{R_1}\,\frac{(R_1/r)^n/n!}{\sum_{k=0}^{n}(R_1/r)^k/k!} \to \sqrt{r}\,\frac{\varphi(\gamma)}{\Phi(\gamma)},$$

as $R_1 \to \infty$. This function of r is plotted in Figure 5.12b as the dashed line.

We observe that in general the probability of blocking increases with r, regardless of the capacity constraints on the needy station. We can explain this by observing that r influences only n in the QED staffing rule. When n reduces, more patients are blocked. Therefore, if patients spend relatively more time in needy state, which usually indicates services that are less interrupted, blocking will increase. Delays, on the other hand, will decrease in such situations—the minimal delay possible can be achieved if service is given in one time ($r = 1$). Returns or interruptions increase delays significantly under QED staffing.

5.6.2 Comparing restricted and unrestricted Erlang-R models

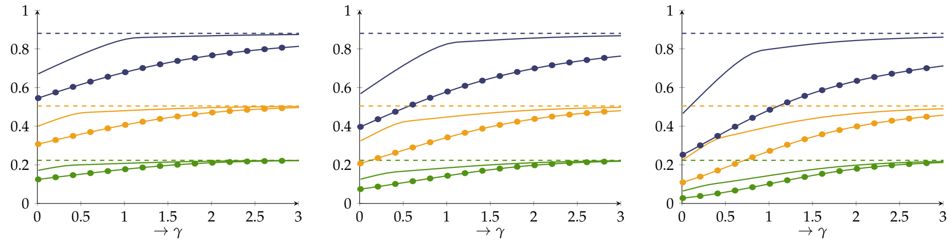
Given the expressions for the asymptotic delay probability in the open Erlang-R model, and its restricted versions with blocking and holding, we compare the three policies for various values of β, γ and r. Figure 5.13 plots the delay probability for blocking ($g^b(\beta,\gamma)$), holding ($g^h(\beta,\gamma)$) and Erlang-R ($g_{\text{HW}}(\beta)$) models, as functions of γ, while keeping β fixed, for three values of r. We make a couple of observations. Notice that

$$g^b(\beta,\gamma) \le g^h(\beta,\gamma) \le g_{\text{HW}}(\beta)$$

for all $\beta,\gamma > 0$ and r. In that sense, the holding model is an interpolation between the blocking and the open model. As expected, the delay probabilities in the restricted models converge to those of the open Erlang-R model, because increasing γ is tantamount to lifting the stringent constraints on the system size. Note that the rate of conversion is fast—one can provide probability of waiting close to that of the open model with small values of γ. Indeed, the fact that when using QED staffing not much of excessive delay results from the beds restriction is important by itself. Also, we observe that the difference between delay probabilities increases with r.

5.6.3 The impact of visit number

We next reflect on the impact of operational capacity decisions on different patient populations. We measure patient's complexity by the number of times she needs to see the nurse or the physician during her stay. In the ED context, simple-to-treat patients will need to see the physician once, while complex ones will need multiple visits. Hence, we divide the patients into complexity groups by the number of visits in the Needy station. Since the number of visits is geometrically distributed, we have a higher proportion of simple patients than complex ones; that fits well the health care environment.



(a) $r = 0.1$.

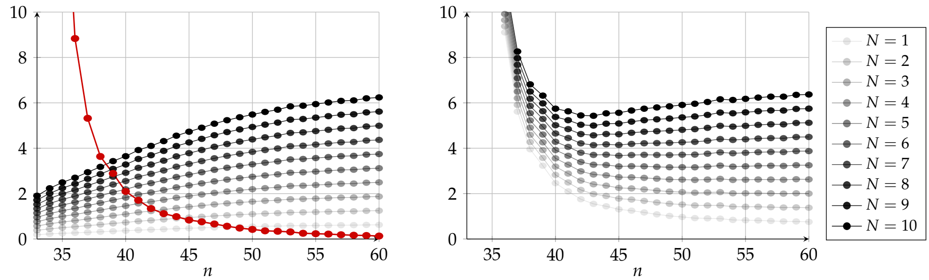
(b) $r = 0.25$.

(c) $r = 0.5$.

Figure 5.13: Asymptotic delay probability in open Erlang-R (dashed), restricted Erlang-R with blocking (marked) and restricted Erlang-R with holding (solid), as function of γ with $\beta = 0.1$ (blue), $\beta = 0.5$ (orange) and $\beta = 1$ (green) fixed.

Figure 5.14 shows the waiting time in the needy and pre-entering queues, and the total waiting time, as a function of n (number of beds), for each complexity group. Obviously, the expected waiting time in the pre-entering queue decreases with n, while the needy waiting time increases. For patients who require a relative large number of visits of the physician, in this case more than 6, the total needy wait is the dominant part of the total waiting time. Therefore, as n grows, the total waiting time first decreases and then increases. In fact, Figure 5.14b suggests that there is an optimal number of beds n that minimizes the total wait for each complexity type. Thus, size restrictions reduce the length-of-stay of patients with complex health conditions (given that the constraint is not too tight). On the other hand, this figure also shows that no such n exists for patients who only require little assistance. Hence, there is no n that improves the sojourn time of all patients in the ED simultaneously. This leaves the decision to the hospital manager to weigh the importance of patients of different complexity levels.

Remark 5.5. From a different perspective, note that in queueing systems such as communication systems, the partitioning of a job to sizable quantities and scheduling those jobs in a similar dynamic to the Erlang-R model became a popular way for increasing throughput. This is because this effectively schedules jobs by their size even though the total job requirements are uncertain. This in fact creates a shortest-job-first policy without prior knowledge of job size [39]. Considering that perspective we note that the Erlang-R model actually prioritizes simple jobs over complex ones. But without restrictions, when load is too high, such procedures may lead to very long LOS of long jobs. The capacity restriction we analyze in this chapter, in both of its versions, limits such delays. Hence, even in cases in which the returns themselves are created by a managerial decision, imposing the additional managerial restriction on entering the system has benefits.



(a) Expected pre-entering waiting (red) and needy waiting times (black)

(b) Total expected waiting times

Figure 5.14: Expected waiting times as a function of n given the number of visits N in the Erlang-R model with holding with $\lambda = 2$ $\mu = 1$, $\delta = 0.25$, $p = 0.75$ and $s = 9$.

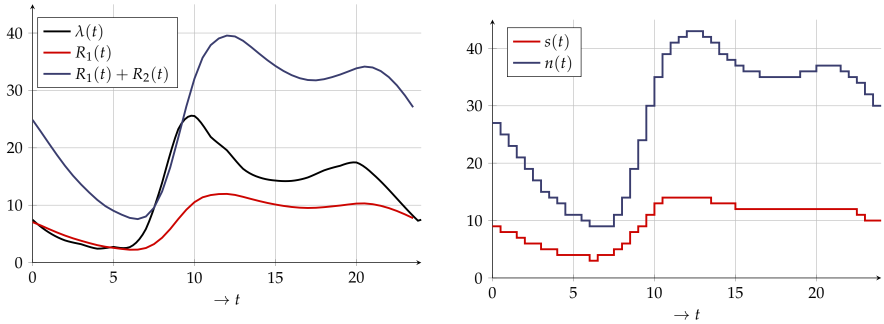
5.6.4 Case study: comparison of operational decisions

We now illustrate how the managerial decision to operate under a specific operational regime affects ED performance in terms of efficiency and quality-of-care, through a case study. The practical environment we investigate is the ED of a moderately-sized hospital, which faces the arrival pattern $\lambda(t)$ plotted in Figure 5.15a on a typical workday. Other parameters of the model are estimated to be $\mu = 6.67$, $\delta = 2.18$ and $p = 0.76$, so that $r = 0.301$. These parameters were taken from [226]. In order to set time-varying staffing levels $s(t)$ and $n(t)$, we adopt the *mean-offered load* (MOL) approximation of the demand process of [125]. This approach initially presumes infinite capacity to obtain the number of patients $R(t)$ in the queueing system as a function of time. This offered load function then replaces the (constant) value of R in the stationary dimensioning scheme under consideration, to determine the adequate number of servers at each point in time. Following this idea in our two-dimensional queueing system, we find the offered load function for the nurses $R_1(t)$ and the offered load function for the beds $R_1(t) + R_2(t)$ as the solution of the system of ODEs,

$$\frac{d}{dt}R_1(t) = \lambda(t) + \delta R_2(t) - \mu R_1(t), \tag{5.24}$$

$$\frac{d}{dt}R_2(t) = p\mu R_1(t) - \delta R_2(t), \tag{5.25}$$

see [226, Thm. 2] for details. For this case study's parameters, these offered load functions are also plotted in Figure 5.15a. While the number of nurses can be adjusted in a relatively flexible manner, the value of n, which echoes a hard restriction on the ED capacity, is naturally less amenable to fluctuations. The reason is that the maximum ED capacity is to a large extent determined by its hardware, such as beds and medical equipment. However, the ED manager might deliberately consider reducing n during more quiet periods of the day, e.g. during the night, by



(a) Dynamic arrival rate function offered load functions

(b) Capacity function for $\beta = \gamma = 0.5$

Figure 5.15: Empirical arrival rate and offered load functions $R_1(t)$ and $R_1(t) + R_2(t)$ in Israeli ED and corresponding capacity functions.

imposing bed-to-physician constraints. This is done, for example, when setting a case management constraint [196, 54]. Therefore, we consider the scenario in which both s and n are time-dependent but we do not force a constant case management quantity, rather let our new methodology recommend an appropriate one.

Extrapolating Algorithm 3 to the time-varying case, Step 4 is replaced by

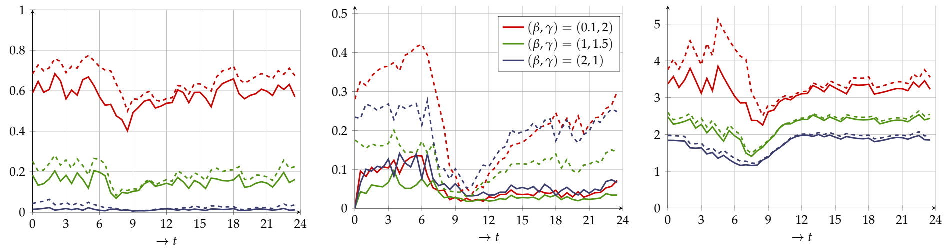
$$s(t) = R_1(t) + \beta\sqrt{R_1(t)},$$
$$n(t) = R_1(t) + R_2(t) + \gamma\sqrt{R_1(t) + R_2(t)},$$

for some $\beta, \gamma > 0$. Since $R_1(t)$ and $R_2(t)$ are given, the QED staffing problem again reduces to finding the pair (β, γ).

Figure 5.15b plots the capacity functions for $\beta = 0.5$ and $\gamma = 0.5$, assuming capacity can only be adjusted every 30 minutes. In this case study, we consider three pairs of parameters (β, γ). For each of these we investigate, using simulation, the differences in the time-varying performance indicators between the policy with blocking and holding.

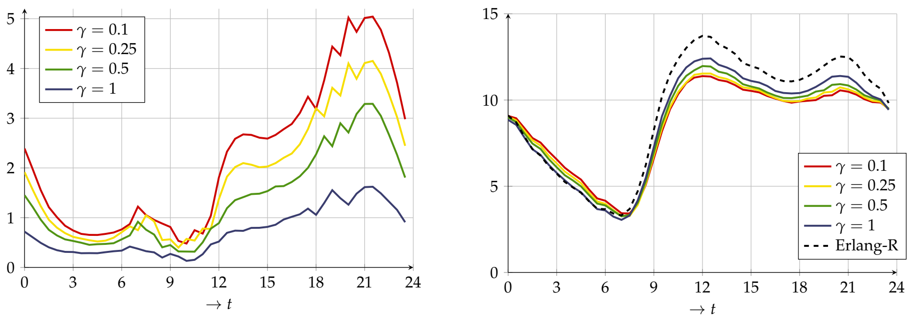
The simulation results are presented in Figure 5.16. Figure 5.16a shows that the MOL approach for capacity allocation roughly stabilizes the delay probability. Figure 5.16b shows that the fraction of patients not entering the ED on arrival in the blocking model is reasonable for all parameter pairs considered and the graphs are ordered according to γ. We also see a significant difference with holding. Observe also that the holding probability drops in the period 8–13, which is exactly the period when the system is experiencing peak offered load. Hence, this temporary reduction is in line with our asymptotic findings that the probability of blocking/holding is $O(1/\sqrt{R_1})$.

Finally note that the three parameter settings lead to different nurse-to-patient ratios. Clearly, larger β leads to small nurse-to-patient ratios (due do larger staffing). Figure 5.16c demonstrates that for $(\beta, \gamma) = (1, 1.5)$ and $(\beta, \gamma) = (2, 1)$ the difference



(a) $\mathbb{P}(\text{delay})$ (b) $\mathbb{P}(\text{block})$ or $\mathbb{P}(\text{hold})$ (c) Nurse-to-patient ratio.

Figure 5.16: Simulation results for case study. Solid and dashed lines represent time-varying performance in the blocking and holding model, respectively.



(a) Mean holding queue length (b) Mean service queue length

Figure 5.17: Simulated queue length of holding model with different values of γ.

between the holding policy and the blocking policy is small. However, for $(\beta, \gamma) = (0.1, 2)$ we see a significant increase in the ratio during night hours. This may be due to the tight nurse schedule, that causes the holding queue to build up just before midnight. This queue then empties latter on, causing an increase in workload per nurse in the period 24–7.

To see the direct effect of the size restriction on the queue lengths, we plotted the mean holding and service queue lengths in the holding model as a function of the parameter γ in Figure 5.17. Note that for all γ considered, the holding queue lengtsh are, as expected, of a smaller order than the number of patients admitted. Also, the holding queue length decreases as we increase γ. The service queue lengths naturally approach the expected queue lengths in the Erlang-R model as γ is increased.

5.7 Conclusion & future research

In this chapter we developed and analyzed a queueing network tailored to a health care environment with finite-size restrictions. Using the asymptotic approximations, numerical analysis and simulation, we gained insight into staffing problems that arise in EDs, and proposed an efficient, flexible, and easy to implement methodology to dimension medical facilities through a two-fold staffing rule.

The dimensioning scheme we developed provides a powerful and elegant way of finding adequate staffing levels in emergency departments. Nonetheless, we see some avenues for further research.

The asymptotic approximations we developed enabled us to take the first step towards characterizing the pre-entering queue behavior in the QED regime. We observed how the holding queue length vanishes at rate $1/\sqrt{R_1}$ as $R_1 \to \infty$. Yet, our analysis did not yield explicit characteristics on the holding queue and holding times. These performance indicators are naturally important to study if one wants to consider the trade-off between waiting time inside the ED and waiting time outside the ED time (pre-entering time).

Furthermore, it is worthwhile to study the robustness of our approximations against the service and content time distributions. Since the content phase of a patient is modeled after an infinite-server queue, we expect our approximations to be useful for content time distributions beyond the exponential distribution as well, due to distributional insensitivity of the service time in infinite-server queues. For the needy phase, modeled after a multi-server queue, this insensitivity result does not hold and hence this needs further research.

Finally, the restricted Erlang-R model obviously gives a highly simplified view of the complex reality of the ED. In practice, distinctive features such as a triage system (with patient priorities), patient boarding time and availability of medical equipment may play a decisive role on ED dynamics. However, we think the analysis and dimensioning algorithms presented in this chapter can serve as a building block for staffing procedures that do account for these case-specific factors.

Appendix

5.A Description of the QBD process

5.A.1 The QBD-process

We consider the QBD-process $X(t) = (N(t), Q_1(t))$ in stationarity. Let $\nu(i) = \min\{i, s\}\mu$. To determine the (outgoing) transition rates of the process X we distinguish between the following cases:

- *Transitions from* $(0,0)$*:* There are no patients in the Emergency Department and thus the only possible occurrence is when a new patient arrives. This results in a transition to $(1,1)$ and 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 results in a transition to $(i,1)$

and occurs at rate $i\delta$, and the second results in a transition to $(i+1,1)$ and occurs with rate λ.

- *Transitions from* $(i,0), i \geq n$*:* Again, the only possible transitions arise from either a newly arrived patient or a patient assigned to a bed becoming needy. However, a newly arrived patient finds all beds occupied and needs to wait. Thus, 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 the holding phase, 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, resulting in a transition to $(n+1,n)$ with rate λ.

- *Transitions from* $(i,n), i > n$*:* We have a transition to $(i+1,n)$ with rate λ and a transition to $(i,n-1)$ with rate $ps\mu$. In case that a patient leaves the Emergency Department there are $i-n > 0$ patients in the holding room waiting for an available bed. Thus, 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. First, with rate λ there is a new arrival which results in a transition to $(i+1,j+1)$. Second, with rate $(i-j)\delta$ a patient in one of the beds becomes needy, which results in a transition to $(i,j+1)$. Third, with rate $p\nu(j)$ a patient turns to the content state after service completion, which results in a transition to $(i,j-1)$. Last, 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 one. 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 the previous one, except when a patient leaves the Emergency Department after service completion.

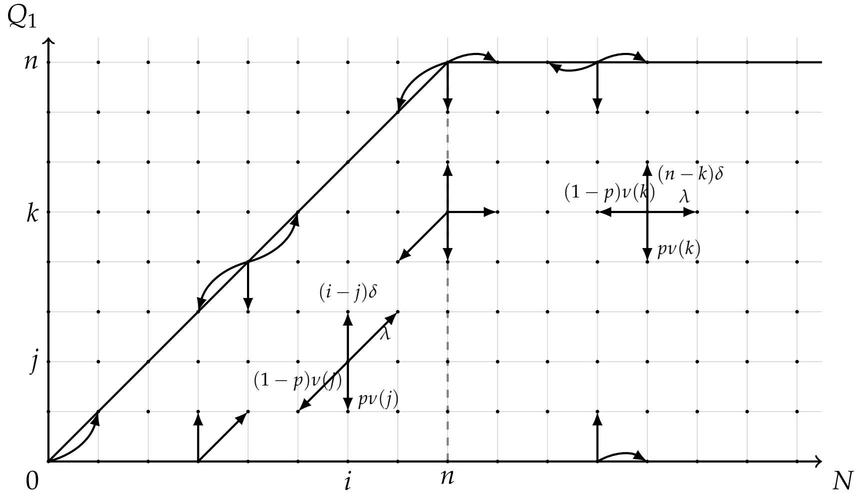


Figure 5.18: Transition diagram for the Erlang-R model with holding.

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 Erlang-R model with holding are illustrated in Figure 5.18. The state space can be partitioned according to its levels, where level i corresponds to a total queue length $N = i$ patients. 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 $(n+1) \times (n+1)$.

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

$$B_{00} = (-\lambda), \qquad B_{i-1\,i} = \begin{pmatrix} 0 & \lambda & & & \\ & \ddots & \lambda & & \\ & & \ddots & \ddots & \\ & & & 0 & \lambda \end{pmatrix},$$

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

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 are given by

$$A_0 = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix}$$

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

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.A.2 Stability condition

From the general theory of QBD processes [169] follows that the Markov process $(N(t), Q_1(t))$ is ergodic (stable) if and only if

$$\pi A_0 e < \pi A_2 e, \tag{5.26}$$

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

$$\pi(A_0 + A_1 + A_2) = 0, \quad \pi e = 1, \tag{5.27}$$

and

$$A_0 + A_1 + A_2 =$$

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

Then π must satisfy the balance equations

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

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

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

It is readily verified that

$$\pi_i = \begin{cases} \pi_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \le i \le s, \\ \pi_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1 \le i \le n \end{cases} \tag{5.28}$$

with

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

satisfies the balance equations and the normalization condition.

Proposition 5.4. *The distribution of the closed two-node Jackson network illustrated in Figure 5.2a is given by*

$$\hat{\pi}_i = \begin{cases} \hat{\pi}_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \leq i \leq s, \\ \hat{\pi}_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1 \leq i \leq n \end{cases} \tag{5.29}$$

with

$$\hat{\pi}_0 = \left[\sum_{i=0}^{s} \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n} \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i \right]^{-1}.$$

Proof. We have a two-node closed Jackson network, with probability transition matrix

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

Let $r_i(m)$ denote the rate of service when there are m patients at queue i, so $r_1(m) = \min\{m, s\}$ and $r_2(m) = m$. The throughput vector $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}_{\mathbb{1}}^2$ must satisfy $\gamma = \gamma P$ and we find that $\gamma = (p, 1)$ suffices. From the general theory of Jackson networks, see [112], it follows that the stationary distribution is given by

$$\pi_i = G^{-1} g_1(i) g_2(n-i)$$

with

$$g_1(i) = \frac{(\gamma_1/\mu)^i}{\prod_{m=1}^{i} r_1(m)}, \quad g_2(n-i) = \frac{(\gamma_2/\delta)^{n-i}}{\prod_{m=1}^{n-i} r_2(m)},$$

and normalization constant $G = \sum_{i=0}^{n} g_1(i) g_2(n-i)$. Then,

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

$$g_2(n-i) = \frac{1}{(n-i)!} \left(\frac{p}{\delta}\right)^n \left(\frac{\delta}{p}\right)^i,$$

and rewriting the expressions yields (5.29). □

5.A.3 Stationary distribution

Assuming that the stability condition is satisfied, we can determine the unique stationary distribution of the Markov process $(N(t), Q_1(t))$. The vector π_i can be written as $\pi_{n+i} = \pi_n G^i$ for $i = 0, 1, ...$, where G is the minimal nonnegative solution of the non-linear matrix equation

$$A_0 + GA_1 + G^2 A_2 = 0. \tag{5.30}$$

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

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

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 (5.30), it is easy to solve numerically by using the following algorithm (matrix-geometric method). Rewriting (5.30) gives

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

where A_1 is invertible, since it is a transient generator matrix. Let

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

starting with $G_0 = 0$. We note that $G_k \uparrow G$ as k grows to infinity [169]. Once $||G_{k+1} - G_k||_2$ is below a certain preset threshold, we approximate G by G_{k+1}.

5.B Proof of Proposition 5.2

First, note that by definition of the Erlang-R model with holding, in which no more than n patients can be admitted in the ED simultaneously, that $Q_1^h(t) + Q_2^h(t) \leq n = Q_1^J(t) + Q_2^J(t)$ follows directly. Therefore, we only consider the relation between the states in the blocking and holding variants Erlang-R model.

As noted Section 5.3.1, the model with holding can be characterized as a three-dimensional Markov chain $X^h(t) = (H(t), Q_1^h(t), Q_2^h(t))$ in which the components denote the number of holding, needy and content patients respectively. The Erlang-R model with blocking similarly admits a Markov process description, but with two dimensions, namely $X^b(t) = (Q_1^b(t), Q_2^b(t))$.

We prove the result by constructing a coupling between the Markov processes X^h and X^b. Let $Z(t) := \big(\hat{X}^h(t), \hat{X}^b(t)\big) = \big(\hat{H}(t), \hat{Q}_1^h(t), \hat{Q}_2^h(t), \hat{Q}_1^b(t), \hat{Q}_2^b(t)\big)$.

We first define the transition rates of this five-dimensional Markov process, which naturally only depend on the current state of the system. After that we show that the transition rates relevant to $\hat{X}^h(t)$ and $\hat{X}^b(t)$ coincide with those of $X^h(t)$ and $X^b(t)$, respectively. The latter implies that the marginal transitions of $\hat{X}^h(t)$ and $X^h(t)$ (and $\hat{X}^b(t)$ and $X^b(t)$) are equal, and hence so are their probability distribution of the Markov processes.

Let $Z(t) = (h, q_1^h, q_2^h, q_1^b, q_2^b)$. While defining the reachable states from this state and associated transition rates, we distinguish four transition types, and further differentiate the transition rates depending on the current state.

Arrival. Arrivals occur in both models simultaneously, but are handled differently according to the current queue lengths.

1. If $q_1^h + q_2^h < n$ and $q_1^b + q_2^b < n$,

$$(h, q_1^h + 1, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{5.31}$$

2. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b < n$,

$$(h + 1, q_1^h, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{5.32}$$

3. if $q_1^h + q_2^h < n$ and $q_1^b + q_2^b = n$,

$$(h, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{5.33}$$

4. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b = n$,

$$(h + 1, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{5.34}$$

Departure. Basically, we align service completions in the two models, but allow a completion occurring solely in either of one of the two models, only if the queue length in this model is strictly larger than in the other one.

1. If $q_1^h \geq q_1^b$ and $h > 0$

$$\begin{cases} (h - 1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1 - p)\mu, \\ (h - 1, q_1^h, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1 - p)\mu. \end{cases} \tag{5.35}$$

2. If $q_1^h < q_1^b$ and $h > 0$

$$\begin{cases} (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \tag{5.36}$$

3. If $q_1^h \geq q_1^b$ and $h = 0$

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (0, q_1^h - 1, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1-p)\mu. \end{cases} \tag{5.37}$$

4. If $q_1^h < q_1^b$ and $h = 0$

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \tag{5.38}$$

Become content. The differentiation between transitions is similar to those in the *departure* transition type.

1. If $q_1^h \geq q_1^b$,

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^b \wedge s)p\mu, \\ (h, q_1^h - 1, q_2^h + 1, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)]p\mu. \end{cases} \tag{5.39}$$

2. If $q_1^h < q_1^b$,

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b + 1) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)]p\mu. \end{cases} \tag{5.40}$$

Become needy.

1. If $q_2^h \geq q_2^b$,

$$\begin{cases} (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^b\delta, \\ (h, q_1^h + 1, q_2^h - 1, q_1^b, q_2^b) & \text{with rate } (q_2^h - q_2^b)\delta, \end{cases} \tag{5.41}$$

2. If $q_2^h < q_2^b$,

$$\begin{cases} (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^h\delta, \\ (h, q_1^h, q_2^h, q_1^b + 1, q_2^b - 1) & \text{with rate } (q_2^b - q_2^h)\delta, \end{cases} \tag{5.42}$$

This set of transitions defines the dynamics of the Markov process $Z(t) = (\hat{X}^h(t), \hat{X}^b(t))$. Let us now restrict our attention to the transitions in which (at least one of) the first three coordinates of $Z(t)$ changes, that is, the marginal transitions of the process $\hat{X}^h$. Let $\hat{X}^h(t) = (h, q_1^h, q_2^h)$, then according to the transition scheme above, $\hat{X}^h$ moves to state

1. If $q_1^h + q_2^h < n$ (and hence necessarily $h = 0$),

$$\begin{cases} (0, q_1^h + 1, q_2^h) & \text{with rate } \lambda, \\ (0, q_1^h - 1, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h - 1, q_2^h + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (0, q_1^h + 1, q_2^h - 1) & \text{with rate } q_2^h\delta. \end{cases}$$

2. if $q_1^h + q_2^h = n$ and $h = 0$,

$$\begin{cases} (1, q_1^h, q_2^h) & \text{with rate } \lambda, \\ (0, q_1^h, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h - 1, q_2^h + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (0, q_1^h + 1, q_2^h - 1) & \text{with rate } q_2^h\delta. \end{cases}$$

3. if $h > 0$ (and hence necessarily $q_1^h + q_2^h = n$),

$$\begin{cases} (h+1, q_1^h, q_2^h) & \text{with rate } \lambda, \\ (h-1, q_1^h, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (h, q_1^h - 1, q_2^h + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (h, q_1^h + 1, q_2^h - 1) & \text{with rate } q_2^h\delta. \end{cases}$$

One can check that these transitions indeed coincide with the transitions in the original holding model, hence $\hat{X}^h(t) \stackrel{d}{=} X^h(t)$.

Similarly, when focusing on transitions of $Z(t)$ that are relevant for $\hat{X}^b(t)$, we deduce the following transition scheme. If $\hat{X}^b(t) = (q_1^b, q_2^b)$, then the next move according to the transitions of $Z(t)$ is

$$\begin{cases} (q_1^b + \mathbb{1}_{\{q_1^b + q_2^b < n\}}, q_2^b) & \text{with rate } \lambda, \\ (q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^b \wedge s)p\mu, \\ (q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^b\delta. \end{cases}$$

These transition rates clearly coincide with the original Erlang-R model with blocking, and also hence $\hat{X}^b(t) \stackrel{d}{=} X^h(t)$.

Next, we show that under this coupling scheme we have that if $\hat{H}(0) = 0$, $\hat{Q}_1^h(0) = \hat{Q}_1^b(0)$ and $\hat{Q}_2^h(0) = \hat{Q}_2^b(0)$ then for all $t \geq 0$, $Z(t)$ satisfies the hypothesis:

(i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) \leq \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$,

(ii) $\hat{Q}_2^b(t) \leq \hat{Q}_2^h(t)$,

(iii) $\hat{Q}_1^b(t) \leq \hat{Q}_1^h(t) + H(t)$.

We do so by induction on the next state reached after a transition of the joint Markov process $Z = (\hat{X}^h, \hat{X}^b)$. First of all, $Z(0)$ clearly satisfies (i)-(iii). Next, assume

$Z(t^-) = (h, q_1^h, q_2^h, q_1^b, q_2^b)$ satisfies the hypothesis and a transition occurs at t. We show that under the specified coupling scheme, the state reached after the next transition, $Z(t)$ must satisfy (i)-(iii) as well. To do so, we differentiate between the four types of transitions that could occur: arrival, departure, become content and become needy.

Arrival. Recall that under our coupling scheme an arrival always occurs in both the holding and blocking model simultaneously, see (5.31)–(5.34). Furthermore, q_2^h and q_2^b are unchanged during this transition, rendering (ii) trivial.

By hypothesis $q_1^b + q_2^b \leq q_1^h + q_2^b$, hence the event $q_1^h + q_2^h < n$ and $q_1^h + q_2^b = n$, with resulting state $(0, q_1^h + 1, q_2^h, q_1^b, q_2^b)$, can be excluded from our analysis. We check the conditions for the remaining three cases.

1. If $Z(t) = (0, q_1^h + 1, q_2^h, q_1^b + 1, q_2^b)$, then $q_1^b + q_2^b < n$ and $q_1^h + q_2^h < n$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \overset{(i)}{\leq} q_1^h + q_2^h + 1 = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\leq} q_1^h + 1 = \hat{Q}_1^h(t) = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h+1, q_1^h, q_2^h, q_1^b + 1, q_2^b)$, then $q_1^b + q_2^b < n$ and $q_1^h + q_2^h = n$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \leq n = q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\leq} q_1^h + 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
3. If $Z(t) = (h+1, q_1^h, q_2^h, q_1^b, q_2^b)$, then $q_1^b + q_2^b = q_1^h + q_2^h = n$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\leq} q_1^h + h < q_1^h + h + 1 = \hat{H}(t)$.

Departure. By carefully examining the possible state transitions of $Z(t)$ following a departure, we list six reachable states. However, by (iii), we have that if $h = 0$, then $q_1^b \leq q_1^h$, which excludes the state $(0, q_1^h, q_2^h, q_1^b, q_2^b)$ in (5.38) from the reachability graph. We check the remaining states for conditions (i)–(iii). Again, during a departure, q_2^b and q_2^h are unchanged, so (ii) is automatically satisfied by the induction hypothesis.

1. If $Z(t) = (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b)$, then $h > 0$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 \overset{(i)}{\leq} q_1^h + q_2^h - 1 < q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h-1, q_1^h, q_2^h, q_1^b, q_2^b)$, then $h > 0$ and $q_1^h \geq q_1^b$ (*).

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(*)}{\leq} q_1^h - 1 \leq q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

3. If $Z(t) = (h, q_1^h, q_2^h, q_1^b - 1, q_2^b)$, then $h > 0$ and $q_1^h < q_1^b$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 < q_1^b + q_2^b \overset{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 < q_1^b \overset{(iii)}{\leq} q_1^h + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

4. If $Z(t) = (h, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b)$, then $h = 0$.

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = (q_1^b - 1) + q_2^b - 1 < \overset{(i)}{\leq} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

5. If $Z(t) = (0, q_1^h - 1, q_2^h, q_1^b, q_2^b)$, then $h = 0$ and $q_1^h > q_1^b$ (*).

   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\leq} (q_1^h - 1) + q_2^b \overset{(ii)}{\leq} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(*)}{\leq} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

Content start. On the event of a patient becoming content, it is clear that the sums $\hat{Q}_1^h(t) + \hat{Q}_2^h(t)$ and $\hat{Q}_1^b(t) + \hat{Q}_2^b(t)$ and $H(t)$ are unaffected. This means that (i) is directly satisfied by the induction hypothesis. According to (5.39)–(5.40), three states can be reached.

1. If $Z(t) = (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1)$,

   (ii) $\hat{Q}_2^b(t) = q_2^b + 1 \overset{(ii)}{\leq} q_2^h + 1 = \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

2. If $Z(t) = (h, q_1^h - 1, q_2^h + 1, q_1^b, q_2^b)$, then $q_1^h > q_1^b$,

   (ii) $\hat{Q}_2^b(t) = q_2^b \overset{(ii)}{\leq} q_2^h < q_2^h + 1 = \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\leq} q_1^h + h < q_1^h + 1 + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

3. If $Z(t) = (h, q_1^h, q_2^h, q_1^b - 1, q_2^b + 1)$, then $q_1^b > q_1^h$ (*) and hence by (iii) $h > 0$. The latter is only possible if $q_1^h + q_2^h = n$,

   (ii) $\hat{Q}_2^b(t) = q_2^b + 1 \leq n - q_1^b + 1 = (q_1^h + q_2^h) - q_1^b + 1 \overset{(*)}{\leq} q_2^h = \hat{Q}_2^h(t)$.

   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 < q_1^h + h - 1 \overset{(*)}{\leq} q_1^h + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

Become needy. Just as in the event of content start, the sums $\hat{Q}_1^h(t) + \hat{Q}_2^h(t)$ and $\hat{Q}_1^b(t) + \hat{Q}_2^b(t)$ and $H(t)$ are unaffected, whereby (i) is directly satisfied by the induction hypothesis. By (ii), we have $q_2^h \geq q_2^b$. This excludes the state $(h, q_1^h, q_2^h, q_1^b + 1, q_2^b - 1)$ from being reached, see (5.42). We check the remaining two possibilities.

1. If $Z(t) = (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1)$.

   (ii) $\hat{Q}_2^b(t) = q_2^b - 1 \overset{(ii)}{\leq} q_2^h - 1 = \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\leq} q_1^h + h + 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

2. If $Z(t) = (h, q_1^h + 1, q_2^h - 1, q_1^b, q_2^b)$, then $q_2^h > q_2^b$ (*).

(ii) $\hat{Q}_2^b(t) = q_2^b \overset{(*)}{\leq} q_2^h - 1 = \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\leq} q_1^h + h < q_1^h + 1 + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

Hence, the state reached after any feasible transition under the coupling scheme satisfies the conditions (i)–(iii). Thus we conclude that the joint process $(\hat{H}(t), \hat{Q}_1^h(t), \hat{Q}_2^h(t), \hat{Q}_1^b(t), \hat{Q}_2^b(t))$ adheres to (i)–(iii) for all t. Consequently, we have that (i) implies

$$
\begin{aligned}
\mathbb{P}\left(Q_1^b(t) + Q_2^b(t) \geq k\right) &= \mathbb{P}\left(\hat{Q}_1^b(t) + \hat{Q}_2^b(t) \geq k\right) \\
&= \sum_{j=0}^{n} \mathbb{P}\left(\hat{Q}_1^b(t) + \hat{Q}_2^b(t) \geq k, \hat{Q}_1^h(t) + \hat{Q}_2^h(t) = j\right) \\
&= \sum_{j=k}^{n} \mathbb{P}\left(\hat{Q}_1^b(t) + \hat{Q}_2^b(t) \geq k, \hat{Q}_1^h(t) + \hat{Q}_2^h(t) = j\right) \\
&\leq \sum_{j=h}^{n} \mathbb{P}\left(\hat{Q}_1^h(t) + \hat{Q}_2^h(t) = j\right) \\
&= \mathbb{P}\left(Q_1^h(t) + Q_2^h(t) \geq k\right) = \mathbb{P}\left(Q_1^h(t) + Q_2^h(t) \geq k\right).
\end{aligned}
$$

The other two orderings follow similarly.

Remark 5.6. Note that under this coupling scheme we cannot get the ordering $\hat{Q}_1^h(t)(t) \geq \hat{Q}_1^b(t)(t)$ for all $t \geq 0$. A minimal counter example occurs for $s = n = 1$. Let $Z(0) = ((0,0,0),(0,0))$. First, two arrivals occur, such that state $((1,1,0),(1,0))$ is reached, followed by a departure transition, yielding $((0,1,0),(0,0))$. Next, the one patient left in the model with holding system becomes content, so that we obtain $((0,0,1),(0,0))$. At this stage, if an arrival occurs, the arriving patient will be put in the holding queue in the model with holding, and admitted to nurse queue in the model with blocking. Hence we end up in state $((1,0,1),(1,0))$, in which $\hat{Q}_1^h(t) < \hat{Q}_1^b(t)$.

