Performance Analysis of Multi-Class Queueing Models

Petra Vis
Performance Analysis of Multi-Class Queueing Models
ISBN: 978-94-6332-213-3

Printed by: GVO drukkers & vormgevers B.V.

VRIJE UNIVERSITEIT

Performance Analysis of Multi-Class Queueing Models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op maandag 18 september 2017 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Petra Vis

geboren te Hoorn

promotor: prof.dr. R.D. van der Mei
copromotor: dr. R. Bekker

Dankwoord

Dit proefschrift is het resultaat van vier jaar onderzoek. Alhoewel alleen mijn naam op de voorkant staat, was dit zeker niet gelukt zonder de hulp van vele anderen, die ik in dit dankwoord graag wil benoemen.

Dit proefschrift had nooit tot stand kunnen komen zonder de hulp van mijn begeleiders Rob van der Mei en René Bekker. Ik wil jullie graag bedanken voor het feit dat ik met problemen altijd bij jullie terecht kon en voor jullie enthousiasme, waarmee ik altijd weer gemotiveerd werd. René wil ik in het bijzonder bedanken voor onze wekelijkse besprekingen, waarna ik altijd weer een stuk verder was.

Ik dank het CWI voor de gastvrijheid, die me in staat stelde deel uit te maken van de Stochastics-groep. Met veel plezier denk ik terug aan het tafeltennissen met de CWI-collega's na de lunch. Op de VU begon ik als enige AiO, maar in de loop van de tijd werden het er steeds meer. Bij dezen wil ik alle collega's op de VU en het CWI bedanken, omdat ik altijd bij ze terecht kon voor hulp of gezelligheid.

De papers waarop de hoofdstukken zijn gebaseerd, heb ik niet alleen geschreven. Rob, René, Jan-Pieter Dorsman, Erik Winands, Sindo Núñez Queija, en Ger Koole wil ik bedanken voor hun bijdragen aan verschillende hoofdstukken. Ik zou ook graag de leden van mijn promotiecommissie, bestaande uit Sandjai Bhulai, Dieter Fiems, Urtzi Ayesta, Onno Boxma en Werner Scheinhardt, bedanken voor het evalueren van mijn proefschrift.

Als laatste bedank ik ook mijn familie, omdat ze mij steunden, interesse toonden in mijn werk en mij hebben geholpen wanneer dat mogelijk was.

Petra Vis
juli 2017

Contents

1 Introduction **1**
- 1.1 Types of multi-class queues . . . 3
- 1.2 Background and motivation . . . 4
  - 1.2.1 Polling model . . . 4
  - 1.2.2 Processor Sharing queue . . . 6
  - 1.2.3 Priority queues . . . 7
- 1.3 Overview of this thesis . . . 10

2 Gated and globally gated polling systems with non-FCFS service **13**
- 2.1 Introduction . . . 13
- 2.2 Model description . . . 15
- 2.3 Analysis of models with gated service . . . 17
  - 2.3.1 First-Come-First-Served . . . 18
  - 2.3.2 Last-Come-First-Served . . . 19
  - 2.3.3 Random Order of Service . . . 21
  - 2.3.4 Processor Sharing . . . 25
  - 2.3.5 Shortest-Job-First . . . 29
- 2.4 Results for models with globally gated service . . . 32
- 2.5 Approximations for systems with arbitrary load . . . 34
- 2.6 Concluding remarks . . . 38

3 Exhaustive polling systems with non-FCFS service **39**
- 3.1 Introduction . . . 39
- 3.2 Model description . . . 40
- 3.3 Preliminaries and method outline . . . 41
  - 3.3.1 Cycle and intervisit times . . . 41
  - 3.3.2 First-Come-First-Served . . . 43
- 3.4 Last-Come-First-Served . . . 44
  - 3.4.1 Non-Preemptive LCFS . . . 45
  - 3.4.2 LCFS with Preemptive Resume . . . 48
- 3.5 Random Order of Service . . . 49
- 3.6 Processor Sharing . . . 55

3.6.1 Conditional waiting-time distribution in heavy traffic 55
3.6.2 Unconditional waiting-time distribution in heavy traffic 57
3.7 n-class priority queues 60
3.7.1 Non-preemptive n-class priority queues 60
3.7.2 Preemptive n-class priority queues 63
3.8 SJF and SRPT 64
3.8.1 Conditional waiting-time distribution in heavy traffic 64
3.8.2 Unconditional waiting-time distribution in heavy traffic 66
3.8.3 SRPT and preemptive SJF 67
3.9 Summary of the results 69
3.10 Closed-form approximations for systems with arbitrary load 71
3.11 Discussion and concluding remarks 73

4 Transient analysis of cycle times in polling systems 77
4.1 Introduction 77
4.2 Model description 78
4.3 Analysis of globally gated service 80
4.3.1 Asymptotic properties 85
4.3.2 Numerical results 87
4.4 Analysis of gated service 89
4.4.1 Asymptotic properties 92
4.4.2 Numerical results 95
4.5 Appendix 95
4.5.1 Proofs 95
4.5.2 Second-order derivatives 98

5 Queue-length distributions in DPS queues with batch arrivals 101
5.1 Introduction 101
5.2 Model description 103
5.3 Main result 104
5.4 Analysis 105
5.4.1 Balance equations and functional equation 106
5.4.2 Heavy-traffic limit 108
5.4.3 Specifying the common distribution 109
5.5 Numerical results 111
5.5.1 State-space collapse 111
5.5.2 Convergence and approximation of moments 111
5.5.3 The impact of batch arrivals 113

6 Analysis of level-dependent MAP/G/1 queues 115
6.1 Introduction 115
6.2 Model description 118
6.2.1 Characteristics of appointment systems 118

6.2.2 Model I: backlog in slots . . . 119
6.2.3 Model II: backlog in days . . . 120
6.3 State-dependent M/G/1 queue . . . 121
6.3.1 Model and method outline . . . 121
6.3.2 Performance analysis . . . 122
6.4 Level-dependent MAP/G/1 queue . . . 126
6.4.1 The arrival process . . . 126
6.4.2 Performance analysis . . . 128
6.4.3 Queue length at arbitrary moments . . . 131
6.5 Numerical experiments . . . 132
6.5.1 Level-independent case . . . 133
6.5.2 Level-dependent case . . . 133
6.5.3 Optimization . . . 136
6.6 Appendix . . . 137
6.6.1 Proof of Theorem 6.1 . . . 137
6.6.2 Proof of Theorem 6.2 . . . 139
6.6.3 Mean queue length for LD-MAP/G/1 . . . 140

7 Waiting-time distributions in call blending models with abandonments 143
7.1 Introduction . . . 143
7.2 Model description . . . 145
7.3 Analysis for equal service requirements . . . 146
7.3.1 Process description . . . 147
7.3.2 Analysis of FIL distribution . . . 149
7.3.3 Performance measures . . . 151
7.4 Analysis for different service requirements . . . 153
7.4.1 Level crossings . . . 153
7.4.2 General impatience . . . 155
7.4.3 Infinite patience . . . 160
7.4.4 Constants and boundary conditions . . . 162
7.4.5 Performance measures . . . 165

Bibliography 169

Summary 181

Chapter 1

Introduction

In this thesis we study multi-class queueing models. In general, queueing models provide a natural means to describe the phenomenon of congestion, and find many applications in everyday life. Examples of such applications range from systems with visible congestion where people are waiting in line (e.g., grocery stores, amusement parks, road networks) to systems with queueing at a more abstract level (e.g., call centers, communication networks, manufacturing, computer systems). The main entities in queueing models are *nodes*, which consist of one or more shared *servers*, some amount of buffer space (finite or infinite), and *customers* (or *jobs*) that arrive at nodes and require some amount of service. We will use the terms customers and jobs interchangeably. Queueing models typically describe the *arrival process* of customers at each node, the *service-time distributions*, the *routing* of customers and servers between the nodes, and the *service process*, describing the way the server capacity is shared among the customers present at a node.

In a *multi-class* queueing model the customers can be of different types. Each type of customers typically has its own arrival process, routing scheme and service process. Figure 1.1 gives a representation of a basic, single-node, single-server, multi-class queueing model. Different customer classes may be treated differently by the server.

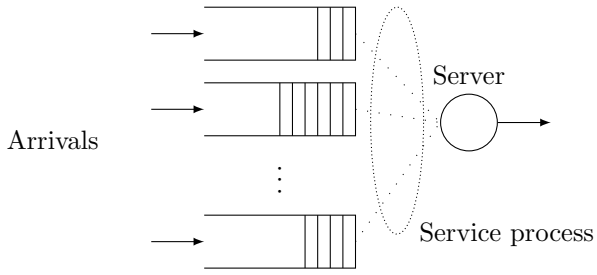


Figure 1.1: A basic single-node multi-class queue.

FCFS	*First-Come-First-Served* serves jobs in the order of arrival.
LCFS	*Last-Come-First-Served* serves the job that arrived most recently, without preemption.
LCFS-PR	*Last-Come-First-Served with preemptive resume* serves the job that arrived most recently preempting the job currently in service.
ROS	*Random Order of Service* randomly selects a job from the jobs that are waiting.
PS	*Processor Sharing* serves all jobs simultaneously at the same rate.
DPS	*Discriminatory Processor Sharing* serves all jobs simultaneously, but at different rates per class.
NPRIOR	*n-class priority regime* serves jobs within the highest priority class first, continuing with other priority classes as long as no jobs with higher priority are present. Jobs within the same priority class are served in the order of arrival.
NPRIOR-PR	*n-class priority regime with preemptive resume* serves jobs with higher priority first, preempting jobs with lower priority which are already in service, jobs within the same priority class are served FCFS.
SJF	*Shortest-Job-First* non-preemptively serves the job in the system with the smallest original service time.
SRPT	*Shortest-Remaining-Processing-Time* preemptively serves the job with the shortest remaining processing time.

Table 1.1: A brief description of the scheduling policies discussed in this thesis.

For example, customers within a 'high' class might have (relative) priority over customers of 'lower' classes. It is also possible that a server needs some amount of (possibly zero) *switch-over time* to be set up for a different class. The server serves one class of customers at a time and switches between the classes.

There are many types of multi-class queues and they have received much attention in the literature. In this thesis we study a variety of multi-class queueing models. We are primarily interested in gaining fundamental understanding of the intrinsic behavior of these queues; some of the multi-class queues that we study are inspired by a specific application.

In the remainder of this chapter, we describe the most common multi-class queues and give some background and motivation specific to the models appearing in this thesis. We will conclude the chapter with an overview of this thesis.

1.1 Types of multi-class queues

Multi-class queueing models arise in modeling processes that involve congestion in many application areas. In general, anything that separates the jobs into groups leads to different job types. This separation could be based on job lengths, arrival rates, priorities, ability of the server to serve the jobs, or a combination of those. The most suitable type of queueing model to use is mainly defined by the way the different job types are handled. Possible ways to handle different job types are:

- Ignore the job types and serve jobs one by one (regular queue).
- Ignore the job types and serve all jobs simultaneously, giving all of them an equal share of the server's capacity (Processor Sharing queue).
- If there are precedence relations, always serve jobs with the highest priority first (priority queue).
- Prioritize based on job lengths, typically shorter jobs get priority (Shortest-Job-First queue).
- Serve all jobs simultaneously, giving more of the server's capacity to certain job types (Generalized/Discriminatory Processor Sharing queue).
- Only serve jobs of one type for a while, then move to the next type (polling model).

Below we describe how these decisions lead to different multi-class queueing models.

Queueing notation We consider a single-node, single-server queueing model with infinite queue size and N different customer classes (see Figure 1.1). Most chapters in this thesis are based on this setting. The *arrival process* of type-i jobs is typically a Poisson process with rate λ_i, $i = 1, \ldots, N$. The *service-time distribution* of type-i jobs is general, with (finite) mean $\mathbb{E}[B_i]$, $i = 1, \ldots, N$. The *scheduling policy* prescribes in what order the arriving jobs are served. Possible scheduling policies are described in Table 1.1. The scheduling policies partly specify the type of multi-class queue. The first four policies do not consider job types, while the last four lead to priority queues. Processor Sharing types of queues are discussed separately.