5.C Proof of Proposition 5.3

Define

$$
A(s,n) = \sum_{k=0}^{s} \frac{k}{s}\binom{n}{k} b^k, \quad B(s,n) = \sum_{k=s+1}^{n} \frac{k!}{s!}\binom{n}{k} s^{s-k} b^k, \quad C(s,n) = \sum_{k=0}^{s}\binom{n}{k} b^k,
$$

where $b = \delta / p\mu = r/(1-r)$. Then

$$\rho_{\max}(s,n) = \frac{A(s,n) + B(s,n)}{C(s,n) + B(s,n)}.$$

Proving that $\rho_{\max}(s,n) \to 1$ as $R_1 \to \infty$ with s and n as in (5.11) is equivalent to showing that

$$1 - \rho_{\max}(s,n) = \frac{C(s,n) - A(s,n)}{C(s,n) + B(s,n)} = \frac{(1+b)^{-n}[C(s,n) - A(s,n)]}{(1+b)^{-n}[C(s,n) + B(s,n)]} \to 0. \tag{5.43}$$

First, we rewrite

$$\begin{aligned}
(1+b)^{-n} A(s,n) &= (1+b)^{-n} \sum_{k=1}^{s} \frac{n}{s} \binom{n-1}{k-1} b^k \\
&= \frac{n}{s} \left(\frac{b}{1+b} \right) \sum_{k=0}^{s-1} \binom{n-1}{k} \left(\frac{b}{1+b} \right)^k \left(\frac{1}{1+b} \right)^{n-1-k} \\
&= \frac{rn}{s} \sum_{k=0}^{s-1} \binom{n-1}{k} r^k (1-r)^{n-1-k} \\
&= \frac{rn}{s} \mathbb{P}(\mathrm{Bin}(n-1,r) \le s-1) \\
&= \frac{rn}{s} \mathbb{P} \left(\frac{\mathrm{Bin}(n-1,r) - (n-1)r}{\sqrt{nr(1-r)}} \le \frac{s-1-(n-1)r}{\sqrt{nr(1-r)}} \right) \\
&\to \Phi \left(\frac{\beta - \gamma\sqrt{r}}{\sqrt{1-r}} \right),
\end{aligned}$$

since $nr/s = 1 + O(1/\sqrt{R_1})$. Also,

$$\begin{aligned}
(1+b)^{-n} C(s,n) &= \sum_{k=0}^{s} \binom{n}{k} \left(\frac{b}{1+b} \right)^k \left(\frac{1}{1+b} \right)^{n-k} \\
&= \sum_{k=0}^{s} \binom{n}{k} r^k (1-r)^{n-k} \\
&= \mathbb{P}(\mathrm{Bin}(n,r) \le s) \to \Phi \left(\frac{\beta - \gamma\sqrt{r}}{\sqrt{1-r}} \right).
\end{aligned}$$

Therefore, we have $(1+b)^{-n}[C(s,n) - A(s,n)] \to 0$ as $\lambda \to \infty$. For the remaining term,

$$
\begin{aligned}
(1+b)^{-n}B(s,n) &= (1+b)^{-n}\sum_{k=s+1}^{n}\binom{n}{k}\frac{k!}{s!}s^{s-k}b^{k}\\
&= (1+b)^{-n}\frac{n!}{s!}\,s^{s}\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{-k}\\
&= (1+b)^{-n}\frac{n!}{s!}\,s^{s}\left(\frac{b}{s}\right)^{n}\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{n-k}\\
&= r^{n}\,\frac{n!}{s!}s^{s-n}\sum_{m=0}^{n-s-1}\frac{1}{m!}\left(\frac{s}{b}\right)^{m}\\
&= \left(\frac{r}{s}\right)^{n}\frac{n!}{s!}s^{s}\,\mathrm{e}^{s/b}\,\mathbb{P}(\mathrm{Pois}(s/b)\le n-s-1),
\end{aligned}
$$

in which

$$
\begin{aligned}
\mathbb{P}(\mathrm{Pois}(s/b)\le n-s-1) &= \mathbb{P}\left(\frac{\mathrm{Pois}(s/b)-s/b}{\sqrt{s/b}}\le\frac{n-s-1-s/b}{\sqrt{s/b}}\right)\\
&\to \Phi\left(\frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}\right),
\end{aligned}
$$

as $\lambda\to\infty$. By Stirling's approximation,

$$
\begin{aligned}
\left(\frac{r}{s}\right)^{n}\frac{n!}{s!}s^{s}\,\mathrm{e}^{s/b} &\sim \left(\frac{r}{s}\right)^{n}\sqrt{\frac{n}{s}}\,\frac{n^{n}\mathrm{e}^{-n}}{s^{s}\mathrm{e}^{-s}}\,s^{s}\,\mathrm{e}^{s/b}\\
&= \left(\frac{rn}{s}\right)^{n}\sqrt{\frac{n}{s}}\mathrm{e}^{-n+s+s/b} = \left(\frac{rn}{s}\right)^{n}\sqrt{\frac{n}{s}}\mathrm{e}^{-n+s/r}.
\end{aligned}
$$

Since,

$$
\frac{rn}{s} = 1+\frac{\gamma\sqrt{r}-\beta}{\sqrt{R_1}}+O(1/R_1),
$$

we find $\sqrt{n/s}=1/\sqrt{r}+O(1/\sqrt{R_1})$ and

$$
\begin{aligned}
\log\left[\left(\frac{rn}{s}\right)^{n}\sqrt{\frac{n}{s}}\mathrm{e}^{-n+\frac{s}{r}}\right] &= n\log\left[\frac{rn}{s}\right]-n+\frac{s}{r}\\
&= -n\left[\left(1-\frac{rn}{s}\right)+\frac{1}{2}\left(1-\frac{rn}{s}\right)^{2}+O(R^{-\frac{3}{2}})\right]+\frac{s}{r}\left(1-\frac{rn}{s}\right)\\
&= \frac{s}{r}\left(1-\frac{rn}{s}\right)^{2}-\frac{n}{2}\left(1-\frac{rn}{s}\right)^{2}+O(1/\sqrt{R_1})\\
&= \frac{(\gamma\sqrt{r}-\beta)^{2}}{2r}+O(1/\sqrt{R_1}),
\end{aligned}
$$

as $\lambda\to\infty$ and hence,

$$
(1+b)^{-n}B(s,n)\to\varphi\left(\frac{\gamma\sqrt{r}-\beta}{\sqrt{r}}\right)\Phi\left(\frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}\right).
$$

Hence, we conclude that the denominator of (5.43) converges to a constant value as R_1 grows, and hence the $1-\rho_{\max}(s,n)\to 0$ as $\lambda\to\infty$.

6

Transient error approximation in a Lévy queue

Motivated by a capacity allocation problem within a finite planning period, we conduct a transient analysis of a single-server queue with Lévy input. From a cost minimization perspective, we investigate the error induced by using stationary congestion measures as opposed to time-dependent measures. Invoking recent results from fluctuation theory of Lévy processes, we derive a refined cost function, that accounts for transient effects. This leads to a corrected capacity allocation rule for the transient single-server queue. Extensive numerical experiments indicate that the cost reductions achieved by this correction can be significant.

Based on
Transient error approximation in a Lévy queue
Britt Mathijsen & Bert Zwart
Queueing Systems, 85(3), 269-304 (2017)

6.1 Introduction

The issue of matching a service system's capacity to stochastic demand induced by its clients arises in many practical settings. Typically, the resources available to satisfy demand are scarce and hence expensive. This forces the manager to consider a trade-off between the system efficiency and the quality of service perceived by its clients. In this chapter, we focus on this trade-off in the context of the $M/G/1$ queue, in which the variable amenable for optimization is the server speed μ.

In general, optimizing the server speed μ in a single-server queue in a time-homogeneous environment, while trading off congestion levels against capacity allocation costs, does not pose any technical challenges. Typically, the objective function to be minimized, the total cost function, has the shape

$$\Pi_\infty(\mu) = \mathbb{E}[Q_\mu(\infty)] + \alpha\mu = \frac{\lambda\mathbb{E}[B^2]}{2(\mu - \lambda\mathbb{E}[B])} + \alpha\mu, \tag{6.1}$$

where $\mathbb{E}[Q_\mu(\infty)]$ denotes the expected steady-state amount of work given server speed μ, and B describes the service requirement per arrival. The parameter $\alpha > 0$ represents the relative capacity allocation costs incurred by deploying service rate μ. This one-dimensional optimization problem yields the optimizer

$$\mu_\infty^\star = \lambda\mathbb{E}[B] + \sqrt{\frac{\lambda\mathbb{E}[B^2]}{2\alpha}}.$$

Despite the simplicity and tractability of the problem described above, the presence of the *steady-state* measure in the cost function in (6.1) should be handled carefully. By employing this particular cost structure, one automatically agrees with the underlying assumption of the system being sufficiently close to its steady state. However, referring to practical applications of the single-server model, system parameters rarely remain constant over time. Moreover, planning periods for the optimization problem are naturally finite. Hence, the *true* expected costs incurred, which we denote by $\Pi_T(\mu)$, in addition depend on the length of the planning period T. Consequently, the usage of steady-state models for decision making needs to be justified by a more elaborate time-dependent or *transient* analysis for these type of settings.

Related literature. The time-dependent behavior of the single-server queue received much attention in queueing theory. First efforts to analyze the time-dependent properties of the $M/G/1$ queue date back to the 1950s and 1960s, e.g. [34, 83, 134, 203, 204]. The analyses in these papers mostly yield implicit expressions for performance characteristics through Laplace transforms, integro-differential equations and infinite convolutions. More specifically, there is vast literature on the transient analysis of the $M/M/1$ queue, with the goal to derive explicit expressions for queue length characteristics, see e.g. [5, 61, 177, 178]. These works provide a variety of explicit expressions for the transient dynamics, although the complexity of the resulting expressions, typically involving Bessel functions, exposes the intricate intractability of the matter. Consequently, approximation methods for insightful quantification of the dynamics based on numerical [168] or asymptotic methods, have become prevalent in more recent literature. The asymptotic methods either exploit knowledge on the evolution of the queueing process as time t grows large [5, 172, 173], or as the arrival rate λ is increased to infinity [3, 4, 84]. It is noteworthy that a substantial contribution to the transient literature is made by Abate and Whitt [3, 4, 5, 7], who exploit the existence of a decomposition of the mean transient queue length and obtain expressions for the moments of the queue length and virtual waiting through probabilistic arguments in several queueing models. More recently, asymptotic methods have been used to justify the application of stationary performance measures in Markovian environments or to refine them, see e.g. [91, 219]. Other approximative methods known as uniform acceleration expansions [162] have been developed to reveal the asymptotic behavior of the single-server queue as a function of t, which are moreover able to capture time-varying arrival rates. The majority of the works mentioned above do reflect on the error imposed by usage of steady-state performance metrics instead of the correct time-dependent counterpart. However, no light has been shed on the accumulation of this error over a finite period of time. To the best of our knowledge, the only work that addresses this issue is the paper by Steckley and Henderson [199], who compute an approximation for the error accumulated between the steady-state and transient delay probability. Our analysis on the other hand is centered around the mean workload, which requires a different approach. In addition, the focus in [199] is on performance measures only, while the main goal of our work is to investigate the quality of staffing rules.

Lévy input. Although the $M/G/1$ queue serves as the leading example in our analysis, we choose to use a more general framework for the arrival process of the queue. Namely, we let the server face a Lévy process. This gives the advantage that once we have obtained the results, we can apply them to broader queue input classes, such as Brownian motion and the Gamma process. To shed light on the influence of the transience of the queueing process on traditional staffing questions, we will study the capacity allocation problem in the context of cost minimization in which the objective function is $\Pi_T(\mu)$, i.e. a function of both μ and T. We investigate how the invalidity of the stationary assumption is echoed through the operational cost accounting for congestion-related penalties. Furthermore, we establish a result on the strict convexity of the function $\Pi_T(\mu)$, for almost all values of T (with a few minor exceptions for certain deterministic initial states), which is an essential property for convergence of both cost function and corresponding minimizer to their stationary counterparts.

Corrected staffing rule. As it will appear that an exact analysis of this disparity is intractable, we will present an explicit approximate correction to the conventional stationary objective function given by $\Psi(\mu)/T$ and prove that

$$\Pi_T(\mu) = \Pi_\infty(\mu) + \frac{\Psi(\mu)}{T} + O(1/T^2),$$

with the help of recent results from the fluctuation theory of Lévy processes. Based on this refinement we ultimately examine how incorporating transient effects changes the optimal capacity level and propose a refinement to the steady-state capacity allocation rule,

$$\mu_T^\star = \mu_\infty^\star + \frac{\mu_\bullet}{T} + o(1/T).$$

We moreover deduce an explicit expression for $\mu_\bullet$ in terms of the initial state and the first three moments of the service requirement per arrival. It is noteworthy that similar refined square-root staffing rules have been proposed for multi-server queues in the Halfin-Whitt regime, see e.g. [118, 117, 120, 183, 230]. In those cases, the relevant decision value is the number of servers and refinements are derived for $\lambda \to \infty$, whereas we consider the regime $T \to \infty$.

Building upon the insights gained through the analysis of this optimality gap, we reflect on the parameter settings of the underlying queueing process in which our refined capacity sizing rule yields significant improvement and in which cases it has little effect. Special emphasis is put on the relationship between the accuracy of the standard procedure and the length of the planning period.

Structure of the chapter. The remainder of this chapter is structured as follows. Section 6.2 is devoted to the model description and presents some preliminary results. The main result will be given in Section 6.3 and results regarding the optimization problem will be discussed in Section 6.4, followed by the validation of our novel techniques through numerical experiments in Section 6.5. We will give some concluding remarks and topics for further research in Section 6.6. We have deferred all proofs to the appendix.

6.2 Model description

6.2.1 A queueing model with Lévy input

The model that inspired our study is the standard $M/G/1$ queue starting out of equilibrium. Customers arrive to the queue according to a Poisson process with rate λ. All arrivals have i.i.d. service requirement B_i, stemming from a common random variable B. Without loss of generality we will assume $\mathbb{E}[B] = 1$ throughout. The server is able to remove μ amounts of work from the system per time unit; a variable we will refer to as the *server speed*. E.g. if $\mu = 3$ and two customers are in the system with remaining service times 4 and 2, then the queue will be empty 2 time units later, provided that no new arrivals occur in the meantime. Let $N_\lambda(t)$ denote the number of arrivals until time t. Accordingly, the total work generated by the customers is given by

$$Z_\lambda(t) = \sum_{i=1}^{N_\lambda(t)} B_i.$$

Furthermore, define $X_{\lambda,\mu}(t) := Z_\lambda(t) - \mu t$. We call $X_{\lambda,\mu}$ the *net-input process*. More generally, we assume throughout the chapter that $X_{\lambda,\mu}$ is a Lévy process. Specifically, we let Z_λ be of the form $Z_\lambda(t) = U(\lambda t)$, where U is a spectrally positive Lévy process generated by the triplet (a, σ, ν) and $\mathbb{E}[U(1)] = 1$. This restriction to spectrally positive processes is equivalent to stating $\nu(-\infty, 0) = 0$ and is a vital assumption to our analysis. Subsequently, we assume the net-input process $X_{\lambda,\mu}$ to be

$$X_{\lambda,\mu}(t) = U(\lambda t) - \mu t, \qquad t \geq 0. \tag{6.2}$$

Note that by setting $a = \sigma = 0$ and $\nu = \lambda\, F_B$, where F_B is the cumulative distribution function of B, we retrieve the original $M/G/1$ queue. The stochastic process central to our analysis is the *workload process* $Q_{\lambda,\mu}(t)$, $t \geq 0$, which describes the amount of work the server is facing at time t. The net-input process $X_{\lambda,\mu}$ completely determines the trajectory of $Q_{\lambda,\mu}$, namely

$$Q_{\lambda,\mu}(t) = \max\left\{ Q(0) + X_{\lambda,\mu}(t), \sup_{s\in[0,t]} [X_{\lambda,\mu}(t) - X_{\lambda,\mu}(s)] \right\}, \qquad t \geq 0, \tag{6.3}$$

where $Q(0)$ is the initial workload in the system. In fact, $Q_{\lambda,\mu}$ is the reflected version of $X_{\lambda,\mu}$ with reflection barrier at zero. Careful inspection of the structure also reveals that $X_{\lambda,\mu}(t) \equiv X_{\lambda/\mu,1}(\mu t) \equiv X_{1,\mu/\lambda}(\lambda t)$, so that

$$Q_{\lambda,\mu}(t) \stackrel{d}{=} Q_{\lambda/\mu,1}(\mu t) \stackrel{d}{=} Q_{1,\mu/\lambda}(\lambda t) \tag{6.4}$$

for all $\lambda, \mu, t > 0$. This identity will prove to be convenient for the numerical analysis in Section 6.5. For reasons of clarity, we omit the subscript λ in our expressions if no ambiguity is possible.

The process Q_μ is a natural indicator of the level of congestion in the system and therefore a good choice for quantifying the Quality of Service (QoS) received by a client. We remark that alternative processes characterizing congestion in the system can be deduced directly from $Q_\mu(t)$. For example, consider the virtual waiting time process $V_\mu(t)$, which is the waiting time a customer would experience if he arrives at time t. This, under the first-come-first-served policy, satisfies the relation $\mathbb{E}[V_\mu(t)] = \mathbb{E}[Q_\mu(t)]/\mu$ for all $t \geq 0$. Likewise, the expected number of the customers in the system $L_\mu(t)$ at time $t \geq 0$ is given by Little's law

$$\mathbb{E}[L_\mu(t)] = \lambda\, \mathbb{E}[V_\mu(t)] = \frac{\lambda}{\mu}\, \mathbb{E}[Q_\mu(t)].$$

To facilitate our investigation of the queueing model, we end this subsection by introducing some notation regarding the net-input and workload process and by stating a useful preliminary result concerning the stationary process $Q_\mu(\infty)$. Throughout the chapter we assume $\mu > \lambda$ to ensure ergodicity of the queue and convergence in distribution to the limit

$$Q_\mu(\infty) := \lim_{t\to\infty} Q_\mu(t),$$

for any initial state $Q(0) < \infty$. The distribution of $Q_\mu(\infty)$ coincides with the stationary distribution of $Q_\mu(t)$. By $\kappa_U(\cdot)$ and $\kappa_\mu(\cdot)$ we denote the Lévy exponents of the processes U and X_μ, respectively:

$$\kappa_\mu(\theta) = \log \mathbb{E}[\mathrm{e}^{\theta X_\mu(1)}] = \log \mathbb{E}[\mathrm{e}^{\theta(U(\lambda)-\mu)}] = \lambda\kappa_U(\theta) - \mu\theta.$$

Furthermore, define $u_k = \mathbb{E}[\{U(1) - \mathbb{E}U(1)\}^k]$ for $k = 2, 3, ...$ Using this representation we obtain the following preliminary result.

Lemma 6.1. *Let $\mathbb{E}|U(1)| < \infty$, $u_2, u_3 < \infty$ and $\mu > \lambda$. If $Q_\mu(\infty)$ represents the steady-state distribution of the workload process, then*

$$\mathbb{E}[Q_\mu(\infty)] = \frac{\lambda u_2}{2(\mu-\lambda)}, \qquad \mathbb{E}[Q_\mu^2(\infty)] = \frac{\lambda^2 u_2^2}{2(\mu-\lambda)^2} + \frac{\lambda u_3}{3(\mu-\lambda)}.$$

The proof of Lemma 6.1 follows directly by differentiation of the Laplace transform of $Q_\mu(\infty)$ and is given in Appendix 6.A.1.

6.2.2 Finite horizon

For the purpose of our research, we are interested in the dynamics of the workload process within a fixed time frame of length $T > 0$. For all $0 \le t \le T$, we assume that the parameters of the queue, λ, μ, u_2, u_3, remain unchanged. If at $t = 0$ the queue is not in steady-state corresponding to the specified parameters of the starting period, the process $\{Q_\mu(t)\}_{t\in[0,T]}$ differs from its stationary counterpart $Q_\mu(\infty)$. To illustrate this, Figure 6.1 depicts the expected value Q_μ in a $M/M/1$ queue as a function of time for several initial workloads $Q(0)$ for a particular setting of λ and μ. Clearly, transient behavior of $\mathbb{E}[Q_\mu(t)]$, for $Q(0) \neq Q_\mu(\infty)$, differs significantly from the steady-state mean with the same system parameters. Note that even if $Q(0) \equiv \mathbb{E}[Q_\mu(\infty)]$, the time-dependent mean does not coincide with the steady-state mean. Moreover, $\mathbb{E}[Q_\mu(t)]$ is not even a strictly increasing nor decreasing function of time. This phenomenon is a consequence of the decomposition of the transient mean into one strictly increasing, and a strictly decreasing term for $Q(0) > 0$, as discussed in [5]. Nonetheless, $Q_\mu(t)$ converges in distribution to $Q_\mu(\infty)$ as $t \to \infty$, if $\mu > \lambda$.

Since the time horizon of our analysis is limited to $t \le T$, the process may not approach the steady-state distribution sufficiently close to appropriately use its steady-state properties for capacity allocation. To overcome this disparity, we propose a way to include the influence of this transient phase in the capacity allocation problem.

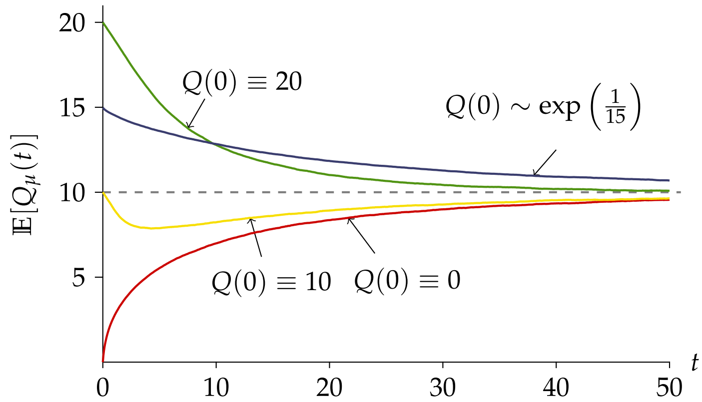


Figure 6.1: Time-dependent mean workload in a $M/M/1$ queue with $\lambda = 10$ and $\mu = 11$ for different initial states $Q(0)$. The dashed line depicts $\mathbb{E}Q_\mu(\infty)$.

6.2.3 Cost structure

As mentioned before, we are interested in balancing the QoS and efficiency of the queue by choosing the optimal server speed μ. The adjective *optimal* indicates that we intend to choose the speed according to some objective function. In our case, we conduct our analysis based on a cost function, which consists of a part accounting for the penalty for congestion in the system and a part for staffing cost. The cost value of both parts is governed by the variable μ. The instantaneous cost incurred at time t equals

$$\mathbb{E}[Q_\mu(t)] + \alpha\mu,$$

where α is a positive constant defining the *relative staffing cost*. Hence, the cost structure we apply is a combination of the transient mean of the workload process and a linear staffing cost. Accumulated and normalized over the period $[0, T]$, the cost function on which the rest of this chapter will be based equals

$$\Pi_T(\mu) := \frac{1}{T}\int_0^T \left(\mathbb{E}[Q_\mu(t)] + \alpha\mu\right) \mathrm{d}t = \frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)]\mathrm{d}t + \alpha\mu. \tag{6.5}$$

We use shorthand notation for the normalized congestion costs:

$$C_T(\mu) := \frac{1}{T}\int_0^T \mathbb{E}[Q_\mu(t)]\mathrm{d}t, \tag{6.6}$$

and $C_\infty(\mu) := \mathbb{E}[Q_\mu(\infty)]$. In order to compare the actual costs incurred over the interval $[0, T]$ to the cost function of the queue in stationary conditions, we define

$$\Pi_\infty(\mu) := C_\infty(\mu) + \alpha\mu = \mathbb{E}[Q_\mu(\infty)] + \alpha\mu, \tag{6.7}$$

which allows an explicit expression by Lemma 6.1. Under mild conditions on the net-input process and the distribution of the initial state, the cost functions coincide for $T \to \infty$.

Proposition 6.1. *Let $\mu > \lambda$ and assume $\mathbb{E}[U(1)]$, $\mathbb{E}[Q(0)] < \infty$. Then*

$$\lim_{T\to\infty} \Pi_T(\mu) = \Pi_\infty(\mu).$$

The proof of Proposition 6.1 can be found in Appendix 6.A.2. Define

$$\Omega_T := \frac{1}{T}\int_0^T \left(\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right) \mathrm{d}t$$

We can then rewriting (6.5) as

$$\Pi_T(\mu) = \frac{1}{T}\int_0^T \left(\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right) \mathrm{d}t + \mathbb{E}[Q_\mu(\infty)] + \alpha\mu = \Omega_T(\mu) + \Pi_\infty(\mu). \tag{6.8}$$

Section 6.3 is concerned with the analysis of the correction factor $\Omega_T(\mu)$.

Ultimately, we are concerned with the additional costs incurred by choosing the server speed through minimization of $\Pi_\infty(\mu)$ instead of $\Pi_T(\mu)$. Therefore, we formulate the exact and approximate optimization problems as follows

$$\mu_T^\star := \arg\min_{\mu\geq 0} \Pi_T(\mu), \qquad \mu_\infty^\star := \arg\min_{\mu\geq 0} \Pi_\infty(\mu), \tag{6.9}$$

$$\Pi_T^\star := \Pi_T(\mu_T^\star), \qquad \Pi_\infty^\star := \Pi_T(\mu_\infty^\star). \tag{6.10}$$

In Section 6.4 we turn to the comparison of $\mu_T^\star$ and $\mu_\infty^\star$ as well as the *optimality gap* $\Pi_\infty^\star - \Pi_T^\star$.

6.3 Analysis of the objective function

From (6.8) it is evident that, for finding an explicit characterization of $\Pi_T(\mu)$, it suffices to study the term $\Omega_T(\mu)$ in more detail. We start by stating the main result of this section, which describes the leading order behavior of $\Omega_T(\mu)$ as T increases.

Theorem 6.1. *Let $X_\mu(t)$ be of the form* (6.2). *If $\mathbb{E}[\max(Q(0), Q_\mu(\infty))^3] < \infty$ and $u_2, u_3 < \infty$, then*

$$\begin{aligned}\Omega_T(\mu) &= \frac{\mathbb{E}[Q(0)^2] - \mathbb{E}[Q_\mu(\infty)^2]}{2T(\mu-\lambda)} + O\left(\frac{1}{T^2}\right)\\ &= \frac{1}{2T(\mu-\lambda)}\left(\mathbb{E}[Q(0)^2] - \frac{\lambda^2 u_2^2}{2(\mu-\lambda)^2} - \frac{\lambda u_3}{3(\mu-\lambda)}\right) + O\left(\frac{1}{T^2}\right),\end{aligned}$$

for $\mu > \lambda$.

Note that this expression provides an *approximation* of the actual cost function $\Pi_T(\mu)$. We elaborate on the implications of this additional information on the optimization problem in Section 6.4.

In the remainder of this section we provide a detailed description of the steps taken to obtain this outcome. We assume a fixed service rate μ throughout the analysis in this section and therefore omit the subscript μ. Proofs of the intermediate results can be found in Appendix 6.B.

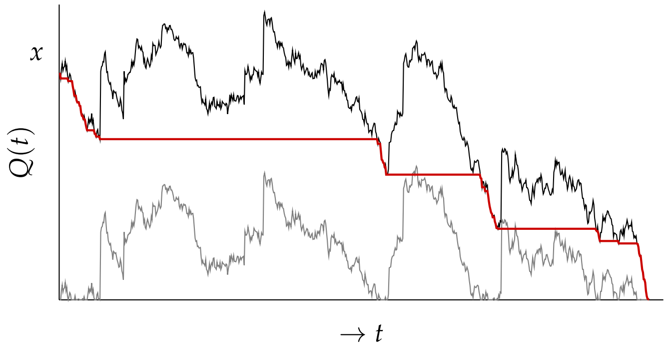


Figure 6.2: Sample path visualization of the processes $Q^x(t)$ (solid), $Q^0(t)$ (gray) and $Y^{x,0}(t)$ (red).

6.3.1 Constructing a coupling

Before starting our analysis of the correction term $\Omega_T(\mu)$ we introduce some auxiliary notation. By $Q^A(t)$ we denote the workload process as described in Subsection 6.2.1 with initial state A and $\mathbb{E}_A$ the expectation with respect to any non-negative random variable A, which is independent of the net-input process X. To be able to compare $\mathbb{E}[Q(t)]$ and $\mathbb{E}[Q(\infty)]$ as in $\Omega_T(\mu)$, we will use a coupling technique. Observe that by definition of the stationary distribution $Q(\infty) \stackrel{d}{=} Q^{Q(\infty)}(t)$ for all $t \geq 0$ and therefore $\mathbb{E}[Q(\infty)] = \mathbb{E}_{Q(\infty)}[Q^{Q(\infty)}(t)]$. Furthermore, $\mathbb{E}[Q(t)] = \mathbb{E}_{Q(0)}[Q^{Q(0)}(t)]$. Hence, quantifying the difference between the transient and stationary mean is equivalent to comparing the workload processes of two queues starting in two different (random) states at $t = 0$.

We begin our analysis for two queues starting in two *deterministic* states $x, y \geq 0$, respectively. At the end of our analysis we will obtain the original form by replacing x with $Q(0)$ and y with $Q(\infty)$.

Equation (6.3) shows that all randomness in the workload process originates from the process $X(t)$. With this in mind, we couple the processes $Q^x(t)$ and $Q^y(t)$ on a sample path level by feeding both queues the same net-input process $X(t)$ for $t \geq 0$. This allows us to compare the processes in the same probability space, so that $\mathbb{E}[Q^x(t)] - \mathbb{E}[Q^y(t)] = \mathbb{E}[Q^x(t) - Q^y(t)]$ for all $t \geq 0$. Define

$$Y^{x,y}(t) := Q^x(t) - Q^y(t)$$

and

$$\Omega_T^{x,y} := \frac{1}{T} \int_0^T \mathbb{E}\left[Y^{x,y}(t)\right] \mathrm{d}t.$$

A possible sample path triple for $Q^x(t)$, $Q^0(t)$ and $Y^{x,0}(t)$ is depicted in Figure 6.2. As we see from this figure, $Y^{x,0}(t)$ has nice structural properties which we will exploit in the next subsection.

6.3.2 Difference process and leading order behavior of the correction term

We further examine the *difference process* $Y^{x,y}(t)$ with $x > y$. Recall from (6.3),

$$Q^z(t) = \max\{z + X(t), \sup_{0<s\leq t}[X(t) - X(s)]\} = X(t) + \max\{z, -\inf_{0\leq s\leq t} X(s)\}, \quad (6.11)$$

for any initial state $z \geq 0$, where $X(t)$ is a Lévy process with no negative jumps. Let $\tau^z(w)$, $0 \leq w < z$ denote the first passage time of level w by the process starting in z, i.e.

$$\tau^z(w) := \inf\{t \geq 0 \,|\, Q^z(t) \leq w\}.$$

Then it is easily seen that for all $z \geq 0$,

$$Q^z(t) = \begin{cases} z + X(t), & \text{if } t < \tau^z(0), \\ \sup_{0<s\leq t}[X(t) - X(s)], & \text{if } t \geq \tau^z(0). \end{cases}$$

Consequently,

$$Y^{x,y}(t) = \begin{cases} x - y, & \text{if } t < \tau^y(0), \\ \inf_{0<s\leq t}\{x + X(s)\}, & \text{if } \tau^y(0) \leq t < \tau^x(0), \\ 0, & \text{if } t \geq \tau^x(0). \end{cases} \quad (6.12)$$

Using this representation we can identify

$$\Omega_T^{x,y} = \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^x(0)\wedge T} Y^{x,y}(t)\mathrm{d}t\right],$$

where $\wedge$ denotes the minimum operator, due to the fact $Y^{x,y}(t) = 0$ for $t \geq \tau^x(0)$. Subsequently, we decompose $\Omega_T^{x,y}$ into two terms

$$\Psi_T^{x,y} := \frac{1}{T}\int_0^\infty \mathbb{E}[Y^{x,y}(t)]\,\mathrm{d}t \qquad \text{and} \qquad \Delta_T^{x,y} := \Omega_T^{x,y} - \Psi_T^{x,y}. \quad (6.13)$$

Note that $\Psi_T^{x,y}$ is obtained by replacing T by ∞ only in the integration bound. It is customary in the literature, particularly in the area of stochastic simulation, to compare the truncated integral to its natural expansion of the integration range to a semi-infinite interval, see e.g. [27, Prop. 2.1].The truncated integral connects to the long-run average estimator of a certain performance metric, whereas the infinite integral reflects its exact expectation. The decomposition in (6.13) is insightful, because $\Psi_T^{x,y}$ prescribes the leading order behavior of $\Omega_T^{x,y}$, while $\Delta_T^{x,y}$ captures the smaller order error term. In this section, we only consider $\Psi_T^{x,y}$. Subsection 6.3.3 investigates the magnitude of $\Delta_T^{x,y}$. The next preliminary result presents a useful property of $\Psi_T^{x,y}$.

Lemma 6.2. *Let $x > y$. If $\mathbb{E}[\tau^x(0)] < \infty$, then*

$$\Psi_T^{x,y} = \frac{1}{T}\,\mathbb{E}[\tau^y(0)](x - y) + \Psi_T^{x-y,0}. \quad (6.14)$$

The proof can be found in Appendix 6.B.1. This leaves us with two unknowns $\mathbb{E}[\tau^y(0)]$ and $\Psi_T^{x-y,0}$. The next lemma gives an equivalent form for the latter.

Lemma 6.3. *If* $\mathbb{E}[\tau^z(0)] < \infty$, *then for all* $z \geq 0$

$$\Psi_T^{z,0} = \int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w. \tag{6.15}$$

The proof can be found in Appendix 6.B.2. Since the term $\mathbb{E}[\tau^z(0)]$, for several values of z, appears in many of the preliminary results, we devote our attention to this in the next subsection.

First passage time. When studying the first passage time of level $0 \leq w < z$, $\tau^z(w)$, of the workload process starting in z, we first observe that $\{\tau^z(z-w)\}_{w=0}^z$ is a spectrally positive Lévy process itself, also visible through Figure 6.2. More precisely, it is a subordinator, i.e. a Lévy process whose paths are almost surely non-decreasing [147]. In order to calculate $\mathbb{E}[\tau^z(z-w)]$ we use theory presented in [190, Section 46], although results presented there are valid for spectrally *negative* Lévy processes, as opposed to the absence of negative jumps in our case. Nonetheless, our setting is easily transformed into this framework by observing that $\hat{X} \equiv -X$, that is $\hat{X}(t) = -X(t)$ for all $t \geq 0$, is spectrally negative. Furthermore, let

$$\hat{\tau}^0(w) := \inf\{t \geq 0 \,:\, \hat{X}(t) \geq w\} = \inf\{t \geq 0 \,:\, z + X(t) \leq z - w\} = \tau^z(z-w). \tag{6.16}$$

For completeness, we cite [190, Thm. 46.3].

Theorem 6.2. *Let* $\hat{X}(t)$ *be a spectrally negative Lévy process with generating triplet* $(-a, \sigma, \hat{\nu})$ *and* $\hat{\tau}^0(y)$ *its corresponding hitting time process. Define* $\Upsilon(\theta)$ *for* $\theta \geq 0$ *as*

$$\Upsilon(\theta) = -a\theta + \tfrac{1}{2}\sigma^2\theta^2 + \int_{-\infty}^0 (\mathrm{e}^{\theta x} - 1 - \theta x \mathbf{1}_{[-1,0)}(x))\,\hat{\nu}(\mathrm{d}x). \tag{6.17}$$

Then $\Upsilon(\theta)$ *is strictly increasing and continuous,* $\Upsilon(0) = 0$, *and* $\Upsilon(\theta) \to \infty$ *as* $\theta \to \infty$. *For* $w \geq 0$ *and* $0 \leq u < \infty$ *we have*

$$\mathbb{E}[\exp(-u\hat{\tau}^0(w))] = \exp(-w\,\Upsilon^{-1}(u)), \tag{6.18}$$

where $\theta = \Upsilon^{-1}(u)$ *is the inverse function of* $u = \Upsilon(\theta)$.

This immediately induces an expression for $\mathbb{E}[\tau^w(0)]$ and henceforth $\Psi^{z,0}$.

Corollary 6.1. *Let* $X(t)$ *be a spectrally positive Lévy process defined as in* (6.2) *with* $\mu > \lambda$. *Let* $\Psi_T^{z,0}$ *as in* (6.15). *Then*

$$\Psi_T^{z,0} = \frac{z^2}{2T(\mu-\lambda)}.$$

Furthermore, if $x, y \geq 0$, *then*

$$\Psi_T^{x,y} = \frac{x^2 - y^2}{2T(\mu-\lambda)}. \tag{6.19}$$

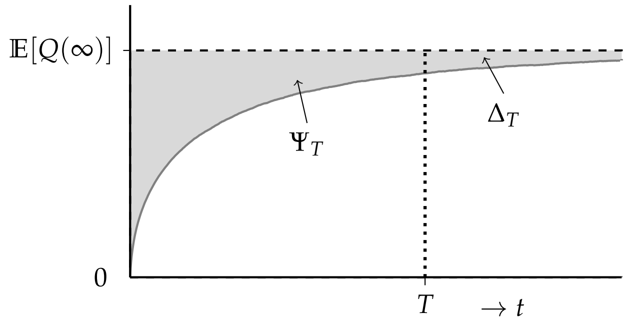


Figure 6.3: Visualization of Ω_T and Ψ_T as the area between the curves $\mathbb{E}[Q(t)]$, $\mathbb{E}[Q(\infty)]$ for $Q(0) = 0$.

The proof of Corollary 6.1 can be found in Appendix 6.B.3. **Randomization**. As we stated before, we easily obtain the original Ω_T from $\Omega_T^{x,y}$ through substitution of x and y by $Q(0)$ and $Q(\infty)$, respectively, and taking the expectation. In the previous paragraph, we deduced an explicit expression for $\Psi_T^{x,y}$, the leading order term for $\Omega_T^{x,y}$. Therefore we equivalently get an approximation for Ω_T, given by

$$\Psi_T := \frac{1}{T}\int_0^\infty \left(\mathbb{E}[Q(t)] - \mathbb{E}[Q(\infty)]\right)\,\mathrm{d}t,$$

through randomization of x and y in $\Psi_T^{x,y}$. By combining the results in Corollary 6.1, Lemma 6.1 and Proposition 6.2, which is given at the end of this section, we directly prove the result in Theorem 6.1.

6.3.3 Truncation error

In order to get a better comprehension of the properties of Ψ_T, we depict the value in terms of the (infinite) region between the curves $\mathbb{E}[Q(t)]$, $\mathbb{E}[Q(\infty)]$ and the vertical axis for the case $Q(0) \equiv 0$ in Figure 6.3. In this figure, Ω_T is given by the area enclosed by the two curves, the vertical axis and the line $t = T$. One can see that the main contribution to the correction term Ω_T is given for small t. As t increases, the difference between transient and stationary mean decreases. Hence for moderate values of T, the contribution to the integral in (6.13) is only minor compared to the contribution over the interval $[0, T]$.

Recall the definition of $\Delta_T^{x,y}$ as in (6.13). As we alluded to in Subsection 6.3.2 we claim the contribution of $\Delta_T^{x,y}$ to $\Omega_T^{x,y}$ is negligible compared to $\Psi_T^{x,y}$. Also note that

$$\Delta_T := \Omega_T - \Psi_T = -\frac{1}{T}\int_T^\infty \big(\mathbb{E}[Q(t)] - \mathbb{E}[Q(\infty)]\big)\,\mathrm{d}t. \tag{6.20}$$

can be derived through $\Delta_T^{x,y}$ in a similar manner as we did for $\Psi_T^{x,y}$ to obtain Ψ_T. To substantiate our claim, we compute an upper bound for $\Delta_T^{x,y}$ of order $1/T^2$.

The existence of such an upper bound poses a limit on the error this tail integral contributed to the cost structure as a whole.

Proposition 6.2. *Let $x, y \geq 0$ and $\mathbb{E}[\max(Q(0), Q_\mu(\infty))^3] < \infty$. Then*

$$|\Delta_T^{x,y}| \leq \frac{1}{T^2}\left(\frac{\max(y,x)^3}{3(\mu-\lambda)^2} + \frac{u_2 \max(y,x)^2}{2(\mu-\lambda)^3}\right)$$

and

$$|\Delta_T| \leq \frac{1}{T^2}\left(\frac{\mathbb{E}[\max(Q(0), Q_\mu(\infty))^3]}{3(\mu-\lambda)^2} + \frac{u_2 \mathbb{E}[\max(Q(0), Q_\mu(\infty))^2]}{2(\mu-\lambda)^3}\right).$$

The proof of Proposition 6.2 is given in Appendix 6.B.4.

Remark 6.1. In case the net-input process X is light-tailed, that is there exists $u > 0$ such that $\mathbb{E}[\mathrm{e}^{uX(1)}] < \infty$, it can be shown that the truncation error is of order $\mathrm{e}^{-\beta T}/T$ for some $\beta > 0$. See Appendix 6.B.4 for details.

6.4 Optimization

The result in Theorem 6.1, characterizing the leading order behavior of $\Omega_T(\mu)$, also reveals the behavior of $\Pi_T(\mu)$ in leading order. Namely,

$$\Pi_T(\mu) = \Pi_\infty(\mu) + \Psi_T(\mu) + O(1/T^2).$$

In fact, this representation naturally gives rise to an *approximation* of the actual cost function:

$$\hat{\Pi}_T(\mu) := \Pi_\infty(\mu) + \Psi_T(\mu) \tag{6.21}$$

Denote the corresponding minimizer of $\hat{\Pi}_T$ by

$$\hat{\mu}_T^\star := \arg\min_{\mu \geq 0} \hat{\Pi}_T(\mu), \qquad \hat{\Pi}_T^\star := \hat{\Pi}_T(\hat{\mu}_T^\star) \tag{6.22}$$

in addition to the definitions in (6.9) and (6.10). This section is devoted to the analysis of the minimizers $\mu_T^\star$, $\hat{\mu}_T^\star$ and $\mu_\infty^\star$, and the optimality gap for the two approximations.

Throughout this section, we assume that $u_2, u_3 < \infty$ and $\mathbb{E}[Q(0)^2] < \infty$.