Priority queue The priority queue is a common way to model systems with different types of jobs that differ in priority. Jobs belonging to a class with higher priority get served before jobs belonging to a lower priority class. This can be done preemptively and non-preemptively. With preemptive priority, an arriving higher priority job can interrupt a lower priority job, and the higher priority job is taken into service immediately. With non-preemptive priority, the service of a job in process is always

finished first, before a new job is taken into service. We note that SJF and SRPT can also be interpreted as priority queues, where priority is based on the (remaining) job size.

Polling model In a polling model, the different customer types are considered to arrive in different queues. The server only serves one queue at a time and switches between the queues to serve all customers. Such a model arises, e.g., when the server needs to be set up before it can serve a certain job type. Modeling decisions are when to switch to the next queue, which queue to serve next, and the order of service within each queue (as in Table 1.1).

Processor Sharing queue In a Processor Sharing (PS) queue, all job types are served *simultaneously*, instead of only serving one job type at the same time (as in e.g. priority queues). In a classical PS queue, all jobs that are present at a node fairly share the available amount of service capacity; this is referred to as *Egalitarian Processor Sharing* (EPS). Processor Sharing is advantageous for the overall system performance since short jobs will not be extremely delayed by long jobs and long jobs will also be served continuously (in contrast to SJF and SRPT). Just like with the priority queue, we can assign priorities to the different classes. Depending on its type, a job could receive more or less of the server's capacity than another job. Important jobs will receive more server capacity and consequently will be served (relatively) faster. Dividing the server's capacity between the different jobs can be done in different ways. The first way is referred to as *Discriminatory Processor Sharing* (DPS), where each job gets a share of the server capacity based on its type and the total number of jobs in the system. Another possibility is giving each job type a share of the capacity based on the total number of types in the system. Within each type, the available capacity is either divided equally between the jobs of that type, or only one job per type gets served. This is referred to as *Generalized Processor Sharing* (GPS). We refer to [142] and references therein for an overview and literature on this topic.

1.2 Background and motivation

In this section we discuss three relevant classes of multi-class queueing models that are directly related to the models in this thesis. For each class of models, we give a description, some applications and some references to relevant literature.

1.2.1 Polling model

The basic polling model is a single-server, multi-queue system, where the server visits the queues in some order to serve customers pending at the queues. This is depicted in

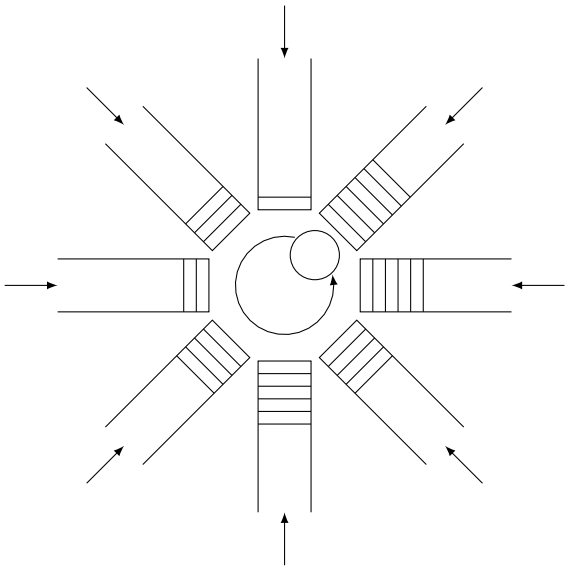


Figure 1.2: A basic polling model.

Figure 1.2. We note that multi-server polling models are also possible, but are much more challenging to analyze [33; 137], so we restrict ourselves to single-server models. The most important features of a polling model are the possibility to serve a single type of job, while 'ignoring' the other types, and the switch-over times. These features distinguish the polling model from the other multi-class queueing models.

There are three decisions that need to be made about the service of the customers present in the system. The *routing policy* describes the order in which the queues are visited by the server. The most commonly used routing policy is cyclic routing, where all queues are visited once and in the same order every cycle. A natural generalization of cyclic routing is *periodic* routing, where the server visits the queues periodically according to a routing table (of finite length). This way some queues can be visited more frequently than others. Other routing policies are probabilistic routing (proceed to queue j with probability p_j), Markovian routing (proceed from queue i to queue j with probability p_{ij}) and dynamic, state-dependent routing policies (e.g., serve longest queue).

The *service discipline* specifies the duration of a visit of the server to a queue. Commonly used examples of service disciplines are:

- **Exhaustive**: Serve the queue until it is empty.
- **Gated**: Serve all jobs that were present in the queue at the 'polling instant', i.e., the beginning of a visit of the server to a queue.

- **Globally gated**: During a cycle, serve all jobs that were present in the system at the 'polling instant' of the first queue.
- **K-limited**: Serve at most K jobs or until the queue is empty (whichever occurs first).
- **Time-limited**: Serve at most t time units or until the queue is empty (whichever occurs first).

We would like to point out, that both the K-limited policy and the time-limited policy are hard to analyze, because they do not satisfy the branching property [69; 121]. Polling models using these policies can be analyzed using heuristics or iterative methods (see, e.g., [4; 68; 143]).

Finally, each queue can have its own local *scheduling policy* that determines the order in which the jobs at the queues are served (see Table 1.1 for an overview).

Polling models find many applications in areas like computer-communication systems, production systems, manufacturing systems, inventory systems and robotics (see [32] for an extensive overview). The classical application is the machine repairman model [107; 108], where the repairman visits a fixed number of machines in cyclic order, checks them and repairs them if necessary. Other examples of applications are Bluetooth and 802.11 protocols, scheduling policies at routers, and I/O subsystems in web servers [81; 138].

Motivated by their wide applicability, polling models have been extensively studied over the past few decades; see [151] for an overview of the state-of-the-art. We refer to [28; 144] and references therein for more recent publications on polling models.

1.2.2 Processor Sharing queue

Originally, the PS queue was introduced as an idealized round robin system, i.e., a system where all jobs present are served cyclically one by one for a small amount of time. In the limit, when the amount of time that each job receives service goes to zero, this leads to a PS system. By adding weights (i.e., weighted round robin), the resulting system can be modeled with a DPS queue. A DPS model is a multi-class queueing model where all the customers are served simultaneously. Customers of classes with higher weights get a larger fraction of the server's capacity than customers of classes with lower weights. It is not guaranteed that high priority jobs will finish before low priority jobs, since this is also influenced by the length of the jobs and the total number of jobs present. For this reason, we have relative priority, instead of strict priority. The advantage over strict priority is that low priority jobs will receive service even when many high priority jobs are present. DPS queues are introduced by Kleinrock [96], who compares it with First-Come-First-Served disciplines. A graphical representation of a DPS queue is given in Figure 1.3.

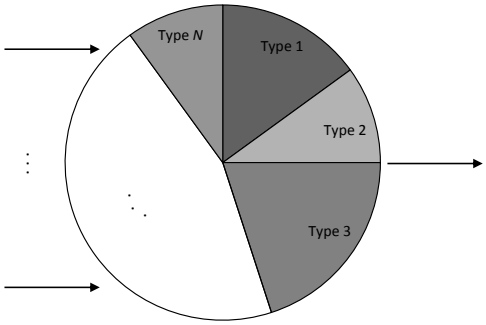


Figure 1.3: A DPS queue.

In a PS queue, the distribution of the queue length is geometric and depends only on the load and is insensitive to the distribution of the service time beyond its mean. In a DPS queue, the queue-length distribution depends on the complete service-time distribution, making the DPS model substantially more difficult to analyze. Analysis of DPS models with general service requirements are hardly found in the literature, and always make use of some limiting regime (e.g., time-scale decomposition, overload).

Applications of DPS models are mainly found in, communication networks and computers. In communication networks, the feedback mechanism in the TCP protocol causes the system to behave like a (D)PS queue, since the acknowledgments are more delayed if there are more jobs in the network. (D)PS models are also frequently used to model the behavior of the CPU in computers, where threads with different priority levels compete for access to the shared processing unit. Applications of DPS in communication networks are also discussed in [6; 46; 48; 76; 83]. Surveys of results on DPS queues are given in [7; 35]. We refer to [63; 112] and references therein for an overview of results and applications for PS queues.

1.2.3 Priority queues

In a pioneering paper, Cobham [49] introduced a queue with multiple priority classes that arrive according to a Poisson process. In case of a single server, the service time distribution is general, but for multiple servers, the service times are exponentially distributed. The result is the mean waiting time for each priority type. The higher the priority of the jobs, the lower their waiting time. Priority queues are hard to analyze in general. Cohen [53, Chapter III.3] uses transforms to analyze some special cases of M/G/1 priority queues, including two-class priority queues with preemptive resume and preemptive repeat and general priority queues with preemptive resume. In this subsection, we describe two special cases of priority queues that are motivated

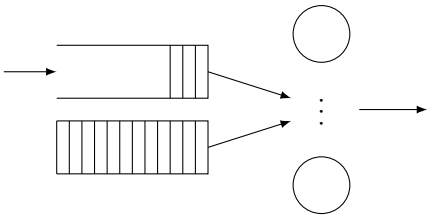


Figure 1.4: A blended system with two classes, the second class has infinite supply.

by applications in call centers and health care. For these models, we are specifically interested in the tail probabilities of the waiting time.

Blended systems The first type of priority queue is a blended system. We consider a system with two job types requiring different service levels (i.e., performance measures); type-1 jobs need to be handled as soon as possible after their arrival, while type-2 jobs need to be handled within a reasonable, but much longer, time frame. Such a differentiation in service levels for the two classes can be achieved in multiple ways. On the one hand, the system can be decoupled such that specialized or dedicated servers handle a single type of traffic. On the other hand, (part of the) servers can be cross-trained or multi-functional to handle both types; this is known as *blending*. The advantage of the latter is that the system may benefit from server flexibility. The benefits of such blended systems are discussed in [116]. Note that the decoupled approach can be analyzed as multiple single-class systems.

The blended model is motivated by call centers. Nowadays, call centers (or, more accurately, contact centers) are facing different sources of customer contacts, such as contacts by phone, email, call backs, and chats. An example of type-1, or 'urgent', jobs are inbound calls. Type-2 ('best effort') jobs could, for example, be emails or outbound calls. We assume that type-1 jobs arrive according to an arrival process, while we assume that there is an infinite supply of type-2 jobs for tractability of the analysis. This model is depicted in Figure 1.4.

Although the blended queueing model is largely motivated by call and contact centers, we envisage that mixing urgent and best effort traffic plays a prominent role in other domains. Within the health care domain, specifically in hospitals, two types of patients are typically distinguished: acute and scheduled (or elective) patients. Acute patients should be treated as soon as possible, whereas scheduled patients can wait for some amount of time and can be classified as best effort (the literature is extensive by now, but see the papers in [59; 79] for an impression). A similar distinction applies to the activity level of patients: when patients are in need of care, a swift response is required [57]. However, many activities have a less time-pressing character, including scheduled care activities and administrative duties. In the nursing home setting, this is referred to as 'care on demand' which is activated by pressing a button

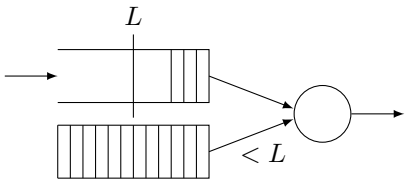


Figure 1.5: A level-dependent queue with two classes, the second class has infinite supply.

and 'care by appointment', including activities as 'giving medicine' and 'help with getting out of bed' [139]. We note that differentiation in service levels are also key in domains such as manufacturing and communication networks. As such, the required capacity and control of servers using a blended system is widely applicable in service systems.

Comprehensive surveys on contact centers can be found in [3; 73; 98]. Since we are only interested in the waiting time of the type-1 jobs and there is an infinite amount of type-2 jobs, the call blending model can also be interpreted as a vacation model. In this case, the server takes a vacation, instead of working on a type-2 job. We refer to [71] for additional references, also related to multi-server queues with vacations.

Level-dependent systems For the second special priority queue, we consider again a system with two job types. Type-2 jobs (e.g., administration tasks) are only taken into service if there are not many type-1 jobs (e.g., treating patients) waiting. Specifically, if the queue length of type-1 jobs (which we refer to as level) is above L, more type-1 jobs will be taken into service, and when the level drops below L, less type-1 jobs will be taken into service (and more type-2 jobs). Depending on the load, this may cause the queue length of type 1 to fluctuate around L. We assume that the number of type-2 jobs in the system is infinite, and we only consider the queue length of type-1 jobs, see Figure 1.5.

This queueing model is motivated by a health care application. In an outpatient clinic, patients arrive to make an appointment as soon as possible. The capacity to treat patients is limited, such that patients that cannot be directly seen move to a virtual waiting list. In our model, the type-1 customers represent the slots filled with patients on the virtual waiting list. If the waiting time is too long, e.g., more than two weeks, the number of available slots per day is increased by planning extra patients. One possibility to create this extra capacity is to perform less other activities, like administration tasks. Another possibility is to treat patients in overtime.

If more type-2 jobs are taken into service, the available service capacity of type-1 jobs decreases. Since we are only interested in the waiting time of type-1 jobs, we can get the same effect by increasing the arrival rate of type-1 jobs. From the perspective

of class 1, we thus have a level-dependent M/G/1 or MAP/G/1 queue with two different arrival processes, where the arrival process below level L differs from the arrival process above level L. In these models, the waiting time is represented by the number of type-1 jobs.

For this model, and for the blended system, we are only interested in the waiting time of type-1 jobs. Therefore we analyze both models as one-dimensional systems, thereby ignoring type-2 jobs.

1.3 Overview of this thesis

The next three chapters of this thesis are about polling models. In Chapter 2, we consider Poisson driven polling models where a single server visits the queues in a cyclic order and with general service and switch-over times. In the vast majority of papers that have appeared on polling models, it is assumed that at each of the queues the customers are served on a First-Come-First-Served (FCFS) basis. In Chapter 2 we study polling models where the local scheduling policy is not FCFS, but instead, is varied as Last-Come-First-Served (LCFS), Random Order of Service (ROS), Processor Sharing (PS) and Shortest-Job-First (SJF). The service policies are assumed to be either *gated* or *globally gated*. The main result of the chapter is the derivation of asymptotic closed-form expressions for the Laplace-Stieltjes Transform (LST) of the scaled waiting-time and sojourn-time distributions under heavy-traffic (HT) assumptions, i.e., when the system tends to saturate. For FCFS service the asymptotic sojourn-time distribution is known to be of the form $U\Gamma$, where U and Γ are uniformly and gamma distributed with known parameters. We show that the asymptotic sojourn-time distribution (1) for LCFS is also of the form $U\Gamma$, (2) for ROS is of the form $\tilde{U}\Gamma$ where $\tilde{U}$ has a *trapezoidal distribution*, and (3) for PS and SJF is of the form $\tilde{U}^*\Gamma$ where $\tilde{U}^*$ has a *generalized trapezoidal distribution*. These results are rather intriguing and lead to new fundamental insight in the impact of the local scheduling policy on the performance of polling models. As a by-product the heavy-traffic results suggest simple closed-form approximations for the complete waiting-time and sojourn-time distributions for stable systems with arbitrary load values. The accuracy of the approximations is evaluated by simulations. This chapter is based on [19].

In Chapter 3, we study the HT asymptotics of the waiting time distribution in cyclic polling models with *exhaustive* service at each queue under a variety of local scheduling policies, including FCFS, LCFS, ROS, the multi-class priority scheduling with and without preemption, SJF and SRPT. For each of these policies, we first express the waiting-time distributions in terms of intervisit-time distributions. Next, we use these expressions to derive the asymptotic waiting-time distributions under heavy-traffic assumptions. The results show that in all cases the asymptotic waiting-time distribution at queue i is fully characterized and of the form $\Theta_i\Gamma$, with Θ_i and Γ

independent, and where Γ is gamma distributed with known parameters (and the same for all scheduling policies). We derive the distribution of the random variable Θ_i which explicitly expresses the impact of the local scheduling policy on the asymptotic waiting-time distribution. Note that in the gated case of Chapter 2 we used the notation U for Θ_i, because the distribution was uniform, or closely related to uniform. When the service policy is exhaustive, this is not the case. With simulations we evaluate the closed-form approximations for the waiting-time distributions in stable systems, suggested by the asymptotic results. This chapter is based on [148].

In Chapter 4, we consider cyclic polling models with gated or globally gated service, and study the *transient* behavior of all cycle lengths. Our aim is to analyze the dependency structure between the different cycles, as this is an intrinsic property making polling models challenging to analyze. Transient performance is of great interest in systems where disruptions or breakdowns may occur, leading to excessive cycle lengths. The time to recover from such events is a primary performance measure. For the analysis we assume that the distribution of the first cycle (globally gated) or N residence times (gated), where N is the number of queues, is known and that the arrivals are Poisson. The joint LST of all x subsequent cycles (globally gated) or all $x > N$ subsequent residence times (gated) is expressed in terms of the LST of the first cycle. From this joint LST, we derive first and second moments and correlation coefficients between different cycles. Finally, a heavy-tailed first cycle length or the heavy-traffic regime provides additional insights into the time-dependent behavior. This chapter is based on [149].

In Chapter 5, we study the performance of Discriminatory Processor Sharing (DPS) systems, with exponential service times and in which batches of customers of different types may arrive simultaneously according to a Poisson process. In a general parameter setting, we show the occurrence of a state-space collapse in HT: as the load ρ goes to 1, the scaled joint queue-length vector $(1-\rho)\mathbf{Q}$ converges in distribution to the product of a known vector and an exponentially distributed random variable with known mean. The results provide new insight in the behavior of DPS systems. They show explicitly how the queue-length distribution depends on the system parameters, and in particular, on the simultaneity of the arrivals. The results also suggest simple and fast approximations for the tail probabilities and the moments of the queue lengths in stable DPS systems, explicitly capturing the impact of the correlation structure in the arrival processes. Numerical experiments indicate that the approximations are accurate for medium and heavily loaded systems. This chapter is based on [150].

In Chapter 6, we consider a level-dependent two-class queue motivated by a health care application. The two job types represent patients and administration tasks. We want to model a phenomenon frequently encountered in practice. In ambulatory care we typically see long access times for making an appointment. Moreover, these access times are often remarkably stable through time. We propose two queueing models that may show such behavior. Specifically, to meet target access times, we allow for

overbookings or flexible capacity by replacing administration tasks by patient care. We argue that access times for appointment systems with overbooking can naturally be modeled as level-dependent M/G/1 or MAP/G/1 queues, depending on the variability in the available capacity. Using transforms, we obtain intuitively appealing results for the distribution of the access time. Based on numerical experiments, we see that appointment systems may efficiently operate at high load, provided that some extra flexible capacity is available. This chapter is based on [147].

Finally, in Chapter 7, we consider a blended multi-server queue with impatient customers that is commonly encountered in call centers. The system receives two types of customers: urgent and best effort. Urgent customers are delay sensitive and we are interested in the service level in the form of the tail distribution of the waiting time. For best effort, we only consider the long-run throughput. Although, such a system is typically called a call blending model, we see applications in many other service settings. We derive the probability to abandon and the full waiting time distribution by considering the waiting time process of the first customer in line, combined with elements of the system point method. When the urgent and best effort classes have the same service rate, the waiting time distribution has a similar structure as in the M/M/s+G queue. For different service rates, the tail of the waiting time distribution can be iteratively solved and satisfies linear second-order differential equations. When customers have infinite patience, the waiting-time distribution can be written as a mixture of exponentials. This chapter is based on [20].

Chapter 2

Gated and globally gated polling systems with non-FCFS service

2.1 Introduction

A polling model is a multi-class single-server queueing model, as described in the previous chapter. In the design of a polling system there are a number of design decisions that have to be made, i.e., cf. Section 1.2.1,

1. The order in which to serve the queues (server routing).
2. How many customers to serve during each visit to a queue (service discipline).
3. The order in which customers within each queue are served (local scheduling policy).

For the first two decisions a wide variety of policies has been proposed and analyzed in the literature. The focus of the current chapter is on the third decision. More specifically, we investigate the influence of the effective local service order on the waiting times of the customers. As a result, we will limit discussion to the most common configurations for the first two decisions: cyclic service order and (globally-)gated service.

It might be natural to assume that the impact of such local scheduling is small, because it only impacts the system performance locally, leaving the amount of time spent outside the targeted queue unaffected. However, the results in [153] illustrate that the impact on the system performance from scheduling within a queue of a polling system can be significant. In many application areas of polling models, such as Bluetooth and 802.11 protocols, scheduling policies at routers and I/O subsystems in web servers, the workloads are known to have high variability and priority-based scheduling could therefore be beneficial. Outside of computer-communication systems, local scheduling proved its worth in the domain of production-inventory control. Our goal

is to explore the impact of the local scheduling in polling systems under heavy traffic (HT) conditions.

The motivation for studying the HT regime is twofold. First of all, it is the most important and challenging regime from a practical scheduling point of view, i.e. the proper operation of the system is particularly critical when the system is heavily loaded. Optimizing the local scheduling is, therefore, an effective mechanism for improving system performance without purchasing additional resources. Second, an attractive feature of HT asymptotics is that in many cases they lead to strikingly simple expressions for the performance measures of interest. This remarkable simplicity of the HT asymptotics leads to structural insights into the dependence of the performance measures on the system parameters and gives fundamental understanding of the behavior of the system in general. As a result, HT asymptotics form an excellent basis for developing simple accurate approximations for the performance measures (distributions, moments, tail probabilities) for stable systems. These closed-form approximations allow for back-of-the-envelope calculations.

Although an enormous number of papers on both polling systems and scheduling policies have appeared, the combination of the two has received very little attention. That is, almost all theoretical studies of scheduling policies are performed in single-queue settings such as the M/G/1 and G/G/1 queue with only a few exceptions studying the effect of local scheduling in multi-queue polling systems. By using the *Mean Value Analysis* (MVA) framework for polling systems [155], Wierman et al. [153] have derived the mean delay in cyclic exhaustive and gated polling systems for various scheduling disciplines such as First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS), Foreground-Background (FB), Processor Sharing (PS), Shortest-Job-First (SJF) and fixed priorities. Building upon these results, Boxma et al. [40] have obtained the waiting-time distribution in cyclic (globally-)gated polling systems for various service orders. As indicated by [40], the derivation of the waiting-time distribution in *exhaustive* polling systems is much more intricate. Waiting-time distributions in HT in exhaustive polling systems are the topic of Chapter 3. Boon et al. [30] study the waiting-time distribution in a two-queue polling model with either the exhaustive, gated or globally-gated service discipline. The first of these two queues contains customers of two priority classes. In [29] these results are generalized to a polling model with N queues and K_i priority levels in queue i. Moreover, Ayesta et al. [12] derive the sojourn-time distribution in polling systems with exhaustive service and where the local scheduling policy is PS. For a general service requirement distribution the analysis is restricted to the mean sojourn time.

In the current chapter we study Poisson-driven cyclic polling models with general service-time and switch-over time distributions, and with two types of service policies: (1) models with gated service at each queue, and (2) models with globally-gated service. For both types of service policies, we consider the following five scheduling policies that determine the local order in which the customers at a given queue are served: FCFS, LCFS, ROS, PS and SJF.

For each of these models we derive exact closed-form expressions for the LST of the (scaled) waiting-time and sojourn-time distributions under HT assumptions. Note that it was shown in [132] that the asymptotic cycle-time distributions converge to a gamma distribution with known parameters. Using this result, for FCFS service it is shown in [113] that the asymptotic sojourn-time distribution is a product of the random variables U and Γ, where U and Γ are uniformly and gamma distributed. In this chapter, we unify and extend this result by presenting rigorous proofs showing that the asymptotic sojourn-time distribution is (1) for LCFS also of the form $U\Gamma$, (2) for ROS of the form $\tilde{U}\Gamma$ where $\tilde{U}$ has a *trapezoidal distribution*, and (3) for PS and SJF of the form $\tilde{U}^*\Gamma$ where $\tilde{U}^*$ has a *generalized trapezoidal distribution*. We would like to stress the unearthed dichotomy between the known HT results on FCFS polling models and our novel asymptotic results for other scheduling disciplines.