By its definition in (6.7) and Lemma 6.1, we have an exact expression for the steady-state cost function:

$$\Pi_\infty(\mu) = \frac{\lambda u_2}{2(\mu-\lambda)} + \alpha\mu.$$

It is easily verified that Π_∞ is strictly convex in μ, for instance by observing that $\Pi_\infty''(\mu) > 0$ for all $\mu > \lambda$. Therefore Π_∞ has a unique global minimizer and

$$\mu_\infty^\star = \lambda + \sqrt{\frac{\lambda u_2}{2\alpha}}, \qquad \Pi_\infty^\star = \alpha\lambda + \sqrt{2\alpha\lambda u_2}. \tag{6.23}$$

We are interested in the relation between $\mu^\star_\infty$ and $\mu^\star_T$, and between $\hat{\mu}^\star_T$ and $\mu^\star_T$. Since $\Pi_T(\mu) = \Pi_\infty(\mu) + O(1/T)$ for all $\mu > \lambda$, we have pointwise convergence of the sequence Π_T, as well as $\hat{\Pi}_T$, to Π_∞ for $T \to \infty$, we also expect $\mu^\star_T \to \mu^\star_\infty$ and $\hat{\mu}^\star_T \to \mu^\star_\infty$ for $T \to \infty$. Before proving that this convergence indeed holds, we present a result on the strict convexity of the function Π_T.

Lemma 6.4. *Let $\mu \geq 0$. The function $\Pi_T(\mu)$ is*

- *convex in μ, if $Q(0) \equiv x$, $T < x/\mu$ and $\sigma = 0$,*
- *strictly convex in μ, otherwise.*

Building upon strict convexity of both $\Pi_T(\mu)$ and $\Pi_\infty(\mu)$ for $\mu > \lambda$, we derive the following convergence result.

Proposition 6.3. *Let $\mu^\star_T$, $\hat{\mu}^\star_T$ and $\mu^\star_\infty$ be as defined in* (6.9) *and* (6.22). *Then*

$$\mu^\star_T \to \mu^\star_\infty \qquad \text{and} \qquad \hat{\mu}^\star_T \to \mu^\star_\infty,$$

for $T \to \infty$.

The next result describes a refinement of $\mu^\star_T$ in terms of $\mu^\star_\infty$.

Proposition 6.4. *For T sufficiently large,*

$$\mu^\star_T = \mu^\star_\infty + \frac{\mu_\bullet}{T} + o(1/T),$$

where

$$\mu_\bullet = \frac{\mathbb{E}[Q(0)^2]}{\sqrt{8\lambda u_2 \alpha}} - \frac{u_3}{3u_2} - 3\sqrt{\frac{\alpha\lambda u_2}{8}}. \tag{6.24}$$

The proofs of the three results above can be found in Appendix 6.C. Based on Proposition 6.4 we propose a *corrected staffing rule*, accounting for the finite horizon

$$\tilde{\mu}^\star_T = \left[\mu^\star_\infty + \frac{\mu_\bullet}{T}\right]^+, \tag{6.25}$$

with $\mu_\bullet$ as in (6.24). Here $[x]^+ := \max\{x, 0\}$, which ensures the value of $\tilde{\mu}^\star_T$ is non-negative and thus is a feasible solution of the optimization problem. This refined capacity allocation rule is expected to reduce the costs incurred in transient settings. However, the value of particular interest to us is the cost penalty for using either one of the approximations rather than the actual minimum $\mu^\star_T$, that is, the *optimality gap*. As it happens, we deduce the order of the optimality gap for $\mu^\star_\infty$ with the help of the explicit form of $\mu_\bullet$ given in (6.24), which is stated in the next proposition. The proof is given in Appendix 6.C.4.

Proposition 6.5. *Let $\mu^\star_\infty$ be as in* (6.23). *Then,*

$$\Pi^\star_\infty - \Pi^\star_T = O(1/T^2).$$

6.5 Numerical experiments

6.5.1 Influence of $\Omega_T(\mu)$

We first assess the contribution of the correction to the cost function provided by Theorem 1. In other words, we investigate whether $\hat{\Pi}_T(\mu)$ as in (6.5) yields a significantly better fit to $\Pi_T(\mu)$, than $\Pi_\infty(\mu)$ does. Note that these three functions only differ in the costs describing the congestion. Therefore, we limit our study in this subsection to the evaluation of $C_T(\mu)$ as in (6.6) with stationary equivalent $C_\infty(\mu) = \mathbb{E}[Q_\mu(\infty)]$. Our novel approximation hence reads

$$\hat{C}_T(\mu) := C_\infty(\mu) + \Omega_T(\mu),$$

with $\Omega_T(\mu)$ given in Theorem 6.1. We conduct our numerical experiments based on three models, namely:

1. $M/M/1$ queue: $U(t)$ is a unit rate compound Poisson process with exponentially distributed increments. We have $u_2 = 2$, $u_3 = 3$, so that

$$\hat{C}_T(\mu) = \frac{\lambda}{\mu-\lambda} + \frac{1}{T(\mu-\lambda)}\left(\frac{x^2}{2} - \frac{\lambda^2}{(\mu-\lambda)^2} - \frac{\lambda}{\mu-\lambda}\right). \tag{6.26}$$

2. $M/$Pareto$/1$ queue: $U(t)$ is a unit rate compound Poisson process with Pareto increments. The Pareto distribution deserves special attention due to its heavy-tailed nature, having tail probability $\bar{F}(x) = (x/k)^{-\gamma}$, if $x \geq k$ and 1 otherwise. It is well-known that heavy-tailed service times lead to long relaxation time. For our purposes, we fix shape parameter $\gamma = 16/5$ and scale parameter $k = 11/16$, so that $\beta = 1$, $u_2 = 121/96$, $u_3 = 1331/256$ and $u_k = \infty$ for all $k > 3$. Hence,

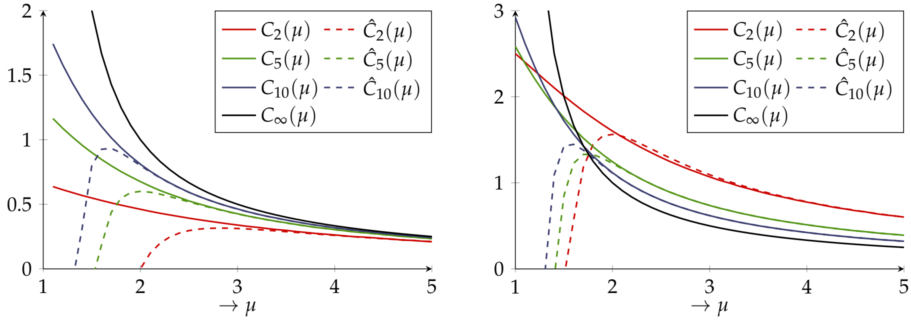
$$\hat{C}_T(\mu) = \frac{121\lambda}{192(\mu-\lambda)} + \frac{1}{2T(\mu-\lambda)}\left(x^2 - \frac{(121\lambda/96)^2}{2(\mu-\lambda)^2} - \frac{1331\lambda/256}{2(\mu-\lambda)}\right) \tag{6.27}$$

3. Reflected Brownian motion: $U(t)$ is Brownian motion with drift 1 and infinitesimal variance σ^2. We have $u_2 = \sigma^2$, $u_3 = 0$, so that

$$\hat{C}_T(\mu) = \frac{\lambda\sigma^2}{2(\mu-\lambda)} + \frac{1}{2T(\mu-\lambda)}\left(x^2 - \frac{\lambda^2\sigma^4}{2(\mu-\lambda)^2}\right). \tag{6.28}$$

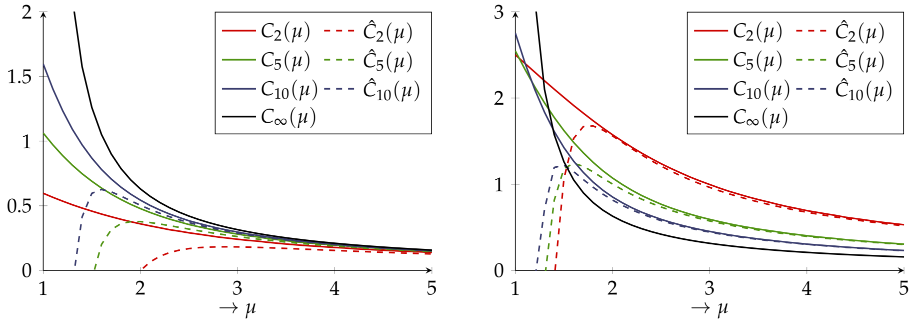
In light of the equivalence relations in (6.4) we only consider the case $\lambda = 1$. The cost values for general values of λ follow by appropriate rescaling of μ and T.

For the $M/M/1$ and $M/$Pareto$/1$ queue, we obtained the function $C_T(\mu)$ through simulation and the results are accurate up until a 95% confidence interval of width 10^{-3}. For reflected Brownian motion, we used the explicit distribution function given in [104] for double numerical integration. The results for several values of



(a) $x = 0$ (b) $x = 2.5$

Figure 6.4: Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for the $M/M/1$ queue with $\lambda = 1$.



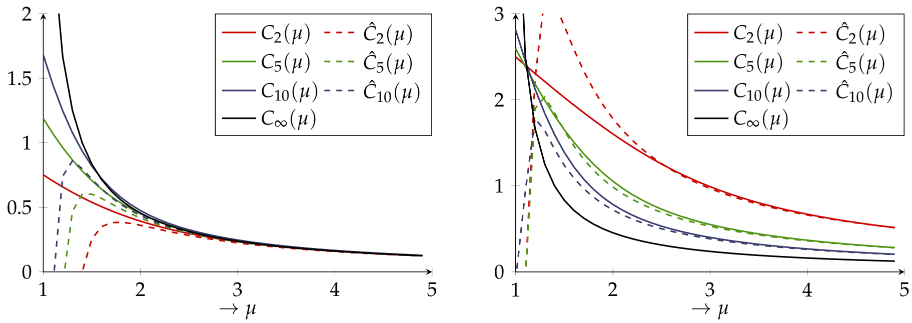
(a) $x = 0$ (b) $x = 2.5$

Figure 6.5: Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for the $M/\text{Pareto}/1$ queue with $\lambda = 1$.

T and two different starting states are depicted in Figures 4-6. These plots also include the approximated functions $\hat{C}_T(\mu)$.

We name a few observations based on these figures. First, we indeed note the pointwise convergence of $\hat{C}_T(\mu)$ to $\hat{C}_\infty(\mu)$ as T grows, for all μ in all three cases. However, the difference between the stationary costs and those for small values of T can be significant. This is most clear in the plots with $x = 2.5$ and when μ is close to λ, i.e. it is in heavy-traffic. In these scenarios, it is evident that refinements to the stationary cost function are needed. $\hat{C}_T(\mu)$ does a fairly good job at providing such correction, especially for moderate values of μ.

Furthermore, we note that $C_T(\mu)$ approaches $C_\infty(\mu)$ from below for $x = 0$ for any value of μ, while this is not strictly the case for $x > 0$. $\hat{C}_T(\mu)$ correctly captures



(a) $x = 0$ (b) $x = 2.5$

Figure 6.6: Comparison of exact waiting cost function $C_T(\mu)$ against corrected cost function $\hat{C}_T(\mu)$ and PSA cost function $C_\infty(\mu)$ for $T = 2, 5$ and 10 for reflected Brownian motion with $\sigma = 1$.

the sign of this correction.

Finally, observe that $\hat{C}_T(\mu) \to -\infty$ as μ approaches λ from above. This divergence is clear from the expressions in (6.26)-(6.28). Our correction term relies on the premise that under the coupling scheme, the sample paths of the two queues starting from different states have hit with high probability. This is equivalent to stating that the 'largest' of the two queues has emptied at least once before time T. However, as μ approaches λ, the system enters heavy traffic, and hence the hitting time of the zero barrier is set to run off to infinity. Consequently, this causes our approximation to be inaccurate for small values of μ.

6.5.2 Validation of corrected staffing rule

In this section, we examine whether the corrected staffing rule $\tilde{\mu}_T^\star$ as in (6.25) indeed yields a significant cost reduction over the choice of $\mu_\infty^\star$ by comparing their true costs $\Pi_T(\tilde{\mu}_T^\star)$ and $\Pi_T(\mu_\infty^\star)$. We conduct this comparison for different values of the parameters, α, T and starting state x through numerical experiments. The three models on which we do our calculations are the $M/M/1$ queue, the $M/\text{Pareto}/1$ queue and the reflected Brownian motion, as introduced in the previous subsection. We again focus on $\lambda = 1$ only.

For each of the three models, we adhere to the following set-up. The quality of both staffing rules is assessed for $\alpha = 0.1, 1$ and 2, resembling three modes of valuation of the QoS in the system. As a benchmark, observe that the expected workload in steady-state conditions with staffing level $\mu_\infty^\star$ equals

$$C_\infty(\mu_\infty^\star) = \sqrt{\frac{\alpha\lambda u_2}{2}}.$$

For each value of α, we consider two scenarios: one in which the system starts empty, i.e. $x = 0$, and one in which the initial state is double this benchmark value, thus $x = \sqrt{2\alpha\lambda u_2}$. The numerics are presented for each model separately. We discuss general conclusions drawn from these results afterwards.

$M/M/1$ **queue.** As we discussed before, if U is a unit rate compound Poisson process with exponentially distributed increments, then Q_μ describes the workload process in an $M/M/1$ queue. For this setting we get

$$\mu_\infty^\star = \lambda + \sqrt{\frac{\lambda}{\alpha}}, \qquad \tilde{\mu}_T^\star = \left[\lambda + \sqrt{\frac{\lambda}{\alpha}} + \frac{1}{T}\left(\frac{x^2}{4\sqrt{\lambda\alpha}} - 1 - \frac{3}{2}\sqrt{\lambda\alpha}\right)\right]^+ .$$

Table 6.1 presents the actual costs corresponding to these two staffing levels for different value of x and α.

		$x=0$					$x=2\sqrt{\alpha}$				
α	T	$\mu_\infty^\star$	$\Pi_T(\mu_\infty^\star)$	$\tilde{\mu}_T^\star$	$\Pi_T(\tilde{\mu}_T^\star)$	r.c.i.	$\mu_\infty^\star$	$\Pi_T(\mu_\infty^\star)$	$\tilde{\mu}_T^\star$	$\Pi_T(\tilde{\mu}_T^\star)$	r.c.i.
0.1	1	4.162	0.620	2.688	0.536	0.136	4.162	0.682	2.688	0.536	0.214
	2	4.162	0.669	3.425	0.641	0.041	4.162	0.700	3.425	0.641	0.085
	5	4.162	0.706	3.867	0.703	0.005	4.162	0.719	3.867	0.703	0.022
	10	4.162	0.719	4.015	0.719	0.001	4.162	0.726	4.015	0.719	0.010
1	1	2.000	2.309	0.000	0.500	0.783	2.000	3.500	0.500	2.750	0.214
	2	2.000	2.461	0.750	1.480	0.398	2.000	3.218	1.250	3.125	0.029
	5	2.000	2.675	1.500	2.400	0.103	2.000	3.043	1.700	2.968	0.025
	10	2.000	2.810	1.750	2.726	0.030	2.000	3.007	1.850	2.980	0.009
2	1	1.707	3.744	0.000	0.500	0.866	1.707	5.889	0.000	3.328	0.435
	2	1.707	3.924	0.146	1.232	0.686	1.707	5.547	0.854	4.682	0.156
	5	1.707	4.209	1.083	3.343	0.206	1.707	5.114	1.366	4.910	0.040
	10	1.707	4.424	1.395	4.108	0.071	1.707	4.945	1.536	4.868	0.016

Table 6.1: Comparison of costs for the $M/M/1$ queue for steady-state and corrected staffing rules and relative cost improvement (r.c.i.).

M**/Pareto/1 queue.** In case the service requirements follow a Pareto distribution with shape parameter $\gamma = 16/5$, the staffing rule becomes

$$\mu_\infty^\star = \lambda + \frac{11}{8}\sqrt{\frac{\lambda}{3\alpha}}, \; \tilde{\mu}_T^\star = \left[\lambda + \frac{11}{8}\sqrt{\frac{\lambda}{3\alpha}} + \frac{1}{T}\left(\frac{2x^2}{11\sqrt{\lambda\alpha/3}} - \frac{11}{8} - \frac{11\sqrt{3\lambda\alpha}}{16}\right)\right]^+ .$$

The numerical results are given in Table 6.2. Just as in the results for the $M/M/1$ queue, we observe a higher reduction for larger value of α and T. Also, again $\tilde{\mu}_T < \mu_\infty^\star$. Hence, the conclusions for the M/Pareto/1 queue are similar to those of the $M/M/1$ queue.

Reflected Brownian motion. In case the input process U is Brownian motion with drift 1 and infinitesimal variance σ^2, the steady-state staffing rule and its corrected version reduce to

		$x=0$					$x = 11/4 \cdot \sqrt{\alpha/3}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	3.510	0.524	1.759	0.461	0.120	3.510	0.573	2.010	0.562	0.019
	2	3.510	0.555	2.635	0.539	0.029	3.510	0.580	2.760	0.574	0.010
	5	3.510	0.580	3.160	0.578	0.003	3.510	0.591	3.210	0.589	0.002
	10	3.510	0.590	3.335	0.590	0.000	3.510	0.596	3.360	0.595	0.001
1	1	1.794	2.076	0.000	0.500	0.759	1.794	2.989	0.000	2.088	0.302
	2	1.794	2.190	0.511	1.291	0.411	1.794	2.790	0.610	2.588	0.072
	5	1.794	2.345	1.281	2.108	0.101	1.794	2.638	1.320	2.607	0.012
	10	1.794	2.441	1.537	2.371	0.029	1.794	2.597	1.557	2.585	0.005
2	1	1.561	3.427	0.000	0.500	0.854	1.561	5.087	0.000	2.745	0.460
	2	1.561	3.567	0.032	1.050	0.706	1.561	4.832	0.172	3.417	0.293
	5	1.561	3.779	0.950	3.012	0.203	1.561	4.499	1.006	4.313	0.041
	10	1.561	3.935	1.255	3.356	0.147	1.561	4.351	1.284	4.304	0.011

Table 6.2: Comparison of costs for the $M/\text{Pareto}/1$ queue for steady-state and corrected staffing rules and relative cost improvement (r.c.i.).

$$\mu^\star_\infty = \lambda + \sqrt{\frac{\lambda\sigma^2}{2\alpha}}, \qquad \tilde{\mu}^\star_T = \left[\lambda + \sqrt{\frac{\lambda\sigma^2}{2\alpha}} + \frac{1}{2\sqrt{2}\,T}\left(\frac{x^2}{\sqrt{\lambda\alpha}\sigma} - 3\sigma\sqrt{\alpha\lambda}\right)\right]^+ .$$

In Tables 6.3 and 6.4, the costs obtained through numerical evaluation are presented for several values of x, T. We also vary σ to examine the influence of the volatility of arrival process on the quality of the staffing rules.

The observations on the influence of α, x and T are similar to those of the $M/M/1$ queue and the $M/\text{Pareto}/1$ queue. However, here we see little improvement by the corrected staffing rule for small values of α for both values of x. The results in Tables 6.3-6.4 also suggest that the reduction is smaller for larger values of σ.

6.5.3 Discussion

Based upon these numerical results in Tables 6.1-6.4, we make a few remarks. The three models roughly exhibit similar behavior as T, x and α are varied.

Non-surprisingly, we note that $\tilde{\mu}_T$ approaches $\mu^\star_\infty$ with increasing T, which also implies that the cost reduction achieved by the corrected staffing rule vanishes as $T \to \infty$. Also, we observe that in all scenarios examined, the cost reduction increases with α. This can be explained through investigation of the objective function Π_T as function of μ. Namely, for α small, the curve is relatively flat around the true optimum $\mu^\star_T$. Hence, in this case a moderate deviation from $\mu^\star_T$ will likely not lead to a significant cost increase. However, as α becomes larger, i.e. server efficiency is valued more than minimization of congestion, the curve becomes more sharp around $\mu^\star_T$, and hence more accurate approximations of $\mu^\star_T$ are required to achieve

		$x = 0$					$x = \sqrt{2\alpha}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	3.236	0.525	2.901	0.518	0.013	3.236	0.565	3.124	0.564	0.001
	2	3.236	0.536	3.068	0.534	0.003	3.236	0.556	3.180	0.556	0.000
	5	3.236	0.543	3.169	0.542	0.000	3.236	0.551	3.214	0.551	0.000
	10	3.236	0.545	3.203	0.545	0.000	3.236	0.549	3.225	0.549	0.000
1	1	1.500	3.420	0.000	0.833	0.756	1.500	4.741	1.000	3.984	0.160
	2	1.500	3.539	0.750	2.386	0.326	1.500	4.579	1.250	4.293	0.063
	5	1.500	3.707	1.200	3.363	0.093	1.500	4.335	1.400	4.274	0.014
	10	1.500	3.820	1.350	3.705	0.030	1.500	4.190	1.450	4.175	0.004
2	1	1.500	3.420	0.000	0.833	0.756	1.500	4.741	1.000	3.984	0.160
	2	1.500	3.539	0.750	2.386	0.326	1.500	4.579	1.250	4.293	0.063
	5	1.500	3.707	1.200	3.363	0.093	1.500	4.335	1.400	4.274	0.014
	10	1.500	3.820	1.350	3.705	0.030	1.500	4.190	1.450	4.175	0.004

Table 6.3: Comparison of costs for RBM with $\sigma = 1$ for steady-state and corrected staffing rules and relative cost improvement (r.c.i.).

		$x = 0$					$x = 2\sqrt{2\alpha}$				
α	T	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.	$\mu^\star_\infty$	$\Pi_T(\mu^\star_\infty)$	$\tilde{\mu}^\star_T$	$\Pi_T(\tilde{\mu}^\star_T)$	r.c.i.
0.1	1	5.472	0.950	4.801	0.936	0.015	5.472	1.030	5.249	1.029	0.001
	2	5.472	0.972	5.137	0.968	0.003	5.472	1.012	5.360	1.012	0.000
	5	5.472	0.985	5.338	0.985	0.000	5.472	1.002	5.427	1.002	0.000
	10	5.472	0.990	5.405	0.990	0.000	5.472	0.998	5.450	0.998	0.000
1	1	2.414	3.176	0.293	1.546	0.513	2.414	4.633	1.707	4.228	0.087
	2	2.414	3.356	1.354	2.690	0.199	2.414	4.375	2.061	4.247	0.029
	5	2.414	3.573	1.990	3.411	0.045	2.414	4.094	2.273	4.073	0.005
	10	2.414	3.689	2.202	3.646	0.012	2.414	3.966	2.344	3.962	0.001
2	1	2.000	4.839	0.000	1.339	0.723	2.000	7.481	1.000	5.967	0.202
	2	2.000	5.078	0.500	2.773	0.454	2.000	7.158	1.500	6.585	0.080
	5	2.000	5.414	1.400	4.726	0.127	2.000	6.670	1.800	6.549	0.018
	10	2.000	5.639	1.700	5.409	0.041	2.000	6.380	1.900	6.349	0.005

Table 6.4: Comparison of costs for RBM with $\sigma = 2$ for steady-state and corrected staffing rules and relative cost improvement (r.c.i.).

an acceptable cost level. Hence, the corrected staffing rule (6.25) proves particularly useful in these cases.

Another point we highlight is that the relative improvement is higher for $x = 0$ than for $x = \sqrt{2\alpha\lambda u_2}$. Moreover, even though the initial state of the system is above the optimal equilibrium, $\tilde{\mu}_T$ is smaller than $\mu_\infty^\star$. This is somewhat counter-intuitive. In fact, from (6.24) it follows that $\mu_\bullet$ positively contributes to the corrected staffing function if

$$\mathbb{E}[Q^2(0)] > 3\alpha\lambda u_2 + \frac{2u_2}{3u_3}\sqrt{2\alpha\lambda u_2}.$$

6.6 Conclusion & further research

Motivated by the time-varying nature of queues in practical applications, we studied the impact that the transient phase has on traditional capacity allocation questions. By defining a cost minimization problem, in which the objective function contains a correction accounting for the transient period, we identified the leading and second-order behavior of the cost function as a function of the interval length T. As a by-product, this result yields an approximation for the actual cost function, which is a refinement to its stationary counterpart. Our numerical experiments in Section 6.5.1 demonstrate the improved accuracy achieved by this approximation in a number of settings. By perturbation analysis of the optimization problem, this furthermore gives rise to a correction to the steady-state optimal capacity allocation of order $1/T$. The necessity of the refined capacity allocation level is substantiated by the numerics in Section 6.5.2, which show the cost reduction that can be achieved in a number of settings, compared to settings in which stationary metrics are used. Especially for small values of T and large values of α this reduction is significant. Additionally, these results also indicate that it is relatively safe to use the stationary cost when T is moderate, or α is small. The latter reflects the scenario in which QoS is much more valued than service efficiency. This observation links to the flat nature of the cost function around its optimal value for α small, a statement on the optimality gap that we formally proved in Proposition 6.5.

Besides the validation of our theoretical results of Sections 6.3 and 6.4, the numerical results also reveal some phenomena that require more investigation. As noted, our corrected capacity allocation level $\tilde{\mu}_T^\star$ is in most studied cases less than the steady-state optimal value $\mu_\infty^\star$. This implies that congestion levels tends to be higher under our staffing scheme then under stationary staffing. A possible explanation for this may be the fact that the planning period under consideration is finite. Clearly, in the setting we analyzed, anything that happens after time T is neglected. Therefore, it might be beneficial from the cost perspective to end the period with a higher expected congestion level, as it does not need to be canceled out in the future. Related to this observation, it would be interesting to look at the setting in which staffing decisions need to be made in consecutive periods of equal length, in which the arrival rate changes at the start of each period. This case requires careful consideration of the correlation among the staffing decisions within the separate periods.

Another question that arises concerns the translation of our (qualitative) findings to more general queues, in particular the $M/G/s$ queue. Whereas in our analysis, the central decision variable is the server speed μ, the variable of interest in multi-server queues is typically the number of servers. It may well be that similar explicit corrections to staffing levels can be deduced to account for transience. Since our analysis heavily relies on the comparibility of the sample paths of two single-server queues, which is due to the equal negative drift for the two processes, another approach must be taken to tackle this extension.

The analysis and findings for the single-server queue with Lévy input presented in this chapter may serve a stepping stone for investigation of these more elaborate problems.

Appendix

6.A Proofs of Section 6.2

6.A.1 Proof of Lemma 6.1

Proof. The conditions of [20, Cor.IX3.4] are satisfied and therefore $Q_\mu(t) \Rightarrow Q_\mu(\infty)$ in distribution for $t \to \infty$. Furthermore, its Laplace transform is for $\mathrm{Re}(s) < 0$

$$\tilde{Q}_\mu(s) = \mathbb{E}\left[\mathrm{e}^{sQ_\mu(\infty)}\right] = \frac{s\kappa'_\mu(0)}{\kappa_\mu(s)} = \frac{s(\lambda\kappa'_U(0) - \mu)}{\lambda\kappa_U(s) - \mu s} = \frac{s(\mu - \lambda)}{\mu s - \lambda\kappa_U(s)}.$$

It can be checked that $\kappa'_U(0) = \mathbb{E}[U(1)] = 1$, $\kappa''_U(0) = u_2$ and $\kappa'''_U(0) = u_3$, and $\kappa'_\mu(0) = \lambda - \mu$, $\kappa''_\mu(0) = \lambda u_2$ and $\kappa'''_\mu(0) = \lambda u_3$. Using l'Hôpital's rule we obtain the first moment of $Q_\mu(\infty)$:

$$\begin{aligned}
\mathbb{E}[Q_\mu(\infty)] &= \lim_{s\to 0} \frac{\mathrm{d}}{\mathrm{d}s}\tilde{Q}_\mu(s) = \lim_{s\to 0} \kappa'_\mu(0)\,\frac{\kappa_\mu(s) - s\kappa'_\mu(s)}{\kappa_\mu(s)^2} \\
&= \kappa'_\mu(0)\,\lim_{s\to 0}\frac{-s\kappa''_\mu(s)}{2\kappa_\mu(s)\kappa'_\mu(s)} = \kappa'_\mu(0)\,\lim_{s\to 0}\frac{-s\kappa'''_\mu(s) - \kappa''_\mu(s)}{2\kappa'_\mu(s)^2 + 2\kappa_\mu(s)\kappa''_\mu(s)} \\
&= -\frac{\kappa''_\mu(0)}{2\kappa'_\mu(0)} = \frac{\lambda u_2}{2(\mu - \lambda)}.
\end{aligned}$$

Similarly, we derive the second moment:

$$\mathbb{E}[Q^2_\mu(\infty)] = \lim_{s\to 0}\frac{\mathrm{d}^2}{\mathrm{d}s^2}\tilde{Q}_\mu(s) = \lim_{s\to 0}\kappa'_\mu(0)\,\frac{2s\kappa'_\mu(s)^2 - 2\kappa'_\mu(s)\kappa_\mu(s) - s\kappa''_\mu(s)\kappa_\mu(s)}{\kappa_\mu(s)^3},$$

We apply l'Hôpital's rule twice, to find

$$
\begin{aligned}
\mathbb{E}[Q_\mu^2(\infty)] &= \kappa_\mu'(0) \lim_{s\to 0} \frac{3s\kappa_\mu''(s)\kappa_\mu'(s) - 3\kappa_\mu''(s)\kappa_\mu(s) - s\kappa_\mu'''(s)\kappa_\mu(s)}{3\kappa_\mu'(s)\kappa_\mu(s)^2} \\
&= \kappa_\mu'(0) \lim_{s\to 0} \frac{2s\kappa_\mu'''(s)\kappa_\mu'(s) + 3s\kappa_\mu''(s)^2 - 4\kappa_\mu'''(s)\kappa_\mu(s) - s\kappa_\mu^{(4)}(s)\kappa_\mu(s)}{6\kappa_\mu'(s)^2\kappa_\mu(s) + 3\kappa_\mu''(s)\kappa_\mu(s)^2} \\
&= \kappa_\mu'(0) \lim_{s\to 0} \frac{s\left[2\kappa_\mu'''(s)\kappa_\mu'(s) + 3\kappa_\mu''(s)^2 - \kappa_\mu^{(4)}(s)\kappa_\mu(s)\right] - 4\kappa_\mu'''(s)\kappa_\mu(s)}{\kappa_\mu(s)\left[6\kappa_\mu'(s)^2 + 3\kappa_\mu''(s)\kappa_\mu(s)\right]} \\
&= \kappa_\mu'(0) \lim_{s\to 0} \frac{s}{\kappa_\mu(s)} \frac{2\kappa_\mu'''(s)\kappa_\mu'(s) + 3\kappa_\mu''(s)^2 - \kappa_\mu^{(4)}(s)\kappa_\mu(s)}{6\kappa_\mu'(s)^2 + 3\kappa_\mu''(s)\kappa_\mu(s)} \\
&\qquad - \kappa_\mu'(0) \lim_{s\to 0} \frac{4\kappa_\mu'''(s)}{6\kappa_\mu'(s)^2 + 3\kappa_\mu''(s)\kappa_\mu(s)}.
\end{aligned}
$$

Since $\kappa_\mu(0) = 0$ and $\lim_{s\to 0} s/\kappa_\mu(s) = 1/\kappa_\mu'(0)$, we have

$$
\begin{aligned}
\kappa_\mu'(0) \lim_{s\to 0} \frac{s}{\kappa_\mu(s)} & \frac{2\kappa_\mu'''(s)\kappa_\mu'(s) + 3\kappa_\mu''(s)^2 - \kappa_\mu^{(4)}(s)\kappa_\mu(s)}{6\kappa_\mu'(s)^2 + 3\kappa_\mu''(s)\kappa_\mu(s)} \\
&= \frac{2\kappa_\mu'''(0)\kappa_\mu'(0) + 3\kappa_\mu''(0)^2}{6\kappa_\mu'(0)^2} = \frac{\kappa_\mu'''(0)}{3\kappa_\mu'(0)} + \frac{\kappa_\mu''(0)^2}{2\kappa_\mu'(0)^2}
\end{aligned} \tag{6.29}
$$

and

$$
\kappa_\mu'(0) \lim_{s\to 0} \frac{4\kappa_\mu'''(s)}{6\kappa_\mu'(s)^2 + 3\kappa_\mu''(s)\kappa_\mu(s)} = \frac{2\kappa_\mu'''(0)}{3\kappa_\mu'(0)}. \tag{6.30}
$$

Combining (6.29) and (6.30) yields

$$
\mathbb{E}[Q_\mu^2(\infty)] = \frac{\kappa_\mu''(0)^2}{2\kappa_\mu'(0)^2} - \frac{\kappa_\mu'''(0)}{3\kappa_\mu'(0)} = \frac{\lambda^2 u_2^2}{2(\mu-\lambda)^2} + \frac{\lambda u_3}{3(\mu-\lambda)}.
$$

□

6.A.2 Proof of Proposition 6.1

Proof. We prove the limit by showing that the difference

$$
\Pi_T(\mu) - \Pi_\infty(\mu) = \frac{1}{T}\int_0^T \Big(\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\Big)\, \mathrm{d}t
$$

converges to zero as $T \to \infty$ for $\mu > \lambda$ fixed. The assumption $\mathbb{E}[U(1)], \mathbb{E}[Q(0)] < \infty$ implies by [7, Prop. 1] that $\mathbb{E}[Q_\mu(t)] < \infty$ for all $t \geq 0$. Following [7], we use the decomposition

$$
\mathbb{E}[Q_\mu(t)] = \mathbb{E}[Q_\mu^0(t)] + \Big\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu^0(t)]\Big\},
$$

where $Q^0_\mu(t)$ represents the workload process if the system starts empty. From this decomposition it is revealed that $\mathbb{E}[Q^0_\mu(t)]$ and $\left\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q^0_\mu(t)]\right\}$ are non-negative monotonically increasing and decreasing functions of t, respectively, see [7, Prop. 2,Thm. 11]. Recall $\mathbb{E}[Q_\mu(t)] \to \mathbb{E}[Q_\mu(\infty)]$ for $t \to \infty$ by ergodicity of the workload process for any initial state $\mathbb{E}[Q(0)] < \infty$, if $\mu > \lambda$. Henceforth,

$$\begin{aligned}\mathbb{E}[Q_\mu(t)] &\le \sup_t \mathbb{E}[Q^0_\mu(t)] + \sup_t \left\{\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q^0_\mu(t)]\right\} \\ &= \mathbb{E}[Q_\mu(\infty)] + \left\{\mathbb{E}[Q_\mu(0)] - \mathbb{E}[Q^0_\mu(0)]\right\} = \mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)],\end{aligned}$$

for all $t \ge 0$, which proves that the expected workload is bounded. Fix $\varepsilon > 0$. By convergence of $\mathbb{E}[Q_\mu(t)]$ for $t \to \infty$, there exists a value $t^* := t^*(\varepsilon)$ such that for all $t \ge t^*$

$$\left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right| < \varepsilon/2. \tag{6.31}$$

Next, set

$$T^* := T^*(\varepsilon) = \frac{2\,t^*(\varepsilon)}{\varepsilon}\,(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]).$$

Then for $T \ge \hat{T} := \max\{t^*, T^*\}$, we have

$$\begin{aligned}\left|\frac{1}{T}\int_0^T \left(\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right)\mathrm{d}t\right| &\le \frac{1}{T}\int_0^{t^*} \left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right|\mathrm{d}t \\ &\quad + \frac{1}{T}\int_{t^*}^T \left|\mathbb{E}[Q_\mu(t)] - \mathbb{E}[Q_\mu(\infty)]\right|\,\mathrm{d}t \\ &\le \frac{1}{T}\int_0^{t^*} \left(\mathbb{E}[Q_\mu(t)] + \mathbb{E}[Q_\mu(\infty)]\right)\mathrm{d}t + \frac{1}{T}\int_{t^*}^T \frac{\varepsilon}{2}\,\mathrm{d}t \\ &< \frac{t^*}{T}\,(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]) + \frac{T-t^*}{T}\,\frac{\varepsilon}{2} \\ &< \frac{t^*}{T^*}\,(2\mathbb{E}[Q_\mu(\infty)] + \mathbb{E}[Q(0)]) + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Hence, for any choice of $\varepsilon > 0$ we can find a value $\hat{T}$ such that $\Pi_{\hat{T}}(\mu)$ approaches $\Pi_\infty(\mu)$ within distance ε, which proves the limit. □

6.B Proofs of Section 6.3

6.B.1 Proof of Lemma 6.2

Proof. Using the representation in (6.12) we write

$$\begin{aligned}
\Psi_T^{x,y} &= \frac{1}{T}\int_0^\infty \mathbb{E}[Y^{x,y}(t)]\mathrm{d}t \\
&= \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^y(0)} Y^{x,y}(t)\right]\mathrm{d}t + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right] + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\infty} Y^{x,y}(t)\,\mathrm{d}t\right] \\
&= \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^y(0)} (x-y)\mathrm{d}t\right] + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right] \\
&= \frac{1}{T}\,\mathbb{E}[\tau^y(0)](x-y) + \frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\mathrm{d}t\right].
\end{aligned}$$

By (6.12) and the Strong Markov property holding for Lévy processes [20], observe that
$Y^{x-y,0}(t) \stackrel{d}{=} Y^{x,y}(\tau^y(0)+t)$, whereby

$$\frac{1}{T}\,\mathbb{E}\left[\int_{\tau^y(0)}^{\tau^x(0)} Y^{x,y}(t)\,\mathrm{d}t\right] = \frac{1}{T}\,\mathbb{E}\left[\int_0^{\tau^{x-y}(0)} Y^{x-y,0}(t)\mathrm{d}t\right] = \Psi_T^{x-y,0},$$

which completes the proof. □

6.B.2 Proof of Lemma 6.3

Proof. Note that $Y^{z,0}(t)$ and $\tau^z(w)$ are intimately related. Namely, due to the fact that X has no negative jumps

$$\{\tau^z(w) \le t\} = \{Y^{z,0}(t) \le w\}.$$

In fact, $Y^{z,0}(\tau^z(w)) = w$, which implies that τ^z is a right inverse for $Y^{z,0}(t)$. Therefore, the following equality holds

$$\int_0^{\tau^z(0)} Y^{z,0}(t)\,\mathrm{d}t = \int_0^z \tau^z(w)\,\mathrm{d}w,$$

which implies with the help of Fubini's theorem

$$\Psi_T^{z,0} = \frac{1}{T}\int_0^z \mathbb{E}[\tau^z(w)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\tau^{z-w}(0)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w.$$

□

6.B.3 Proof of Corollary 6.1

Proof. From (6.18),

$$\mathbb{E}[\hat{\tau}^0(w)] = -\tfrac{\mathrm{d}}{\mathrm{d}u}\,\mathbb{E}[\exp(-u\,\hat{\tau}^0(w))]\Big|_{u=0} = w\,\frac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0}. \tag{6.32}$$

Since $\Upsilon(\theta)$ is strictly increasing and $\Upsilon(0)=0$, we get $\Upsilon^{-1}(0)=0$ and

$$\tfrac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0} = \frac{1}{\Upsilon'(\Upsilon^{-1}(0))} = \{\Upsilon'(0)\}^{-1}.$$

Furthermore,

$$\begin{aligned}\Upsilon'(\theta) &= -a+\sigma^2\theta+\int_{-\infty}^{0}(x\,\mathrm{e}^{\theta x}-x\mathbf{1}_{[-1,0)}(x))\hat{\nu}(\mathrm{d}x)\\ &= -a+\sigma^2\theta-\int_{0}^{\infty}(y\,\mathrm{e}^{-\theta y}-y1_{(0,1]}(y))\nu(\mathrm{d}y).\end{aligned}$$

Thus, $\Upsilon'(0)=-\mathbb{E}[X(1)]=\mu-\lambda$ and $\mathbb{E}[\hat{\tau}^0(w)]=w/(\mu-\lambda)$. By (6.15) and (6.16), we deduce that

$$\Psi_T^{z,0} = \frac{1}{T}\int_0^z \mathbb{E}[\tau^w(0)]\,\mathrm{d}w = \frac{1}{T}\int_0^z \mathbb{E}[\hat{\tau}^0(w)]\mathrm{d}w = \frac{z^2}{2T(\mu-\lambda)}.$$