These results are rather intriguing and provide new fundamental insight in the impact of the local scheduling policy on the performance of polling models. Our results lead not only to unification but also to extension of the literature studying scheduling policies, polling systems and HT asymptotics. As a by-product the HT results suggest simple closed-form approximations for the complete waiting-time and sojourn-time distributions for stable systems with arbitrary load values and *general renewal arrival processes*. Numerical results show that these approximations perform well for a wide range of parameter combinations.

The remainder of the chapter is organized as follows. In Section 2.2, the model is described and the notation required is introduced. In Section 2.3, we derive the HT asymptotics for the model with gated service at each queue under various local scheduling policies. Section 2.4 presents similar results for the case of globally gated service. Furthermore, Section 2.5 proposes a simple approximation for the sojourn-time distributions for arbitrary load values and present numerical results to evaluate the accuracy of the approximations. Section 2.6 contains a number of concluding remarks.

2.2 Model description

We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. Customers arrive at Q_i according to a Poisson process with rate λ_i. These customers are referred to as type-i customers. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with LST $B_i^*(\cdot)$ and finite kth moment $b_i^{(k)} = \mathbb{E}[B_i^k]$, $k = 1, 2$. The kth moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \mathbb{E}[B^k] = \sum_{i=1}^{N} \lambda_i \, \mathbb{E}[B_i^k]/\Lambda$, $k = 1, 2$. The load offered to Q_i is $\rho_i = \lambda_i \, \mathbb{E}[B_i]$ and the total load offered to the system is equal to $\rho = \sum_{i=i}^{N} \rho_i$. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is an independent random variable S_i with mean $r_i := \mathbb{E}[S_i]$. Let $S = \sum_{i=1}^{N} S_i$ denote the total switch-

over time in a cycle and let $r := \mathbb{E}[S]$ denote its mean. A necessary and sufficient condition for stability of the system is $\rho < 1$. C_i denotes the cycle time at queue i, defined as the time between two successive arrivals of the server at queue i; it is well known that $\mathbb{E}[C_i] = r/(1-\rho)$ for each i (cf. [125, Equation (5.39b)]).

The *service policy* determines *which* customers are served during a visit of the server to a queue. In this chapter we assume two variants: (1) the model with gated service at each of the queues, and (2) the globally-gated model. For gated service, all customers are served that were present at polling instant, i.e., at the moment when the server arrives at the queue. For globally gated, during a cycle, all customers are served that were present at polling instant of the first queue. The *local scheduling policy* determines the *order* in which the customers are served during a visit period at a queue. We consider the following five local scheduling policies: FCFS, LCFS, ROS, PS or SJF. For policy $P \in \{\text{FCFS}, \text{LCFS}, \text{ROS}, \text{PS}, \text{SJF}\}$, we denote $i \in I_P$ if Q_i receives scheduling policy P. For example, I_{FCFS} is the (index) set of queues i that are served on a FCFS basis, I_{LCFS} is the (index) set of queues i that are served on a LCFS basis, and so on.

In this chapter we study heavy-traffic limits, i.e., the limiting behavior as ρ approaches 1. The heavy-traffic limits, denoted $\rho \uparrow 1$, taken in this chapter are such that the arrival rates are increased, while keeping both the service-time distributions and the ratios between the arrival rates fixed. Light-traffic limits, denoted $\rho \downarrow 0$, are defined similarly. The notation $\to_d$ means convergence in distribution. For each variable x that is a function of ρ, we denote its value *evaluated at* $\rho = 1$ by $\hat{x}$. In particular we have $\hat{\rho}_i = \frac{\rho_i}{\rho}$ and $\hat{\lambda}_i = \frac{\hat{\rho}_i}{\mathbb{E}[B_i]}$.

Let W_i denote the waiting time of an arbitrary customer at Q_i, defined as the time between the arrival of a customer and the moment at which he enters service. The sojourn time of an arbitrary customer at Q_i, represented by T_i, is defined as the time between the arrival of a customer and the moment at which he departs from the system. The LSTs of W_i and T_i are denoted by $W_i^*(s)$ and $T_i^*(s)$, respectively. When $\rho \uparrow 1$, all queues become unstable, therefore the focus lies on the random variables $(1-\rho)W_i$ and $(1-\rho)T_i$ as $\rho \uparrow 1$, referred to as the *scaled* waiting times and sojourn times at Q_i, respectively. A summary of the notation with respect to a random variable X is given in Table 2.1.

A key role is played by the gamma distribution and the uniform distribution. A non-negative continuous random variable $\Gamma(\alpha, \mu)$ is said to have a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\mu > 0$ if it has the probability density function

$$f_\Gamma(x) = \frac{\mu^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\mu x} \quad (x > 0), \quad \text{with} \quad \Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t} \mathrm{d}t \tag{2.1}$$

and LST

$$\Gamma^*(s) = \left(\frac{\mu}{\mu + s}\right)^\alpha \quad (Re(s) > 0). \tag{2.2}$$

$f_X(\cdot)$	Probability density function (pdf) of X
$F_X(\cdot)$	Cumulative distribution function (cdf) of X
$X^*(\cdot)$	Laplace-Stieltjes transform (LST) of X, i.e., $X^*(s) = \mathbb{E}[e^{-sX}]$
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}[X^k]$	kth moment of X
c_X^2	Squared coefficient of variation (SCV) of X
X^{res}	Residual length of X
	with $\mathbb{E}[X^{res}] = \frac{\mathbb{E}[X^2]}{2\,\mathbb{E}[X]}$ and LST $\mathbb{E}[e^{-sX^{res}}] = \frac{1-\mathbb{E}[e^{-sX}]}{s\,\mathbb{E}[X]}$
$\mathbf{X}$	Length-biased version of X
	with $f_{\mathbf{X}}(x) = \frac{x f_X(x)}{\mathbb{E}[X]}$

Table 2.1: Notation with respect to a random variable X.

Note that in the definition of the gamma distribution μ is a scaling parameter, and that $\Gamma(\alpha, \mu)$ has the same distribution as $\mu^{-1}\Gamma(\alpha, 1)$. Moreover, we denote by $U[a, b]$ ($a < b$) a random variable that is uniformly distributed over the interval $[a, b]$. For later reference note, that the LST of the random variable $\Gamma(\alpha{+}1, \mu)U[a, b]$, where $\Gamma(\alpha{+}1, \mu)$ and $U[a, b]$ are independent, is given by, for $Re(s) > 0$,

$$\mathbb{E}\left[e^{-sU[a,b]\Gamma(\alpha+1,\mu)}\right] = \frac{\mu}{\alpha s(b-a)}\left\{\left(\frac{\mu}{\mu+sa}\right)^{\alpha} - \left(\frac{\mu}{\mu+sb}\right)^{\alpha}\right\}. \tag{2.3}$$

2.3 Analysis of models with gated service

In this section we consider the case of gated service at all queues. In Subsection 2.3.1 we review some known preliminary results for FCFS disciplines to be used for later reference. In Subsections 2.3.2–2.3.5 we use the results in Subsection 2.3.1 to derive HT limits for LCFS, ROS, PS and SJF, respectively. In Section 2.5 we use these results to propose and validate approximations for the distributions of the waiting times and sojourn times for arbitrary load values and renewal arrivals.

It is easy to see that for FCFS, LCFS and ROS service the sojourn time is simply the convolution of the waiting time and the service time, i.e., for $Re(s) \geq 0$,

$$T_i^*(s) = W_i^*(s)B_i^*(s) \quad (i \in I_{FCFS}, I_{LCFS}, I_{ROS}). \tag{2.4}$$

For this reason, in Subsections 2.3.1–2.3.3 we focus on the waiting-time distributions. The results for the sojourn-time distributions then follow directly from (2.4). Note that for $i \in I_{PS}$ and $i \in I_{SJF}$ relation (2.4) is generally not true, because in those cases the waiting times and the service times are not independent. Relation (2.4) is used for the approximation in Section 2.5, since sojourn times and waiting times are equal in HT.

To start, let us consider the distribution of the cycle time C_i, defined as the time between two successive arrivals of the server at queue i. A simple but important observation is that the distribution of C_i is independent of the local scheduling policy (i.e., FCFS, LCFS, ROS, PS and SJF). To this end, recall that the *service policy* (e.g., gated or globally gated) determines *which* customers are served during a visit V of the server to a queue, and that the local *scheduling policies* determine the *order* in which these customers are served during V. For this reason the cycle-time distributions are the same for all local scheduling policies under consideration, provided that they are work-conserving.

The following result gives a characterization for the limiting behavior of the cycle-time distributions, stating that the (scaled) cycle times $(1-\rho)C_i$ in HT converge to a gamma distribution with known parameters (proven in [132; 136]).

Property 2.1 (Convergence of the cycle times)**.** *For the model with gated service at each queue we have, for $i = 1, \ldots, N$, as $\rho \uparrow 1$,*

$$(1-\rho)C_i \to_d \tilde{\Gamma},$$

where $\tilde{\Gamma}$ has a gamma distribution with parameters

$$\alpha := \frac{r\delta}{\sigma^2}, \quad \mu := \frac{\delta}{\sigma^2}, \tag{2.5}$$

with

$$\sigma^2 := \frac{b^{(2)}}{b^{(1)}}, \quad \textit{and} \quad \delta := \sum_{i=1}^{N} \hat{\rho}_i(1+\hat{\rho}_i). \tag{2.6}$$

2.3.1 First-Come-First-Served

In this section we review several known results for the case of FCFS service at queue i. In Subsections 2.3.2–2.3.5 these results are used to derive new results for LCFS, ROS, PS and SJF, respectively. For FCFS service, the following result gives an expression for the LST of the waiting time W_i in terms of the distribution of the cycle time C_i (proven in [40]):

Property 2.2 (Cycle-time expression for the waiting times)**.** *For the gated service model, we have for $Re(s) > 0$ and $\rho < 1$,*

$$W_i^*(s) = \frac{C_i^*(\lambda_i(1-B_i^*(s))) - C_i^*(s)}{\mathbb{E}\left[C_i\right](s - \lambda_i(1-B_i^*(s)))} \qquad (i \in I_{FCFS}). \tag{2.7}$$

The following result, which was shown in [132], characterizes the limiting behavior of the waiting-time distribution in heavy traffic.

Property 2.3 (Convergence of the waiting times)**.** *For the gated service model, we have for* $\rho \uparrow 1$*,*

$$(1-\rho)W_i \to_d U_i \tilde{\mathbf{C}}_i \quad (i \in I_{FCFS}), \tag{2.8}$$

where U_i *is uniformly distributed on the interval* $[\hat{\rho}_i, 1]$*, and where* $\tilde{\mathbf{C}}_i$ *has a gamma distribution with parameters* $\alpha+1$ *and* μ*, where* α *and* μ *are given in Equation* (2.5)*. The random variables* U_i *and* $\tilde{\mathbf{C}}_i$ *are independent.*

Note that here $\tilde{\mathbf{C}}_i$ is the *length-biased* version of $\tilde{C}_i$, a gamma-distributed random variable with parameters α and μ as in Equation (2.5). It is well known that if a gamma random variable has parameters α and μ then its length-biased version has parameters $\alpha+1$ and μ. The following result gives an expression for the higher moments of the waiting times in heavy traffic (proven in [133; 134]):

Property 2.4 (Convergence of moments of the waiting time)**.** *For* $k = 1, 2, \ldots$*,*

$$\omega_i^{(k)} := \lim_{\rho\uparrow 1} (1-\rho)^k \, \mathbb{E}\left[W_i^k\right] = \frac{1-\hat{\rho}_i^{k+1}}{1-\hat{\rho}_i} \frac{\prod_{j=1}^{k}(\alpha+j)}{(k+1)\mu^k} \quad (i \in I_{FCFS}), \tag{2.9}$$

assuming that the $(k+1)$*-st moments of the service-time distributions and the* k*th moments of the switch-over time distributions are finite.*

2.3.2 Last-Come-First-Served

The LST for the waiting-time distribution for the LCFS service is expressed in terms of the cycle-time distributions as follows (cf. [40]): For $Re(s) > 0$ and $\rho < 1$,

$$W_i^*(s) = \frac{1 - C_i^*(s + \lambda_i(1 - B_i^*(s)))}{\mathbb{E}\left[C_i\right](s + \lambda_i(1 - B_i^*(s)))} \quad (i \in I_{LCFS}). \tag{2.10}$$

The following result gives an expression for the asymptotic waiting-time distribution for LCFS service in heavy traffic.

Theorem 2.1. *For* $\rho \uparrow 1$*,*

$$(1-\rho)W_i \to_d U_i \tilde{\mathbf{C}}_i \quad (i \in I_{LCFS}), \tag{2.11}$$

where U_i *is uniformly distributed on the interval* $[0, 1+\hat{\rho}_i]$ *and* $\tilde{\mathbf{C}}_i$ *has a gamma distribution with parameters* $\alpha+1$ *and* μ*, where* α *and* μ *are given in Equation* (2.5)*. The random variables* U_i *and* $\tilde{\mathbf{C}}_i$ *are independent.*

Proof. Take $i \in I_{LCFS}$. Then combining (2.10) with Property 2.1 gives the following expressions for the LST of the (scaled) waiting-time distribution. For $i \in I_{LCFS}$,

$Re(s) > 0$, we have

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \lim_{\rho\uparrow 1} \frac{1 - C_i^*(s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho))))}{\mathbb{E}\left[C\right](s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho))))} \tag{2.12}$$

$$= \lim_{\rho\uparrow 1} \frac{1 - \left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s(1-\rho)+\lambda_i(1-B_i^*(s(1-\rho)))}\right)^{\alpha}}{\frac{r}{(1-\rho)}\left(s(1-\rho) + \lambda_i(1 - B_i^*(s(1-\rho)))\right)} \tag{2.13}$$

$$= \lim_{\rho\uparrow 1} \frac{1 - \left(\frac{\mu}{\mu+s+\lambda_i(1-B_i^*(s(1-\rho)))/(1-\rho)}\right)^{\alpha}}{r\left(s + \frac{\lambda_i(1-B_i^*(s(1-\rho)))}{1-\rho}\right)}. \tag{2.14}$$

Using l'Hôpital's rule and the fact that $-B_i^{*'}(0) = \mathbb{E}[B_i]$ we see that:

$$\lim_{\rho\uparrow 1} \frac{\lambda_i(1 - B_i^*(s(1-\rho)))}{1-\rho} = \lim_{\rho\uparrow 1} \frac{0 - \lambda_i B_i^{*'}(s(1-\rho))s}{1} = \hat{\rho}_i s,$$

which immediately implies that, for $Re(s) > 0$ and $i \in I_{LCFS}$,

$$\tilde{W}_i^*(s) = \frac{1 - \left(\frac{\mu}{\mu+s+\hat{\rho}_i s}\right)^{\alpha}}{r(s + \hat{\rho}_i s)} = \frac{1}{rs(1+\hat{\rho}_i)}\left\{1 - \left(\frac{\mu}{\mu + s(1+\hat{\rho}_i)}\right)^{\alpha}\right\}, \tag{2.15}$$

where α and μ are given in (2.5). Using (2.3) and $\mu/\alpha = 1/r$, it now follows that (2.15) corresponds to the LST of a uniform random variable on $[0, 1+\hat{\rho}_i]$ times a gamma distribution. Application of Levy's Continuity Theorem [154] completes the proof. □

Using Theorem 2.1, it is easily verified that the moments of the asymptotic delay distribution are given by the following expression.

Corollary 2.1 (Moments of the asymptotic delay)**.** *For $k = 1, 2, \ldots$,*

$$\omega_i^{(k)} := \lim_{\rho\uparrow 1} (1-\rho)^k \, \mathbb{E}[W_i^k] = \frac{(1+\hat{\rho}_i)^k \prod_{j=1}^{k}(\alpha + j)}{(k+1)\mu^k} \quad (i \in I_{LCFS}), \tag{2.16}$$

where α and μ are defined in Equation (2.5)*, assuming that the $(k+1)$-st moments of the service-time distributions and the kth moments of the switch-over time distributions are finite.*

We end this section with a number of remarks.

Remark 2.1 (Comparison between FCFS and LCFS case using the heavy-traffic averaging principle)**.** Property 2.3 and Theorem 2.1 reveal an interesting difference in the waiting-time distributions of the FCFS case and the LCFS case. More precisely, for the FCFS case the limiting behavior of W_i is of the form $U_{FCFS}\Gamma$, where U_{FCFS} is uniformly distributed on the interval $[\hat{\rho}_i, 1]$, whereas for the LCFS case the limiting

distribution of W_i is of the form $U_{LCFS}\Gamma$, where U_{LCFS} is uniformly distributed on the interval $[0, 1+\hat{\rho}_i]$, which is multiplied by the *same* gamma distribution. To provide an intuitive explanation for this, we use the insights that can be obtained by the so-called Heavy-Traffic Averaging Principle (HTAP), see e.g. [51; 52] and [113; 114]. Loosely speaking, HTAP for polling models means that the total scaled workload may be considered as a constant during a cycle, whereas the workloads of the individual queues change much faster according to deterministic trajectories, or a fluid model. Due to the HTAP, we let the constant c denote the cycle length. Let us first consider the fluid model for FCFS. Note that the waiting time consists of two parts. First, a customer has to wait for the residual cycle time, which is $(1-U)c$ for U uniformly distributed on $[0, 1]$. Second, a customer has to wait for all customers that have arrived before him during the course of the ongoing cycle. Hence, this equals $\hat{\rho}_i U c$. The total waiting time in the fluid model then equals $(1-U+\hat{\rho}_i U)c$, which has a uniform distribution on $[\hat{\rho}_i c, c]$. Using that the cycle time follows a gamma distribution explains the shape of the waiting-time distribution in heavy traffic. For LCFS, as for FCFS, an arriving customer still has to wait for the residual cycle length $(1-U)c$, with U a uniform random variable on $[0, 1]$. In addition, the arriving customer has to wait for all customers that arrived after him during the same cycle, which is of length $\hat{\rho}_i(1-U)c$. Hence, the waiting time in the fluid model is $(1+\hat{\rho}_i)(1-U)c$, which is a uniform distribution on $[0, (1+\hat{\rho}_i)c]$. This interpretation gives much insight in the heavy-traffic asymptotics.

Remark 2.2 (Alignment with asymptotics with large switch-over times)**.** Further support can be given for the distribution in Theorem 2.1 by considering a different asymptotic regime as in [131]. Let the switch-over times be deterministic with length r_i. We consider the behavior of W_i when the switch-over times tend to infinity. Because the waiting times are known to grow without bound when the switch-over times increase to infinity, the analysis is oriented towards the limiting distribution of $\frac{W_i}{r}$ as $r \to \infty$. Using similar techniques as in [131], it may be shown that

$$\frac{W_i}{r} \to_d \hat{W}_i \quad (r \to \infty), \tag{2.17}$$

where $\hat{W}_i$ is uniformly distributed over the interval $[\tilde{a}_i, \tilde{b}_i]$, with, for $i \in I_{FCFS}$,

$$\tilde{a}_i = \frac{\rho_i}{1-\rho}, \quad \tilde{b}_i = \frac{1}{1-\rho}, \quad \text{and} \quad \tilde{a}_i = 0, \quad \tilde{b}_i = \frac{\rho_i + 1}{1-\rho} \text{ for } i \in I_{LCFS}. \tag{2.18}$$

Note that the uniform distribution is the same as in the HT regime. However, in HT the cycle times follow a gamma distribution whereas here the cycle times become deterministic as the switch-over times grow large.

2.3.3 Random Order of Service

In this section we derive heavy-traffic limits for the Random Order of Service (ROS) local scheduling policy. ROS is represented by ordering marks. Each customer that

arrives gets an ordering mark x, a realization from a uniform distribution on $[0,1]$. When the server arrives at the queue, the gate closes and the customers before the gate are served in order of their marks. It is convenient to condition with respect to x and then uncondition. Let $W_i(x)$ be the waiting time of a customer in queue i with ordering mark x, with $i \in I_{ROS}$, and let $W_i^*(s|x)$ be the corresponding LST. The following result was shown in [40]: for $Re(s) > 0$, $0 < x < 1$ and $\rho < 1$,

$$W_i^*(s|x) = \frac{C_i^*(\lambda_i x(1 - B_i^*(s))) - C_i^*(s + \lambda_i x(1 - B_i^*(s)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{ROS}). \tag{2.19}$$

The next result gives the heavy-traffic limit of the distribution of $W_i(x)$.

Theorem 2.2 (Conditional waiting time)**.** *For $\rho \uparrow 1$, $0 < x < 1$,*

$$(1-\rho)W_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{ROS}), \tag{2.20}$$

where $U_i(x)$ is uniformly distributed over the interval $[\hat{\rho}_i x, 1 + \hat{\rho}_i x]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (2.5). The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.

Proof. Combining (2.19) and Property 2.1 and using l'Hôpital's rule, we find the following LST of the waiting time conditional on the ordering mark x: for $Re(s) > 0$, $0 < x < 1$,

$$\begin{aligned}
\tilde{W}_i^*(s|x) &= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)|x) \\
&= \lim_{\rho\uparrow 1} \frac{C_i^*(\lambda_i x(1 - B_i^*(s(1-\rho)))) - C_i^*(s(1-\rho) + \lambda_i x(1 - B_i^*(s(1-\rho))))}{s(1-\rho)\,\mathbb{E}[C_i]} \\
&= \frac{1}{rs}\left\{\left(\frac{\mu}{\mu + \hat{\rho}_i x s}\right)^{\alpha} - \left(\frac{\mu}{\mu + (1+\hat{\rho}_i x)s}\right)^{\alpha}\right\} \quad (i \in I_{ROS}).
\end{aligned} \tag{2.21}$$

Applying Levy's Continuity Theorem completes the proof. □

To obtain the *unconditional* distribution of the waiting time, we first consider a more general setting that also covers 'unconditioning' for PS and SJF. For this, let $a(\cdot)$ be a continuous and strictly increasing function on some interval $\mathcal{X} = [x_{min}, x_{max}]$, where we allow x_{max} to be infinite. Suppose we have a conditional random variable, denoted $T|x$, that is uniformly distributed on the interval $[a(x), a(x)+1]$. We want to find the unconditional distribution $\tilde{T}$. Here, x is a realization of the random variable X with support $\mathcal{X} \subseteq \mathbb{R}^+$ having distribution function $F_X(\cdot)$ and density $f_X(\cdot)$. We have the following lemma.

Lemma 2.1. *Assume that the conditional random variable $T|x$ is uniformly distributed on $[a(x), a(x)+1]$, where $x \in \mathcal{X} = [x_{min}, x_{max}]$. Suppose that $a(x_{min}) = m$, $a(x_{max}) = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing in x, such that $a(\cdot)$ has*

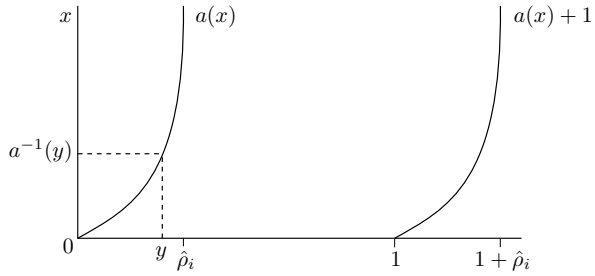


Figure 2.1: Boundaries of the uniform distribution.

an inverse denoted by $a^{-1}(\cdot)$*. Then, the unconditional distribution of* $T|x$*, denoted by* $\tilde{T}$*, has probability density function*

$$f_{\tilde{T}}(y) = \begin{cases} F_X(a^{-1}(y)) & y \in [m, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1+m] \\ 1 - F_X(a^{-1}(y-1)) & y \in (1+m, 1+\hat{\rho}_i]. \end{cases} \tag{2.22}$$

Proof. Note that $T|x$ has the following probability density function

$$f_{T|x}(y) = \begin{cases} 1 & y \in (a(x), a(x)+1) \\ 0 & \text{Otherwise} \end{cases} \qquad \forall\, x \in \mathcal{X}.$$

Figure 2.1 shows an example of the boundaries of the uniform distribution, by plotting $a(x)$ and $a(x)+1$ with x on the vertical axis. The possible values of $T|x$ then lie between the two lines. To find $f_{\tilde{T}}(y)$, we need to integrate out x with respect to its density function. First, take $y \in (m, \hat{\rho}_i)$, in which case the probability density $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$. This gives, for $y \in (m, \hat{\rho}_i)$,

$$f_{\tilde{T}}(y) = \int_{x_{min}}^{a^{-1}(y)} f_X(x) f_{T|x}(y)\, \mathrm{d}x = F_X(a^{-1}(y)).$$

If $y \in (\hat{\rho}_i, 1+m)$ then y is between the boundaries of the uniform distribution for every $x \in \mathcal{X}$. Hence, we get

$$f_{\tilde{T}}(y) = \int_{x_{min}}^{x_{max}} f_X(x) f_{T|x}(y)\, \mathrm{d}x = 1.$$

Finally, for $y \in (1+m, 1+\hat{\rho}_i)$, we can use that the boundaries are described by similar curves, i.e., x needs to be larger than $a^{-1}(y-1)$, so

$$f_{\tilde{T}}(y) = \int_{a^{-1}(y-1)}^{x_{max}} f_X(x) f_{T|x}(y)\, \mathrm{d}x = 1 - F_X(a^{-1}(y-1)).$$

Finally, it follows from the properties of $a(\cdot)$ that $f_{\tilde{T}}(\cdot)$ is a density function. This completes the proof. □

Remark 2.3. For convenience it is assumed in Lemma 2.1 that the underlying random variable X has a density. For, e.g., PS it can be of interest to consider the case that X is a discrete random variable. This is directly related to the properties of $a(\cdot)$, i.e., that $a(\cdot)$ is continuous and strictly increasing. It is not difficult to modify Lemma 2.1 to the case of discrete random variables by either redefining the inverse of $a(\cdot)$ as $a^{-1}(y) = \sup\{x \in \mathcal{X} : a(x) \leq y\}$, or by extending the function $a(\cdot)$ from the range of X to an interval $[x_{min}, x_{max}]$, such that $a(\cdot)$ is continuous and strictly increasing.