For $x>y$, we use Lemma 6.2 to conclude

$$\Psi_T^{x,y} = \frac{y(x-y)}{T(\mu-\lambda)}+\frac{(x-y)^2}{2T(\mu-\lambda)} = \frac{x^2-y^2}{2T(\mu-\lambda)}.$$

The result for $x<y$ follows directly by the observation $\Psi_T^{y,x}=-\Psi_T^{x,y}$. □

6.B.4 Proof of Proposition 6.2

Proof. To derive the upper bound for $\Delta_T^{x,y}$, we apply the same coupling argument as described in Section 6.3. Let us assume without loss of generality $x>y$. In this case,

$$|\Delta_T^{x,y}| = \frac{1}{T}\int_T^\infty \mathbb{E}[Q^x(t)-Q^y(t)]\mathrm{d}t \le \frac{1}{T}\int_T^\infty \mathbb{E}[Q^x(t)-Q^0(t)]\mathrm{d}t.$$

By the decomposition in (6.12),

$$\begin{aligned}\int_T^\infty \mathbb{E}[Q^x(t)-Q^0(t)]\mathrm{d}t &= \int_T^\infty \mathbb{E}[(x+\inf_{s\le t}X(s))\mathbb{1}_{\{\tau^x(0)>t\}}]\mathrm{d}t\\ &= \int_T^\infty\int_0^x P(x-u+\inf_{s\le t}X(s)>0)\mathrm{d}u\mathrm{d}t\\ &= \int_T^\infty\int_0^x P(\tau^{x-u}(0)>t)\mathrm{d}u\mathrm{d}t \qquad (6.33)\\ &\le \int_T^\infty\int_0^x \frac{\mathbb{E}[\tau^{x-u}(0)^2]}{t^2}\mathrm{d}u\mathrm{d}t\\ &= \int_0^x\int_T^\infty \frac{\mathbb{E}[\tau^{x-u}(0)^2]}{t^2}\mathrm{d}t\mathrm{d}u = \int_0^x \frac{\mathbb{E}[\tau^w(0)^2]}{T}\,\mathrm{d}w.\end{aligned}$$

We obtain $\mathbb{E}[\tau^w(0)^2]$ with the help of its Laplace transform in (6.18). Namely,

$$\begin{aligned}\mathbb{E}[\tau^w(0)^2] &= \tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\mathbb{E}[\exp(-u\tau^w(0))]\Big|_{u=0} \\ &= w^2\left(\tfrac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0}\right)^2 - w\,\tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\Upsilon^{-1}(u)\Big|_{u=0}.\end{aligned}$$

As in the previous subsection we have $\frac{\mathrm{d}}{\mathrm{d}u}\Upsilon^{-1}(u)\Big|_{u=0} = (\mu-\lambda)^{-1}$, and

$$\tfrac{\mathrm{d}^2}{\mathrm{d}u^2}\Upsilon^{-1}(u)\Big|_{u=0} = -\frac{\Upsilon''(\Upsilon^{-1}(0))}{\Upsilon'(\Upsilon^{-1}(0))^3} = -\frac{\Upsilon''(0)}{\Upsilon'(0)^3}.$$

Since $\Upsilon'(0) = \mu - \lambda$ and

$$\Upsilon''(0) = \sigma^2 + \int_0^\infty x^2\,\nu(\mathrm{d}x) = u_2,$$

we conclude

$$\mathbb{E}[\tau^w(0)^2] = \frac{w^2}{(\mu-\lambda)^2} + \frac{u_2 w}{(\mu-\lambda)^3},$$

so that

$$|\Delta_T^{x,y}| \le \frac{1}{T^2}\int_0^x \left(\frac{w^2}{(\mu-\lambda)^2} + \frac{u_2 w}{(\mu-\lambda)^3}\right)\mathrm{d}w = \frac{1}{T^2}\left(\frac{x^3}{3(\mu-\lambda)^2} + \frac{u_2 x^2}{2(\mu-\lambda)^3}\right). \tag{6.34}$$

For general $x, y \ge 0$,

$$|\Delta_T^{x,y}| \le \frac{1}{T^2}\left(\frac{\max(y,x)^3}{3(\mu-\lambda)^2} + \frac{u_2\max(y,x)^2}{2(\mu-\lambda)^3}\right).$$

As a direct consequence,

$$|\Delta_T| \le \frac{1}{T^2}\left(\frac{\mathbb{E}[\max(Q(0),Q_\mu(\infty))^3]}{3(\mu-\lambda)^2} + \frac{u_2\mathbb{E}[\max(Q(0),Q_\mu(\infty))^2]}{2(\mu-\lambda)^3}\right).$$

□

Remark 6.2. Observe that if X is light-tailed, that is $\mathbb{E}[\exp\{-\theta X(1)\}] = \mathbb{E}[\exp\{\kappa(\theta)\}] < \infty$ for some $\theta < 0$, then $\Upsilon(\theta)$ as in (6.18) has an analytic continuation in the negative half-plane, and in this region $\Upsilon(\theta) < 0$. Consequently, we can replace the upper bound on the tail probability of $\tau^{x-u}(0)$ by

$$\mathbb{P}\left(\tau^{x-u}(0) > t\right) = \mathbb{P}\left(\mathrm{e}^{\beta\tau^{x-u}(0)} > \mathrm{e}^{\beta t}\right) \le \mathrm{e}^{-\beta t}\,\mathrm{e}^{(x-u)\Upsilon^{-1}(-\beta)},$$

for some $\beta > 0$, so that

$$\int_T^\infty \mathbb{E}[Q^x(t) - Q^0(t)]\,\mathrm{d}t \le \mathrm{e}^{-\beta T}\,\frac{\mathrm{e}^{x\Upsilon^{-1}(-\beta)} - 1}{\beta\,\Upsilon^{-1}(-\beta)}.$$

Along similar lines we deduce

$$|\Delta_T^{x,y}| \leq \frac{\mathrm{e}^{-\beta T}}{T} \frac{\mathrm{e}^{x\Upsilon^{-1}(-\beta)} + \mathrm{e}^{y\Upsilon^{-1}(-\beta)} - 2}{\beta\, \Upsilon^{-1}(-\beta)}$$

and

$$|\Delta_T| \leq \frac{\mathrm{e}^{-\beta T}}{T} \frac{\mathbb{E}[\mathrm{e}^{Q(0)\Upsilon^{-1}(-\beta)}] + \mathbb{E}[\mathrm{e}^{Q_\mu(\infty)\Upsilon^{-1}(-\beta)}] - 2}{\beta\, \Upsilon^{-1}(-\beta)},$$

assuming that $\mathbb{E}[\mathrm{e}^{-yQ(0)}] < \infty$ for all $y > 0$. The condition $\mathbb{E}[\mathrm{e}^{Q_\mu(\infty)\Upsilon^{-1}(-\beta)}] < \infty$ follows from Lemma 6.1. Hence, the error decays exponentially fast for light-tailed input processes.

6.C Proofs of Section 6.4

6.C.1 Proof of Lemma 6.4

Proof. Since the term $\alpha\mu$ is convex, the strictness should come from the term $C_T(\mu)$. Furthermore, observe that if a function $f_\mu(t)$ is convex for all $t \geq 0$, and strictly convex for all $t \geq \varepsilon$ for some $\varepsilon \in [0, T)$, i.e. for any $\mu_1, \mu_2 > 0$ and $a \in (0, 1)$

$$a\, f_{\mu_1}(t) + (1-a) f_{\mu_2}(t) > f_{a\mu_1 + (1-a)\mu_2}(t),$$

then,

$$a \int_0^T f_{\mu_1}(t)\,\mathrm{d}t + (1-a)\int_0^T f_{\mu_2}(t)\mathrm{d}t = \int_0^T \left(a f_{\mu_1}(t) + (1-a) f_{\mu_2}(t)\right)\mathrm{d}t$$

$$\begin{aligned} &= \int_0^\varepsilon \left(a f_{\mu_1}(t) + (1-a) f_{\mu_2}(t)\right)\mathrm{d}t + \int_\varepsilon^T \left(a f_{\mu_1}(t) + (1-a) f_{\mu_2}(t)\right)\mathrm{d}t \\ &> \int_0^\varepsilon f_{a\mu_1+(1-a)\mu_2}(t)\mathrm{d}t + \int_\varepsilon^T f_{a\mu_1+(1-a)\mu_2}(t)\mathrm{d}t. \\ &= \int_0^T f_{a\mu_1+(1-a)\mu_2}(t)\mathrm{d}t. \end{aligned}$$

Hence, it suffices to prove the convexity of $\mathbb{E}[Q_\mu(t)]$ as a function of μ for all $t \geq 0$, and strict convexity for $t \geq \varepsilon$ for some $\varepsilon \in [0, T)$. Let $\tau_\mu^x(0)$ denote the first passage time of level 0 in the process Q_μ with $Q(0) = x$. Then,

$$Q_\mu(t) = U(t) - \mu t + \max\left\{x, -\inf_{s\leq t}[U(s) - \mu s]\right\} \tag{6.35}$$

$$= \begin{cases} x + U(t) - \mu t, & \text{if } t < \tau_\mu^x(0), \\ U(t) - \mu t - \inf_{s\leq t}[U(s) - \mu s], & \text{if } t \geq \tau_\mu^x(0), \end{cases} \tag{6.36}$$

where

$$\tau_\mu^x(0) := \inf\{t \geq 0 \,:\, x + U(t) - \mu t \leq 0\}$$

and $U(t)$ is a spectrally positive Lévy process. Fix $\mu_1, \mu_2 > 0$ and $a \in (0,1)$. Define $\mu_3 := a\mu_1 + (1-a)\mu_2$, and

$$D(t) := aQ_{\mu_1}(t) + (1-a)Q_{\mu_2}(t) - Q_{\mu_3}(t).$$

In order to prove strict convexity we have to show that $D(t) \geq 0$ for all $t \geq 0$, thereby implying $\mathbb{E}[D(t)] \geq 0$, i.e. convexity, for all $t \geq 0$, and $D(t) > 0$ with positive probability for $t \in [\varepsilon, T]$, for some $\varepsilon \in [0, T)$. We distinguish two cases: $x > 0$ and $x = 0$.

The case $x > 0$**.** We start by noticing that if Q_{μ_1}, Q_{μ_2} and Q_{μ_3} experience the same input process $U(t)$, then by absence of negative jumps in $U(t)$, it holds that

$$\tau^x_{\mu_2}(0) < \tau^x_{\mu_3}(0) < \tau^x_{\mu_1}(0). \tag{6.37}$$

We use shorthand notation

$$I_k(t) := \inf_{0 \leq s \leq t}[U(s) - \mu_k s],$$

for $k = 1, 2, 3$. Using representation (6.36) of the workload process, we obtain

$$D(t) = \begin{cases} 0, & \text{if } t < \tau^x_{\mu_2}(0), \\ -(1-a)\left(x + I_2(t)\right), & \text{if } \tau^x_{\mu_2}(0) \leq t < \tau^x_{\mu_3}(0), \\ ax - (1-a)I_2(t) + I_3(t), & \text{if } \tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0), \\ -aI_1(t) - (1-a)I_2(t) + I_3(t), & \text{if } t \geq \tau^x_{\mu_1}(0). \end{cases}$$

This partition allows us to spot when strict convexity can occur. Note that by definition $t \geq \tau^x_{\mu_2}(0)$, $I_2(t) = \inf_{0\leq s\leq t}[U(s) - \mu_2 s] \leq -x$, so that $D(t) \geq 0$ if $\tau^x_{\mu_2}(0) \leq t < \tau^x_{\mu_3}(0)$. Moreover, by subadditivity of the infimum,

$$\begin{aligned} I_3(t) &= \inf_{0 \leq s \leq t}[U(s) - \mu_3 s] = \inf_{0 \leq s \leq t}[a(U(s) - \mu_1 s) + (1-a)(U(s) - \mu_2 s)] \\ &\geq a \inf_{0 \leq s \leq t}[U(s) - \mu_1 s] + (1-a) \inf_{0 \leq s \leq t}[U(s) - \mu_2 s] = aI_1(t) + (1-a)I_2(t), \end{aligned}$$

and hence $D(t) \geq 0$ for $t \geq \tau^x_{\mu_1}(0)$. Using the same argument, we deduce

$$ax - (1-a)I_2(t) + I_3(t) \geq ax - (1-a)I_2(t) + aI_1(t) + (1-a)I_2(t) = a(x + I_1(t)).$$

In particular for $t < \tau^x_{\mu_1}(0)$, this value is strictly positive. As a result, $D(t) \geq 0$ for all $t \geq 0$. On top of that $D(t) > 0$ for $t \in [\tau^x_{\mu_3}(0), \tau^x_{\mu_1}(0))$. Accordingly, the latter implies strict positivity of $\mathbb{E}D(t)$, and therefore strict convexity of $\mathbb{E}Q_\mu(t)$, if the event $\{\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\}$ occurs with positive probability. That is,

$$\begin{aligned} P(D(t) > 0) &\geq P\left(a(x + I_1(t))\mathbb{1}_{\{\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\}} > 0\right) \\ &= P\left(x + I_1(t) > 0, \tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) \\ &= P\left(x + I_1(t) > 0 | \tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) P\left(\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) \\ &= P\left(\tau^x_{\mu_3}(0) \leq t < \tau^x_{\mu_1}(0)\right) = P(\tau^x_{\mu_3}(0) \leq t) - P(\tau^x_{\mu_1}(0) \leq t) > 0, \end{aligned} \tag{6.38}$$

by the stochastic dominance in (6.37). To ensure the strict inequality in (6.38) we have to enforce the condition

$$P(\tau^x_{\mu_1}(0) < T) > 0. \tag{6.39}$$

Remark 6.3. An example illustrating the need for this condition is the case in which $U(t)$ is a compound Poisson process and $T < x/\mu_2 < x/\mu_1$. Then

$$Q_{\mu_k}(t) = x + U(t) - \mu_k t,$$

for all $t \in [0, T]$, since $U(t) \geq 0$ and therefore $\tau^x_{\mu_1}(0) > T$. Consequently, for all $a \in (0, 1)$,

$$a\, Q_{\mu_1} + (1-a)\, Q_{\mu_2}(t) = Q_{\mu_3}(t),$$

proving only convexity of $\mathbb{E}Q_\mu(t)$ and subsequently convexity of $\int_0^T \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t$. In case $\sigma > 0$, the probability in (6.39) is necessarily positive.

The case $x = 0$**.** By the fact that $\tau_\mu(0) = 0$ for all $\mu > 0$, proving that $D(t) > 0$ in the case $x = 0$ reduces to showing that the probability of

$$D(t) = aI_1(t) + (1-a)I_2(t) - I_3(t) > 0$$

happening is positive for all $t > 0$. Define

$$t_0 := \inf\{t > 0 \,:\, U(t) > 0\},$$

and

$$\tilde{\tau}_\mu := \inf\{t > t_0 \,: U(t) - \mu t \leq 0\}.$$

We note that t_0 as defined above, also defines the epoch of the start of a new excursion of the reflection Q_μ for all $\mu > 0$. Namely,

$$U(s) \leq 0 \quad \Rightarrow \quad U(s) - \mu s \leq -\mu s \qquad \text{for all } 0 \leq s < t_0$$

$$\Rightarrow \inf_{0 \leq s < t_0} [U(s) - \mu s] \leq -\mu t_0 \quad \Rightarrow U(t_0) - \mu t_0 - \inf_{0 \leq s < t_0} [U(s) - \mu s] \geq U(t_0) > 0.$$

Then $Q_\mu(t_0-) = 0$ for all $\mu > 0$. By virtue of the Strong Markov Property, note that $Q_\mu(t_0 + t) \stackrel{d}{=} Q_\mu(t)$. Hence we assume without loss of generality $t_0 = 0$. Again, we have a stochastic dominance relation similar to (6.37):

$$\tilde{\tau}_{\mu_2} < \tilde{\tau}_{\mu_3} < \tilde{\tau}_{\mu_1},$$

for all $\mu_1 < \mu_3 < \mu_2$. Then

$$D(t) \stackrel{d}{=} \begin{cases} 0, & \text{if } t < \tilde{\tau}_{\mu_2}, \\ -(1-a)I_2(t), & \text{if } \tilde{\tau}_{\mu_2} \leq t < \tilde{\tau}_{\mu_3}, \\ (1-a)I_2(t) + I_3(t), & \text{if } \tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}, \\ -aI_1(t) - (1-a)I_2(t) + I_3(t), & \text{if } t \geq \tilde{\tau}_{\mu_1}. \end{cases}$$

Clearly, $D(t) \geq 0$ for all $t \geq 0$ and

$$-(1-a)I_2(t) + I_3(t) \geq aI_1(t) > 0,$$

for $\tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}$. Hence, in a similar manner to (6.38),

$$\begin{aligned}
P(D(t) > 0) &\geq P\left(aI_1(t)\mathbb{1}_{\{\tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}\}} > 0\right) \\
&= P\left(I_1(t) > 0, \tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}\right) \\
&= P\left(I_1(t) > 0 | \tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}\right) P\left(\tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}\right) \\
&= P\left(\tilde{\tau}_{\mu_3} \leq t < \tilde{\tau}_{\mu_1}\right) = P(\tilde{\tau}_{\mu_3} \leq t) - P(\tilde{\tau}_{\mu_1} \leq t) > 0. \qquad (6.40)
\end{aligned}$$

The last inequality is satisfied it $P(\tilde{\tau}_{\mu_1} < T) > 0$, which is equivalent to $P(U(T) - \mu T \leq 0) > 0$, a condition that is clearly true for all our choices of U. In conclusion, for $x = 0$, $\mathbb{E}[D(t)] > 0$ and therefore $\mathbb{E}[Q_\mu(t)]$ is a strictly convex function of μ. □

6.C.2 Proof of Proposition 6.3

The proof of the proposition relies on the following auxiliary lemma, of which we include the proof for completeness.

Lemma 6.5. *Consider the sequence of functions $f_n : [x_0, \infty) \to \mathbb{R}$ and let $f : [x_0, \infty) \to \mathbb{R}$ be the pointwise limit for some $x_0 \in \mathbb{R}$. Assume f and f_n are strictly convex for all n. Furthermore, let $f(y) \to \infty$ for both $y \to x_0^+$ and $y \to \infty$. If x_n and x are the minimizers for f_n and f, respectively, then $x_n \to x$ for $n \to \infty$.*

Proof. We start by showing that the sequence x_n is bounded. Fix u_l, u_r such that $x_0 < u_l < x < u_r$. We claim that there exists a $N \in \mathbb{N}$ such that $x_n \in [u_l, u_r]$ for all $n \geq N$. First, we prove the upper bound on x_n. For any strictly convex function h with minimizer x_h, the following statement holds true:

$$x_h < u_r \quad \Leftrightarrow \quad h \text{ is strictly increasing at } u_r. \qquad (6.41)$$

The first implication follows from observing that $h(x_h) < h(y)$ for all $y > x^*$ and definition of convexity:

$$0 < \frac{h(u_r) - h(x_h)}{u_r - x_h} \leq \frac{h(u_r + \delta) - h(u_r)}{\delta},$$

for all $\delta > 0$. Hence $h(u_r) < h(u_r + \delta)$, i.e. h is increasing at u_r. The converse follows immediately by observing that $h(u_r) < h(u_r + \delta)$ for all $\delta > 0$, so that $x_h < u_r$. Next, we show that f_n must be increasing at u_r for n sufficiently large. By pointwise convergence of f_n we have

$$\lim_{n\to\infty} [f_n(u_r + \delta) - f_n(u_r)] = f(u_r + \delta) - f(u_r).$$

Let $w_r := f(u_r + \delta) - f(u_r) > 0$. Then

$$\exists N_r \in \mathbb{N} : \forall n \geq N_r : |[f_n(u_r + \delta) - f_n(u_r)] - [f(u_r + \delta) - f(u_r)]| < w_r/2.$$

Hence for $n \geq N_r$,

$$f(u_r+\delta)-f(u_r)-w_r/2 < f_n(u_r+\delta)-f_n(u_r) < f(u_r+\delta)-f(u_r)+w_r/2$$
$$\Rightarrow 0 < w_r/2 < f_n(u_r+\delta)-f_n(u_r).$$

Hence by (6.41), $x_n < u_r$ for sufficiently large n. Similarly, we argue

$$x_h > u_l \quad \Leftrightarrow \quad h \text{ is strictly decreasing at } u_l,$$

for any strictly convex function h with minimizer x_h. Note that $x_h > u_l$ implies $h(x_h) - h(u_l) < 0$ and for all $\delta > 0$ we get by strict convexity

$$\frac{h(u_l)-h(u_l-\delta)}{\delta} < \frac{h(x_h)-h(u_l)}{x_h-u_l} < 0,$$

by which $h(u_l-\delta) > h(u_l)$, i.e. h is decreasing in u_l. Moreover, if h is decreasing at u_l, then it is decreasing for all $y < u_l$, by arguments similar to the above. Therefore, $h(u_l-\delta) > h(u_l)$ for all $\delta > 0$ and it must hold that $x_h > u_l$. Define $f(u_l) - f(u_l - \delta) := w_l < 0$, then again by pointwise convergence, we have that

$$\exists N_l \in \mathbb{N} : \forall n \geq N_l : |[f_n(u_l)-f_n(u_l-\delta)]-[f(u_l)-f(u_l-\delta)]| < w_l,$$

whereupon

$$f_n(u_l)-f_n(u_l-\delta) < f(u_l)-f(u_l-\delta)+w_l = 2w_l < 0.$$

Hence, for sufficiently large n, we also have $x_n > u_l$. Fix $N = \max\{N_l, N_r\}$, then for $n \geq N$, $x_n \in (u_l, u_r)$. That is, the sequence x_n is bounded. Therefore, by the theorem of Bolzano-Weierstrass, x_n has to have a convergent subsequence. That is, there exists a sequence n_k such that $n_k \to \infty$ and $x_{n_k} \to a$ as $k \to \infty$ for some $a \in [u_l, u_r]$. We prove that every subsequence must converge to x by contradiction. Suppose there exists a subsequence n_k such that $x_{n_k} \to a \neq x$. Since, $x_n \in [u_l, u_r]$ for $n \geq N$, we may restrict our attention to the sequence of functions $\hat{f}_n : [u_l, u_r] \to \mathbb{R}^+$, consisting of the original function f_n restricted to the domain $[u_l, u_r]$. To be precise $x_n = \arg\min_y f_n(y) = \arg\min_y \hat{f}_n(y)$ for $n \geq N$. Because $\hat{f}_n$ and $\hat{f}$ are bounded, we furthermore have that $\hat{f}_n \to \hat{f}$ uniformly.

Fix $\varepsilon > 0$. By uniform convergence, there exists an $K_0 \in \mathbb{N}$ such that

$$|\hat{f}_{n_k}(y) - \hat{f}(y)| < \varepsilon/2, \quad \forall k \geq K_0,\ y \in [u_l, u_r].$$

Also, because $\hat{f}$ is convex, it is continuous, so that there exists a $\delta := \delta(\varepsilon)$ so that

$$|z-y| < \delta \quad \Rightarrow \quad |\hat{f}(z) - \hat{f}(y)| < \varepsilon/2.$$

Let K_1 be such that $|x_{n_k} - a| < \delta$ for all $k \geq K_1$. Then for $k \geq K = \max\{K_0, K_1\}$ this implies

$$\begin{aligned} |f_{n_k}(x_{n_k}) - f(a)| &= |\hat{f}_{n_k}(x_{n_k}) - \hat{f}(a)| \\ &\leq |\hat{f}_{n_k}(x_{n_k}) - \hat{f}(x_{n_k})| + |\hat{f}(x_{n_k}) - f(a)| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Hence we conclude $\lim_{k\to\infty} \hat{f}_{n_k}(x_{n_k}) = f(a)$. Therefore,

$$\limsup_{n\to\infty} f_n(x_n) \geq f(a) > f(x),$$

by minimality of x. However, $f_n(x_n) \leq f_n(x)$, which implies $\limsup_{n\to\infty} f_n(x_n) \leq \lim_{n\to\infty} f_n(x) = f(x)$, contradicting the strict inequality above. Hence we deduce $x = a$. Consequently, every subsequence of x_n converges to x and therefore $x_n \to x$ as $n \to \infty$. Applying Lemma 6.5 to the functions Π_T and Π_∞ with $x_0 = \lambda$, together with Lemma 6.4, we obtain the result immediately.

□

6.C.3 Proof of Proposition 6.4

Proof. Note that Π_∞ is a smooth function. By the first optimality condition $\Pi'_\infty(\mu^\star_\infty) = 0$. We first prove that also $\Pi_T(\mu)$ is differentiable with respect to μ for all $\mu \geq 0$. Recall (6.5), which defines the cost function as a combination of the accumulated expected transient queue length, and linear staffing costs. The latter term is clearly differentiable, hence it remains to be proved that

$$C_T(\mu) = \frac{1}{T}\int_0^\infty \mathbb{E}[Q_\mu(t)]\,\mathrm{d}t,$$

admits a derivative for all $\mu \geq 0$ with T fixed. This holds if and only if $\mathbb{E}[Q_\mu(t)]$ is differentiable for all $t \geq 0$. Let $Q(0) = x \geq 0$. Following (6.3),

$$\begin{aligned}\mathbb{E}[Q_\mu(t)] &= \mathbb{E}[X_\mu(t)] + \mathbb{E}\Big[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}\Big] \\ &= (\lambda-\mu)t + \mathbb{E}\Big[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}\Big],\end{aligned}$$

where the first term is differentiable. Furthermore,

$$\begin{aligned}\mathbb{E}[\max\{x, \sup_{s\in[0,t]}\{-X_\mu(s)\}\}] &= x + \int_x^\infty P(\sup_{s\in[0,t]}\{-X_\mu(s)\} > u)\mathrm{d}u \\ &= x + \int_x^\infty P(\hat{\tau}^0(u) \leq t)\mathrm{d}u,\end{aligned}$$

with $\hat{\tau}^0(u)$ as defined in (6.16).

Since $-X_\mu$ is a process with no positive jumps, we may apply [36, Cor. VII.3], which states that the following equivalence between measures holds:

$$s\,P(\hat{\tau}^0(u) \in ds)du = u\,P(-X_\mu(s) \in du)ds, \tag{6.42}$$

so that

$$
\begin{aligned}
\int_{u=x}^{\infty} P(\hat{\tau}^0(u) \le t)\,\mathrm{d}u &= \int_{u=x}^{\infty}\int_{s=0}^{t} P(\hat{\tau}^0(u) \in ds)\mathrm{d}u \\
&= \int_{u=x}^{\infty}\int_{s=0}^{t} s^{-1}u\; P(-X_\mu(s) \in \mathrm{d}u)\mathrm{d}s \\
&= \int_{u=x}^{\infty}\int_{s=0}^{t} s^{-1}u\, P(X_\mu(s) \in \mathrm{d}u)\mathrm{d}s \\
&= \int_{s=0}^{t} s^{-1}\mathbb{E}[\max\{x, X_\mu(s)\}]\mathrm{d}s \\
&= \int_{s=0}^{t}\int_{v=x/s}^{\infty} P(X_\mu(s)/s > v)\mathrm{d}v\mathrm{d}s \\
&= \int_{s=0}^{t}\int_{v=x/s}^{\infty} P(U(\lambda s)/s > v+\mu)\mathrm{d}v\mathrm{d}s \\
&= \int_{s=0}^{t}\int_{w=x/s+\mu}^{\infty} P(U(\lambda s)/s > w)\mathrm{d}w\mathrm{d}s, \qquad (6.43)
\end{aligned}
$$

where the interchange of integrals is justified by Fubini's theorem and this last form is differentiable with respect to μ. Substituting $Q(0)$ for x straightforwardly yields differentiability of the complete cost function Π_T for all T.

Consequently we invoke the first optimality condition for $\mu_T^\star$ to find

$$
\begin{aligned}
0 = \Pi_T'(\mu_T^\star) &= \Pi_\infty'(\mu_T^\star) + \Psi_T'(\mu_T^\star) + O(1/T^2) \\
&= \Pi_\infty'(\mu_\infty^\star) + \Psi_T'(\mu_\infty^\star) + (\mu_T^\star - \mu_\infty^\star)\left[\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star)\right] \\
&\quad + \frac{1}{2}(\mu_T - \mu_\infty^\star)^2\left[\Pi_T'''(\xi) + \Psi_T'''(\xi)\right] + O(1/T^2) \\
&= \Psi_T'(\mu_\infty^\star) + (\mu_T^\star - \mu_\infty^\star)\left[\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star)\right] \\
&\quad + \frac{1}{2}(\mu_T - \mu_\infty^\star)^2\left[\Pi'''(\xi) + \Psi_T'''(\xi)\right] + O(1/T^2),
\end{aligned}
$$

for some $\xi \in [\mu_T^\star, \mu_\infty^\star]$. Rearranging this gives

$$
\begin{aligned}
\mu_T^\star - \mu_\infty^\star &= \frac{-\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star) + \Psi_T''(\mu_\infty^\star) + \frac{1}{2}(\mu_T^\star - \mu_\infty^\star)(\Pi_\infty'''(\mu_T^\star) + \Psi_T'''(\xi))} + O(1/T) \\
&= -\frac{\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}\left[1 - \frac{\Psi_T''(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)} - \frac{\mu_T^\star - \mu_\infty}{2}\frac{\Pi_\infty'''(\mu_\infty^\star) + \Psi_T'''(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}\right] + O(1/T) \\
&= -\frac{\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}[1 + o(1)]
\end{aligned}
$$

for $T \to \infty$, since both $\mu_T - \mu_\infty$ and $\Psi_T''(\mu_\infty^\star)$ are $o(1)$. Let

$$\mu_\bullet := \lim_{T\to\infty} \frac{T\Psi_T'(\mu_\infty^\star)}{\Pi_\infty''(\mu_\infty^\star)}.$$

By (6.19) we have

$$T\Psi_T'(\mu) = -\frac{\mathbb{E}[Q(0)^2]}{2(\mu-\lambda)^2} + \frac{\lambda u_3}{3(\mu-\lambda)^3} + \frac{3\lambda^2 u_2^2}{4(\mu-\lambda)^4}.$$

Together with

$$\Pi_\infty''(\mu) = \frac{\lambda u_2}{(\mu-\lambda)^3}$$

and (6.23) we obtain the expression for $\mu_\bullet$ in (6.24). □

6.C.4 Proof of Proposition 6.5

Proof. We upper bound the optimality gap by using the decomposition in (6.21).

$$\begin{aligned}|\Pi_\infty^\star - \Pi_T^\star| &= \left|\hat{\Pi}_T(\mu_\infty) + \Delta_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star) - \Delta_T(\mu_T^\star)\right| \\ &\leq |\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| + |\Delta_T(\mu_\infty^\star)| + |\Delta_T(\mu_T^\star)| \\ &= |\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| + O(1/T^2), \end{aligned} \tag{6.44}$$

since $\Delta_T(\mu) = O(1/T^2)$ by Proposition 6.2. Next, we find an upper bound for $|\hat{\Pi}_T(\gamma) - \hat{\Pi}_T(\beta)|$, with $\hat{\Pi}_T(\cdot)$ as in (6.21), in terms of the difference between γ and β. For simplicity, denote $\hat{\gamma} = \gamma - \lambda$ and $\hat{\beta} = \beta - \lambda$, implying $\hat{\gamma} - \hat{\beta} = \gamma - \beta$. Then, using (6.19), we get

$$\begin{aligned}|\hat{\Pi}_T(\mu_\infty^\star) - \hat{\Pi}_T(\mu_T^\star)| = \Bigg|&\alpha(\hat{\gamma}-\hat{\beta}) + \left(\frac{\lambda u_2}{2} + \frac{\mathbb{E}[Q(0)^2]}{2T}\right)\left(\frac{1}{\hat{\gamma}} - \frac{1}{\hat{\beta}}\right) \\ &- \frac{\lambda^2 u_2^2}{4T}\left(\frac{1}{\hat{\gamma}^3} - \frac{1}{\hat{\beta}^3}\right) - \frac{\lambda u_3}{6T}\left(\frac{1}{\hat{\gamma}^2} - \frac{1}{\hat{\beta}^2}\right)\Bigg|.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\frac{1}{\hat{\gamma}} - \frac{1}{\hat{\beta}} &= -\frac{\hat{\gamma}-\hat{\beta}}{\hat{\beta}^2} + \frac{(\hat{\gamma}-\hat{\beta})^2}{\hat{\beta}^3} + O\left((\gamma-\beta)^3\right), \\ \frac{1}{\hat{\gamma}^2} - \frac{1}{\hat{\beta}^2} &= -\frac{2(\hat{\gamma}-\hat{\beta})}{\hat{\beta}^3} + \frac{3(\hat{\gamma}-\hat{\beta})^2}{\hat{\beta}^4} + O\left((\gamma-\beta)^3\right), \\ \frac{1}{\hat{\gamma}^3} - \frac{1}{\hat{\beta}^3} &= -\frac{3(\hat{\gamma}-\hat{\beta})}{\hat{\beta}^4} + \frac{6(\hat{\gamma}-\hat{\beta})^2}{\hat{\beta}^5} + O\left((\gamma-\beta)^3\right).\end{aligned}$$

Substituting these yields

$$\begin{aligned}|\hat{\Pi}_T(\gamma)-\hat{\Pi}_T(\beta)| = \Bigg| &(\gamma-\beta)\left[\alpha-\frac{\lambda u_2}{2\hat{\beta}^2}+\frac{1}{2T\hat{\beta}^2}\left(\mathbb{E}[Q(0)^2]+\frac{3\lambda^2u_2^2}{2\hat{\beta}^2}+\frac{2\lambda u_3}{3\hat{\beta}}\right)\right] \\ &-(\gamma-\beta)^2\left[\frac{\lambda u_2}{2\hat{\beta}^3}+\frac{1}{2T\hat{\beta}^3}\left(\mathbb{E}[Q(0)^2]-\frac{3\lambda^2u_2^2}{\hat{\beta}^2}-\frac{\lambda u_3}{\hat{\beta}}\right)\right]\Bigg| \\ &+O\left((\gamma-\beta)^3\right).\end{aligned}$$

Given that $\mu_T^\star = \mu_\infty^\star + \mu_\bullet/T + o(1/T)$, we find

$$\begin{aligned}|\hat{\Pi}_T(\mu_\infty^\star)-\hat{\Pi}_T(\mu_T^\star)| &= \frac{|\mu_\bullet|}{T}\left(\alpha-\frac{\lambda u_2}{2(\mu_\infty^\star-\lambda)^2}\right)+O(1/T^2) \\ &= \frac{|\mu_\bullet|}{T}\left(\alpha-\frac{\lambda u_2}{2(\sqrt{\lambda u_2/2\alpha})^2}\right)+O(1/T^2)=O(1/T^2),\end{aligned}$$

which concludes the proof. □

7
A blood bank model

We consider a stochastic model for a blood bank, in which amounts of blood are offered and demanded according to independent compound Poisson processes. Blood is perishable, that is, blood can only be kept in storage for a limited amount of time. Furthermore, demand for blood is impatient, that is, a demand for blood may be canceled if it cannot be satisfied soon enough. For a range of perishability functions and demand impatience functions, we derive the steady-state distribution of the blood inventory level. Moreover, we deduce fluid and diffusion limits for the inventory process as the arrival rates of of the compound Poisson processes grow indefinitely. These scaling limits in turn provide normal approximations for the performance of large-scale systems.

Based on
A blood bank model with perishable blood and demand impatience
Shaul Bar-Lev, Onno Boxma, Britt Mathijsen & David Perry
Submitted to *Stochastic Systems*

7.1 Introduction

This chapter is devoted to the study of a stochastic blood bank model in which amounts of blood are offered and demanded according to stochastic processes, and in which blood is perishable (that is, blood can only be kept for a limited amount of time) and demand for blood is impatient (that is, a demand request for blood may be canceled if it cannot be satisfied soon enough). Let us first provide some background, and subsequently sketch the blood bank model in some more detail.

Practical background. One of the major issues in securing blood supply to patients worldwide is to provide blood of the best achievable quality, in the needed quantities. In most countries, blood, which is collected as whole blood units from human donors, is separated into different components which are subsequently stored under different storage conditions according to their biological characteristics, functions and respective expiration dates. Blood units and components are ordered by local hospital blood banks (LBB) from the Central Blood Bank (CBB) according to their operational needs. The CBB has to run its inventory and supply according to these requests and to the need to keep sufficient stock for immediate release in emergency situations. It also has to perform tests to determine the unit's blood type and to detect the presence of various pathogens which are able to cause transfusion-transmitted diseases, such as Hepatitis B, Hepatitis C, Human Immunodeficiency Virus (HIV) and Syphilis, see e.g. Steiner et al. [201].

Blood consists of several components: red blood cells, plasma and plate-lets. In addition, there are 8 blood groups (types): O^+, O^-, A^+, A^-, B^+ ,B^-, AB^+, AB^- ($-$ means Rh negative) where the interrelationship between the transfusion issuing policies among the 8 types is quite intricate. It turns out that each of the negative types can satisfy the corresponding $+$ type, but not vice versa. Blood components are perishable as red blood cells can be used for only 35 to 42 days and platelets for only 5 days (plasma, however, can be frozen and kept for one year). Accordingly, if red blood cells and particularly platelets are not used for blood transfusion within their expiration dates, then they perish.

In most developed countries demand requirements of about 50.000 blood donations are needed per one million persons per year. About 95% of these donations are aggregated by CBBs and the remaining 5% by LBBs. Blood units stored at the CBB are usually ordered by LBBs for planned elective surgeries. However, as it happens rather frequently, elective surgeries turn out to become emergency ones due to various conditions of the patient involved. In such cases, hospitals use their own local blood banks to supply the demand, and they cancel the required demand from the CBB; this is what we refer to as demand impatience. A good review on supply chain management in blood products appears in Beliën & Forcé [33] and the references cited therein. Other relevant studies are Ghandforoush and Sen [85] & Stanger et al. [198].

Inventory model. In this chapter we consider the analysis of blood perishability and demand impatience, concentrating on only one blood type. We do this by considering the stochastic inventory processes $\{X_b(t)\}_{t\geq 0}$, with $X_b(t)$ the amount of blood kept in storage at time t, and $\{X_d(t)\}_{t\geq 0}$, with $X_d(t)$ the amount of demand for blood (the shortage) at time t. If $X_b(t) > 0$ then $X_d(t) = 0$, and if $X_d(t) > 0$ then $X_b(t) = 0$. We assume that amounts of blood arrive according to a Poisson process, and that requests for blood arrive according to another, independent, Poisson process. The delivered and requested amounts of blood are assumed to be random variables. We represent the perishability of blood by letting the amount of blood, when positive, decrease in a state-dependent way: if the amount is v, then the decrement rate is $\xi_b v + \alpha_b$. The ξ_b factor is motivated by the fact that a large amount of blood suggests that some of the blood has been present for quite a while – and hence there is a relatively high perishability rate when much blood is in inventory. The α_b factor provides additional modeling flexibility. One can in this way represent the blood perishability more accurately; but the α_b term could also, e.g., represent a fluid demand rate of individuals or organizations, which contact the CBB directly, and that is only satisfied when there is inventory. Similarly, we represent the demand impatience by a decrement rate $\xi_d v + \alpha_d$. The ξ_d factor is motivated by the following fact. When there is a large shortage (demand) of blood, there are probably many patients waiting for blood, so many patients that might become impatient (that is, they could recover, or die, or become in need of emergency surgery) leading to a cancellation of the required demand from the CBB. Again, the α_d factor provides additional modeling flexibility; it not only allows us to represent demand impatience more accurately, but it could also, e.g., represent additional donations of individuals in times of blood shortage.

The inclusion of both the perishability factor $\xi_b v + \alpha_b$ and the demand impatience factor $\xi_d v + \alpha_d$ makes the analysis of the ensuing model mathematically quite challenging, but leads to a very general model that contains many well-known models as special cases. Our two-sided stochastic process, with both upward and downward jumps, and with the rather general slope factors $\xi v + \alpha$, could represent a quite large class of stochastic phenomena. It should for example be noted that this model is a two-sided generalization of the well-known shot-noise model that describes certain physical phenomena, see [132]). In some of our calculations we remove either the ξ factors or the α factors, and this results in easier calculations and more explicit results.