Note that the density function in (2.22) is continuous, increasing on $[m, \hat{\rho}_i)$, constant on $[\hat{\rho}_i, 1+m]$ and decreasing on $(1+m, 1+\hat{\rho}_i]$, which closely resembles the traditional trapezoidal distribution. In line with [61], we refer to (2.22) as a *generalized trapezoidal distribution* consisting of stages of growth, stability, and decay, i.e., the function is increasing, constant, and decreasing, respectively.

For further references, it is of interest to determine the mean of this generalized trapezoidal distribution. There are different ways to represent this mean, for instance,

$$\mathbb{E}[\tilde{T}] = \int_m^{1+\hat{\rho}_i} x f_{\tilde{T}}(x)\,\mathrm{d}x = \frac{1}{2} + \hat{\rho}_i - \int_m^{\hat{\rho}_i} F_X(a^{-1}(y))\,\mathrm{d}y = \frac{1}{2} + \int_{u\in\mathcal{X}} a(u) f_X(u)\,\mathrm{d}u,$$

where the second step follows after some rewriting. Substituting $y = a(u)$, the third step follows after partial integration. In Subsections 2.3.4 and 2.3.5, the mean of the generalized distribution $\mathbb{E}[\tilde{T}]$ is specified for PS and SJF.

We now apply the lemma above to the case $i \in I_{ROS}$, in which case $a(x) = \hat{\rho}_i x$, with $x \in [0, 1]$. The asymptotic scaled unconditional delay is presented in the following theorem.

Theorem 2.3 (Unconditional waiting time)**.** *For $\rho \uparrow 1$,*

$$(1-\rho)W_i \to_d \tilde{U}_i^* \tilde{\mathbf{C}}_i \quad (i \in I_{ROS}),$$

where $\tilde{U}_i^*$ has a trapezoidal distribution with density function

$$f_{\tilde{U}_i^*}(y) := \begin{cases} y/\hat{\rho}_i & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ (1+\hat{\rho}_i - y)/\hat{\rho}_i & y \in (1, 1+\hat{\rho}_i]. \end{cases} \tag{2.23}$$

and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ, where α and μ are given in Equation (2.5)*. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Take $i \in I_{ROS}$. Then Theorem 2.2 implies that $a(x) = \hat{\rho}_i x$ and $\mathcal{X} = [0, 1]$. Note that this function has the desired properties: $a(0) = 0$, $a(1) = \hat{\rho}_i$ and $a(x)$ continuous and strictly increasing in $x \in \mathcal{X}$. The cumulative distribution function of X is given by $F_X(x) = x$ and the inverse function of $a(\cdot)$ is $a^{-1}(y) = y/\hat{\rho}_i$. The use of Lemma 2.1 now yields the result. □

Remark 2.4 (HTAP)**.** Interestingly, Theorem 2.3 shows that the uniform distribution that appears in the heavy-traffic limit for FCFS and LCFS is replaced by a trapezoidal distribution for ROS. The shape of this distribution can be explained by the fact that the waiting time of a customer does not only depend on the time that the customer enters the system, but also on an independent random mechanism that determines the moment that the customer is served. More specifically, exploiting the HTAP we let the constant c denote the cycle length again and consider the fluid model for the conditional waiting time of a customer with mark x. An arriving customer has to wait for the residual cycle length $(1-U)c$, with U uniformly distributed on $[0,1]$, and the time required to serve the customers that arrived during the same cycle and have a mark smaller than x, i.e., $\hat{\rho}_i xc$. Clearly, the conditional waiting time in the fluid model is uniformly distributed on $[\hat{\rho}_i xc, (1+\hat{\rho}_i)xc]$. Since x is an arbitrary order mark, the unconditional waiting time in the fluid model is $(U_1+U_2)c$, with U_1 and U_2 independent uniform distribution on the intervals $[0,1]$ and $[0,\hat{\rho}_i]$, respectively. Note that such a convolution gives rise to a trapezoidal distribution as obtained in Theorem 2.3.

Remark 2.5 (First moments of waiting times)**.** Observe that it follows from Property 2.4, Corollary 2.1 and Theorem 2.3 that the first moments of the asymptotic waiting-time distributions for the FCFS, LCFS and the ROS scheduling disciplines are the same. This is in line with the observation in [40] that the mean waiting times for these disciplines coincide for a general value of $\rho < 1$. To this end, it is easy to see that $\mathbb{E}[W_i] = (1+\rho_i)\frac{\mathbb{E}[C_i^2]}{2\,\mathbb{E}[C_i]}$ and that the cycle-time distributions are independent of the local scheduling policy.

2.3.4 Processor Sharing

When the scheduling discipline is PS, the LST of the *conditional* sojourn time (denoted $T_i^*(s|x)$) can also be expressed in terms of the LST of the cycle time. When x is the amount of work that a tagged customer brings into the system, it holds that (cf. [40]), for $\rho < 1$, $Re(s) > 0$, $x > 0$,

$$T_i^*(s|x) = e^{-sx}\frac{C_i^*(\lambda_i(1-\varphi(s,x))) - C_i^*(s+\lambda_i(1-\varphi(s,x)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{PS}), \tag{2.24}$$

where $\varphi(s,x) = \mathbb{E}\left[e^{-s\min(B_i,x)}\right]$, the LST of the minimum of B_i and x. The next theorem gives an expression for the asymptotic distribution of the conditional sojourn time $T_i(x)$.

Theorem 2.4 (Conditional sojourn time)**.** *For $\rho \uparrow 1$, $x > 0$,*

$$(1-\rho)T_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{PS}), \tag{2.25}$$

where $U_i(x)$ is uniformly distributed between $\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ and $1+\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (2.5)*. The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Combining (2.24) with Property 2.1, we obtain, for $Re(s) > 0$ and $x > 0$,

$$\begin{aligned}\tilde{T}_i(s|x) &:= \lim_{\rho\uparrow 1} T_i^*(s(1-\rho)|x) \qquad (2.26)\\ &= \frac{1}{rs}\left\{\left(\frac{\mu}{\mu+\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]s}\right)^{\alpha} - \left(\frac{\mu}{\mu+(1+\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)])s}\right)^{\alpha}\right\},\end{aligned}$$

with $\alpha+1$ and μ as given in (2.5). An application of Levy's Continuity Theorem yields the result. □

We now proceed with the unconditional sojourn time. For notational convenience, we assume here that the service-time distributions are absolutely continuous (see however Remark 2.3).

Theorem 2.5 (Unconditional sojourn time)**.** *For $\rho\uparrow 1$,*

$$(1-\rho)T_i \to_d U_i^*\tilde{\mathbf{C}}_i \quad (i\in I_{PS}),$$

where U_i^* is a type of generalized trapezoidal distribution as characterized in Equation (2.27) *with $a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ and x_{min} the lowest possible value of B_i. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Take $a(x) = \hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]$ such that $U_i(x)$ is uniformly distributed on $[a(x), a(x)+1]$, see Theorem 2.4. Clearly, $a(x_{min}) = \hat{\lambda}_i x_{min}$, $a(x_{max}) = \hat{\lambda}_i\,\mathbb{E}[B_i] = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing. Using Lemma 2.1, we obtain the unconditioned distribution

$$f_{U_i^*}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y\in[\hat{\lambda}_i x_{min},\hat{\rho}_i)\\ 1 & y\in[\hat{\rho}_i, 1+\hat{\lambda}_i x_{min}]\\ 1-F_{B_i}(a^{-1}(y-1)) & y\in(1+\hat{\lambda}_i x_{min}, 1+\hat{\rho}_i],\end{cases} \qquad (2.27)$$

where x_{min} is the minimum value that B_i can take. This completes the proof. □

Remark 2.6 (HTAP)**.** Theorem 2.5 shows that the conditional sojourn time in heavy traffic still is a uniform times a gamma distribution. This can again be intuitively explained from the HTAP. Now, in a cycle of length c, arriving customers have to wait for the residual cycle length $(1-U)c$ and their departure is delayed by all traffic in queue i that arrives during the same cycle and has been served before the tagged customer leaves. The latter equals $\hat{\lambda}_i\,\mathbb{E}[\min(B_i,x)]c$ in the fluid model. The distribution of the unconditional sojourn time not only depends on the first two moments of the service time, but depends on the complete service-time distribution. In particular, the curve $F_{B_i}(a^{-1}(y))$, with $y\in[\hat{\lambda}_i x_{min},\hat{\rho}_i)$, can be interpreted as the fluid model of the cumulative number of departures from queue i from the moment that the gate opens at queue i. To interpret this, note that in the fluid model $a(x)$ represents the amount of work served since the gate is open to a customer with service

requirement x, and $a^{-1}(\cdot)$ can thus be seen as the time to accumulate such an amount of service. Hence, $F_{B_i}(a^{-1}(y))$ counts the number of customers for which $a^{-1}(y)$ is sufficient to leave.

Remark 2.7 (Deterministic service times)**.** In most queueing models, high variability leads generally to longer waiting times. However, for the polling model under consideration, note that Theorem 2.4 implies that for deterministic service times, the waiting time in heavy traffic is also a uniform times a gamma distribution. Here, the boundaries of the uniform distribution are $\hat{\rho}_i$ and $1 + \hat{\rho}_i$. We note that this is the worst possible case for U_i^* among all service-time distributions, in the sense that it has the largest tail $\mathbb{P}(U_i^* > x)$ for all x. This is caused by the fact that all customers are served simultaneously and, in the end, they all jointly leave.

Below we give some examples of the type of generalized trapezoidal distribution U_i^* for some specific service-time distributions. Together with the gamma distribution, representing the cycle time, this fully specifies the scaled sojourn time in heavy traffic.

Exponential service times
Suppose B_i is exponentially distributed with parameter b_i. Then

$$\mathbb{E}[\min(B_i, x)] = \int_0^x y b_i e^{-b_i y} \, \mathrm{d}y + x e^{-b_i x} = \frac{1}{b_i}(1 - e^{-b_i x}),$$

so $a(x) = \hat{\rho}_i(1 - e^{-b_i x})$. Solve $a(x) = y$ for x to find $a^{-1}(y) = \ln(1 - y/\hat{\rho}_i)/(-b_i)$. Now substituting this in Equation (2.27) it follows after some simplification that the generalized trapezoidal distribution U_i^* coincides with the density function of U_i^* for ROS given in (2.23). This means that for the case of exponential service-time distributions, the sojourn-time distributions for ROS and PS coincide.