Our main results are: (i) Determination of the steady-state distributions of the amounts of blood and of demand in inventory; in particular, we present a detailed analysis of the case in which the delivered and requested amounts of blood are both exponentially distributed. (ii) Expressions for mean amounts of blood and demand in storage, and for the probability of not being able to satisfy demand. (iii) We obtain the fluid and diffusion limits of the blood inventory process, providing in particular sufficient conditions for the limit process to be an Ornstein-Uhlenbeck process.

Structure of the chapter. The chapter is organized as follows: Section 7.2 presents a detailed model description. A steady-state analysis of the densities of demand and of blood amount in storage is contained in Section 7.3, including the special case of exponentially distributed delivered and requested blood amounts when $\alpha_b = \alpha_d = 0$ (i.e., pure proportionality). The fluid and diffusion scalings are discussed in Section 7.4, and in Section 7.5 we present numerical results for certain performance measures like mean net amount of blood and the probability that there is a shortage of blood. These results indicate, among other things, that the probability that there is a shortage of blood can be accurately approximated via a normal approximation, based on the Ornstein-Uhlenbeck process appearing in the diffusion scaling. Section 7.6 contains some conclusions and suggestions for further research.

7.2 Model description

We consider the following highly simplified model of a blood bank, restricting ourselves to only one type of blood.

Blood amounts arrive according to a Poisson process with rate λ_b. The amounts which successively arrive are independent, identically distributed random variables $B_1, B_2, \dots$ with distribution $F_b(\cdot)$; $\bar{F}_b(x) = 1 - F_b(x)$. Demands for blood arrive according to a Poisson process with rate λ_d. The successive demand amounts are independent, identically distributed random variables $D_1, D_2, \dots$ with distribution $F_d(\cdot)$; $\bar{F}_d(x) = 1 - F_d(x)$. We view these amounts as continuous quantities, measured in, for instance liters. If there is enough blood for a demand, then that demand is immediately satisfied. If there is some blood, but not enough to fully satisfy a demand, then that demand is partially satisfied, using all the available blood. The remainder of the demand may be satisfied later.

Blood has a finite expiration date. We make the assumption that if the total amount of blood present is $v > 0$, then blood is discarded – because of its finite expiration date – at a rate $\xi_b v + \alpha_b$, so linear in v. Blood demands have a finite patience. We make the assumption that if the total amount of demand present is $v > 0$, then demand disappears – because of its finite patience – at a rate $\xi_d v + \alpha_d$, so linear in v.

Notice that *either* the total amount of blood present, *or* the total amount of demands, is zero, *or* both are zero; they cannot be both positive. Hence we can easily in one figure depict the two-sided process $\{X(t)\}_{t\geq 0} = \{(X_b(t), X_d(t))\}_{t\geq 0}$ of total blood and total demand amounts present at any time t, as we have done in Figure 7.1. For our purposes, we are mainly interested in the characteristics of the process described above in stationarity. Let us denote by X_d the steady-state total amount of demand and by X_b the steady-state total amount of blood present, with corresponding density functions $f(\cdot)$ and $g(\cdot)$, respectively. Notice that these are defective densities; we have $\int_{0+}^{\infty} f(v)\mathrm{d}v = \pi_d = \mathbb{P}(\text{demand} > 0)$ and $\int_{0+}^{\infty} g(v)\mathrm{d}v = \pi_b = \mathbb{P}(\text{blood} > 0)$. If $\alpha_b = \alpha_d = 0$, then neither X_b nor X_d has

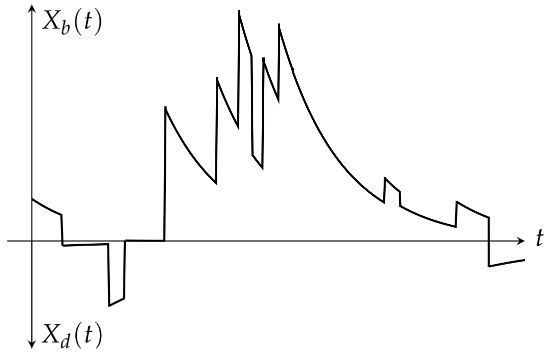


Figure 7.1: Sample path of net amount of blood available as a function of time.

probability mass at zero, and $\pi_b + \pi_d = 1$ (when there is only a very small amount v present, the "decay" rate $\xi_b v$ or $\xi_d v$ is very small). However, if α_b and/or α_d is positive, then there is a positive probability π_0 of being in 0.

When ξ_d and ξ_b are positive, existence of these steady-state densities is obvious; otherwise, the conditions for the existence of the steady-state distributions require some discussion, see Section 7.3.3.

7.3 Steady-state analysis

In this section we present a global approach towards determining $f(\cdot)$ and $g(\cdot)$ in the most general form of our model. Using the Level Crossing Technique (LCT), we derive two integral equations in $f(\cdot)$ and $g(\cdot)$. Before attempting to solve these equations, we consider a few important performance measures which can be expressed in $f(\cdot)$ and $g(\cdot)$, π_0 and the mean length of time during which, uninterruptedly, there is a positive amount of blood (respectively demand). The latter could be viewed as the busy period of the X_b process (respectively of the X_d process).

First we consider the density $g(\cdot)$ of the amount of blood. We equate the rate at which some positive blood level v is upcrossed and downcrossed, respectively. LCT leads to the following integral equation: for $v > 0$,

$$\lambda_b \int_0^v g(y)\bar{F}_b(v-y)\mathrm{d}y + \lambda_b \int_0^\infty f(y)\bar{F}_b(v+y)\mathrm{d}y + \pi_0\lambda_b\bar{F}_b(v)$$
$$= \lambda_d \int_v^\infty g(y)\bar{F}_d(y-v)\mathrm{d}y + (\xi_b v + \alpha_b)g(v). \tag{7.1}$$

Here the three terms in the left-hand side represent the rate of crossing level v from below; the first term corresponds to a jump from a blood inventory level between 0 and v, whereas the second term corresponds to a jump from a shortage level, and the third term corresponds to a jump from level 0. The two terms in the right-hand side represent the rate of crossing level v from above; the first term corresponds to a jump from above v, and the second term to a smooth crossing.

Next, we consider the density $f(\cdot)$ of the amount of demand (shortage). We equate the rate at which some positive demand level v is upcrossed and downcrossed, respectively. LCT leads to the following integral equation: for $v > 0$,

$$\lambda_d \int_0^v f(y)\bar{F}_d(v-y)\mathrm{d}y + \lambda_d \int_0^\infty g(y)\bar{F}_d(v+y)\mathrm{d}y + \pi_0\lambda_d\bar{F}_d(v)$$
$$= \lambda_b \int_v^\infty f(y)\bar{F}_b(y-v)\mathrm{d}y + (\xi_d v + \alpha_d)f(v). \tag{7.2}$$

It should be noted that these two, coupled, equations are symmetric (swap f and g, and the b and d parameters).

In general, it appears to be very difficult to solve these integral equations. In Section 7.3.1 we assume that both $F_b(\cdot)$ and $F_d(\cdot)$ are exponential. In that case, we are able to obtain explicit expressions of $f(\cdot)$ and $g(\cdot)$, in terms of hypergeometric functions. In Section 7.3.2 we consider the case that $F_b(\cdot)$ and $F_d(\cdot)$ are Coxian distributions, a class of distributions that lies dense in the class of all distributions of non-negative random variables, and that is suitable for handling the above coupled integral equations via Laplace transforms (LT). We are able to transform (7.1) and (7.2) into inhomogeneous first-order differential equations in the LTs of $f(\cdot)$ and $g(\cdot)$, and thus to obtain those LTs.

A few simple performance measures. Without solving (7.1)-(7.2) explicitly, we are able to deduce some characteristics of the steady-state inventory level.

First, we can relate π_0 to the densities $f(\cdot)$ and $g(\cdot)$; see Proposition 7.1 below. Subsequently we express the mean length of time during which there is, uninterruptedly, a positive amount of blood present (we call this the non-emptiness period of the inventory system), into $f(\cdot)$, $g(\cdot)$ and π_0. We do the same for the mean length of time during which there is, uninterruptedly, a positive demand, i.e., the non-emptiness period of the demand process, see Proposition 7.2.

Proposition 7.1. *Let π_0 be the steady-state atom probability of the zero period. Then*

$$\pi_0 = \frac{\alpha_d f(0) + \alpha_b g(0)}{\lambda_d + \lambda_b}.$$

Proof. Substitute $v = 0$ in (7.1) and (7.2) and take the sum. The result is obtained after several steps of elementary algebra. □

The result introduced in the proposition above is very intuitive. By LCT, $\alpha_d f(0) + \alpha_b g(0)$ is the rate at which level 0 is reached (i.e., the process will now really stay at 0 for a while), so that $[\alpha_d f(0) + \alpha_b g(0)]^{-1}$ is the expected length of time between two successive times level 0 is reached by the fluid. More precisely, the *zero periods* and *non-zero periods* generate an alternating renewal process whose expected cycle length is $[\alpha_d f(0) + \alpha_b g(0)]^{-1}$. The expected length of the zero period is $[\lambda_d + \lambda_b]^{-1}$, since the end of the zero period is terminated at the moment of the next jump. But the jump process is a Poisson process with rate $\lambda_d + \lambda_b$. Now the renewal reward theorem simply says that

$$\pi_0 = \frac{\mathbb{E}[\text{zero period}]}{\mathbb{E}[\text{cycle}]}.$$

In preparation of the next proposition, for the process $\{X(t)\}_{t\geq 0}$ we define a modified process $\{X_m(t)\}_{t\geq 0}$, where X_m is constructed by deleting the zero-periods (only the zero periods, not the emptiness periods) from X and gluing together the *non-zero periods*. The modified process is X_m such that $X_m(t) = X_d(t)\mathbb{1}_{\{X_d(t)>0\}} + X_b(t)\mathbb{1}_{\{X_b(t)>0\}}$ where by definition of the model $\{X_d(t) > 0\} \Rightarrow \{X_b(t) = 0\}$ and $\{X_b(t) > 0\} \Rightarrow \{X_d(t) = 0\}$.

Proposition 7.2. *Let B_b and I_b be the generic non-emptiness period and the emptiness period, respectively, of the inventory system. Similarly, let B_d and I_d be the generic non-emptiness period and the emptiness period, respectively, of the demand process. Then*

$$\text{(i)} \begin{cases} \mathbb{E}[B_b] = \frac{1-\pi_0}{\alpha_b g(0)+\lambda_d \int_0^\infty \bar{F}_d(y)g(y)\mathrm{d}y}, \\ \mathbb{E}[B_d] = \frac{1-\pi_0}{\alpha_d f(0)+\lambda_b \int_0^\infty \bar{F}_b(y)f(y)\mathrm{d}y} \end{cases}$$

and

$$\text{(ii)} \begin{cases} \mathbb{E}[I_b] = \frac{1}{\lambda_b \int_0^\infty \bar{F}_b(y)f(y)dy+\lambda_b\pi_0} - \mathbb{E}[B_b], \\ \mathbb{E}[I_d] = \frac{1}{\lambda_d \int_0^\infty \bar{F}_d(y)g(y)dy+\lambda_d\pi_0} - \mathbb{E}[B_d]. \end{cases}$$

Proof. (i) Consider the non-emptiness period of the inventory system. The steady-state densities of the inventory system and the demand process of X_m are given by

$$g_m(x) = \frac{g(x)}{1-\pi_0}, \qquad f_m(x) = \frac{f(x)}{1-\pi_0},$$

respectively. At the end of the non-emptiness period of the inventory system there are two disjoint ways (disjoint events) to downcross level 0+. Either level 0 is downcrossed by a negative jump or level 0+ is reached by the fluid reduction (both in X_m). The rate of the first event is $\lambda_d \int_0^\infty \bar{F}_d(y)g_m(y)\mathrm{d}y$ and by LCT the rate of the second event is $\alpha_b g_m(0)$. Since the events are disjoint, the rate of downcrossings of level 0+ is $\lambda_d \int_0^\infty \bar{F}_d(y)g_m(y)\mathrm{d}y + \alpha_b g_m(0)$. That means that the expected length of the non-emptiness period is given by $[\lambda_d \int_0^\infty \bar{F}_d(y)g_m(y)\mathrm{d}y + \alpha_b g_m(0)]^{-1}$. Thus

$$\mathbb{E}[B_b] = \frac{1-\pi_0}{\alpha_b g(0) + \lambda_d \int_0^\infty \bar{F}_d(y)g(y)\mathrm{d}y}.$$

The expression for $\mathbb{E}[B_d]$ is obtained by symmetry.
(ii) Define a *cycle* in the real process X (not the modified process X_m) as the time between two upcrossings of level 0+. By definition, the emptiness period plus the non-emptiness period is a cycle in X. That means that the expected length of the emptiness period is the expected length of the cycle minus the expected

length of the non-emptiness period. The non-emptiness period in X and in X_m are identical and the length of the expected cycle is $[\lambda_b \int_0^\infty \bar{F}_b(y)f(y)\mathrm{d}y + \lambda_b\pi_0]^{-1}$, since $\lambda_b \int_0^\infty \bar{F}_b(y)f(y)\mathrm{d}y + \lambda_b\pi_0$ is the rate of the upcrossings of level 0+. We obtain

$$\mathbb{E}[I_b] + \mathbb{E}[B_b] = \frac{1}{\lambda_b \int_0^\infty \bar{F}_b(y)f(y)dy + \lambda_b\pi_0},$$

yielding $\mathbb{E}[I_b]$. $\mathbb{E}[I_d]$ is obtained by symmetry. □

For the special case in which $\xi_b = \xi_d = \xi$ and $\alpha_b = \alpha_d = 0$, we are able to deduce that the expected steady-state inventory level $\mathbb{E}[X]$ has a simple form.

Proposition 7.3. *If $\xi_b = \xi_d = \xi$ and $\alpha_b = \alpha_d = 0$, then*

$$\mathbb{E}[X] = m/\xi, \tag{7.3}$$

where $m = \lambda_b\mathbb{E}[B] - \lambda_d\mathbb{E}[D]$.

Proof. We study the discrete-time embedding of the blood inventory process $\{X_k\}_{k\geq 1}$, where X_k denotes the blood inventory level *just before* the k^{th} arrival (either blood or demand). Suppose the process is in steady state. By the PASTA property, we have that $X_k \stackrel{d}{=} X$ for all $k \geq 1$. Also, the process $\{X_k\}_{k\geq 1}$ constitutes a Markov chain, of which the evolution is characterized by the recursion

$$X_{k+1} = (X_k + \mathbb{1}_{k,b}B_k - \mathbb{1}_{k,d}D_k) \cdot \mathrm{e}^{-\xi A_k}, \tag{7.4}$$

where $\mathbb{1}_{k,b}$ and $\mathbb{1}_{k,d}$ denote the indicator function of the event that the k^{th} arrival is a blood or demand arrival, respectively. Remark that the relation holds for both $X_k \geq 0$ and $X_k < 0$. Furthermore, B_k and D_k denote the amount of blood or demand in the k^{th} jump, respectively, and A_k denotes the interarrival time between the k^{th} and $(k+1)^{th}$ arrival. Note that A_k is the minimum of two exponentially distributed random variables with rate λ_b and λ_d, so that A_k is exponentially distributed with rate $\lambda_b + \lambda_d$. Next, we take the expectation on both sides of (7.4), which gives

$$\mathbb{E}[X_{k+1}] = \big(\mathbb{E}[X_k] + p_{k,b}\mathbb{E}[B] - p_{k,d}\mathbb{E}[D]\big)\,\mathbb{E}\big[\mathrm{e}^{-\xi A_k}\big]. \tag{7.5}$$

Here, we used independence between Poisson processes and their jump sizes, and their memoriless property, and $p_{k,b} = \lambda_b/(\lambda_b+\lambda_d)$ and $p_{k,d} = \lambda_d/(\lambda_b+\lambda_d)$ denote probability of the k^{th} jump being either a blood delivery or demand, respectively. Since $X_k \stackrel{d}{=} X$, we have $\mathbb{E}[X_{k+1}] = \mathbb{E}[X_k] = \mathbb{E}[X]$, and thus we may rewrite (7.5) as

$$\mathbb{E}[X] = \left(\mathbb{E}[X] + \frac{\lambda_b\mathbb{E}[B] - \lambda_d\mathbb{E}[D]}{\lambda_b+\lambda_d}\right) \cdot \frac{\lambda_b+\lambda_d}{\lambda_b+\lambda_d+\xi}, \tag{7.6}$$

from which we easily deduce $\mathbb{E}[X] = (\lambda_b\mathbb{E}[B] - \lambda_d\mathbb{E}[D])/\xi = m/\xi$. □

7.3.1 The exponential case

Density functions. We assume in this section that $\bar{F}_b(x) = \mathrm{e}^{-\mu_b x}$ and $\bar{F}_d(x) = \mathrm{e}^{-\mu_d x}$. Let $\rho_d := \lambda_d/\mu_d$ and $\rho_b := \lambda_b/\mu_b$ denote the expected amount of demand requested, and amount of blood delivered into the system, per time unit. Moreover, we take $\alpha_b = \alpha_d = 0$. Under these assumptions, we can solve (7.2) and (7.1) explicitly.

Equations (7.1) and (7.2) reduce to:

$$\begin{aligned}\lambda_d \int_0^v f(y)\mathrm{e}^{-\mu_d(v-y)}\mathrm{d}y + \lambda_d \mathrm{e}^{-\mu_d v} \int_0^\infty g(y)\mathrm{e}^{-\mu_d y}\mathrm{d}y \\ = \lambda_b \int_v^\infty f(y)\mathrm{e}^{-\mu_b(y-v)}\mathrm{d}y + \xi_d v f(v),\end{aligned} \tag{7.7}$$

$$\begin{aligned}\lambda_b \int_0^v g(y)\mathrm{e}^{-\mu_b(v-y)}\mathrm{d}y + \lambda_b \mathrm{e}^{-\mu_b v} \int_0^\infty f(y)\mathrm{e}^{-\mu_b y}\mathrm{d}y \\ = \lambda_d \int_v^\infty g(y)\mathrm{e}^{-\mu_d(y-v)}\mathrm{d}y + \xi_b v g(v),\end{aligned} \tag{7.8}$$

for $v > 0$. In our analysis, we concentrate on the derivation of $f(v)$. Notice that, once $f(\cdot)$ has been determined, $g(\cdot)$ follows by swapping parameters (symmetry).

In Appendix 7.A we show how the integral equations (7.7)-(7.8) can be translated into the following decoupled second order differential equations:

$$\begin{aligned}\xi_d v f''(v) + (2\xi_d - \lambda_d - \lambda_b + \mu_d \xi_d v - \mu_b \xi_d v)\, f'(v) \\ + (\mu_d \xi_d - \mu_b \xi_d - \mu_d \lambda_b + \mu_b \lambda_d - \mu_b \mu_d \xi_d v)\, f(v) = 0\end{aligned} \tag{7.9}$$

and

$$\begin{aligned}\xi_b v g''(v) + (2\xi_b - \lambda_d - \lambda_b + \mu_b \xi_d v - \mu_d \xi_b v)\, g'(v) \\ + (\mu_b \xi_b - \mu_d \xi_b - \mu_b \lambda_d + \mu_d \lambda_b - \mu_d \mu_b \xi_b v)\, g(v) = 0,\end{aligned} \tag{7.10}$$

with the additional constraint (obtained by applying the level crossing identity for level $v = 0$ in either (7.7) or (7.8)):

$$\lambda_b \int_0^\infty f(y)\mathrm{e}^{-\mu_b y}\mathrm{d}y = \lambda_d \int_0^\infty g(y)\mathrm{e}^{-\mu_d y}\mathrm{d}y. \tag{7.11}$$

Equation (7.9) describes a known type of second order differential equation, namely the *extended confluent hypergeometric equation* [195], which allows an explicit solution. A detailed deduction of the solution to (7.9) is given in Appendix 7.B, and yields the following result.

Proposition 7.4. *The probability density functions of the amount of demand X_d and the* *amount of blood present X_b are given by*

$$f(v) = \pi_d \frac{\Gamma\left(1+\frac{\lambda_b}{\xi_d}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}\right)} \frac{\mathrm{e}^{-\mu_d v} U\left(1-\frac{\lambda_d}{\xi_d}, 2-\frac{\lambda_b+\lambda_d}{\xi_d}, (\mu_b+\mu_d)v\right)}{{}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)}, \tag{7.12}$$

$$g(v) = \pi_b \frac{\Gamma\left(1+\frac{\lambda_d}{\xi_b}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_b}\right)} \frac{\mathrm{e}^{-\mu_b v} U\left(1-\frac{\lambda_b}{\xi_b}, 2-\frac{\lambda_b+\lambda_d}{\xi_b}, (\mu_b+\mu_d)v\right)}{{}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)}, \tag{7.13}$$

for $v > 0$, respectively.

Here, $\Gamma(\cdot)$ denotes the gamma function, ${}_2F_1(a,b,c,z)$ is the Gaussian hypergeometric function, defined as

$$ {}_2F_1(a,b,c,z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n\, n!} z^n \tag{7.14}$$

and $U(a,b,z)$ is Tricomi's confluent hypergeometric function, see [195],

$$U(a,b,x) = \frac{\Gamma(b-1)}{\Gamma(1+a-b)} \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} x^n + \frac{\Gamma(b-1)}{\Gamma(a)} x^{1-b} \sum_{n=0}^{\infty} \frac{(1+a-b)_n}{(2-b)_n n!} x^n, \tag{7.15}$$

in which $(a)_n$ is the Pochhammer symbol, defined as $(a)_n = a \cdot (a+1) \cdots (a+n-1)$. As a direct consequence of Proposition 7.4, we obtain expressions for the LTs $\phi(s) = \int_0^\infty \mathrm{e}^{-sv} f(v) \mathrm{d}v$ and $\gamma(s) = \int_0^\infty \mathrm{e}^{-sv} g(v) \mathrm{d}v$ for $\mathrm{Re}\, s \geq 0$ through [195, Eq. (3.2.51)], which we state here for future use.

Corollary 7.1. *The Laplace transforms for X_d and X_b, for $\mathrm{Re}\, s \geq 0$, are given by*

$$\phi(s) = \pi_d \frac{\mu_d}{\mu_d+s} \frac{{}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, \frac{s-\mu_b}{s+\mu_d}\right)}{{}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)}, \tag{7.16}$$

$$\gamma(s) = \pi_b \frac{\mu_b}{\mu_b+s} \frac{{}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, \frac{s-\mu_d}{s+\mu_b}\right)}{{}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)}, \tag{7.17}$$

respectively.

Last, we obtain expressions for π_d and π_b. These follow immediately by using the normalization equation $\pi_b + \pi_d = 1$ and (7.11), or equivalently, $\lambda_b \phi(\mu_b) = \lambda_d \gamma(\mu_d)$. By filling in $s = \mu_b$ in (7.16),

$$\begin{aligned} &\pi_d \frac{\lambda_b \mu_d}{\mu_b+\mu_d} {}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)^{-1} \\ &\qquad = \pi_b \frac{\lambda_d \mu_b}{\mu_b+\mu_d} {}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)^{-1}, \end{aligned} \tag{7.18}$$

where we used that ${}_2F_1(a,b,c,0) = 1$. Using the normalization equation, we obtain

$$\pi_d = \frac{\rho_{b,2}F_1\left(1-\frac{\lambda_d}{\xi_d},1,1+\frac{\lambda_b}{\xi_d},-\frac{\mu_b}{\mu_d}\right)}{\rho_{d,2}F_1\left(1-\frac{\lambda_d}{\xi_d},1,1+\frac{\lambda_b}{\xi_d},-\frac{\mu_b}{\mu_d}\right)+\rho_b\,{}_2F_1\left(1-\frac{\lambda_b}{\xi_b},1,1+\frac{\lambda_d}{\xi_b},-\frac{\mu_d}{\mu_b}\right)}. \tag{7.19}$$

By substituting this result into both (7.12) and (7.16), we obtain the full pdf for the blood inventory process in steady-state.

Theorem 7.1. *The steady-state pdf of the net inventory level X is given by*

$$h(v) = \begin{cases} f(-v), & \text{if } v < 0, \\ g(v), & \text{if } v \geq 0, \end{cases} \tag{7.20}$$

where

$$f(v) = \bar{C}^{-1}\,\frac{\Gamma\left(1+\frac{\lambda_b}{\xi_d}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}\right)}\,\rho_d\,\mathrm{e}^{-\mu_d v}\,U\left(1-\frac{\lambda_d}{\xi_d},2-\frac{\lambda_b+\lambda_d}{\xi_d},(\mu_b+\mu_d)v\right), \tag{7.21}$$

$$g(v) = \bar{C}^{-1}\,\frac{\Gamma\left(1+\frac{\lambda_d}{\xi_b}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_b}\right)}\,\rho_b\,\mathrm{e}^{-\mu_b v}\,U\left(1-\frac{\lambda_b}{\xi_b},2-\frac{\lambda_b+\lambda_d}{\xi_b},(\mu_b+\mu_d)v\right), \tag{7.22}$$

with

$$\bar{C} = \rho_d\,{}_2F_1\left(1-\frac{\lambda_d}{\xi_d},1,1+\frac{\lambda_b}{\xi_d},-\frac{\mu_b}{\mu_d}\right)+\rho_b\,{}_2F_1\left(1-\frac{\lambda_b}{\xi_b},1,1+\frac{\lambda_d}{\xi_b},-\frac{\mu_d}{\mu_b}\right). \tag{7.23}$$

Remark 7.1. By applying the Pfaff transformation ${}_2F_1(a,b,c,z) = (1-z)^{-b}\,{}_2F_1\left(c-a,b,c,\frac{z}{1-z}\right)$, we may reformulate

$${}_2F_1\left(1-\frac{\lambda_d}{\xi_d},1,1+\frac{\lambda_b}{\xi_d},-\frac{\mu_b}{\mu_d}\right) = \frac{\mu_d}{\mu_b+\mu_d}\,{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d},1,\frac{\lambda_b}{\xi_d},\frac{\mu_b}{\mu_b+\mu_d}\right), \tag{7.24}$$

so that

$$\pi_d = \frac{\lambda_d\,{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d},1,\frac{\lambda_b}{\xi_d},\frac{\mu_b}{\mu_b+\mu_d}\right)}{\lambda_d\,{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d},1,\frac{\lambda_b}{\xi_d},\frac{\mu_b}{\mu_b+\mu_d}\right)+\lambda_b\,{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_b},1,\frac{\lambda_d}{\xi_b},\frac{\mu_d}{\mu_b+\mu_d}\right)}. \tag{7.25}$$

By also transforming the hypergeometric term in the numerator of (7.12), we get an equivalent form of (7.21), namely

$$f(v) = \bar{C}_{\text{alt}}^{-1}\frac{\Gamma\left(1+\frac{\lambda_b}{\xi_d}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}\right)}\,\rho_b\mu_b(\mu_b+\mu_d)\,\mathrm{e}^{-\mu_d v}\,U\left(1-\frac{\lambda_d}{\xi_d},2-\frac{\lambda_b+\lambda_d}{\xi_d},(\mu_b+\mu_d)v\right), \tag{7.26}$$

with

$$\bar{C}_{\text{alt}} = \lambda_d\, {}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d}, 1, \frac{\lambda_b}{\xi_d}, \frac{\mu_b}{\mu_b+\mu_d}\right) + \lambda_b\, {}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_b}, 1, \frac{\lambda_d}{\xi_b}, \frac{\mu_d}{\mu_b+\mu_d}\right). \quad (7.27)$$

As a consequence, (7.16) is given by

$$\phi(s) = \pi_d \frac{{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d}, 1, \frac{\lambda_b}{\xi_d}, \frac{\mu_b-s}{\mu_b+\mu_d}\right)}{{}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d}, 1, \frac{\lambda_b}{\xi_d}, \frac{\mu_b}{\mu_b+\mu_d}\right)} = \bar{C}_{\text{alt}}^{-1}\, \lambda_d\, {}_2F_1\left(\frac{\lambda_b+\lambda_d}{\xi_d}, 1, \frac{\lambda_b}{\xi_d}, \frac{\mu_b-s}{\mu_b+\mu_d}\right). \quad (7.28)$$

Based on the density functions in Theorem 7.1, we make some comments on its properties, and discuss parameter settings that leads to special cases.

By close inspection of these derived density functions, we can observe the following on the distribution shape around $z = 0$. The confluent hypergeometric function $U(a,b,z)$ has limiting form as $z \to 0$,

$$U(a,b,z) = \frac{\Gamma(1-b)}{\Gamma(a-b+1)} + \frac{\Gamma(b-1)}{\Gamma(a)}\, z^{1-b} + O(z^{2-b}), \qquad b \leq 2, \quad (7.29)$$

see [174, Sub. 13.2]. Note that in our model, $b = 2 - (\lambda_b + \lambda_d)/\xi_d < 2$ for all parameter settings. Equation (7.29) shows that $U(a,b,z)$ has a singularity at $z = 0$ if $\text{Re}(b) > 1$, which in our case translates to $f(v)$ and $g(v)$ being analytic at $v = 0$ if $\lambda_b + \lambda_d > \xi_d$ and $\lambda_b + \lambda_d > \xi_b$, respectively. Assuming $\lambda_b + \lambda_d > \max\{\xi_b, \xi_d\}$, (7.29) also implies that

$$\begin{aligned}\lim_{v\to 0} f(v) &= \bar{C}^{-1} \frac{\Gamma\left(1+\frac{\lambda_b}{\xi_d}\right)}{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}\right)} \lambda_d \mu_b \cdot \frac{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}-1\right)}{\Gamma\left(\frac{\lambda_b}{\xi_d}\right)} \\ &= \bar{C}^{-1} \frac{\frac{\lambda_b}{\xi_d}}{\frac{\lambda_b+\lambda_d}{\xi_d}-1} \lambda_d \mu_b = \bar{C}^{-1} \frac{\lambda_b \lambda_d \mu_b \mu_d}{\lambda_b+\lambda_d-\xi_d}.\end{aligned} \quad (7.30)$$

Similarly,

$$\lim_{v\to 0} g(v) = \bar{C}^{-1} \frac{\lambda_b \lambda_d \mu_b \mu_d}{\lambda_b+\lambda_d-\xi_b}. \quad (7.31)$$

By equating these two expressions, we conclude that $\lim_{v\to 0} f(v) = \lim_{v\to 0} g(v) < \infty$, i.e. the overall density function $h(v)$ is continuous at $v = 0$, if and only if $\xi_b = \xi_d$. The asymptotic behavior of U as $z \to \infty$ is given by [195, p. 60],

$$U(a,b,z) \sim z^{-a}, \qquad z \to \infty, \quad (7.32)$$

which implies that the density function tail decays as

$$f(v) \sim C^* e^{-\mu_d v} v^{\lambda_d/\xi_d - 1}, \qquad v \to \infty, \quad (7.33)$$

for some constant C^*.

Special cases. Equation (7.33) suggests that the case $\lambda_d = \xi_d$ is special. Indeed, then (7.16) reduces to

$$\phi(s) = \bar{C}^{-1} \lambda_d \mu_b \frac{\mu_d}{\mu_d + s} = \pi_d \frac{\mu_d}{\mu_d + s}, \tag{7.34}$$

where we used that ${}_2F_1(0, a, b, z) = 1$ for all a, b, z. Hence, conditioned on being positive, the amount of demand present is exponentially distributed with parameter μ_d, regardless of the values of $\lambda_d = \xi_d$, as well as λ_b, ξ_b, and μ_b. If we moreover let $\lambda_b = \xi_b$, then

$$\pi_d = \frac{\lambda_d/\mu_d}{\lambda_b/\mu_b + \lambda_d/\mu_d} = \frac{\rho_d}{\rho_b + \rho_d},$$

and X has exponential distribution both above and below 0, with parameters μ_b and μ_d, respectively.

A second special case arises when the process is symmetric, that is, $\lambda_b = \lambda_d = \lambda$, $\mu_b = \mu_d = \mu$ and $\xi_b = \xi_d = \xi$. Obviously, we get $\pi_b = \pi_d = \frac{1}{2}$ due to the symmetry. If we define $\eta := \lambda/\xi$,

$$\begin{aligned} f(v) &= \frac{\Gamma(1+\eta)\, \mu e^{-\mu v}\, U\left(1-\eta, 2(1-\eta), 2\mu v\right)}{2\,\Gamma(2\eta)\, {}_2F_1\left(2\eta, 1, 1+\eta, \frac{1}{2}\right)} \\ &= \frac{\Gamma(1+\eta)}{2\,\Gamma(2\eta)\, {}_2F_1\left(2\eta, 1, 1+\eta, \frac{1}{2}\right)} \frac{\mu}{2\sqrt{\pi}} \left(2\mu v\right)^{\eta - \frac{1}{2}} K_{\frac{1}{2}-\eta}\left(\mu v\right), \end{aligned} \tag{7.35}$$

where $K_\alpha(\cdot)$ is the modified Bessel function of the second kind, see [174, Eq. (13.6.10)].

Performance measures. Based on Theorem 7.1, we can directly derive a couple of characteristics of the process. First, we consider the mean inventory level

Corollary 7.2. *The expected amount of demand (blood) present, given that it is positive equals*

$$\mathbb{E}[X_d | X_d > 0] = \frac{1}{\xi_d} \left[\rho_d - \rho_b + \rho_b \, {}_2F_1\left(1 - \frac{\lambda_d}{\xi_d}, 1, 1 + \frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)^{-1} \right], \tag{7.36}$$

$$\mathbb{E}[X_b | X_b > 0] = \frac{1}{\xi_b} \left[\rho_b - \rho_d + \rho_d \, {}_2F_1\left(1 - \frac{\lambda_b}{\xi_b}, 1, 1 + \frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)^{-1} \right]. \tag{7.37}$$

Accordingly, the expected net amount of blood present equals

$$\mathbb{E}[X] = (\rho_b - \rho_d) \left(\frac{\pi_b}{\xi_b} + \frac{\pi_d}{\xi_d} \right) + \frac{\lambda_b \lambda_d}{\bar{C}} \left(\frac{1}{\xi_b} - \frac{1}{\xi_d} \right). \tag{7.38}$$

Proof. Let us use shorthand notation

$$F(s) = \left(1 - \frac{\lambda_d}{\xi_d}, 1, 1 + \frac{\lambda_b}{\xi_d}, \frac{s - \mu_b}{s + \mu_d}\right),$$

so that

$$\phi(s) = \pi_d \frac{\mu_b}{\mu_b + s} \frac{F(s)}{F(0)}.$$

Through [174, Eq. (15.5.20)],

$$\frac{\mathrm{d}}{\mathrm{d}z} {}_2F_1(a,1,c,z) = \frac{c-1}{z(1-z)} + \frac{1-c+az}{z(1-z)} {}_2F_1(a,1,c,z), \tag{7.39}$$

where we also used that ${}_2F_1(a,1,c,z) = 1$. Then,

$$\frac{\phi'(0)}{\pi_d} = \left[\frac{-\mu_d}{(\mu_d+s)^2} \frac{F(s)}{F(0)} + \frac{\mu_d}{\mu_d+s} \frac{F'(s)}{F(0)}\right]_{s=0} = -\frac{1}{\mu_d} + \frac{F'(0)}{F(0)}.$$

By (7.39), we find

$$\begin{aligned}
F'(s) &= \left(\frac{\lambda_b/\xi_d}{\frac{s-\mu_b}{s+\mu_d}\cdot\frac{\mu_b+\mu_d}{s+\mu_d}} + \frac{-\lambda_b/\xi_d + (1-\lambda_d/\xi_d)\frac{s-\mu_b}{s+\mu_d}}{\frac{s-\mu_b}{s+\mu_d}\cdot\frac{\mu_b+\mu_d}{s+\mu_d}} F(s)\right) \frac{\mathrm{d}}{\mathrm{d}s}\left[\frac{s-\mu_b}{s+\mu_d}\right] \\
&= \left(\frac{\lambda_b}{\xi_d} + \left[\frac{-\lambda_b}{\xi_d} + \left(1 - \frac{\lambda_d}{\xi_d}\right)\frac{s-\mu_b}{s+\mu_d}\right] F(s)\right) \frac{(s+\mu_d)^2}{(s-\mu_b)(\mu_b+\mu_d)} \cdot \frac{\mu_b+\mu_d}{(s+\mu_d)^2} \\
&= \left(\frac{\lambda_b}{\xi_d} + \left[-\frac{\lambda_b}{\xi_d} + \left(1 - \frac{\lambda_d}{\xi_d}\right)\frac{s-\mu_b}{s+\mu_d}\right] F(s)\right) \frac{1}{s-\mu_b},
\end{aligned}$$

so that

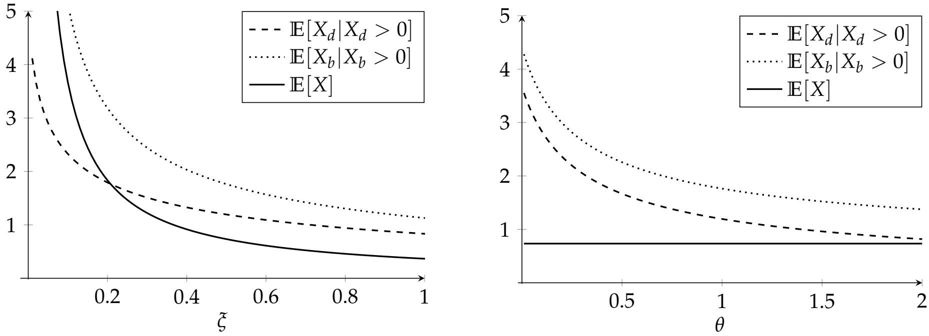
$$\begin{aligned}
F'(0) &= -\frac{\lambda_d/\mu_b}{\xi_d} + \left(\frac{\lambda_d/\mu_b}{\xi_d} + \frac{1}{\mu_d} - \frac{\lambda_d/\mu_d}{\xi_d}\right) F(0) \\
&= -\frac{\rho_b}{\xi_d} + \left(\frac{\rho_b - \rho_d}{\xi_d} + \frac{1}{\mu_d}\right) F(0).
\end{aligned}$$

Hence, we find

$$\begin{aligned}
\mathbb{E}[X_d | X_d > 0] &= -\frac{\phi'(0)}{\pi_d} = \frac{1}{\mu_d} - \frac{1}{F(0)}\left[-\frac{\rho_b}{\xi_d} + \left(\frac{\rho_b-\rho_d}{\xi_b} + \frac{1}{\mu_d}\right) F(0)\right] \\
&= \frac{1}{\xi_d}\left(\rho_d - \rho_b + \rho_b/F(0)\right) = \frac{1}{\xi_d}\left(-m + \rho_b/F(0)\right),
\end{aligned}$$

which equals (7.36). The expression for (7.37) follows by symmetry. Furthermore,

$$\begin{aligned}
\mathbb{E}[X] &= \pi_b \mathbb{E}[X_b | X_b > 0] + \pi_d \mathbb{E}[-X_d | X_d > 0] \\
&= m\left[\frac{\pi_b}{\xi_b} + \frac{\pi_d}{\xi_d}\right] + \frac{\lambda_d}{\mu_d\,\xi_b} \frac{\pi_b}{{}_2F_1\left(1 - \frac{\lambda_b}{\xi_b}, 1, 1 + \frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)} \\
&\quad - \frac{\lambda_b}{\mu_b\,\xi_d} \frac{\pi_d}{{}_2F_1\left(1 - \frac{\lambda_d}{\xi_d}, 1, 1 + \frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)}.
\end{aligned}$$



(a) As a function of ξ

(b) As a function of θ

Figure 7.2: Expected mean amount of blood, demand, and net blood present.

Note that $\pi_d\, {}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)^{-1} = \lambda_d \mu_b \bar{C}^{-1}$. Hence,

$$\mathbb{E}[X] = m\left[\frac{\pi_b}{\xi_b} + \frac{\pi_d}{\xi_d}\right] + \frac{\lambda_b \lambda_d}{\bar{C}}\left(\frac{1}{\xi_b} - \frac{1}{\xi_d}\right),$$

which completes the proof. □

Remark 7.2. Note that if $\xi_b = \xi_d = \xi$, we get $\mathbb{E}[X] = m(\pi_b + \pi_d)/\xi = m/\xi$, which is consistent with Proposition 7.3. The expression in (7.36) contains no ξ_b. Indeed, while the value of ξ_b influences the probability that $X_d > 0$, it does not influence the mean of X_d given that $X_d > 0$.

In Figure 7.2, we plot the behavior of the three performance metrics in Corollary 7.2 while keeping m fixed. In Figure 7.2(a) we set $\lambda_b = 1.2$, $\lambda_d = 1$, $\mu_b = 1$, $\mu_d = 1.2$, so that $m = 11/30$ and vary $\xi_b = \xi_d = \xi$ between 0 and 1. In Figure 7.2b, we fix $\xi_b = \xi_d = 0.5$ and take $\lambda_b = 1.2\theta$, $\lambda_d = \theta$, $\mu_b = \theta$, $\mu_d = 1.2\theta$, so that still $m = 11/30$, and vary θ. Observe that in Figure 7.2b, $\mathbb{E}[X]$ is constant, since the value of m/ξ if unaffected by the parameter θ.

Secondly, we present the probability of positive (cq. negative) inventory.

Corollary 7.3. *The probability of positive (cq. negative) inventory is given by,*

$$\pi_b = \frac{\rho_b\, {}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)}{\rho_d\, {}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right) + \rho_b\, {}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)}, \tag{7.40}$$

$$\pi_d = \frac{\rho_d\, {}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right)}{\rho_d\, {}_2F_1\left(1-\frac{\lambda_d}{\xi_d}, 1, 1+\frac{\lambda_b}{\xi_d}, -\frac{\mu_b}{\mu_d}\right) + \rho_b\, {}_2F_1\left(1-\frac{\lambda_b}{\xi_b}, 1, 1+\frac{\lambda_d}{\xi_b}, -\frac{\mu_d}{\mu_b}\right)}, \tag{7.41}$$

respectively.

Proof. The expressions follow directly from (7.19) and $\pi_b = 1 - \pi_d$. □

The last relevant performance indicator we consider is the fraction of demand that is immediately satisfied from stock.

Corollary 7.4. *The probability that a demand request can be fully satisfied from stock is given by*

$$\mathbb{P}(\text{demand satisfied}) = \bar{C}^{-1}\rho_b \left({}_2F_1\left(1 - \tfrac{\lambda_d}{\xi_d}, 1, 1 + \tfrac{\lambda_b}{\xi_d}, -\tfrac{\mu_b}{\mu_d}\right) - \frac{\mu_b}{\mu_b + \mu_d} \right). \tag{7.42}$$

Proof. Using the PASTA property of the Poisson process, we get

$$\begin{aligned}\mathbb{P}(\text{demand satisfied}) &= \mathbb{P}(X > D) = \mathbb{P}(X_b > D) \\ &= \int_0^\infty g(u)(1 - \mathrm{e}^{-\mu_d u})\,\mathrm{d}u = \pi_b - \gamma(\mu_d).\end{aligned}$$

Substituting the expressions for π_b as in Corollary 7.3 and $\gamma(\mu_b)$ as in (7.17) yields the result. □

7.3.2 The general case

In this section we outline how the integral equations (7.1) and (7.2) can be solved using Laplace transforms, when we make the restriction that $F_b(\cdot)$ and $F_d(\cdot)$ are Coxian distributions. This is not a major restriction, because the class of Coxian distributions lies dense in the class of all distributions of non-negative random variables, see e.g. [20, Sec. III.4]. Hence, one can approximate $F_b(\cdot)$ arbitrarily closely by a Coxian distribution.

If X_i, $i = 1, 2, \ldots, K$ are independent, exponentially distributed random variables, and $\mathbb{E}[X_i] = \frac{1}{\beta_i}$, $i = 1, 2, \ldots, K$, then a Coxian amount of blood B can be represented as:

$$B = \sum_{j=1}^{i} X_j \quad \text{with probability} \quad p_i \prod_{j=1}^{i-1}(1 - p_j), \quad i = 1, 2, \ldots, K. \tag{7.43}$$

In the above case, it is easily verified that one can represent $\bar{F}_b(x)$ as follows:

$$\bar{F}_b(x) = \mathbb{P}(B > x) = \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1 - p_h) \sum_{j=1}^{i} \mathrm{e}^{-\beta_j x} \prod_{l=1; l\neq j}^{i} \frac{\beta_l}{\beta_l - \beta_j}, \tag{7.44}$$

if all β_j are different. If two β_j coincide, then a term with $x\mathrm{e}^{-\beta_j x}$ (Erlang-2) must be added. We leave this to the reader, but in Remark 7.8 below we outline how Erlang terms can be handled in solving the integral equations (7.2) and (7.1). The counterpart of (7.44) for the case that $F_d(\cdot)$ is Coxian, is

$$\bar{F}_d(x) = \mathbb{P}(D > x) = \sum_{i=1}^{L} q_i \prod_{h=1}^{i-1}(1 - q_h) \sum_{j=1}^{i} \mathrm{e}^{-\delta_j x} \prod_{l=1; l\neq j}^{i} \frac{\delta_l}{\delta_l - \delta_j}. \tag{7.45}$$

Taking Laplace transforms $\phi(s) = \int_0^\infty e^{-sy} f(y) dy$ and $\gamma(s) = \int_0^\infty e^{-sy} g(y) dy$ in (7.1) and (7.2) results in first-order inhomogeneous differential equations in $\phi(s)$ and $\gamma(s)$, respectively, which can be solved in a straightforward way.

$$\phi'(s) = A_H(s)\phi(s) + A_I(s), \tag{7.46}$$

with the homogeneous term $A_H(s)$ being given by

$$\begin{aligned} A_H(s) := -\frac{1}{\xi_d} \Bigg[& \lambda_d \sum_{i=1}^{L} q_i \prod_{h=1}^{i-1} (1 - q_h) \sum_{j=1}^{i} \frac{1}{\delta_j + s} \prod_{l=1; l \neq j}^{i} \frac{\delta_l}{\delta_l - \delta_j} \\ & - \lambda_b \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1} (1 - p_h) \sum_{j=1}^{i} \frac{1}{\beta_j - s} \prod_{l=1; l \neq j}^{i} \frac{\beta_l}{\beta_l - \beta_j} - \alpha_d \Bigg], \end{aligned} \tag{7.47}$$

and the inhomogeneous term $A_I(s)$ being given by

$$\begin{aligned} A_I(s) := -\frac{1}{\xi_d} \Bigg[& \lambda_d \sum_{i=1}^{L} q_i \prod_{h=1}^{i-1} (1 - q_h) \sum_{j=1}^{i} \frac{1}{\delta_j + s} [\gamma(\delta_j) + \pi_0] \prod_{l=1; l \neq j}^{i} \frac{\delta_l}{\delta_l - \delta_j} \\ & + \lambda_b \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1} (1 - p_h) \sum_{j=1}^{i} \frac{1}{\beta_j - s} \phi(\beta_j) \prod_{l=1; l \neq j}^{i} \frac{\beta_l}{\beta_l - \beta_j} \Bigg]. \end{aligned} \tag{7.48}$$

The solution of (7.46) is given by the following expression:

$$\phi(s) = \phi(0) e^{\int_0^s A_H(z) dz} + \int_0^s A_I(u) e^{\int_u^s A_H(z) dz} du, \quad s \geq 0. \tag{7.49}$$

$\gamma(s)$ is given by a mirror expression, where $\phi(0)$ is replaced by $\gamma(0)$ and where $A_H(s)$ and $A_I(s)$ are replaced by expressions in which K and L are swapped, and p and q, and β_i and δ_i.

It should be noticed that $\phi(0)$, $\gamma(0)$ and π_0 still have to be determined. Furthermore, it should be noticed that $A_H(s)$ and $A_I(s)$ have singularities at $s = \beta_1, \ldots, \beta_K$. These singularities are removable, but handling Equation (7.49) clearly requires some care. Instead of working out the details, we shall below return to the case of exponentially distributed amounts of blood and demand – so $K = L = 1$. For that case, we shall not only work out the solution of the differential equation for $\phi(s)$ in detail, including the determination of the missing constants, but we also relate the results to those obtained in Section 7.3.1 without resorting to Laplace transforms. Taking $K = 1, p_1 = 1, \delta_1 = \mu_d$, and $L = 1, q_1 = 1, \beta_1 = \mu_b$, we obtain the following two inhomogeneous first order differential equations in the LTs $\phi(s)$ and $\gamma(s)$:

$$\phi'(s) = \phi(s) \left[\frac{\lambda_b}{\xi_d} \frac{1}{\mu_b - s} - \frac{\lambda_d}{\xi_d} \frac{1}{\mu_d + s} \right] - \frac{\lambda_b}{\xi_d} \frac{\phi(\mu_b)}{\mu_b - s} - \frac{\lambda_d}{\xi_d} \frac{\gamma(\mu_d)}{\mu_d + s}, \tag{7.50}$$

$$\gamma'(s) = \gamma(s) \left[\frac{\lambda_d}{\xi_b} \frac{1}{\mu_d - s} - \frac{\lambda_b}{\xi_b} \frac{1}{\mu_b + s} \right] - \frac{\lambda_d}{\xi_b} \frac{\gamma(\mu_d)}{\mu_d - s} - \frac{\lambda_b}{\xi_b} \frac{\phi(\mu_b)}{\mu_b + s}. \tag{7.51}$$

They are routinely solved:

$$\begin{aligned}
\phi(s) = \left(\frac{\mu_b}{\mu_b - s}\right)^{\frac{\lambda_b}{\xi_d}} \left(\frac{\mu_d}{\mu_d + s}\right)^{\frac{\lambda_d}{\xi_d}} \Bigg[\phi(0) \\
- \frac{\lambda_d}{\xi_d}\gamma(\mu_d) \int_0^s \left(\frac{\mu_b - z}{\mu_b}\right)^{\frac{\lambda_b}{\xi_d}} \left(\frac{\mu_d + z}{\mu_d}\right)^{\frac{\lambda_d}{\xi_d}-1} \frac{\mathrm{d}z}{\mu_d} \\
- \frac{\lambda_b}{\xi_d}\phi(\mu_b) \int_0^s \left(\frac{\mu_b - z}{\mu_b}\right)^{\frac{\lambda_b}{\xi_d}-1} \left(\frac{\mu_d + z}{\mu_d}\right)^{\frac{\lambda_d}{\xi_d}} \frac{\mathrm{d}z}{\mu_b} \Bigg].
\end{aligned} \tag{7.52}$$

Similarly,

$$\begin{aligned}
\gamma(s) = \left(\frac{\mu_d}{\mu_d - s}\right)^{\frac{\lambda_d}{\xi_b}} \left(\frac{\mu_b}{\mu_b + s}\right)^{\frac{\lambda_b}{\xi_b}} \Bigg[\gamma(0) \\
- \frac{\lambda_b}{\xi_b}\phi(\mu_b) \int_0^s \left(\frac{\mu_d - z}{\mu_d}\right)^{\frac{\lambda_d}{\xi_b}} \left(\frac{\mu_b + z}{\mu_b}\right)^{\frac{\lambda_b}{\xi_b}-1} \frac{\mathrm{d}z}{\mu_b} \\
- \frac{\lambda_d}{\xi_b}\gamma(\mu_d) \int_0^s \left(\frac{\mu_d - z}{\mu_d}\right)^{\frac{\lambda_d}{\xi_b}-1} \left(\frac{\mu_b + z}{\mu_b}\right)^{\frac{\lambda_b}{\xi_b}} \frac{\mathrm{d}z}{\mu_d} \Bigg].
\end{aligned} \tag{7.53}$$

Notice that the exponents in the above integrals have powers which are larger than -1 (e.g., $\frac{\lambda_d}{\xi_d} - 1$), so that these integrals do not lead to singularities. We still need to determine the two constants $\phi(0) = \pi_d$ and $\gamma(0) = \pi_b$. Together with $\phi(\mu_b)$ and $\gamma(\mu_d)$, we have four unknowns. We determine these unknowns using the following four equations: (i) From (7.11), we get $\lambda_b\phi(\mu_b) = \lambda_d\gamma(\mu_d)$, while (ii) $\pi_d + \pi_b = 1$. Finally, we take (iii) $s = \mu_b$ in (7.52) and (iv) $s = \mu_d$ in (7.53).

Notice that the identity $\lambda_b\phi(\mu_b) = \lambda_d\gamma(\mu_d)$ allows us to reduce the two integrals in (7.52) to one integral (and similarly in (7.53)):

$$\begin{aligned}
\phi(s) = \left(\frac{\mu_b}{\mu_b - s}\right)^{\frac{\lambda_b}{\xi_d}} \left(\frac{\mu_d}{\mu_d + s}\right)^{\frac{\lambda_d}{\xi_d}} \Bigg[\phi(0) \\
- \frac{\lambda_d}{\xi_d}\gamma(\mu_d) \frac{\mu_b + \mu_d}{\mu_b\mu_d} \int_0^s \left(\frac{\mu_b - z}{\mu_b}\right)^{\frac{\lambda_b}{\xi_d}-1} \left(\frac{\mu_d + z}{\mu_d}\right)^{\frac{\lambda_d}{\xi_d}-1} \mathrm{d}z \Bigg].
\end{aligned} \tag{7.54}$$

Remark 7.3. We have numerically verified that the expressions in (7.52) and (7.16) coincide.

Remark 7.4. If $\lambda_b = 0$ then we have a known queueing model or shot-noise model with state-dependent service rate, see Keilson & Mermin [132] and Bekker et al. [31] for the so-called shot noise model.

Remark 7.5. The case $\lambda_d = \xi_d$ is special. Equation (7.52) now reduces to

$$\phi(s) = \left(\frac{\mu_b}{\mu_b - s}\right)^{\frac{\lambda_b}{\lambda_d}} \frac{\mu_d}{\mu_d + s} \left[\phi(0) \;\; - \gamma(\mu_d) \int_0^s \left(\frac{\mu_b - z}{\mu_b}\right)^{\frac{\lambda_b}{\lambda_d}} \frac{\mathrm{d}z}{\mu_d} \right.$$
$$\left. - \frac{\lambda_b}{\lambda_d} \phi(\mu_b) \int_0^s \left(\frac{\mu_b - z}{\mu_b}\right)^{\frac{\lambda_b}{\lambda_d} - 1} \frac{\mu_d + z}{\mu_d} \frac{\mathrm{d}z}{\mu_b} \;\; \right].$$

Both integrals are easily evaluated (rewrite, in the last integral, $\mu_d + z = \mu_d + \mu_b - (\mu_b - z)$). We find

$$\begin{aligned}\phi(s) = {} & \left(\frac{\mu_b}{\mu_b - s}\right)^{\frac{\lambda_b}{\lambda_d}} \frac{\mu_d}{\mu_d + s} \\ & \cdot \left[\phi(0) + \frac{\gamma(\mu_d)}{\mu_d} \frac{\lambda_d}{\lambda_b + \lambda_d} \mu_b - \phi(\mu_b) \frac{\mu_d + \mu_b}{\mu_d} - \frac{\phi(\mu_b)}{\mu_d} \frac{\lambda_b}{\lambda_b + \lambda_d} \mu_b \right] \\ & + \frac{\mu_d}{\mu_d + s} \left[\frac{\gamma(\mu_d)}{\mu_d} \frac{\lambda_d}{\lambda_b + \lambda_d} (\mu_b - s) + \phi(\mu_b) \frac{\mu_d + \mu_b}{\mu_d} - \frac{\phi(\mu_b)}{\mu_d} \frac{\lambda_b}{\lambda_b + \lambda_d} (\mu_b - s)\right].\end{aligned} \tag{7.55}$$

Now observe through (7.11), that $\lambda_b \phi(\mu_b) = \lambda_d \gamma(\mu_d)$. Hence, in both lines of the above formula, two terms cancel. Moreover, $\phi(s)$ should be analytic for $s = \mu_b$, yielding

$$\phi(0) = \phi(\mu_b) \frac{\mu_d + \mu_b}{\mu_d}. \tag{7.56}$$

Finally we obtain, see also (7.34),

$$\phi(s) = \frac{\mu_d}{\mu_d + s} \phi(\mu_b) \frac{\mu_d + \mu_b}{\mu_d} = \phi(0) \frac{\mu_d}{\mu_d + s} = \pi_d \frac{\mu_d}{\mu_d + s}, \tag{7.57}$$

and hence

$$f(x) = \pi_d \mu_d \mathrm{e}^{-\mu_d x}, \quad x > 0; \tag{7.58}$$

the shortage (amount of demand present) is exponentially distributed when $\lambda_d = \xi_d$.

It should be noticed that, if $\lambda_d = \xi_d$, then the first and last term of (7.7) are equal when (7.58) holds; and using (7.11) it is also readily verified that the second and third term of (7.7) are equal. The constant π_d will in general still depend on the parameters $\lambda_d = \xi_d$, λ_b, μ_b and ξ_b.

We end this remark with the observation that in the one-sided shot-noise process (so $\lambda_b = 0$), Bekker et al. [31] also observe that $\lambda_d = \xi_d$ results in an exponential density.

7.3.3 A variant

In this section, we assume that the expiration rate of blood and the patience rate of demand are constant. So, we take $\xi_b = \xi_d = 0$. A visualization of a possible sample path is depicted in Figure 7.3.

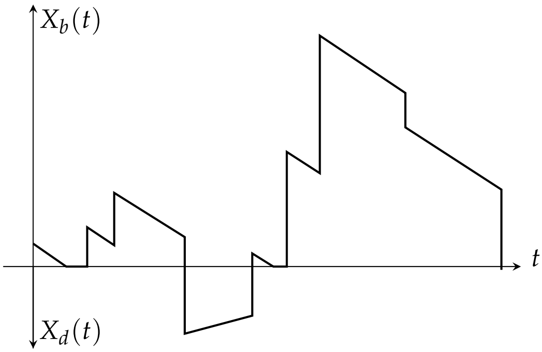


Figure 7.3: Sample path of the net amount of blood present if $\xi_b = \xi_d = 0$.

We again restrict ourselves to the case of exponentially distributed amounts of demand and of blood deliveries. We now need to impose stability conditions. In the case of positive demand, the drift is towards zero if $\lambda_d \mathbb{E}[D] < \alpha_d + \lambda_b \mathbb{E}[B]$, while in the case of a positive amount of blood, the drift is towards zero if $\lambda_b \mathbb{E}[B] < \alpha_b + \lambda_d \mathbb{E}[D]$. If these two conditions are violated, either the amount of demand or the amount of blood present increases without bound (see also below). In this case, (7.9) reduces to

$$\alpha_d f''(v) + (-\lambda_d - \lambda_b + \mu_d \alpha_d - \mu_b \alpha_d) f'(v) + (-\mu_d \lambda_b + \mu_b \lambda_d - \mu_b \mu_d \alpha_d) f(v) = 0. \tag{7.59}$$

Hence $f(\cdot)$ is a mixture of two exponential terms: $f(v) = R_+ \mathrm{e}^{-x_+ v} + R_- \mathrm{e}^{-x_- v}$, where x_+ and x_- are the positive and negative root of the equation

$$\alpha_d x^2 - (\mu_d \alpha_d - \mu_b \alpha_d - \lambda_d - \lambda_b) x + (-\mu_d \lambda_b + \mu_b \lambda_d - \mu_b \mu_d \alpha_d) = 0. \tag{7.60}$$

Notice that the last term in the left-hand side of (7.60) is negative if the stability condition $\lambda_d \mathbb{E}[D] < \alpha_d + \lambda_b \mathbb{E}[B]$ holds, that is, if $\mu_b \lambda_d < \mu_d \lambda_b + \mu_b \mu_d \alpha_d$, thus guaranteeing that the product of the two roots x_+ and x_- is negative, and hence that there is a positive and a negative root. One should subsequently observe that R_- must be zero to have a probability density. Hence $f(v)$ is simply (a constant times) an exponential; similarly for $g(v)$. In addition, the steady-state amounts of demand and of blood have an atom at 0 (since ξ_d and ξ_b are no longer zero, the demand and blood processes can reach 0).

Interestingly, the model of this section is closely related to the model with workload removal that is considered in [44]. There an $M/G/1$ queue is studied with the extra feature that, at Poisson epochs, a stochastic amount of work is removed. In the $M/M/1$ case with removal of exponential amounts of work, see [44, Sec. 5.1], one has the model of the present section when we concentrate on the amount of demand present. One difference with the model in [44] is that, when the workload in that model has become zero, the work becomes positive at rate λ_d, whereas in the present model the amount of blood can become positive (so zero demand is present) and the amount of demand does not have to become positive when demands arrive (because they are immediately satisfied, see Figure 7.3). So the atom at zero is in the present model larger than in the model of [44]. In our model a positive demand level may be reached from below zero (by a jump, i.e., a demand arriving at an epoch that there is some, but not enough, blood present). The memoryless property of the exponential demand requirement distribution implies that this jump results in a demand level that is $\exp(\mu_d)$, just as if the initial demand level had been zero. In the case of non-exponential demand requirements, our model becomes equivalent with an $M/G/1$ queue with exponential amounts of work removed, and with the special feature that the first service requirement of a busy period has a different distribution. Lemmas 4.1 and 4.2 of [44] present the stability condition of that $M/G/1$ queue with work removal; it amounts to $\lambda_d \mathbb{E}[D] < \alpha_d + \lambda_b \mathbb{E}[B]$, which indeed is one of the two stability conditions of the present demand/blood model.

Finally we observe that Equation (5.1) of [44] coincides with (7.60) (take $\alpha_d = 1$, $\lambda_d = \lambda_+$, $\lambda_b = \lambda_-$, $\mu_d = 1/\beta$ and $\mu_b = 1/\gamma$).

7.4 Asymptotic analysis

We finally study the model with $\alpha_b = \alpha_d = 0$ from an asymptotic perspective, by obtaining the fluid and diffusion limits of the blood inventory process. That is, we will create a sequence of processes, indexed by $n = 1, 2, \ldots$, in which we let the rates of blood and demand arrivals grow large. If we then scale the process in a proper manner, we are able to deduce a non-degenerate limiting process, that provides insight in the overall behavior of the arrival volume when the system grows large, which only relies on the first two moments of the blood and demand distributions.

7.4.1 Identification of the limiting process

First, we introduce some additional notation. Let $X_b(t)$ and $X_d(t)$ denote the amount of blood and demand, respectively, at time $t > 0$. Let

$$X(t) := X_b(t) - X_d(t), \tag{7.61}$$

be the net amount of blood available at time t. Remember that $X_b(t), X_d(t) \geq 0$, and $X_b(t) > 0$ *or* $X_d(t) > 0$ for all t, since $\alpha_d = \alpha_b = 0$. Let $N_b(t)$, $N_d(t)$ be the two independent Poisson processes counting the number of arrivals of blood and demand, respectively. Then the following integral representation holds for $X(t)$,

$$X(t) = X(0) - \xi_b \int_0^t X_b(s)\, ds + \xi_d \int_0^t X_d(s)\, ds + \sum_{i=1}^{N_b(t)} B_i - \sum_{i=1}^{N_d(t)} D_i. \tag{7.62}$$

For the sake of exhibition, we will concentrate on the case $\xi_b = \xi_d =: \xi$. Our analysis may be extended to the general case. A sketch of this generalization is given at the end of this section without going into the technical difficulties that arise when rigorously proving these limits.

Define

$$Z(t) = \sum_{i=1}^{N_b(t)} B_i - \sum_{i=1}^{N_d(t)} D_i, \tag{7.63}$$

that is, the difference between two compound Poisson processes, so that (7.62) reduces to

$$X(t) = X(0) - \xi \int_0^t X(s)\,\mathrm{d}s + Z(t). \tag{7.64}$$

The first step in the definition of the sequence of processes under investigation is defining the asymptotic scheme we are interested in. As mentioned above, we intend to let the arrival rates grow to infinity. Therefore, in the n^{th} process $X_n(t)$, we replace the rates of the arrival processes by $n\lambda_b$ and $n\lambda_d$. This induces Poisson processes $N_b^{(n)}(t)$ and $N_d^{(n)}(t)$ with arrival rates $n\lambda_b$ and $n\lambda_d$, respectively. However, we have

$$N_b^{(n)}(t) \stackrel{d}{=} N_b(nt) \qquad \text{and} \qquad N_d^{(n)}(t) \stackrel{d}{=} N_d(nt), \tag{7.65}$$

so that the term $Z(t)$ in (7.64) in this asymptotic scheme can be replaced by

$$Z_n(t) = \sum_{i=1}^{N_b(nt)} B_i - \sum_{i=1}^{N_d(nt)} D_i. \tag{7.66}$$

The first step in our analysis is obtaining the fluid limit of the process. Bearing in mind application of the Functional Law of Large Numbers (FLLN), we scale the process as $\bar{X}_n(t) = X_n(t)/n$, so that with (7.64)

$$\bar{X}_n(t) = \bar{X}_n(0) - \xi \int_0^t \bar{X}_n(s)\,ds + \bar{Z}_n(t), \tag{7.67}$$

where $\bar{Z}_n(t) = Z_n(t)/n$.

The essential step in establishing a result on the convergence of $\bar{X}_n$ is the application of [176, Thm. 4.1], which we cite here for completeness, slightly rewritten to fit our setting.

Theorem 7.2 ([176, Thm. 4.1])**.** *Let $D[0,\infty)$ be the space of all one-dimensional real-valued càdlàg functions defined on $[0,\infty)$, endowed with the usual J_1-Skorohod topology. Consider the integral representation*

$$x(t) = y(t) + \int_0^t u(x(s))\,ds, \qquad t \geq 0, \tag{7.68}$$

where $u : \mathbb{R} \to \mathbb{R}$ satisfies $u(0) = 0$ and is Lipschitz continuous. The integral representation in (7.68) *has a unique solution x, so that the integral representation constitutes a function $H_u : D[0,\infty) \to D[0,\infty)$ mapping y into $x \equiv H_u(y)$. In addition, the function H_u is continuous, and if y is continuous, then so is x.*

In our case, we set $u(x) = -\xi x$, to be able to write $\bar{X}_n = H_u\left(\bar{X}_n(0) + \bar{Z}_n\right)$. Since u is clearly Lipschitz continuous, the mapping H_u is indeed continuous. Let us rewrite (7.67), by observing

$$\mathbb{E}\bar{Z}_n(t) = \frac{1}{n}\Big(\mathbb{E}[N_b(nt)]\mathbb{E}[B] - \mathbb{E}[N_d(nt)]\mathbb{E}[D]\Big) = \lambda_b\mathbb{E}[B]t - \lambda_d\mathbb{E}[D]t, \tag{7.69}$$

where the expectation is taken with respect to the compound Poisson processes. Since $m = \lambda_b\mathbb{E}[B] - \lambda_d\mathbb{E}[D]$,

$$\bar{X}_n(t) = \bar{X}_n(0) - \xi\int_0^t \left(\bar{X}_n(s) - \frac{m}{\xi}\right) \mathrm{d}s + \bar{Y}_n(t), \tag{7.70}$$

where $\bar{Y}_n(t) := \bar{Z}_n(t) - mt$ is now a centered process. This allows us to state the next result.

Proposition 7.5 (Fluid limit). *Let $\mathbb{E}[B]$, $\mathbb{E}[D] < \infty$ and $\bar{X}_n(0) = X_n(0)/n \to q_0 \in \mathbb{R}$, as $n \to \infty$. Then for $n \to \infty$,*

$$\bar{X}_n \overset{d}{\Rightarrow} q, \tag{7.71}$$

where

$$q(t) = \frac{m}{\xi} + \left(q_0 - \frac{m}{\xi}\right)\mathrm{e}^{-\xi t}. \tag{7.72}$$

Proof. First, we concentrate on the process $\bar{Y}_n$. Observe that, by the FLLN for renewal-reward processes, which follows from [221, Thm. 7.4.1], we have

$$\frac{1}{nt}\sum_{i=1}^{N_b(nt)} B_i \overset{d}{\Rightarrow} \lambda_b\mathbb{E}[B], \qquad \frac{1}{nt}\sum_{i=1}^{N_d(nt)} D_i \overset{d}{\Rightarrow} \lambda_d\mathbb{E}[D], \tag{7.73}$$

for $n \to \infty$ and for all $t > 0$. Hence, $\bar{Z}_n(t) \overset{d}{\Rightarrow} \lambda_b\mathbb{E}[B]t - \lambda_d\mathbb{E}[D]t = mt$. By definition of $\bar{Y}_n$ and the assumption of convergence of $\bar{X}_n(0)$, this implies

$$\bar{Y}_n + \bar{X}_n \overset{d}{\Rightarrow} q_0 \tag{7.74}$$

as $n \to \infty$. Next, note $\bar{X}_n = H_u\left(\bar{X}_n(0) + \bar{Z}_n\right) = H_u\left(\bar{X}_n(0) + \bar{Y}_n + It\right)$, where I denotes the identity map, i.e. $I(t) = t$ for all $t \geq 0$. Due to Lipschitz continuity of u, H_u constitutes a continuous mapping, and hence we can apply the Continuous Mapping Theorem (CMT), to find

$$\bar{X}_n = H_u\left(\bar{X}_n(0) + \bar{Y}_n + mI\right) \Rightarrow H_u\left(q_0 + mI\right) \equiv q, \tag{7.75}$$

for all $t \geq 0$, where $q(\cdot)$ is the solution of

$$\begin{aligned} q(t) &= q_0 + \int_0^t u(q(s))\,\mathrm{d}s = q_0 + mt - \xi\int_0^t q(s)\,\mathrm{d}s \\ &= q_0 - \xi\int_0^t \left(q(s) - \frac{m}{\xi}\right)\mathrm{d}s. \end{aligned}$$

The unique solution of this integral equation is given in (7.72). □

According to Proposition 7.5, the fluid limit approaches $\mathbb{E}[X] = m/\xi$ exponentially fast. To obtain an expression for the *diffusion limit* of the process, we analyze the fluctuations of the process around the fluid limit in (7.71), again by scaling the process in a proper manner. First, we subtract $q(t)$ on both sides of (7.70), and multiply by $\sqrt{n}$:

$$\sqrt{n}\,(\bar{X}_n(t) - q(t)) = \sqrt{n}\,(\bar{X}_n(0) - q_0) - \xi \int_0^t \sqrt{n}\,(\bar{X}_n(s) - q(s))\; ds + \sqrt{n}\,\bar{Y}_n(t). \tag{7.76}$$

Let $\hat{X}_n \equiv \sqrt{n}\,(\bar{X}_n - q)$ and $\hat{Y}_n \equiv \sqrt{n}\,\bar{Y}_n$, then this reduces to

$$\hat{X}_n(t) = \hat{X}_n(0) - \xi \int_0^t \hat{X}_n(s)\, ds + \hat{Y}_n(t). \tag{7.77}$$

Again the process $\hat{Y}_n$ needs special attention.

Lemma 7.1. *Let $\mathbb{E}[B^2], \mathbb{E}[D^2] < \infty$. Then $\hat{Y}_n \overset{d}{\Rightarrow} \sigma W$ as $n \to \infty$, where $\sigma^2 := \lambda_b \mathbb{E}[B^2] + \lambda_d \mathbb{E}[D^2]$ and W is a standard Brownian motion.*

Proof. Recall that

$$\hat{Y}_n(t) \overset{d}{=} \sqrt{n}\left[\Big(\frac{1}{n}\sum_{i=1}^{N_b(nt)} B_i - \lambda_b \mathbb{E}[B]t\Big) - \Big(\frac{1}{n}\sum_{i=1}^{N_d(nt)} D_i - \lambda_d \mathbb{E}[D]t\Big)\right]. \tag{7.78}$$

By the Functional Central Limit Theorem (FCLT) for renewal-reward processes given in [221, Thm. 7.4.1], the process

$$\hat{Y}_n^b(t) = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{N_b(nt)} B_i - \lambda_b \mathbb{E}[B]t\Big), \tag{7.79}$$

converges weakly to $\sigma_b W_b$, where W_b is a standard Brownian motion, and

$$\sigma_b^2 = \lambda_b \operatorname{Var} B + \lambda_b (E[B])^2 = \lambda_b \mathbb{E}[B^2]. \tag{7.80}$$

Similarly, $\hat{Y}_n^d \Rightarrow \sigma_d W_d$ as $n \to \infty$, with the obvious parameter switches and W_d is standard Brownian motion. Since the processes $\hat{Y}_n^b$ and $\hat{Y}_n^d$ are independent, so are their limits, and

$$\hat{Y}_n \Rightarrow \sqrt{\lambda_b \mathbb{E}[B^2]}\, W_b + \sqrt{\lambda_d \mathbb{E}[D^2]}\, W_d \overset{d}{=} \sqrt{\lambda_b \mathbb{E}[B^2] + \lambda_d \mathbb{E}[D^2]}\, W, \tag{7.81}$$

for $n \to \infty$ and W a standard Brownian motion. □

Now, we are ready to prove the diffusion counterpart of Proposition 7.5.

Proposition 7.6 (Diffusion limit). *Let $\mathbb{E}[B^2], \mathbb{E}[D^2] < \infty$. If $\hat{X}_n(0) \to \hat{X}(0)$, then $\hat{X}_n \Rightarrow \hat{X}$ as $n \to \infty$, where $\hat{X}$ satisfies the integral equation*

$$\hat{X}(t) = \hat{X}(0) - \xi \int_0^t \hat{X}(s)\, \mathrm{d}s + \sigma W(t). \tag{7.82}$$

In other words, $\hat{X}$ is an Ornstein-Uhlenbeck diffusion process with infinitesimal mean ξ and infinitesimal variance $\sigma^2 := \lambda_b \mathbb{E}[B^2] + \lambda_d \mathbb{E}[D^2]$.

Proof. We again rely on the result that the mapping H_u as in the proof of Proposition 7.5 is continuous if u is Lipschitz continuous. Here, we set $u(x) = -\xi x$ which again clearly satisfies this condition. We have $\hat{X}_n \equiv H_u(\hat{X}_n(0) + \hat{Y}_n)$. From Lemma 7.1, we know

$$\hat{X}_n(0) + \hat{Y}_n \Rightarrow \hat{X}(0) + \sigma W, \tag{7.83}$$

for $n \to \infty$. As a consequence of the CMT, we conclude

$$\hat{X}_n = H_u\left(\hat{X}_n(0) + \hat{Y}_n\right) \Rightarrow H_u\left(\hat{X}(0) + \sigma W\right) \equiv \hat{X}, \tag{7.84}$$

where $\hat{X}$ solves (7.82). $\square$

7.4.2 Generalization for $\xi_b \neq \xi_d$

We now sketch the scaling approach towards fluid and diffusion limits for the general case in which ξ_b may differ from ξ_d. In case $\xi_b \neq \xi_d$, the integral equation for $\bar{X}_n$ as in (7.67) becomes

$$\begin{aligned} \bar{X}_n(t) &= \bar{X}_n(0) + \int_0^t (-\xi_b \bar{X}_n^+(s) + \xi_d \bar{X}_n^-(s) - m)\, \mathrm{d}s + \bar{Y}_n(t) \\ &= \bar{X}_n(0) - \int_0^t (\left[\xi_b \mathbb{1}_{\{\bar{X}_n(s)\geq 0\}} + \xi_d \mathbb{1}_{\{\bar{X}_n(s)<0\}}\right] \bar{X}_n(s) + m)\, \mathrm{d}s + \bar{Y}_n(t), \end{aligned} \tag{7.85}$$

where $\bar{Y}_n(t)$ is defined as before. Note that $\hat{X}_n \equiv H_u(\bar{X}_n(0) + \bar{Y}_n)$, where we now have

$$u(x) = -\left[\xi_b \mathbb{1}_{\{x\geq 0\}} + \xi_d \mathbb{1}_{\{x<0\}}\right] x + m, \tag{7.86}$$

which is still Lipschitz continuous. Following the same reasoning of the proof of Proposition 7.5, we obtain the fluid limit $\bar{X}_n \overset{d}{\Rightarrow} q$, where q is the solution of

$$q(t) = q_0 - \int_0^t (\left[\xi_b \mathbb{1}_{\{q(s)\geq 0\}} + \xi_d \mathbb{1}_{\{q(s)<0\}}\right] q(s) - m)\, \mathrm{d}s. \tag{7.87}$$

The solution to this integral equation is more elaborate than (7.71) and depends on the sign of m and q_0. Assuming $m \geq 0$, one can check that,

$$q(t) = \frac{m}{\xi_b} + \left(q_0 - \frac{m}{\xi_b}\right) e^{-\xi_b t}, \qquad \text{if } q_0 \geq 0, \tag{7.88}$$

$$q(t) = \begin{cases} \frac{m}{\xi_d} + \left(q_0 - \frac{m}{\xi_d}\right) e^{-\xi_d t}, & \text{if } 0 \leq t < t_d^*, \\ \frac{m}{\xi_b}\left(1 - e^{-\xi_b (t - t_d^*)}\right), & \text{if } t \geq t_d^*, \end{cases} \qquad \text{if } q_0 < 0, \tag{7.89}$$

where

$$t_d^* = -\frac{1}{\xi_d} \log\left(\frac{m/\xi_d}{m/\xi_d - q_0}\right). \tag{7.90}$$

If $m < 0$,

$$q(t) = \frac{m}{\xi_d} + \left(q_0 - \frac{m}{\xi_d}\right) e^{-\xi_d t}, \qquad \text{if } q_0 \leq 0, \tag{7.91}$$

$$q(t) = \begin{cases} \frac{m}{\xi_b} + \left(q_0 - \frac{m}{\xi_b}\right) e^{-\xi_b t}, & \text{if } 0 \leq t < t_b^*, \\ \frac{m}{\xi_d}\left(1 - e^{-\xi_d (t - t_b^*)}\right), & \text{if } t \geq t_b^*, \end{cases} \qquad \text{if } q_0 > 0, \tag{7.92}$$

where

$$t_b^* = -\frac{1}{\xi_b} \log\left(\frac{m/\xi_b}{m/\xi_b - q_0}\right). \tag{7.93}$$

Note that the equilibrium of the fluid limit also depends on the sign of m:

$$\lim_{t\to\infty} q(t) = \begin{cases} m/\xi_b, & \text{if } m \geq 0, \\ m/\xi_d, & \text{if } m < 0. \end{cases} \tag{7.94}$$

In the remainder, without loss of generality $m \geq 0$. Furthermore, set $q_0 = m/\xi_b$ so that $q \equiv m/\xi_b$. Subtracting $q(t)$ on both sides of (7.85) yields,

$$\begin{aligned} (\bar{X}_n(t) - q(t)) &= (\bar{X}_n(0) - q_0) - \int_0^t \left\{ \left[\xi_b \mathbb{1}_{\{\bar{X}_n(s) \geq 0\}} + \xi_d \mathbb{1}_{\{\bar{X}_n(s) < 0\}}\right] \bar{X}_n(s) \right. \\ &\qquad \left. - \xi_b\, q(s) \right\} \mathrm{d}s + \bar{Y}_n(t) & (7.95) \\ &= (\bar{X}_n(0) - q_0) - \int_0^t \xi_b \left(\bar{X}_n(s) - q(s)\right) \mathrm{d}s \\ &\qquad + \int_0^t \mathbb{1}_{\{\bar{X}_n(s) < 0\}} (\xi_b - \xi_d) \bar{X}_n(s)\, \mathrm{d}s. & (7.96) \end{aligned}$$

Let $\hat{X}_n(t) = \sqrt{n}\,(\bar{X}_n(t) - q(t))$. Then

$$\hat{X}_n(t) = \hat{X}_n(0) - \xi_b \int_0^t \hat{X}_n(s)\, \mathrm{d}s + \int_0^t \mathbb{1}_{\{\bar{X}_n(s) < 0\}} (\xi_b - \xi_d) \bar{X}_n(s)\, \mathrm{d}s + \hat{Y}_n(t) \tag{7.97}$$

Now, we argue non-rigorously that the one-but-last term vanishes as $n \to \infty$. Namely, by defining the function $G : D[0,\infty) \to D[0,\infty)$ by the integration operator:

$$G(u) = \int_0^t \mathbb{1}_{\{u(s) < 0\}} (\xi_b - \xi_d) u(s)\, \mathrm{d}s, \tag{7.98}$$

this term can be expressed as $G(\bar{X}_n)$. Hence by the fact that $\hat{X}_n \overset{d}{\Rightarrow} m/\xi_b$ and the CMT we see $G(\hat{X}_n) \Rightarrow 0$.