Uniform service times
Suppose B_i is a uniformly distributed random variable on the interval $[a_i, b_i]$. Then

$$\begin{aligned} a(x) &= \hat{\lambda}_i \, \mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i \left(\int_{a_i}^x y/(b_i - a_i) \, \mathrm{d}y + x \int_x^{b_i} 1/(b_i - a_i) \, \mathrm{d}y \right) \\ &= \frac{-\hat{\lambda}_i}{2(b_i - a_i)} \left(a_i^2 - 2b_i x + x^2 \right) = \frac{-\hat{\rho}_i}{b_i^2 - a_i^2} \left(a_i^2 - 2b_i x + x^2 \right). \end{aligned}$$

Now $a^{-1}(y)$ can be found using the quadratic formula:

$$a^{-1}(y) = \left(2b_i \pm \sqrt{4b_i^2 - 4(a_i^2 + y/\hat{\rho}_i(b_i^2 - a_i^2))} \right) /2 = b_i - \sqrt{(1 - y/\hat{\rho}_i)(b_i^2 - a_i^2)},$$

where the final equality follows from $x \in [a_i, b_i]$.

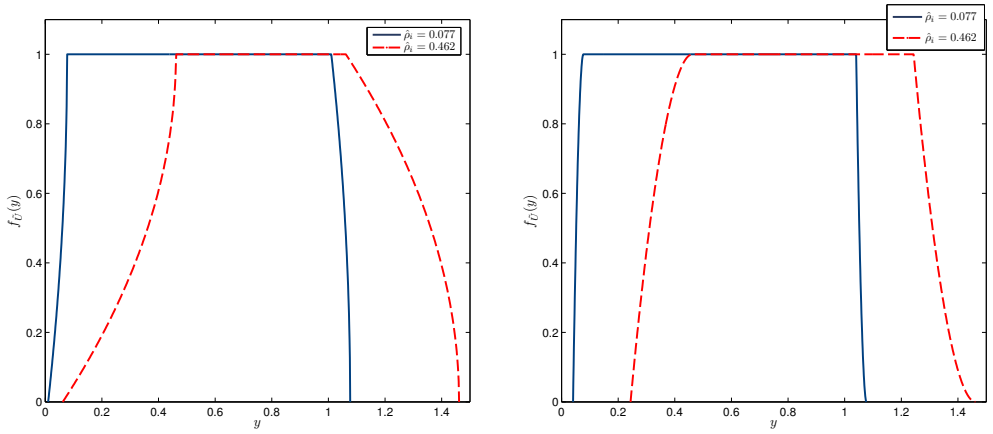


Figure 2.2: Probability density function of U_i^* with uniform (left) and Pareto (right) service times in a PS polling system.

In this case $\mathcal{X} = [a_i, b_i]$, which means that the minimum value for y is $a(a_i) = \hat{\lambda}_i a_i$. On the other side of the boundaries of the conditional uniform distribution, y needs to be greater than $1 + \hat{\lambda}_i a_i$, using this we get

$$f_{U_i^*}(y) = \begin{cases} 1 - \frac{\sqrt{(1-y/\hat{\rho}_i)(b_i^2 - a_i^2)}}{b_i - a_i} & y \in [\hat{\lambda}_i a_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i a_i] \\ \frac{\sqrt{(1-(y-1)/\hat{\rho}_i)(b_i^2 - a_i^2)}}{b_i - a_i} & y \in (1 + \hat{\lambda}_i a_i, \hat{\rho}_i + 1]. \end{cases}$$

Figure 2.2 illustrates the shape of the pdf of U_i^*, when the $c_{B_i}^2$ of the uniform service-time distribution is equal to 0.25 and for two different values of $\hat{\rho}_i$.

Pareto service times

Assume that the service time has a Pareto distribution with parameters a_i and b_i, i.e. we assume that the density of the service time, for $x \geq b_i$, is

$$f_{B_i}(x) = a_i b_i^{a_i} x^{-(a_i+1)}.$$

We assume that $a_i > 2$ such that the second moment is finite. In line with [52; 115] this is sufficient for the HT limit to hold.

Now, we have

$$\begin{aligned} a(x) &= \hat{\lambda}_i \, \mathbb{E}[\min(B_i, x)] = \hat{\lambda}_i \left(\frac{a_i b_i}{a_i - 1} \left(1 - b_i^{a_i - 1} x^{1-a_i}\right) + b_i^{a_i} x^{1-a_i} \right) \\ &= \hat{\rho}_i \left(1 - b_i^{a_i - 1} x^{1-a_i} a_i^{-1}\right). \end{aligned}$$

Some basic calculations lead to

$$a^{-1}(y) = b_i(a_i(1 - y/\hat{\rho}_i))^{\frac{1}{1-a_i}}.$$

Here y needs to be larger than $a(b_i) = \hat{\rho}_i(1 - a_i^{-1}) = \hat{\lambda}_i b_i$. We have

$$f_{U_i^*}(y) = \begin{cases} 1 - (a_i(1 - y/\hat{\rho}_i))^{\frac{-a_i}{1-a_i}} & y \in [\hat{\lambda}_i b_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i b_i] \\ (a_i(1 - (y-1)/\hat{\rho}_i))^{\frac{-a_i}{1-a_i}} & y \in (1 + \hat{\lambda}_i b_i, 1 + \hat{\rho}_i]. \end{cases}$$

Figure 2.2 shows the pdf of U_i^* if the Pareto service-time distribution has a squared coefficient of variation equal to 4, for two different values of $\hat{\rho}_i$.

For the special case in which $a_i \to \infty$ the squared coefficient of variation (SCV) of the Pareto distribution goes to zero. In that case, it can be seen that U_i^* has a uniform distribution on the interval $[\hat{\rho}_i, 1 + \hat{\rho}_i]$, which is in line with the case of deterministic service times.

Discrete service times
Using Remark 2.3, Theorem 2.5 still applies by extending the range of $a(\cdot)$ (or modifying the inverse $a^{-1}(\cdot)$). An interesting example is when the service time has probability mass at two points. Assume that B_i equals a small value a_i with probability p_i, or a large value b_i with probability $1 - p_i$. Now, letting $x \in [a_i, b_i]$, we have $\mathbb{E}[\min(B_i, x)] = (1 - p_i)x + p_i a_i$, giving $a(x) = \hat{\lambda}_i((1 - p_i)x + p_i a_i)$. With $x \in [a_i, b_i]$, we note that $a(x)$ is thus continuous and strictly increasing. Hence, we obtain

$$f_{U_i^*}(y) = \begin{cases} p_i & y \in [a_i\hat{\lambda}_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + a_i\hat{\lambda}_i] \\ 1 - p_i & y \in (1 + a_i\hat{\lambda}_i, 1 + \hat{\rho}_i]. \end{cases}$$

2.3.5 Shortest-Job-First

For the SJF policy, it is convenient to condition on x, the amount of work that a tagged customer brings into the system. For SJF, the service-time distribution is assumed to be absolutely continuous. The following results gives an expression for the LST of the conditional sojourn time $T_i^*(x)$ in terms of the cycle-time distributions (cf. [40]): for $\rho < 1$, $Re(s) > 0$, $x > 0$,

$$T_i^*(s|x) = e^{-sx}\frac{C_i^*(\lambda_i(1 - \varphi(s, x))) - C_i^*(s + \lambda_i(1 - \varphi(s, x)))}{s\,\mathbb{E}[C_i]} \quad (i \in I_{SJF}), \quad (2.28)$$

where $\varphi(s, x) := \mathbb{E}\left[e^{-sB_i\mathbf{1}_{\{B_i \le x\}}}\right]$. This leads to the following theorem for the limiting distribution of the conditional sojourn time $T_i(x)$.

Theorem 2.6 (Conditional sojourn time)**.** *For $\rho \uparrow 1$, $x > 0$,*

$$(1-\rho)T_i(x) \to_d U_i(x)\tilde{\mathbf{C}}_i \quad (i \in I_{SJF}),$$

where $U_i(x)$ is a uniform$[\hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i\le x\}}], 1+\hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i\le x\}}]]$ random variable and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (2.5)*. The random variables $U_i(x)$ and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. The result follows directly by combining Equation (2.28) and Property 2.1 along lines similar to those in the proof of Theorem 2.4. □

The unconditional sojourn time is presented in the following theorem.

Theorem 2.7 (Unconditional sojourn time)**.** *For $\rho \uparrow 1$,*

$$(1-\rho)T_i \to_d U_i^*\tilde{\mathbf{C}}_i \quad (i \in I_{SJF}),$$

where U_i^* is a generalized trapezoidal distribution as characterized in Equation (2.29) *with $a(x) = \hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i\le x\}}]$ and $\tilde{\mathbf{C}}_i$ as given in Theorem 2.6. The random variables U_i^* and $\tilde{\mathbf{C}}_i$ are independent.*

Proof. Take $a(x) = \hat{\lambda}_i\,\mathbb{E}[B_i\mathbf{1}_{\{B_i\le x\}}]$ such that $U_i(x)$ is uniformly distributed on $[a(x), a(x)+1]$, see Theorem 2.6. Clearly, $a(x_{min}) = 0$, $a(x_{max}) = \hat{\rho}_i$ and $a(x)$ is continuous and strictly increasing. Using Lemma 2.1, we obtain the unconditioned distribution

$$f_{U_i^*}(y) = \begin{cases} F_{B_i}(a^{-1}(y)) & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ 1 - F_{B_i}(a^{-1}(y-1)) & y \in (1, 1+\hat{\rho}_i]. \end{cases} \tag{2.29}$$

This completes the proof. □

Note that, similar to the PS case, the trapezoidal distribution U_i^* depends on the complete service-time distribution. Below, we present some special cases.

Exponential service times
Suppose B_i is exponentially distributed with parameter b_i. First calculate

$$\mathbb{E}[B_i\mathbf{1}_{\{B_i\le x\}}] = \int_0^x y b_i e^{-b_i y}\,\mathrm{d}y = \frac{1}{b_i}(1 - e^{-b_i x}(1+b_i x)).$$

Hence, $a(x) = \hat{\rho}_i(1 - e^{-b_i x}(1+b_i x))$. To determine $a^{-1}(y)$, we solve $a(x) = y$ for x and, after some rewriting, obtain the following equation

$$-e^{-1}(1 - y/\hat{\rho}_i) = te^t, \tag{2.30}$$

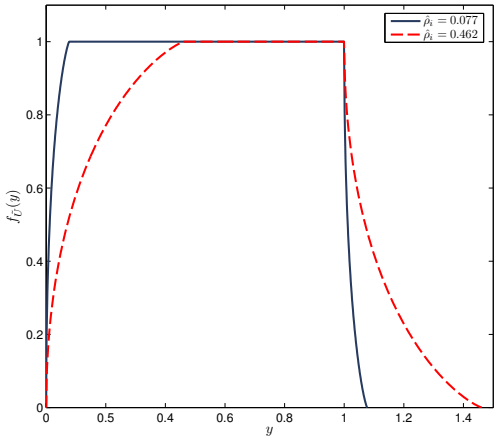


Figure 2.3: Probability density function of $\tilde{U}$ with exponential service times in a SJF polling system.

where $t = -(1 + b_i x)$. We thus need the solution of (2.30), which is known to be given in terms of the Lambert W function. Observe that the equation te^t may have multiple solutions, but we need the solutions for real $t \leq -1$, denoted by $W_{-1}(\cdot)$. This function decreases from $W_{-1}(-1/e) = -1$ to $W_{-1}(0^-) = -\infty$. From the above we derive $a^{-1}(y) = -(W_{-1}(-e^{-1}(1 - y/\hat{\rho}_i)) + 1)/b_i$.

Since $F_{B_i}(x) = 1 - e^{-b_i x}$, the probability density function $f_{U_i^*}(y)$ of the generalized trapezoidal distribution U_i^* becomes

$$f_{U_i^*}(y) = \begin{cases} 1 - e^{W_{-1}(-e^{-1}(1-y/\hat{\rho}_i))+1} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ e^{W_{-1}(-e^{-1}(1-(y-1)/\hat{\rho}_i))+1} & y \in (1, 1 + \hat{\rho}_i]. \end{cases} \tag{2.31}$$

The form of this distribution only depends on $\hat{\rho}_i$, this means that it only depends on the ratio between the mean interarrival time and the mean service time. In Figure 2.3, the probability density function is plotted for two different values of $\hat{\rho}_i$. The figure shows that for small $\hat{\rho}_i$, the distribution is close to a uniform distribution. When $\hat{\rho}_i$ increases, the distribution gets more skewed to the right.