Under this claim, we deduce by the approach of Proposition 7.6, that if $\hat{X}_n \Rightarrow \hat{X}$ for $n \to \infty$, then $\hat{X}$ satisfies the stochastic integral equation

$$\hat{X}(t) = \hat{X}(0) - \xi_b \int_0^t \hat{X}(s)\, \mathrm{d}s + \sigma W(t), \tag{7.99}$$

which implies that $\hat{X}$ is an Ornstein-Uhlenbeck process with infinitesimal mean ξ_b and variance $\sigma^2 := \lambda_b \mathbb{E}[B^2] + \lambda_d \mathbb{E}[D^2]$.

The result that the scaled process converges to an Ornstein-Uhlenbeck process can be intuitively justified by the so-called *mean-reverting* behavior of the original process. That is, the further the process is away from its mean, the greater the drift towards that equilibrium. This is the defining feature of the OU diffusion process. The decay rates ξ_b and ξ_d are responsible for the original process being 'forced' towards 0 and therefore the similarities should not be surprising. However, note that in the diffusion limit X_n has drift ξ_b (cq. ξ_d) towards nm/ξ_b (cq. nm/ξ_d), if $m > 0$ (cq. < 0) at *any* position of the process. This implies that if $X_n \in (0, nm/\xi_b)$, it has an upward drift equal to ξ_b, which is at first sight counter-intuitive. However, we can argue that in case $X_n(t) = v \in (0, nm/\xi_b)$, the mean upward drift of the process X_n equals $n\lambda_b \mathbb{E}[B]$, and the mean downward drift equals $n\lambda_d \mathbb{E}[D] + \xi_b v$, since $v > 0$. Rewrite $v = nm/\xi_b - w\sqrt{n}$ for some $w \in (0, \sqrt{n}m/\xi_b)$. Then, the mean net drift equals

$$n\lambda_b \mathbb{E}[B] - n\lambda_d \mathbb{E}[D] - \xi_b \left(\frac{nm}{\xi_b} - w\sqrt{n} \right) = \xi_b w \sqrt{n} > 0,$$

which explains both the sign and magnitude of the drift factor in the scaled process.

7.4.3 Related literature

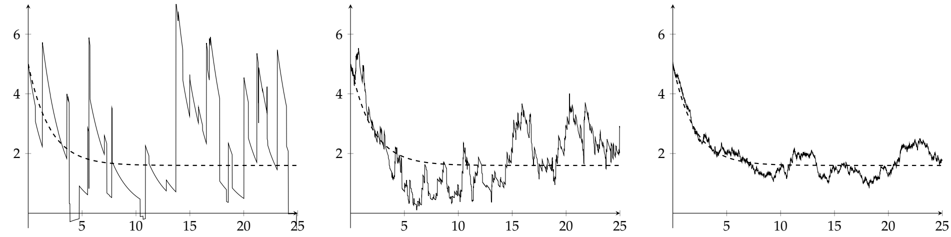
The Ornstein-Uhlenbeck process is a diffusion process that often arises as the limit of a sequence of stochastic systems, in which the system size tends to infinity. Particularly in queueing settings with mean reverting behavior, the OU process appears in so-called heavy traffic, i.e. the arrival rate grows without bound. We mention a couple of models that exhibit limiting behavior that is similar to ours.

First, it is well-known that the properly normalized $M/M/\infty$ queue length process converges weakly to a OU process as the arrival rate tends to infinity, see e.g. [221, Sec. 10.3]. This limiting behavior continues to hold in case the queueing process is modulated by a Markovian background process, see [13].

Another well-known queueing model in which a (piecewise) OU process appears in the limit is the multi-server queue with abandonments. For the $M/M/s + M$ queue, where $+M$ denotes the exponentially distributed patience of customers, Garnett et al. [82] showed that in the Halfin-Whitt regime, the queue length process, centered and scaled around the number of servers s, approaches a hybrid OU process, of which the drift parameter depends on the current state: If the queue length is larger (cq. smaller) than zero, then the drift is governed by the abandonment rate (cq. service rate). Dai et al. [64] find a similar piecewise diffusion process under more general assumptions on the model primitives.

For the single-server queue with abandoning customers, Ward & Glynn [215, 216] showed that in conventional heavy traffic, the queue length process converges to a OU process with reflecting barrier 0.

Since we in our setting assumed both demand impatience and perishability of inventory (which can be seen as a kind of impatience as well), it should not come as a surprise that we also find our limiting process to be a OU process. Observe



(a) $n = 1$ (b) $n = 10$ (c) $n = 100$

Figure 7.4: Sample paths of the process $\bar{X}_n(t) = X_n(t)/n$ with $\bar{X}_n(0) = 5$, $\lambda_b = 1.2$, $\lambda_d = 1$, $\xi_b = \xi_d = 0.5$ and $\mu_b = 0.5$ and $\mu_d = 1$. The fluid limit is depicted by the dashed line.

however that in our model, unless $m = 0$, we find a OU process with constant, rather than piecewise, parameters, and no reflection barrier, since our (scaled) inventory process can go both positive and negative.

Last, we mention that there is a connection between our blood inventory process and the work of Reed & Zwart [186]. Rather than looking at the OU process as the limit of a sequence of stochastic processes, Reed and Zwart [186] study a stochastic differential equation that is closely related to Equation (7.62), in the sense that the process has a different (constant) drift term in the upper and lower half plane. Under the assumption that the input process is a Lévy process with only one-sided jumps, they develop a methodology to derive the invariant distribution of the solution of the SDE. Unfortunately, the input in our scenario exhibits both positive and negative jumps, which prevents us from applying their results directly to (7.62).

7.5 Numerical evaluation

7.5.1 Approximation scheme

The asymptotic results of the previous section regarding the fluid and diffusion limits allude to the fact that for large arrival rates, the normalized inventory process $\{\hat{X}_n(t) \,|\, t \geq 0\}$, resembles that of the Ornstein-Uhlenbeck process. Indeed, the sample paths of the scaled process $\bar{X}_n$ for increasing values of n in Figures 7.4 and 7.5 show that the mean-reverting behavior around m/ξ^*, that is typical of OU processes, kicks in rather quickly. Moreover, the fluid limits $q(t)$ as presented by Proposition 7.5 and (7.89)-(7.92) predict the mean well for both $\xi_b = \xi_d$ and $\xi_b \neq \xi_d$. Furthermore, we observe that steady state is attained fairly quickly. This is suggestive of the claim that the steady-state distribution of the normalized process $\hat{X}_n$ is well-described by the steady-state distribution of the OU process $\hat{X}$. Since the OU process with mean 0, infinitesimal variance σ^2 and drift ξ^* is known to be normally distributed with mean 0 and variance $\sigma^2/2\xi^*$ in steady-state, this leads to a simpler approximation scheme based on the first two moments of B and D only.

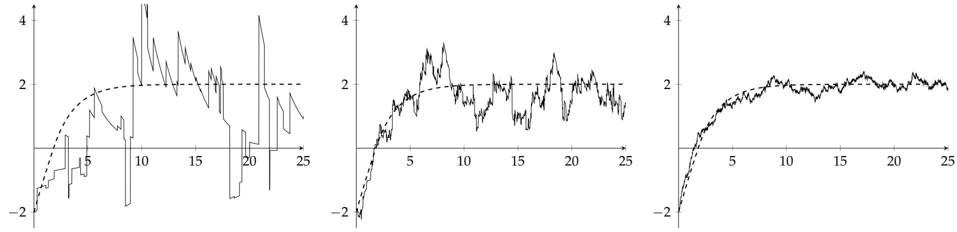


Figure 7.5: Sample paths of the process $\bar{X}_n(t) = X_n(t)/n$ with $\bar{X}_n(0) = -2$, $\lambda_b = 2$, $\lambda_d = 1$, $\xi_b = 0.5, \xi_d = 0.1$ and $\mu_b = 1$ and $\mu_d = 1$. The fluid limit is depicted by the dashed line.

In non-rigorous mathematical terms, we use the approximation that

$$\hat{X}_n = \frac{X_n - nm/\xi^*}{\sqrt{n}} \overset{d}{\approx} Z^*, \tag{7.100}$$

where Z^* is a normally distributed random variable with mean 0 and variance $\sigma^2/2\xi^*$.

Note that justification of the conjecture that the normal approximation is indeed an asymptotically correct approximation for systems with large arrival rates requires proof that the interchange-of-limits between $t \to \infty$ and $n \to \infty$ is indeed valid. Rather than going into the technical details, we provide in the remainder of this section numerical evidence that this interchange indeed holds, and that the normal approximation is able to capture characteristics of processes with exponential jumps as well as generally distributed jumps.

7.5.2 Distribution functions

Since we obtained an explicit expression for the steady-state density function of the net inventory process X in case B and D are exponential, see Theorem 7.1, we will exploit this formula for numerical comparison to the normal approximation arising from the OU process.

Let $h(\cdot)$ as in Theorem 7.1 be the pdf of X with parameters λ_b, λ_d, μ_b, μ_d, ξ_b and ξ_d, and the corresponding cdf H, defined as $H(v) = \int_{-\infty}^{v} h(x)\mathrm{d}x$. We denote by h_n and H_n the pdf and cdf, respectively, of the inventory process X_n with arrival rates $n\lambda_b$ and $n\lambda_d$, and the remaining parameters unchanged. Then, the pdf and cdf of the normalized process are given by $\hat{h}_n(v) = \sqrt{n}\, h_n(v_n)$ and $\hat{H}_n(v) = H_n(v_n)$, respectively, with $v_n = nm/\xi^* + v\sqrt{n}$ for all $v \in \mathbb{R}$. By the normal approximation scheme, we expect

$$\hat{h}_n(v) \approx \frac{\sqrt{2\xi^*}}{\sigma}\, \varphi\left(\frac{\sqrt{2\xi^*}}{\sigma} v\right), \quad \text{and} \quad \hat{H}_n(v) \approx \Phi\left(\frac{\sqrt{2\xi^*}}{\sigma} v\right). \tag{7.101}$$

We perform this numerical comparison of probability functions in Figure 7.6 for three cases: $\xi_b = \xi_d$, $\xi_b > \xi_d$ and $\xi_b < \xi_d$.

From Figure 7.6, in which $m = 1$, so that $\xi^* = \xi_b$, the convergence of the pdf and cdf is evident. For $n = 10$, the distribution functions of the scaled processes are almost aligned with the normal distribution already. For $\xi_b = \xi_d$, the convergence is fastest. This can be explained by observing that in cases where $\xi_b \neq \xi_d$, the parameter ξ_d still plays a role in pre-limit systems, whereas it does not appear in the normal limit. In the cases where $\xi_b \neq \xi_d$ we furthermore see that the functions are not smooth around $v_n = 0$ or $v^* = -\sqrt{nm/\xi^*}$, which is the zero-inventory level in the original (unscaled) process. As n increases, this point of irregularity goes to $-\infty$ and therefore disappears.

7.5.3 Approximations to performance metrics

The plots in the previous section indicate that the normal approximation gives simple yet accurate approximations to the stationary distribution of the inventory process. We now assess if this also translates to the performance measures. Again, we choose to fix the parameters λ_b and λ_d, and evaluate the system with arrival rates $n\lambda_b$ and $n\lambda_d$ for increasing n. First, the normal approximation in (7.100) yields the following approximation for the expected inventory level:

$$\mathbb{E}[X_n] \approx \frac{nm}{\xi^*} = \frac{n(\lambda_b \mathbb{E}[B] - \lambda_d \mathbb{E}[D])}{\xi^*}. \tag{7.102}$$

For the probability of negative inventory, we have

$$\pi_d = \mathbb{P}(X_n < 0) \approx \mathbb{P}\left(Z^* < -\sqrt{n}\, m/\xi^*\right) = \Phi\left(-\sqrt{n/2\xi^*}\, m/\sigma\right). \tag{7.103}$$

Last, the probability of demand being satisfied immediately is approximately

$$\mathbb{P}(\text{demand satisfied}) = \mathbb{P}(X_n > D) \approx 1 - \int_0^\infty \Phi\left(-\frac{\sqrt{2\xi^*}}{\sigma}\,\frac{x - nm/\xi^*}{\sqrt{n}}\right) \mathrm{d}F_d(x). \tag{7.104}$$

Remark 7.6. Note that if λ_b and λ_d are large themselves, the parameter n can be eliminated from (7.102)-(7.104), so that

$$\mathbb{E}[X] \approx \frac{m}{\xi^*}, \qquad \pi_d \approx \Phi\left(-m/(\sigma\sqrt{2\xi^*})\right),$$

$$\mathbb{P}(\text{demand satisfied}) \approx 1 - \int_0^\infty \Phi\left(-\sqrt{2\xi^*}\,\frac{x - m/\xi^*}{\sigma}\right) \mathrm{d}F_d(x),$$

where $m = \lambda_b \mathbb{E}[B] - \lambda_d \mathbb{E}[D]$ and $\sigma^2 = \lambda_b \mathbb{E}[B^2] + \lambda_d \mathbb{E}[D^2]$.

We will now test these approximations under various assumptions on the distribution of B and D. In Tables 7.1-7.3 we compare the values obtained through the

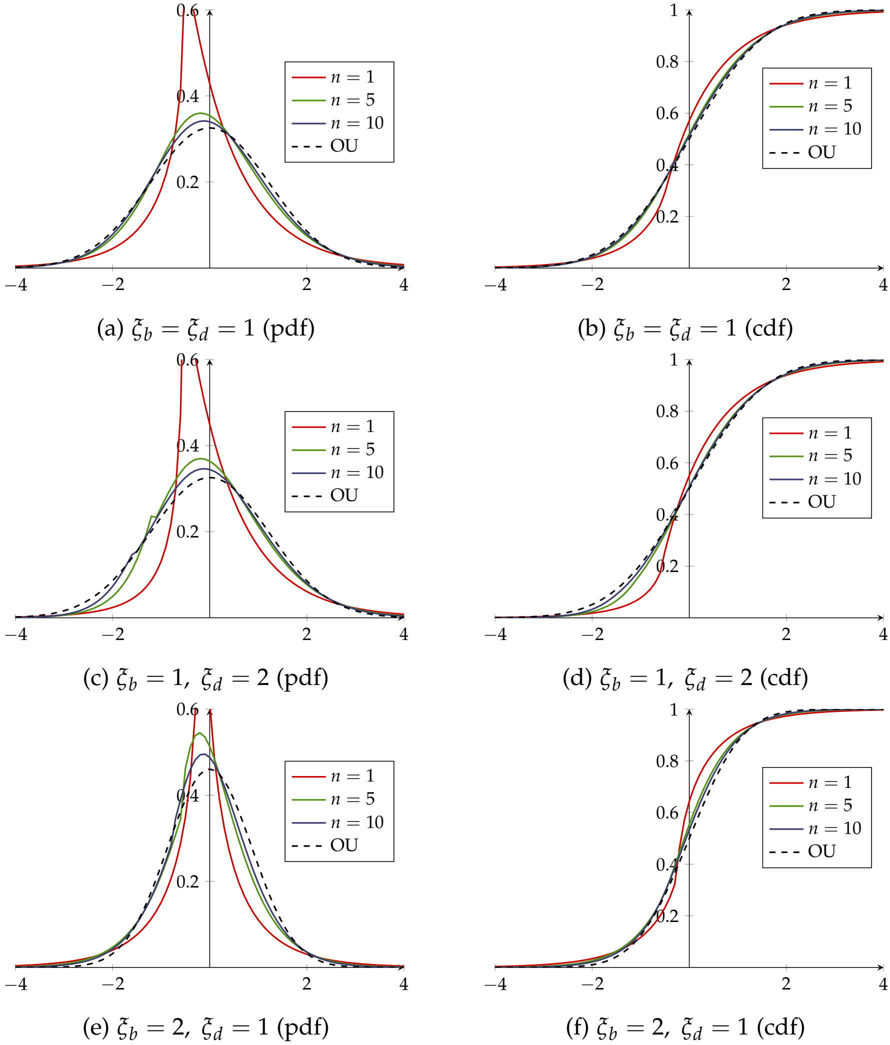


Figure 7.6: Probability functions of $\hat{X}_n$ for $n = 1, 5$ and 10 with $\lambda_b = 1, \lambda_d = 0.5, \mu_b = \mu_d = 1$, and the probability function of the OU process.

normal approximation against the true values obtained through numerical evaluation (for exponential jump sizes only) and simulation. All simulation results are accurate up to a 95% confidence interval of width 10^{-4}. We set $\lambda_b = 1$ and $\lambda_d = 0.5$ and let the mean jump sizes be equal to 1, i.e. $\mathbb{E}[B] = 1$ and $\mathbb{E}[B] = 1$ in all numerical experiments. In Table 7.1, we let the jump sizes be deterministic, so that $\operatorname{Var} B = \operatorname{Var} D = 0$. Table 7.2 shows the results in case of exponential jump sizes, so that $\operatorname{Var} B = \operatorname{Var} D = 1$. Last, in Table 7.3 we investigate the quality of the approximation for jump sizes that follow a Gamma$(0.25, 0.25)$ distribution, yielding $\operatorname{Var} B = \operatorname{Var} D = 4$. With this set-up we cover jump distributions of increasing variance, so that we are able to study the impact of increased variability on the accuracy of the approximations. Moreover, we investigate the influence of the decay parameters ξ_b and ξ_d by considering the scenarios $\xi_b = \xi_d$, $\xi_b < \xi_d$ and $\xi_b > \xi_d$.

We make a couple of observations based on the numbers in Tables 7.1-7.3. First, we see that the approximation for the mean blood inventory level $\mathbb{E}[X_n]$ is exact if $\xi_b = \xi_d$, see Proposition 7.3. This obviously does not extend to π_d and $\mathbb{P}(\text{demand satisfied})$, since these performance measures are based on the entire distribution of X_n rather than the mean. Nonetheless, the normal approximation appears to be very accurate in the case $\xi_b = \xi_d$. We may explain this by observing that in the approximations (7.102)-(7.104), only ξ^* appears. In our setting, we have $m = \lambda_b - \lambda_d = 0.5$, so that $\xi^* = \xi_b$. If $\xi_b \neq \xi_d$, then the value of ξ_d plays a role in pre-limit systems, which induces inaccuracies in the approximation of performance measures. In case $\xi_b = \xi_d$, we have $\xi^* = \xi_b = \xi_d$, so that this discrepancy is overcome.

Moreover, since $m > 0$, we see that $\pi_d \to 0$ and $\mathbb{P}(\text{demand satisfied}) \to 1$ as n increases. This is due to the observation that as n grows large, the inventory process concentrates around the level nm with fluctuations of order $\sqrt{n}$, so that the process stays away from level zero, see Figure 7.4. The approximations (7.103)-(7.104) adequately capture this convergence.

As expected, the accuracy of the approximations increases with n. Moreover, increased variability in the jump distributions appears to cause a decrease in accuracy. However, for all cases considered in Tables 7.1-7.3, the normal approximations (7.102)-(7.104) seem to yield relatively sharp estimates for the relevant performance measures under various assumptions on the distributions of the jump sizes.

7.6 Conclusions & suggestions for further research

In this chapter, we studied a stochastic model for a blood bank. We have presented a global approach to the model in its full generality, and obtained very detailed exact expressions for the densities of amount of inventory and amount of demand (shortage) in special cases (exponential amounts of donated and requested blood; and either $\xi_b = \xi_d = 0$ or $\alpha_b = \alpha_d = 0$). Moreover, we have shown how an appropriate scaling, for the model in full generality, leads to an Ornstein-Uhlenbeck diffusion process, which can be used as a tool to obtain simple yet accurate approximations

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.500	0.500	0.2702	0.2819	0.2598	0.2819
2	1.000	1.000	0.2014	0.2071	0.4859	0.5000
5	2.500	2.500	0.0943	0.0984	0.7814	0.7807
10	5.000	5.000	0.0316	0.0339	0.9306	0.9279
20	10.000	10.000	0.0043	0.0049	0.9908	0.9899
50	25.000	25.000	0.0000	0.0000	1.0000	1.0000

(a) $\xi_b = 1$, $\xi_d = 1$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.584	0.500	0.2522	0.2819	0.2712	0.2819
2	1.086	1.000	0.1809	0.2071	0.5020	0.5000
5	2.558	2.500	0.0837	0.0984	0.7911	0.7807
10	5.024	5.000	0.0286	0.0339	0.9335	0.9279
20	10.006	10.000	0.0040	0.0049	0.9912	0.9899
50	25.000	25.000	0.0000	0.0000	1.0000	1.0000

(b) $\xi_b = 1$, $\xi_d = 2$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.158	0.250	0.3308	0.3415	0.1006	0.1103
2	0.397	0.500	0.2973	0.2819	0.2465	0.2819
5	1.164	1.250	0.1952	0.1807	0.5482	0.5724
10	2.447	2.500	0.1036	0.0984	0.7729	0.7807
20	4.980	5.000	0.0340	0.0339	0.9283	0.9279
50	12.497	12.500	0.0017	0.0019	0.9964	0.9960

(c) $\xi_b = 2$, $\xi_d = 1$.

Table 7.1: Accuracy of diffusion approximation for the blood inventory process $\mathbb{E}[X_n]$, the probability of negative inventory π_d and the probability of demand being fully satisfied $\mathbb{P}(dem.sat)$, with arrival rates $n\lambda_b = n$ and $n\lambda_d = 0.5n$ and deterministic jump sizes, $B \equiv 1$ and $D \equiv 1$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}$(dem.sat.)	
n	Exact	(7.102)	Exact	(7.103)	Exact	(7.104)
1	0.500	0.500	0.2929	0.3415	0.3536	0.3925
2	1.000	1.000	0.2500	0.2819	0.5000	0.5135
5	2.500	2.500	0.1642	0.1807	0.7062	0.7009
10	5.000	5.000	0.0898	0.0984	0.8491	0.8418
20	10.000	10.000	0.0307	0.0339	0.9506	0.9467
50	25.000	25.000	0.0017	0.0019	0.9974	0.9970

(a) $\xi_b = 1$, $\xi_d = 1$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}$(dem.sat.)	
n	Exact	(7.102)	Exact	(7.103)	Exact	(7.104)
1	0.621	0.500	0.2589	0.3415	0.3705	0.3925
2	1.153	1.000	0.2164	0.2819	0.5224	0.5135
5	2.656	2.500	0.1414	0.1807	0.7254	0.7009
10	5.113	5.000	0.0784	0.0984	0.8598	0.8418
20	10.050	10.000	0.0275	0.0339	0.9538	0.9467
50	25.004	25.000	0.0016	0.0019	0.9975	0.9970

(b) $\xi_b = 1$, $\xi_d = 2$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}$(dem.sat.)	
n	Exact	(7.102)	Exact	(7.103)	Exact	(7.104)
1	0.125	0.250	0.3548	0.3864	0.2168	0.2942
2	0.333	0.500	0.3333	0.3415	0.3333	0.3925
5	1.059	1.250	0.2647	0.2593	0.5264	0.5570
10	2.333	2.500	0.1856	0.1807	0.6881	0.7009
20	4.893	5.000	0.0995	0.0984	0.8400	0.8418
50	12.475	12.500	0.0198	0.0206	0.9692	0.9678

(c) $\xi_b = 2$, $\xi_d = 1$.

Table 7.2: Accuracy of diffusion approximation for the blood inventory process $\mathbb{E}[X_n]$, the probability of negative inventory π_d and the probability of demand being fully satisfied $\mathbb{P}(dem.sat)$, with arrival rates $n\lambda_b = n$ and $n\lambda_d = 0.5n$ and exponentially distributed jump sizes, $B \sim \exp(1)$ and $D \sim \exp(1)$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.500	0.500	0.3118	0.3981	0.4412	0.4636
2	1.000	1.000	0.2894	0.3575	0.5343	0.5288
5	2.500	2.500	0.2375	0.2819	0.6590	0.6381
10	5.000	5.000	0.1785	0.2071	0.7592	0.7385
20	10.000	10.000	0.1090	0.1241	0.8593	0.8454
50	25.000	25.000	0.0303	0.0339	0.9624	0.9583

(a) $\xi_b = 1$, $\xi_d = 1$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.667	0.500	0.2695	0.3981	0.4636	0.4636
2	1.253	1.000	0.2469	0.3575	0.5632	0.5288
5	2.863	2.500	0.2009	0.2819	0.6895	0.6381
10	5.385	5.000	0.1518	0.2071	0.7834	0.7385
20	10.328	10.000	0.0938	0.1241	0.8739	0.8454
50	25.124	25.000	0.0269	0.0339	0.9658	0.9583

(b) $\xi_b = 1$, $\xi_d = 2$.

	$\mathbb{E}[X_n]$		π_d		$\mathbb{P}(\text{dem.sat.})$	
n	Sim.	(7.102)	Sim.	(7.103)	Sim.	(7.104)
1	0.081	0.250	0.3694	0.4276	0.3270	0.4104
2	0.238	0.500	0.3593	0.3981	0.4137	0.4636
5	0.857	1.250	0.3237	0.3415	0.5311	0.5528
10	2.045	2.500	0.2739	0.2819	0.6282	0.6381
20	4.568	5.000	0.2039	0.2071	0.7361	0.7385
50	12.231	12.500	0.0966	0.0984	0.8797	0.8779

(c) $\xi_b = 2$, $\xi_d = 1$.

Table 7.3: Accuracy of diffusion approximation for the blood inventory process $\mathbb{E}[X_n]$, the probability of negative inventory π_d and the probability of demand being fully satisfied $\mathbb{P}(dem.sat)$, with arrival rates $n\lambda_b = n$ and $n\lambda_d = 0.5n$ and Gamma distributed jump sizes, $B \sim \text{Gamma}(0.25, 0.25)$ and $D \sim \text{Gamma}(0.25, 0.25)$.

for some key performance measures.

Our model is a two-sided model, in the sense that we simultaneously consider the amount of blood in inventory and the amount of demand (shortage), one of the two at any time being zero. Such two-sided processes arise in many different settings, and thus are of considerable interest. The present setting is reminiscent of an organ transplantation problem, where there is either a queue of persons waiting to receive an organ, or a queue of donor organs. The perishability/impatience aspect features there too [46]. A quite different setting is that of insurance risk. We refer to Albrecher & Lautscham [11] who extend the classical Cramér-Lundberg insurance risk model by allowing the capital of an insurance company to become negative – a situation that is usually indicated by "ruin" in the insurance literature. Their process thus becomes two-sided. The capital might become positive again; however, at a rate $\omega(x)$ when the capital has a negative value $-x$, bankruptcy is declared and the process ends. Interestingly, similar special functions (like hypergeometric functions) play a role in [11] and in the present study.

The analyses performed in this chapter, which evolved around a simplified version of the inventory process of a blood bank, revealed some interesting avenues for further research. We name a couple of them.

First, we remark that our results are restricted to one type of blood. Naturally, it would be very interesting to extend the analysis to multiple types of blood.

Another important extension would be to use our results to facilitate the decision process that is faced by the CBB on a daily basis: Which amounts of blood, and of which types, should today be sent to the local blood banks (hospitals)? Knowing that, e.g., blood types O^-, A^-, B^-, AB^- can satisfy the corresponding $+$ type (but not vice versa), one may try to optimize the blood allocation process on the basis of actual amounts of blood present.

Finally, we mention a significant open research question regarding the process limits that we derived in Section 7.4, of which the steady-state distributions were used to approximate steady-state performance measures in pre-limit systems. As we pointed out earlier, the justification that the steady-state distribution of the scaled inventory process indeed converges to the steady-state distribution of the fluid (cq. diffusion) limit requires a rigorous argument why the order of limits $n \to \infty$ and $t \to \infty$ may be interchanged. Proving interchange-of-limits statements typically raises many technical challenges, see e.g. [63, 77, 96, 79] for works tackling this issue in the context of queues in heavy traffic. The usual approach is to prove tightness of the sequence of steady-state distributions of pre-limit, followed by applying Prokhorov's theorem, see e.g. [37, Sec. 1.5]. For our model, such an approach seems to be straightforward for the fluid scaling, since our inventory process can be upper (cq. lower) bounded by a shot-noise process with only positive (cq. negative) jumps. Of the latter, the steady-state behavior is known. This allows us to derive a uniform bound on the absolute mean of the stationary fluid-scaled process, which gives tightness. The final step uses the deterministic nature of the differential equation governing the dynamics of the fluid limit, by which the steady-state distribution must be unique. For the diffusion-scaled process, the steps towards proving the interchange-of-limits are not obvious and hence this needs further investigation. Our numerical results for various jump size distributions, however, support the conjecture that this interchange is indeed valid.

Appendix

7.A Transformation integral equation

In this appendix we show how integral equation (7.7) can be transformed into a second-order differential equation, in the case of exponential $F_b(\cdot)$ and $F_d(\cdot)$. Differentiate (7.7) w.r.t. v:

$$\begin{aligned}\lambda_d f(v) - \mu_d \left[\lambda_d \int_0^v f(y)\mathrm{e}^{-\mu_d(v-y)}\mathrm{d}y + \lambda_d \int_0^\infty g(y)\mathrm{e}^{-\mu_d(v+y)}\mathrm{d}y\right] \\ = -\lambda_b f(v) + \lambda_b \mu_b \int_v^\infty f(y)\mathrm{e}^{-\mu_b(y-v)}\mathrm{d}y + \xi_d f(v) + \xi_d v f'(v).\end{aligned} \tag{7.105}$$

Using (7.7) once more, now to replace the term between square brackets in (7.105), we get:

$$\begin{aligned}\xi_d v f'(v) = {}& (\lambda_d + \lambda_b - \xi_d) f(v) \\ & - \mu_d \left(\lambda_b \int_v^\infty f(y)\mathrm{e}^{-\mu_b(y-v)}\mathrm{d}y + \xi_d v f(v)\right) \\ & - \mu_b \lambda_b \int_v^\infty f(y)\mathrm{e}^{-\mu_b(y-v)}\mathrm{d}y,\end{aligned} \tag{7.106}$$

and once more differentiating w.r.t. v then gives:

$$\begin{aligned}\xi_d v f''(v) + \xi_d f'(v) - (\lambda_d + \lambda_b - \xi_d - \mu_d \xi_d v) f'(v) \\ = -\mu_d \xi_d f(v) + (\mu_b + \mu_d)\lambda_b f(v) - \mu_b(\mu_b + \mu_d)\lambda_b \int_v^\infty f(y)\mathrm{e}^{-\mu_b(y-v)}\mathrm{d}y.\end{aligned} \tag{7.107}$$

The integral that appears in (7.106) can be eliminated by using (7.107), and we thus finally obtain the following second order homogeneous differential equation:

$$\begin{aligned}&\xi_d v f''(v) + (2\xi_d - \lambda_d - \lambda_b + \mu_d \xi_d v - \mu_b \xi_d v)\, f'(v) \\ &+ (\mu_d \xi_d - \mu_b \xi_d - \mu_d \lambda_b + \mu_b \lambda_d - \mu_b \mu_d \xi_d v)\, f(v) = 0.\end{aligned} \tag{7.108}$$

7.B Proof of Proposition 7.4

In the proof, we concentrate on the derivation of $f(v)$, which is the solution to

$$\begin{aligned}\xi_d v f''(v) + (2\xi_d - \lambda_d - \lambda_b + \mu_d \xi_d v - \mu_b \xi_d v)\, f'(v) \\ + (\mu_d \xi_d - \mu_b \xi_d - \mu_d \lambda_b + \mu_b \lambda_d - \mu_b \mu_d \xi_d v)\, f(v) = 0\end{aligned} \tag{7.109}$$

The expression for $g(v)$ follows directly from exchanging λ_b with λ_d, μ_b with μ_d, ξ_b with ξ_d, and π_b with π_d in $f(v)$. We rewrite (7.109) as follows:

$$vf''(v) + (A + Bv)f'(v) + (C + Dv)f(v) = 0, \tag{7.110}$$

where

$$A = 2 - \frac{\lambda_b + \lambda_d}{\xi_d}, \quad B = \mu_d - \mu_b, \quad C = \mu_d - \mu_b + \frac{\lambda_d \mu_b - \lambda_b \mu_d}{\xi_d}, \quad D = -\mu_b \mu_d.$$

Note that we divided both sides of equation (7.109) by ξ_d here. We will try to transform the differential equation into one for which the solution is easily derived. In order to do so, we first guess f to be of the form $f(v) = \mathrm{e}^{\beta v} h(v)$, where β is a constant and h another real-valued function. Substituting this into (7.110) gives

$$vh''(v) + [(2\beta + B)v + A]\, h'(v) + \Big[(\beta^2 + B\beta + D)v + A\beta + C\Big]\, h(v) = 0. \tag{7.111}$$

Next, we would like to choose β such that $\beta^2 + B\beta + D = 0$, that is

$$\beta = \frac{-B \pm \sqrt{B^2 - 4D}}{2},$$

which equals either $-\mu_d$ or μ_b. Since the solution of (7.110) we are looking for is a density, and necessarily $f(v) = \mathrm{e}^{\beta v} h(v) \to 0$ as $v \to \infty$, we set β equal to the negative root $-\mu_d$. Lastly, we apply a change of variable, $x = \delta v$, and $h(v) = w(x)$, so that (7.111) is transformed into

$$xw''(x) + \Big[(2\beta + B)\delta^{-1} x + A\Big]\, w'(x) + \delta^{-1}\, [A\beta + C]\, w(x) = 0.$$

By choosing $(2\beta + B)\delta^{-1} = -1$, i.e.

$$\delta = -(2\beta + B) = \mu_b + \mu_d,$$

we obtain

$$xw''(x) + [A - x]w'(x) + \delta^{-1}\, [A\beta + C]\, w(x) = 0,$$

which is known as Kummer's equation, $xw''(x) + (b - x)w'(x) - aw(x) = 0$, see [195], with parameters

$$a = -\delta^{-1}\, [A\beta + C] = 1 - \frac{\lambda_d}{\xi_d},$$
$$b = A = 2 - \frac{\lambda_b + \lambda_d}{\xi_d}.$$

Kummer's equation has two linearly independent solutions, namely $w(x) = M(a, b, x)$, where M is Kummer's hypergeometric function, also denoted by

${}_1F_1(a,b,x)$, and $U(a,b,x)$, Tricomi's hypergeometric function. These are defined as, see [195, Eq. (1.3.1)],

$$M(a,b,x) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} x^n,$$
$$U(a,b,x) = \frac{\Gamma(b-1)}{\Gamma(1+a-b)}\, M(a,b,x) + \frac{\Gamma(b-1)}{\Gamma(a)}\, x^{1-b}\, M(1+a-b, 2-b, x),$$

where $(.)_n$ is the Pochhammer symbol, which is used to represent $(y)_n = y \cdot (y+1) \cdot \ldots \cdot (y+n-1)$. We can therefore deduce that $f(v)$ is of the form

$$\mathrm{e}^{\beta v}\left[c_1\, M(a,b,\delta v) + c_2\, U(a,b,\delta v)\right],$$

or

$$\mathrm{e}^{-\mu_d v}\left[c_1 M\left(1-\tfrac{\lambda_d}{\xi_d}, 2-\tfrac{\lambda_b+\lambda_d}{\xi_d}, (\mu_b+\mu_d)v\right) + c_2 U\left(1-\tfrac{\lambda_d}{\xi_d}, 2-\tfrac{\lambda_b+\lambda_d}{\xi_d}, (\mu_b+\mu_d)v\right)\right],$$

where c_1 and c_2 are constants. From [195, p. 60], we have

$$M(a,b,x) \sim \frac{\Gamma(b)}{\Gamma(a)} \mathrm{e}^x x^{a-b}, \qquad \text{as } x \to \infty.$$

Hence,

$$\begin{aligned}\mathrm{e}^{-\mu_d v} M\left(1-\frac{\lambda_d}{\xi_d}, 2-\frac{\lambda_b+\lambda_d}{\xi_d}, (\mu_b+\mu_d)v\right)& \\ \sim \frac{\Gamma(2-\frac{\lambda_b+\lambda_d}{\xi_d})}{\Gamma(1-\frac{\lambda_d}{\xi_d})} \mathrm{e}^{\mu_b v}\left((\mu_b+\mu_d)v\right)^{\lambda_b/\xi_d - 1} &\to \infty\end{aligned}$$

for all $\mu_b > 0$, which leads us to conclude $c_1 = 0$. We deduce c_2 by exploiting the restriction that

$$\int_0^\infty f(v)\,\mathrm{d}v = \pi_d,$$

where π_d is the probability of positive demand. Hence

$$\pi_d c_2^{-1} = \int_0^\infty \mathrm{e}^{-\mu_d v} U\left(1-\tfrac{\lambda_d}{\xi_d}, 2-\tfrac{\lambda_b+\lambda_d}{\xi_d}, (\mu_b+\mu_d)v\right)\, dv.$$

By slightly transforming [195, Eq. (3.2.51)], we find

$$c_2^{-1} = \frac{1}{\pi_d}\, \frac{\Gamma\left(\frac{\lambda_b+\lambda_d}{\xi_d}\right)}{\Gamma\left(1+\frac{\lambda_b}{\xi_d}\right)}\, {}_2F_1\left(1-\tfrac{\lambda_d}{\xi_d}, 1, 1+\tfrac{\lambda_b}{\xi_d}, -\tfrac{\mu_b}{\mu_d}\right),$$

where ${}_2F_1(a_1,a_2,a_3,x) := \sum_{n=0}^{\infty} \frac{(a_1)_n (a_2)_n}{(a_3)_n n!} x^n$ is the hypergeometric function of Gauss.

7.C Laplace Transforms for Coxian jumps

We outline how the differential equation (7.46) is obtained. We take Laplace transforms in (7.2), considering its five terms and calling them T_1, T_2, T_3, T_4 and T_5, successively. Equation (7.2) then translates into

$$T_1 + T_2 + T_3 = T_4 + T_5,$$

where

$$\begin{aligned} T_1 &= \lambda_d \int_{v=0}^{\infty} \mathrm{e}^{-sv} \int_{y=0}^{v} f(y)\bar{F}_d(v-y)\mathrm{d}y\mathrm{d}v \\ &= \lambda_d \phi(s) \frac{1-\mathbb{E}[\mathrm{e}^{-sD}]}{s}, && (7.112) \\ T_2 &= \lambda_d \int_{v=0}^{\infty} \mathrm{e}^{-sv} \int_{y=0}^{\infty} g(y)\bar{F}_d(v+y)\mathrm{d}y\mathrm{d}v \\ &= \lambda_d \int_{y=0}^{\infty} \mathrm{e}^{sy} g(y) \int_{z=y}^{\infty} \mathrm{e}^{-sz}\bar{F}_d(z)\mathrm{d}z\mathrm{d}y, && (7.113) \\ T_3 &= \pi_0 \lambda_d \int_0^{\infty} \mathrm{e}^{-sy}\bar{F}_d(y)\mathrm{d}y, && (7.114) \\ T_4 &= \lambda_b \int_{v=0}^{\infty} \mathrm{e}^{-sv} \int_{y=v}^{\infty} f(y)\bar{F}_b(y-v)\mathrm{d}y\mathrm{d}v \\ &= \lambda_b \int_{y=0}^{\infty} \mathrm{e}^{-sy} f(y) \int_{z=0}^{y} \mathrm{e}^{sz}\bar{F}_b(z)\mathrm{d}z\mathrm{d}y, && (7.115) \\ T_5 &= \xi_d \int_{v=0}^{\infty} v\mathrm{e}^{-sv} f(v)\mathrm{d}v + \alpha_d \phi(s) \\ &= -\xi_d \phi'(s) + \alpha_d \phi(s). && (7.116) \end{aligned}$$

We now evaluate the terms appearing in the righthand sides of (7.112)-(7.115) for the Coxian case of (7.44) and (7.45):

$$\int_{z=0}^{y} \mathrm{e}^{sz}\bar{F}_b(z)\mathrm{d}z = \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1-p_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\beta_l}{\beta_l-\beta_j} \frac{1}{\beta_j - s}(1-\mathrm{e}^{(s-\beta_j)y}), \quad (7.117)$$

$$\int_{z=y}^{\infty} \mathrm{e}^{-sz}\bar{F}_b(z)\mathrm{d}z = \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1-p_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\beta_l}{\beta_l-\beta_j} \frac{1}{\beta_j + s}\mathrm{e}^{-(s+\beta_j)y}, \quad (7.118)$$

$$\mathbb{E}[\mathrm{e}^{-sB}] = \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1-p_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\beta_l}{\beta_l-\beta_j} \frac{\beta_j}{\beta_j + s}, \quad (7.119)$$

and hence

$$\frac{1-\mathbb{E}[\mathrm{e}^{-sB}]}{s} = \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1-p_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\beta_l}{\beta_l-\beta_j} \frac{1}{\beta_j + s}. \quad (7.120)$$

Combining (7.C) with (7.112)-(7.116), and using (7.117) and the counterparts of (7.118) and (7.120) for $\bar{F}_d(\cdot)$, we find:

$$\begin{aligned}
&\lambda_d \phi(s) \sum_{i=1}^{K} q_i \prod_{h=1}^{i-1}(1-q_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\delta_l}{\delta_l-\delta_j} \frac{1}{\delta_j+s} \\
&\quad + \lambda_d \sum_{i=1}^{L} q_i \prod_{h=1}^{i-1}(1-q_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\delta_l}{\delta_l-\delta_j} \frac{1}{\delta_j+s} [\gamma(\delta_j)+\pi_0] \\
&= \lambda_b \sum_{i=1}^{K} p_i \prod_{h=1}^{i-1}(1-p_h) \sum_{j=1}^{i} \prod_{l=1;l\neq j}^{i} \frac{\beta_l}{\beta_l-\beta_j} \frac{1}{\beta_j-s} (\phi(s)-\phi(\beta_j)) \\
&\quad - \xi_d \phi'(s) + \alpha_d \phi(s),
\end{aligned} \tag{7.121}$$

which is readily rewritten into (7.46).

Remark 7.7. If $\xi_d = 0$, then $\phi(s)$ is obtained from (7.121) in a standard manner, see also Section 7.3.3.

Remark 7.8. We now outline how (7.118) and (7.119) change when the B_i have an Erlang-$(l+1,\beta)$ distribution, and when the D_i have an Erlang-$(k+1,\delta)$ distribution (see also (7.44) and the line below it); (7.117) and (7.120) do not change (but of course $\mathbb{E}[e^{-sD}]$ changes). Firstly,

$$\int_{z=0}^{y} e^{sz} \bar{F}_b(z) dz = \sum_{j=0}^{l} \frac{\beta^j}{(\beta-s)^{j+1}} \left[1 - \sum_{i=0}^{j} e^{-(\beta-s)y} \frac{((\beta-s)y)^i}{i!} \right].$$

Term T_4 now becomes:

$$\begin{aligned}
T_4 &= \lambda_b \int_{v=0}^{\infty} e^{-sv} \int_{y=v}^{\infty} f(y) \bar{F}_b(y-v) dy dv \\
&= \lambda_b \sum_{j=0}^{l} \frac{\beta^j}{(\beta-s)^{j+1}} \left[\phi(s) - \sum_{i=0}^{j} \frac{(\beta-s)^i}{i!} \int_{y=0}^{\infty} y^i e^{-\beta y} f(y) dy \right].
\end{aligned}$$

It should be noted that $s = \beta$ is a removable singularity. E.g., for $l = 0$ one has $T_4 = \lambda_b \frac{\phi(s)-\phi(\beta)}{\beta-s}$.

Secondly,

$$\int_{z=y}^{\infty} e^{-sz} \bar{F}_b(z) dz = \sum_{j=0}^{k} \frac{\delta^j}{(s+\delta)^{j+1}} \sum_{i=0}^{j} e^{-(s+\delta)y} \frac{((s+\delta)y)^i}{i!}.$$

Term T_2 now becomes:

$$\begin{aligned}
T_2 &= \lambda_d \int_{v=0}^{\infty} e^{-sv} \int_{y=0}^{\infty} g(y) \bar{F}_d(v+y) dy \, dv \\
&= \lambda_d \sum_{j=0}^{k} \frac{\delta^j}{(s+\delta)^{j+1}} \sum_{i=0}^{j} \frac{(s+\delta)^i}{i!} \int_{y=0}^{\infty} y^i e^{-\delta y} g(y) \, dy.
\end{aligned}$$

It is readily seen that the resulting counterpart of (7.121) can again be written in the form (7.46), and hence the solution is formally still given by (7.49).

Bibliography

[1] Amazon auto-scaling. https://aws.amazon.com/autoscaling.

[2] J. Abate, G.L. Choudhury, and W. Whitt. Calculation of the $GI/G/1$ waiting-time distribution and its cumulants from Pollaczek's formulas. *Archiv fur Elektronik und Ubertragungstechnik (International Journal of Electronics and Communication)*, 47(5/6):311–321, 1993.

[3] J. Abate and W. Whitt. Transient behavior of regulated Brownian motion, I: Starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.

[4] J. Abate and W. Whitt. Transient behavior of regulated Brownian motion, II: Non-zero initial conditions. *Advances in Applied Probability*, 19(3):599–631, 1987.

[5] J. Abate and W. Whitt. Transient behavior of the $M/M/1$ queue: Starting at the origin. *Queueing Systems*, 2(1):41–65, 1987.

[6] J. Abate and W. Whitt. Transient behavior of the $M/M/1$ queue via Laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1988.

[7] J. Abate and W. Whitt. Transient behavior of the $M/G/1$ workload process. *Operations Research*, 42(4):750–764, 1994.

[8] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[9] R. Aghajani and K. Ramanan. The limit of stationary distributions of many-server queues in the Halfin-Whitt regime. https://arxiv.org/abs/1610.01118, 2016.

[10] M.S. Aguir, O.Z. Aksin, F. Karaesmen, and Y. Dallery. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, 191(2):398–408, 2008.

[11] H. Albrecher and V. Lautscham. From ruin to bankruptcy for compound Poisson surplus processes. *ASTIN Bulletin*, 2013.

[12] American Hospital Association. Survey of hospital leaders, 2007.

[13] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck. A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability*, 18(1):151–168, 2016.

[14] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871–1894, 1982.

[15] M. Armbrust, A. Fox, R. Griffith, A.D.. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. *Commun. ACM*, 53(4):50–58, April 2010.

[16] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3):287–329, 2005.

[17] M. Armony, S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, and G.B. Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

[18] M. Armony and C. Maglaras. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

[19] M. Armony and A.R. Ward. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637, 2010.

[20] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, second edition, 2003.

[21] R. Atar and A. Biswas. Control of the multiclass $G/G/1$ queue in the moderate deviation regime. *The Annals of Applied Probability*, 24(5):2033–2069, 2012.

[22] R. Atar and A. Cohen. A differential game for a multiclass queueing model in the moderate-deviation heavy-traffic regime. *Mathematics of Operations Research*, 41(4):1354–1380, 2016.

[23] R. Atar, A. Mandelbaum, and M. Reiman. Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 14(3):1084–1134, 2014.

[24] R. Atar and S. Saha. Optimality of the generalized $c\mu$ rule in the moderate deviation regime. http://webee.technion.ac.il/people/atar/ata-sah-3-2.pdf, 2015.

[25] F. Avram, A.J.E.M. Janssen, and J.S.H. van Leeuwaarden. Loss systems with slow retrials in the Halfin-Whitt regime. *Advances in Applied Probability*, 45(1):274–294, 2013.

[26] A.N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Rate-based daily arrival process models with application to call centers. *Management Science*, 50(7):893–908, 2004.

[27] H.P. Awad and P.W. Glynn. On the theoretical comparison of low-bias steady-state estimators. *ACM Transactions on Modeling and Computer Simulation*, 17(1):1–30, 2007.

[28] S.K. Bar-Lev, O.J. Boxma, B.W.J. Mathijsen, and D. Perry. A blood bank model with perishable blood and demand impatience. 2016.

[29] A. Bassamboo, R.S. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686, 2010.

[30] A. Bassamboo and A. Zeevi. On a data-driven method for staffing large call centers. *Operations Research*, 57(3):714–726, 2009.

[31] R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella. Queues with workload-dependent arrival and service rates. *Queueing Systems*, 46(3):537–556, 2004.

[32] R. Bekker and A.M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, 2009.

[33] J. Beliën and H. Forcé. Supply chain management of blood products: A literature review. *European Journal of Operational Research*, 217(1):1–16, 2012.

[34] V.E. Benes. On queues with Poisson arrivals. *The Annals of Mathematical Statistics*, 28(3):670–677, 1957.

[35] R. Bennidor and S.H. Israelit. Emergency department intermediate stay unit - a failed model. Unpublished manuscript, 2015.

[36] J. Bertoin. *Lévy Processes*. Cambridge University Press, 1996.

[37] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2 edition, 1995.

[38] J. Blanchet and P.W. Glynn. Complete corrected diffusion approximations for the maximum of a random walk. *The Annals of Applied Probability*, 16(2):951–983, 2006.

[39] T. Bonald and C. Comté. Networks of multi-server queues with parallel processing. https://arxiv.org/abs/1604.06763, 2016.

[40] M.A.A. Boon, A.J.E.M. Janssen, and J.S.H. van Leeuwaarden. Heavy-traffic limits for dimensioning fixed-cycle intersections. *Working paper*, 2017.

[41] M.A.A. Boon, A.J.E.M. Janssen, and J.S.H. van Leeuwaarden. Pollaczek contour integrals for the fixed-cycle traffic-light queue. *arXiv:1701.02872 (preprint)*, 2017.

[42] A.A. Borovkov. Some limit theorems in the theory of mass service, II multiple channels systems. *Theory of Probability & its Applications*, 10(3):375–400, 1965.

[43] S.C. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, January-February 2004.

[44] R.J. Boucherie and O.J. Boxma. The workload in the $M/G/1$ queue with work removal. *Probability in the Engineering and Informational Sciences*, 10(2):261–277, 1996.

[45] P.E. Boudreau, J.S. Griffin Jr., and M. Kac. An elementary queueing problem. *The American Mathematical Monthly*, 69(8):713–724, 1962.

[46] O.J. Boxma, I. David, D. Perry, and W. Stadje. A new look at organ transplant models and double matching queues. *Probability in the Engineering and Informational Sciences*, 25(2):135–155, 2011.

[47] A. Braverman and J.G. Dai. Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems. https://arxiv.org/abs/1503.00774, 2015.

[48] A. Braverman, J.G. Dai, and J. Feng. Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. https://arxiv.org/abs/1512.09364, 2015.

[49] L.D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[50] L.D. Brown and L.H. Zhao. A test for the Poisson distribution. *The Indian Journal of Statistics*, 64(3):611–625, 2002.

[51] S. Browne and W. Whitt. *Advances in Queueing: Theory, Methods, and Open Problems*, chapter Piecewise-linear diffusion processes, pages 463–480. CRC Press, Boca Raton, FL, 1995.

[52] S.L. Brumelle. Some inequalities for parallel-server queues. *Operations Research*, 19(2):402–413, 1971.

[53] H. Bruneel and B.G. Kim. *Discrete-time Models for Communication Systems Including ATM*. Kluwer Academic Publishers, Boston, 1993.

[54] F. Campello, A. Ingolfsson, and R.A. Shumsky. Queueing models of case managers. *Management Science, Articles in Advance*, 2016.

[55] R. Carmen and I. van Nieuwenhuyse. How inpatient boarding impacts ED performance: A queueing analysis. 2016. KU Leuven working paper.

[56] C.-S. Chang, D.D. Yao, and T. Zajic. Moderate deviations for queues with long-range dependent input. *Stochastic Networks, Lecture Notes in Statistics*, 117:275–298, 1996.

[57] J.T. Chang and Y. Peres. Ladder heights, Gaussian random walks and the Riemann zeta function. *Annals of Probability*, 25(2):787–802, 1997.

[58] B.P.K. Chen and S. G. Henderson. Two issues in setting call center staffing levels. *Annals of Operations Research*, 108(1):175–192, 2001.

[59] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*. Number 46 in Springer Series: Stochastic Modelling and Applied Probability. Springer-Verlag, 2001.

[60] J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Review*, 18(2):49–100, 1957.

[61] J.W. Cohen. *The Single Server Queue*, volume 8 of *North-Holland Series in Applied Mathematics and Mechanics*. North-Holland Publishing Co., Amsterdam, second edition, 1982.

[62] D.R. Cox. Some statistical models connected with series of events. *Journal of the Royal Statistical Society*, 17(2):129–164, 1955.

[63] J.G. Dai, A.B. Dieker, and X. Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems*, 78(1):1–29, 2014.

[64] J.G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[65] J.G. Dai and S. He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, 21(1):1–36, 2012.

[66] J.G. Dai and P. Shi. A two-time-scale approach to time-varying queues for hospital inpatient flow management. *Operations Research, Articles in Advance*, 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2489533.

[67] N.G. de Bruijn. *Asymptotic Methods in Analysis*. Dover Publications Inc., New York, third edition, 1981.

[68] A.M. de Bruin, R. Bekker, L. van Zanten, and G.M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2009.

[69] B.T. Denton, editor. *Handbook of Healthcare Operations Management: Methods and Applications*. Springer, second edition, 2013.

[70] KC.S. Diwas. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2014.

[71] S.G. Eick, W.A. Massey, and W. Whitt. The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742, 1993.

[72] A.K Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektroteknikeren*, 13, 1917.

[73] G. Falin and J. Templeton. *Retrial Queues*. Chapman & Hall, 1997.

[74] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, first edition, 2009.

[75] B.H. Fralix, C. Knessl, and J.S.H. van Leeuwaarden. First passage times to congested states of many-server systems in the Halfin-Whitt regime. *Stochastic Models*, 30(2):162–186, 2014.

[76] D. Gamarnik and D.A. Goldberg. On the rate of convergence to stationarity of the $M/M/N$ queue in the Halfin-Whitt regime. *The Annals of Applied Probability*, 23(5):1879–1912, 2013.

[77] D. Gamarnik and D.A. Goldberg. Steady-state $GI/G/N$ queue in the Halfin-Whitt regime. *The Annals of Applied Probability*, 23(6):2382–2419, 2013.

[78] D. Gamarnik and P. Momcilovic. Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 40(2):548–577, 2008.

[79] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, 16(1):56–90, 2006.

[80] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[81] N. Gans, H. Shen, Y. Zhou, N. Korolev, A. McCord, and H. Ristock. Parametric stochastic programming models for call-center workforce scheduling. *Manufacturing & Service Operations Management*, 17(4):571–588, 2015.

[82] O. Garnett, A. Mandelbaum, and M.I. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[83] D.P. Gaver. Imbedded Markov chain analysis of a waiting-line process in continuous time. *The Annals of Mathematical Statistics*, 30(3):698–720, 1959.

[84] D.P. Gaver. Diffusion approximations and models for certain congestion problems. *Journal of Applied Probability*, 5(3):607–623, 1968.

[85] P. Ghandforoush and T.K. Sen. A DSS to manage platelet production supply chain for regional blood centers. *Decision Support Systems*, 50(1):32–42, 2010.

[86] D.A. Goldberg. On the steady-state probability of delay and large negative deviations for the $GI/GI/n$ queue in the Halfin-Whitt regime. https://arxiv.org/abs/1307.0241v2, 2013.

[87] J. Grandell. *Mixed Poisson Processes*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1997.

[88] L. Green and N. Yankovic. Identifying good nursing levels: A queuing approach. *Operations Research*, 59(4):942–955, 2011.

[89] L.V. Green. *Operations Research and Health Care: A Handbook of Methods and Applications*, chapter 1. Kluwer, 2004.

[90] L.V. Green. Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[91] L.V. Green and P. Kolesar. The pointwise stationary approximation for queues with non-stationary arrivals. *Management Science*, 37(1):84–97, 1991.

[92] L.V. Green, P. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549 – 564, 2001.

[93] L.V. Green and S. Savin. Reducing delays for medical appointments: a queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[94] L.V. Green, S.S. Savin, and M. Murray. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety*, 33(4):211–218, 2007.

[95] V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 46(9):1062–1081, 2007.

[96] I. Gurvich. Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines. *Mathematics of Operations Research*, 39(1):121–162, 2013.

[97] I. Gurvich, M. Armony, and A. Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[98] I. Gurvich, J. Huang, and A. Mandelbaum. Excursion-based universal approximations for the Erlang-A queue in steady-state. *Mathematics of Operations Research*, 39(2):325–373, 2014.

[99] I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call-centers with uncertain demand forecasts: a chance-constrained optimization approach. *Management Science*, 56(7):1093–1115, 2010.

[100] I. Gurvich and W. Whitt. Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34(2):363–396, 2009.

[101] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[102] R.W. Hall, editor. *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer, 2006.

[103] R.W. Hall, editor. *Handbook of Healthcare System Scheduling*. Springer, 2012.

[104] J.M. Harrison. *Brownian Motion and Stochastic Flow Systems*. John Wiley & Sons, 1985.

[105] J.M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 52(2):243–257, 2004.

[106] J. Huang, B. Carmeli, and A. Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

[107] J. Huang and I. Gurvich. Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784752, 2016.

[108] D.L. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2(2):429–441, 1965.

[109] D.L. Iglehart. Weak convergence in queueing theory. *Advances in Applied Mathematics*, 5(3):570–594, 1973.

[110] D.L. Iglehart. Weak convergence of compound stochastic processes, I. *Stochastic Processes and their Applications*, 1(1):11–31, 1973.

[111] D.L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

[112] J.R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

[113] A.J.E.M. Janssen and J.S.H. van Leeuwaarden. Analytic computation schemes for the discrete-time bulk service queue. *Queueing Systems*, 50(2):141–163, January 2005.

[114] A.J.E.M. Janssen and J.S.H. van Leeuwaarden. Relaxation time for the discrete $D/G/1$ queue. *Queueing Systems*, 50(1):53–80, 2005.

[115] A.J.E.M. Janssen and J.S.H. van Leeuwaarden. On Lerch's transcendent and the Gaussian random walk. *The Annals of Applied Probability*, 17(2):421–439, 2006.

[116] A.J.E.M. Janssen and J.S.H. van Leeuwaarden. Cumulants of the maximum of the Gaussian random walk. *Stochastic Processes and their Applications*, 117(12):1928–1959, 2007.

[117] A.J.E.M. Janssen and J.S.H. van Leeuwaarden. Back to the roots of the $M/D/s$ queue and the works of Erlang, Crommelin, and Pollaczek. *Statistica Neerlandica*, 62(3):299–313, 2008.

[118] A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and B.W.J. Mathijsen. Novel heavy-traffic regimes for large-scale service systems. *SIAM Journal of Applied Mathematics*, 75(2):787–812, 2015.

[119] A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and J. Sanders. Scaled control in the QED regime. *Performance Evaluation*, 70(10):750–769, 2013.

[120] A.J.E.M Janssen, J.S.H. van Leeuwaarden, and A.P. Zwart. Refining square-root safety staffing by expanding Erlang-C. *Operations Research*, 59(6):1512–1522, 2011.

[121] A.J.E.M. Janssen, J.S.H. vvan Leeuwaarden, and A.P. Zwart. Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. *Queueing Systems*, 58(4):261–301, 2008.

[122] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems*, 47:53–69, 2004.

[123] O.B. Jennings and F. de Véricourt. Dimensioning large-scale membership services. *Operations Research*, 55(1):173–187, 2008.

[124] O.B. Jennings and F. de Véricourt. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[125] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[126] O.B. Jennings and J.E. Reed. An overloaded multiclass FIFO queue with abandonments. *Operations Research*, 60(5):1282–1295, 2012.

[127] G. Jongbloed and G. Koole. Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, March 2001.

[128] W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability*, 22(2):477–521, 2012.

[129] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *The Annals of Applied Probability*, 21(1):33–114, 2011.

[130] H. Kaspi and K. Ramanan. SPDE limits of many-server queues. *The Annals of Applied Probability*, 23(1):145–229, 2013.

[131] G. Kazahaya. Harnessing technology to redesign labor cost management reports. *Healthcare Financial Management*, 59(4):94–100, 2005.

[132] J. Keilson and N.D. Mermin. The second-order distribution of integrand shot noise. *IRE Transactions on Information Theory*, 5:75–77, 1959.

[133] F.P. Kelly. Stochastic models of computer communication systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3):379–395, 1985.

[134] D.G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society*, 113(2):151–185, 1951.

[135] P. Khudyakov. Designing a call center with an IVR (Interactive Voice Response). Master's thesis, Technion, 2006.

[136] P. Khudyakov, P.D. Feigin, and A. Mandelbaum. Designing a call center with an IVR (Interactive Voice Response). *Queueing Systems*, 66(3):215–237, 2010.

[137] S-H. Kim, P. Vel, W. Whitt, and W.C. Cha. Poisson and non-Poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters*, 43(3):247–253, 2015.

[138] S-H. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

[139] S-H. Kim, W. Whitt, and W.C. Cha. A data-driven model of an appointment-generated arrival process at an outpatient clinic. Unpublished manuscript, 2015.

[140] J.F.C. Kingman. The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57(4):902–904, 1961.

[141] J.F.C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B*, 24(2):383–392, 1962.

[142] L. Kleinrock. *Queueing Systems, Volume 2: Computer Applications*. Wiley, 1976.

[143] L. Kleinrock. *Communication Nets: Stochastic Message Flow and Delay*. Dover, 2007.

[144] Y.L. Koçaga, M. Armony, and A.R. Ward. Staffing call centers with uncertain arrival rate and co-sourcing. *Production and Operations Management*, 24(7):1101–1117, 2015.

[145] J. Köllerström. Heavy traffic theory for queues with several servers. I. *Journal of Applied Probability*, 11(3):544–552, 1974.

[146] J. Köllerström. Heavy traffic theory for queues with several servers. II. *Journal of Applied Probability*, 16(2):393–401, 1979.

[147] A.E. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, 2006.

[148] D.V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):277–289, 1952.

[149] S. Lundgren and K. Segesten. Nurses use of time in a medical-surgical ward with all-RN staffing. *Journal of Nursing Management*, 9(1):13–20, 2001.

[150] S. Maman. Uncertainty in the demand for service: The case of call centers and emergency departments. Master's thesis, Technion – Israel Institute of Technology, 2009.

[151] A. Mandelbaum, W.A. Massey, M. Reiman, B. Rider, and A. Stolyar. Queue lengths and waiting times for multiserver queues with abandonment and retrials. In *Selected Proceedings of the Fifth INFORMS Telecommunications Conference*, 2000.

[152] A. Mandelbaum, W.A. Massey, M.I. Reiman, and B. Rider. Time varying multiserver queues with abandonment and retrials. In *ICT-16*, 1999.

[153] A. Mandelbaum, W.A. Massey, M.I. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2–4):149–171, 2002.

[154] A. Mandelbaum, W.A. Massey, M.I. Reiman, and A.L. Stolyar. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. In *Proceedings of the Allerton Conference*, 1999.

[155] A. Mandelbaum and P. Momčilovic. Queues with many servers: The virtual waiting-time process in the QED regime. *Mathematics of Operations Research*, 33(3):561–586, 2008.

[156] A. Mandelbaum and P. Momčilovic. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

[157] A. Mandelbaum, P. Momčilovic, and Y. Tseytlin. On fair routing from eds to hospital wards. *Management Science*, 58(7):1273–1291, 2012.

[158] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

[159] A. Mandelbaum and S. Zeltyn. Data-stories about (im)patient customers in tele-queues. *Queueing Systems*, 75(2):115–146, 2013.

[160] W.A. Massey and R.B. Wallace. An asymptotically optimal design of the $M/M/c/k$ queue. *Unpublished report*, 2004.

[161] W.A. Massey and W. Whitt. An analysis of the modified offered-load approximation for the non-stationary Erlang loss model. *The Annals of Applied Probability*, 4(4):1145–1160, 1994.

[162] W.A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *The Annals of Applied Probability*, 1998.

[163] B.W.J. Mathijsen, A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and A.P. Zwart. Robust heavy-traffic approximations for service systems facing overdispersed demand. https://arxiv.org/abs/1512.05581, 2017.

[164] B.W.J. Mathijsen and A.P. Zwart. Transient error approximation in a Lévy queue. *Queueing Systems*, 85(3):269–304, 2017.

[165] V. Mehrotra, O. Ozlük, and R. Saltzmann. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management*, 19(3):353–367, 2010.

[166] P. Mell and T. Grance. The NIST definition of cloud computing. Technical report, NIST, 2011.

[167] S.V. Nagaev. Large deviations of sums of independent random variables. *Annals of Probability*, 7(5):745–789, 1979.

[168] M.F. Neuts. The single server queue with Poisson input and semi-Markov service times. *Journal of Applied Probability*, 3(1):202–230, 1966.

[169] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. The Johns Hopkins University Press, 1981.

[170] G.F. Newell. Queues for a fixed-cycle traffic light. *The Annals of Mathematical Statistics*, 31(3):589–597, 1960.

[171] G.F. Newell. *Approximate Stochastic Behavior of n-Server Service Systems with Large n*. Number 87 in Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, 1973.

[172] G.F. Newell. *Applications of Queueing Theory*. Chapman & Hall, 1982.

[173] A.R. Odoni and E. Roth. An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research*, 31(3):432–455, 1983.

[174] F.W. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[175] R.C.A. Palm. *Research on telephone traffic carried by full availability groups*. Tele, 1957.

[176] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.

[177] C.D. Pegden and M. Rosenshine. Some new results for the $M/M/1$ queue. *Management Science*, 28(7):821–828, 1982.

[178] N.U. Prabhu. Time-dependent results in storage theory. *Journal of Applied Probability*, 1(1):1–46, 1964.

[179] A.A. Puhalskii. Moderate deviations for queues in critical loading. *Queueing Systems*, 31(3):359–392, 1999.

[180] A.A. Puhalskii and J.E. Reed. On many-server queues in heavy traffic. *The Annals of Applied Probability*, 20(1):129–195, 2010.

[181] A.A. Puhalskii and M.I. Reiman. The multiclass $GI/Ph/N$ queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):564–595, 2000.

[182] A.A. Puhalskii and W. Whitt. Functional large deviation principles for waiting and departure processes. *Probability in the Engineering and Informational Sciences*, 12(4):479–507, 1998.

[183] R.S. Randhawa. Optimality gap of asymptotically derived prescriptions in queueing systems: $o(1)$-optimality. *Queueing Systems*, 83(1):131–155, 2016.

[184] J.E. Reed. The $G/GI/N$ queue in the Halfin-Whitt regime. *The Annals of Applied Probability*, 19(6):2211–2269, 2009.

[185] J.E. Reed and T. Tezcan. Hazard rate scaling for the $GI/M/n + GI$ queue. *Operations Research*, 70(3):1–34, 2012.

[186] J.E. Reed and A.P. Zwart. A piecewise linear stochastic differential equation driven by a Lévy process. *Journal of Applied Probability*, 48A:109–119, 2011.

[187] T.R. Robbins, D.J. Medeiros, and T.P. Harrison. Does the Erlang C model fit in real call centers? In *Proceedings of the 2010 Winter Simulation Conference*, 2010.

[188] S.M. Ross. *Stochastic Processes*. John Wiley & Sons, 1996.

[189] M.R. Sampford. Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics*, 24(1):130–132, 1953.

[190] K.-I. Sato. *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press, 1999.

[191] D. Siegmund. Corrected diffusion approximations in certain random walk problems. Technical report, Stanford University, 1978.

[192] K. Sigman and W. Whitt. Heavy-traffic limits for nearly deterministic queues. *Journal of Applied Probability*, 48(3):657–678, 2011.

[193] K. Sigman and W. Whitt. Heavy-traffic limits for nearly deterministic queues: Stationary distributions. *Queueing Systems*, 69:145–173, 2011.

[194] D. Sinreich and Y.N. Marmor. Emergency department operations: the basis for developing a simulation tool. *IIE transactions*, 37(3):233–345, 2005.

[195] L.J. Slater. *Confluent Hypergeometric Functions*. Cambridge University Press, 1960.

[196] H. Song, A.L. Tucker, and K.L. Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053, 2015.

[197] F. Spitzer. *Principles of Random Walk.* D. van Nostrand Company, Inc., 1964.

[198] S.H.W. Stanger, N. Yates, R. Wilding, and S. Cotton. Blood inventory management: Hospital best practice. *Transfusion Medicine Reviews*, 26(2):153–163, 2012.

[199] S.G. Steckley and S.G. Henderson. The error in steady-state approximations for the time-dependent waiting time distribution. *Stochastic Models*, 23(2):307–332, 2007.

[200] S.G. Steckley, S.G. Henderson, and V. Mehrotra. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences*, 23(2):305–332, 2009.

[201] M.E. Steiner, S.F. Assmann, J.H. Levy, J. Marshall, S. Pulkrabek, S.R. Sloan, D. Triulzi, and C.P. Stowell. Addressing the question of the effect of RBC storage on clinical outcomes: The red cell storage duration study. *Transfusion and Apheresis Science*, 43(1):107–116, 2010.

[202] A.L. Stolyar and T. Tezcan. Control of systems with flexible multi-server pools: a shadow routing approach. *Queueing Systems*, 66(1):1–51, 2010.

[203] L. Takács. Investigation of waiting time problems by reduction to markov processes. *Acta Mathematica Academiae Scientiarum Hungarica*, 6(1):101–129, 1955.

[204] L. Takács. The time dependence of a single-server queue with Poisson input and general service times. *The Annals of Mathematical Statistics*, 33(4):1340–1348, 1962.

[205] J. Tan, H. Feng, X. Meng, and L. Zhang. Heavy-traffic analysis of cloud provisioning. In *Proceedings of the 24th International Teletraffic Congress*, 2012.

[206] T. Tezcan and J.G. Dai. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*, 58(1):94–110, 2010.

[207] J.S.H. van Leeuwaarden. *Queueing Models for Cable Access Networks*. PhD thesis, Eindhoven University of Technology, 2005.

[208] J.S.H. van Leeuwaarden. Delay analysis for the fixed-cycle traffic light queue. *Transportation Science*, 40(2):189–199, 2006.

[209] J.S.H. van Leeuwaarden and C. Knessl. Transient behavior of the Halfin-Whitt diffusion. *Stochastic Processes and their Applications*, 21(7):1524–1545, 2011.

[210] J.S.H. van Leeuwaarden and C. Knessl. Spectral gap of the Erlang-A model in the Halfin-Whitt regime. *Stochastic Systems*, 2(1):149–207, 2012.

[211] J.S.H. van Leeuwaarden, B.W.J. Mathijsen, and F. Sloothaak. Delayed workload shifting in many-server systems. *ACM SIGMETRICS Performance Evaluation Review*, 43(2):10–12, 2015.

[212] J.S.H. van Leeuwaarden, B.W.J. Mathijsen, and F. Sloothaak. Cloud provisioning in the QED regime. *Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 180–187, 2016.

[213] J.S.H. van Leeuwaarden, B.W.J. Mathijsen, F. Sloothaak, and G.B. Yom-Tov. The restricted Erlang-R queue: Finite-size effects in service systems with returning customers. https://arxiv.org/abs/1612.07088, 2016.

[214] A.R. Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO single class models. *Surveys in Operations Research and Management Science*, 17(1):1–14, 2012.

[215] A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43(1):103–128, 2003.

[216] A.R. Ward and P.W. Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[217] W. Whitt. *Heavy Traffic Limit Theorems for Queues: A Survey*, pages 307–350. Springer Berlin Heidelberg, 1974.

[218] W. Whitt. On the heavy-traffic limit theorem for $\mathrm{GI/G/\infty}$ queues. *Advances in Applied Probability*, 14(1):171–190, 1982.

[219] W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314, 1991.

[220] W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212, 1999.

[221] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

[222] W. Whitt. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research*, 30(1):1–27, 2005.

[223] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102, 2006.

[224] R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[225] G.B. Yom-Tov. *Queues in hospitals: Queueing networks with re-entering customers in the QED regime*. PhD thesis, Technion, 2010.

[226] G.B Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

[227] C. Zacharias and M. Armony. Joint panel sizing and appointment scheduling in outpatient care. *Management Science: Articles in Advance*, 2016.

[228] J. Zan. *Staffing service centers under arrival-rate uncertainty*. PhD thesis, University of Texas, 2012.

[229] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems*, 51(3):361–402, 2005.

[230] B. Zhang, J.S.H. van Leeuwaarden, and A.P. Zwart. Staffing call centers with impatient customers: refinements to many-server asymptotics. *Operations Research*, 60(2):461–474, 2012.

[231] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Systems*, 73(2):147–193, 2013.

[232] N. Zychlinski, A. Mandelbaum, P. Momčilović, and I. Cohen. Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. Unpubllished manuscript, 2016.

Summary

Asymptotic dimensioning of stochastic service systems

Stochastic service systems describe situations in which customers compete for service from scarce resources. Think of check-in lines at airports, waiting rooms in hospitals or queues in supermarkets, where the scarce resource is human manpower. Next to these traditional settings, resource sharing is also important in large-scale service systems such as the internet, wireless networks and cloud computing facilities. In these virtual environments, geographical conditions do not restrict the system size, paving the way for the emergence of large-scale resource sharing networks. This thesis investigates how to design large-scale systems in order to achieve the dual goal of operational efficiency and quality-of-service, by which we mean that the system is highly occupied and hence efficiently utilizes the expensive resources, while at the same time, the level of service, experienced by customers, remains high.

The intrinsic stochastic variability of arrival and service processes is the predominant cause of delays experienced by customers. Queueing theory and stochastics provide the tools to describe and evaluate congestion in these systems. An important insight obtained through queueing analysis is the effect of resource pooling for systems with many servers and corresponding economies-of-scale that can be achieved by increasing the scale of the system. Although classical queueing theory allows for exact evaluation of the performance of queueing systems of moderate size, exact analysis becomes intractable as demand R and capacity s become large. In those cases, one typically resorts to asymptotic approximation techniques, such as heavy-traffic diffusion approximations: the analysis of a sequence of queueing processes, scaled in space, in which the server utilization level approaches 100%. The resulting probabilistic limiting processes are easier to analyze. Moreover, the diffusion approximations have direct interpretations in terms of the original systems and lead to tractable characterizations of their performance.

The heavy-traffic regime that plays a central role in this thesis is the Halfin-Whitt regime, also known as the Quality-and-Efficiency Driven (QED) regime, which dictates that capacity should be equal to the nominal demand plus an additional variability hedge which is proportional to the square-root of the nominal load, i.e. $s = R + \beta\sqrt{R}$ for some $\beta > 0$. The driving force behind this scaling regime is the central limit theorem (CLT). The rule $s = R + \beta\sqrt{R}$, commonly known as the square-root staffing principle, has been proved to secure both efficiency (utilization approaches 100%) and quality-of-service, since the mean waiting time is negligible under this scaling as the system grows large. Since the QED regime allows coexistence of the two seemingly conflicting objectives in large-scale service systems, the paradigm has been implemented in a wide variety of operational settings. However, the standard QED regime fails to acknowledge features that play a dominant role in practice. This thesis contributes to the existing literature by identifying these distinctive traits and showing how to account for them in a modified QED framework.

In Chapters 2 & 3, we study how the limiting behavior of many-server queues is affected when one deviates from the standard square-root staffing principle. In Chapter 2 we investigate a novel family of scaling regimes, in which the amount of overcapacity $s - R$ is not necessarily of the order $\sqrt{R}$, which gives rise to a novel family of heavy-traffic regimes and corresponding scaling limits. Continuing our study of alternative scaling regimes, we investigate in Chapter 3 how to adapt the square-root staffing paradigm in case the system faces demand patterns that are stochastically more volatile than anticipated. This phenomenon is known as overdispersion and can be caused by e.g. the existence of correlation between the sources generating demand, or uncertainty about the arrival volume.

In Chapters 4 & 5, we review a family of queueing models in the QED regime in which the total number of customers that can reside in the system simultaneously is limited. As a result, customers may be denied access in case they find a full system on arrival. This fraction of arrivals may either reattempt later or leave the system directly. The impact of retrials on scaling rules in the QED regime is the focus of Chapter 4. Since the volume of initially blocked customers is proportional to $\sqrt{R}$, that is, the same order as the variability hedge in the staffing rule, retrials are prone to have a non-negligible effect on performance. We propose a heuristic method for the performance analysis of these types of queueing models with finite-size restrictions, which is based on a fixed-point equation. As a by-product this yields a two-fold square-root staffing principle, which prescribes a synchronous scaling for both the system capacity and waiting space. Chapter 5 describes how these ideas can be applied in the context of an emergency department.

Chapter 6 studies a cost minimization problem in a single-server queue with non-stationary input. The bulk of the queueing literature concerns performance analysis assuming that steady state is reached. However, the validity of this assumption in practice is questionable, for the simple fact that no service system runs infinitely long. Moreover, system parameters, such as the arrival volume, are likely to change over time. In this chapter, we characterize the error in performance metrics that follows from this transient nature of queues, and present a correction to the original staffing rule to account for the finite time horizon.

Finally, we analyze in Chapter 7 a specific stochastic service system: an inventory model of a blood bank with backlogs, perishable goods and consumer impatience. We obtain the stationary distribution of the inventory level, and deduce under appropriate scaling the stochastic process limit in terms of a diffusion process. This process limit allows for a more tractable approximate analysis of the model in case the number of blood deliveries and demand is large.

About the author

Britt Mathijsen was born in 's-Hertogenbosch, The Netherlands, on May 12, 1990. After completing secondary education at Gymnasium Bernrode in Heeswijk-Dinther in 2008, she went to study Industrial & Applied Mathematics at Eindhoven University of Technology (TU/e). In 2011, she received her Bachelor's degree with highest honors (cum laude). Following a final project in collaboration with Philips Research in Eindhoven, she obtained her Master's degree with highest honors (cum laude) in September 2013.

Britt then continued as a PhD student within the Stochastic Operations Research section at TU/e under the supervision of Johan van Leeuwaarden and Bert Zwart on the topic of the asymptotic analysis of large-scale service systems. Next to doing research, she has been involved in teaching activities regarding probability theory and stochastic simulation. During her PhD, she also acted as the treasurer of the department's PhD Council (2014-2016) and co-organized the tenth edition of the Young European Queueing Theorists (YEQT) workshop in 2016.

Britt presented her work at several international conferences, among which the European Conference on Queueing Theory 2014 (Gent, Belgium), Applied Probability Society Conference 2015 (Istanbul, Turkey), Performance 2015 (Sydney, Australia), INFORMS Annual Meeting 2015 (Philadelphia, United States), VALUETOOLS (Berlin, Germany), Operations Research Society of Israel Conference 2016 (Jeruzalem, Israel), European Conference on Queueing Theory 2016 (Toulouse, France) and INFORMS Annual Meeting 2016 (Nashville, United States). Furthermore, she paid multiple visits to the group of Avishai Mandelbaum at the Technion in Haifa, Israel.

Britt will defend her PhD thesis at TU/e on May 18, 2017. As of April 2017, she starts working at Consultants in Quantitative Methods (CQM) in Eindhoven.