Uniform service times

Suppose B_i has a uniform distribution with parameters a_i and b_i, with $a_i < b_i$. We have

$$\mathbb{E}[B_i \mathbf{1}_{\{B_i \leq x\}}] = \int_{a_i}^{x} \frac{u}{b_i - a_i} \, \mathrm{d}u = \frac{x^2 - a_i^2}{2(b_i - a_i)} = \mathbb{E}[B_i] \frac{x^2 - a_i^2}{b_i^2 - a_i^2}, \quad \text{for } a_i \leq x \leq b_i.$$

This gives $a(x) = \hat{\rho}_i \frac{x^2 - a_i^2}{b_i^2 - a_i^2}$. Some basic calculus yields the inverse of $a(\cdot)$: $a^{-1}(y) = \sqrt{y\,(b_i^2 - a_i^2)\,/\hat{\rho}_i + a_i^2}$. Because $F_{B_i}(x) = (x - a_i)/(b_i - a_i)$,

$$f_{U_i^*}(y) = \begin{cases} \frac{\sqrt{y/\left(b_i^2 - a_i^2\right)/\hat{\rho}_i + a_i^2} - a_i}{b_i - a_i} & y \in [\hat{\lambda}_i a_i, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1 + \hat{\lambda}_i a_i] \\ 1 - \frac{\sqrt{(y-1)\left(b_i^2 - a_i^2\right)/\hat{\rho}_i + a_i^2} - a_i}{b_i - a_i} & y \in (1 + \hat{\lambda}_i a_i, 1 + \hat{\rho}_i]. \end{cases}$$

Pareto service times
Suppose B_i is Pareto distributed with parameters $a_i > 2$ and b_i. Note that $a_i > 2$ ensures that the second moment is finite, such that the HT limit exists (see e.g. [52; 115]). It is easy to show that $a(x) = \hat{\rho}_i(1 - b_i^{a_i - 1} x^{1 - a_i})$ and $a^{-1}(y) = b_i(1 - y/\hat{\rho}_i)^{\frac{1}{1-a_i}}$. Using that $F_{B_i}(x) = 1 - (b_i/x)^{a_i},\ x \geq b_i$ gives

$$f_{U_i^*}(y) = \begin{cases} 1 - (1 - y/\hat{\rho}_i)^{\frac{-1}{1-a_i}} & y \in [0, \hat{\rho}_i) \\ 1 & y \in [\hat{\rho}_i, 1] \\ (1 - (y-1)/\hat{\rho}_i)^{\frac{-1}{1-a_i}} & y \in (1, 1 + \hat{\rho}_i]. \end{cases}$$

2.4 Results for models with globally gated service

In this section we consider the case of a globally gated service. Recall that (without loss of generality) we assume that the global gate closes at successive polling instants at Q_1 (see for example [40] for a description of the globally gated model). As in Section 2.3, we analyze LCFS, ROS, PS and SJF in addition to FCFS. Since the derivations for globally gated are largely similar to the case gated service at all queues, we only present the final results and omit the proofs.

The following result (proven in [136]) gives an asymptotic expression for the distribution of the cycle times C_i, defined in Section 2.2. Note again that the cycle times do not depend on the local scheduling policy.

Property 2.5 (Convergence of cycle times for globally gated service discipline)**.** *For the globally-gated system we have: for $i = 1, \ldots, N$, as $\rho \uparrow 1$,*

$$(1 - \rho) C_i \to_d \tilde{\Gamma},$$

where $\tilde{\Gamma}$ has a gamma distribution with parameters

$$\alpha := \frac{2r}{\sigma^2}, \quad \mu := \frac{2}{\sigma^2}. \tag{2.32}$$

with σ^2 given by (2.6)*.*

Following the same line of reasoning as in Section 2.3, we obtain the waiting-time distributions for all considered scheduling disciplines for globally gated service in heavy traffic. For convenience, we define $\mathrm{P}_i := \sum_{j=1}^{i} \hat{\rho}_j$ for $i = 1, \ldots N$ and by convention $\mathrm{P}_0 := 0$.

Theorem 2.8. *For globally gated service and $\rho \uparrow 1$, the following properties hold:*

(i) For $i \in I_{FCFS}, I_{LCFS}$,

$$(1-\rho)W_i \to_d U_i \tilde{\mathbf{C}}_i,$$

where U_i is a uniform$[\mathrm{P}_i, 1+\mathrm{P}_{i-1}]$ random variable if $i \in I_{FCFS}$ and U_i is a uniform random variable on the interval $[\mathrm{P}_{i-1}, 1+\mathrm{P}_i]$ if $i \in I_{LCFS}$.

(ii) For $i \in I_{ROS}$,

$$(1-\rho)W_i \to_d \tilde{U}_i^* \tilde{\mathbf{C}}_i,$$

where $\tilde{U}_i^*$ has a trapezoidal distribution with probability density function

$$f_{\tilde{U}_i^*}(y) = \begin{cases} (y-\mathrm{P}_{i-1})/\hat{\rho}_i & y \in [\mathrm{P}_{i-1}, \mathrm{P}_i) \\ 1 & y \in [\mathrm{P}_i, \mathrm{P}_{i-1}+1] \\ (\mathrm{P}_i+1-y)/\hat{\rho}_i & y \in (\mathrm{P}_{i-1}+1, \mathrm{P}_i+1]. \end{cases}$$

(iii) For $i \in I_{PS}, I_{SJF}$,

$$(1-\rho)T_i(x) \to_d U_i(x) \tilde{\mathbf{C}}_i,$$

where $U_i(x)$ is a uniform$[\mathrm{P}_i + \hat{\lambda}_i \kappa_{i,x}, 1 + \mathrm{P}_i + \hat{\lambda}_i \kappa_{i,x}]$ random variable, with $\kappa_{i,x} := \mathbb{E}[\min\{B_i, x\}]$ for $i \in I_{PS}$ and $\kappa_{i,x} := \mathbb{E}[B_i \mathbf{1}_{\{B_i \le x\}}]$ for $i \in I_{SJF}$.

In all cases, $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (2.32) and it is independent of the uniform distributions.

In the above theorem we only presented the conditional waiting times for PS and SJF. Using Lemma 2.1, this results in a generalized trapezoidal times a gamma distribution for the unconditional waiting time, as in Subsections 2.3.4 and 2.3.5. Finally, also the intuitive interpretation of the heavy-traffic limit using HTAP is directly in line with that of Section 2.3.

Remark 2.8 (Renewal arrival processes)**.** For the model under consideration with Poisson arrivals, Theorems 2.1–2.8 give the asymptotic waiting-time and sojourn-time distributions for the LCFS, ROS, SJF and PS scheduling disciplines. Following the well-established line of argumentation found in [51; 52; 114], we conjecture that similar results hold for renewal arrival processes. In particular, in [114] a strong conjecture is given that in heavy traffic the same results for the scaled cycle-time and waiting-time distributions for FCFS hold as those in Properties 2.1 and 2.3, respectively, but where the parameter σ^2 is now replaced by

$$\sigma^2_{renewal} = \sum_{i=1}^{N} \hat{\lambda}_i \left(\mathrm{Var}[B_i] + c^2_{A_i} \, \mathbb{E}[B_i]^2 \right). \tag{2.33}$$

Here A_i represents the interarrival times between arriving customers at queue i and $c^2_{A_i}$ is its squared coefficient of variation. Note for the special case of Poisson arrivals, the expression for $\sigma^2_{renewal}$ coincides with σ^2. Based on the HTAP, we derive the following conjecture for renewal arrivals.

Conjecture 2.1. *For independent renewal arrival processes, Theorems 2.1–2.8 are also valid when* σ^2 *defined in* (2.6) *is replaced by* $\sigma^2_{renewal}$ *defined in* (2.33).

In the next section, we use this conjecture to derive and validate approximations for the waiting-time and sojourn-time distributions for renewal arrivals.

2.5 Closed-form approximations for systems with arbitrary load

In Sections 2.3 and 2.4, we have derived heavy-traffic limits for the (scaled) waiting-time and sojourn-time distributions under several scheduling disciplines. These results not only give valuable insights into polling models operating under a critical load, but are also useful in the study of polling models that are arbitrary loaded (i.e. $\rho < 1$). Below, we describe how the results derived in this chapter can be used to obtain closed-form approximations for the waiting-time and sojourn-time distributions in polling models with renewal arrivals and arbitrary load.

For systems with FCFS service at all queues, Boon et al. [31] derive a closed-form approximation, denoted by $\mathbb{E}[W_i^{(app)}]$, for the *mean* waiting time by interpolating between known light-traffic and heavy-traffic limits. Based on this approximation, Dorsman et al. [62] propose to approximate the *complete* waiting-time distribution by, for $x > 0$,

$$\mathbb{P}(W_i < x) \approx \mathbb{P}(U_i C_i < (1-\rho)x) \quad (i \in I_{FCFS}), \tag{2.34}$$

where U_i is uniformly $[\hat{\rho}_i, 1]$ distributed (as defined in Property 2.1) and where C_i is gamma-distributed with shape parameter $\alpha + 1$ and scale parameter

$$\mu_i^{(app)} := \frac{1+\hat{\rho}_i}{1-\rho} \frac{r\delta + \sigma^2_{renewal}}{2\sigma^2_{renewal}\,\mathbb{E}[W_i^{(app)}]^2}, \tag{2.35}$$

where α, δ and $\sigma^2_{renewal}$ are defined in Property 2.1 and (2.33).

To develop an approximation for the other scheduling disciplines under consideration, recall that the cycle-time distribution is insensitive to the scheduling discipline. Based on this observation, for $i \in I_{LCFS}$, we approximate the waiting-time distribution by (2.34), where the distribution of C_i is kept the same, but with U_i uniformly distributed on $[0, 1+\hat{\rho}_i]$ (cf. Theorem 2.1). Likewise, for $i \in I_{ROS}$, the waiting-time distribution can be approximated by (2.34) with U_i replaced by a trapezoidal distribution defined in Theorem 2.3. Approximations for the sojourn-time distributions can be obtained by using (2.4). For PS and SJF, approximations can be obtained directly for the

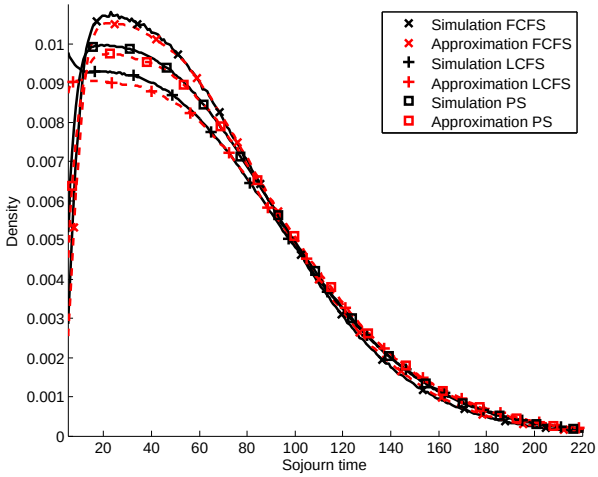


Figure 2.4: Simulated and approximated sojourn-time distributions at queue 1 for the gated model with $\rho = 0.95$ for FCFS, LCFS and PS.

sojourn-time distributions, using Theorems 2.5 and 2.7, respectively. For the case of globally gated service, waiting-time distributions are approximated in a similar way using the results in Section 2.4; details are omitted here.

Throughout this section we will show numerical results based on simulations to illustrate the usefulness and accuracy of the closed-form approximations. We consider a three-queue polling model with gated service at each queue and with the following parameters. The service times at queues 1, 2 and 3 are uniformly distributed with means 1, 2, 3, respectively, and with squared coefficient of variation 1/4. The switch-over time distributions are exponentially distributed with means $r_1 = r_2 = 1$ and $r_3 = 3$. The arrival processes at each of the queues are renewal and mutually independent. The ratios between the arrival rates are 1:3:2, and interarrival-time distributions are uniformly distributed with squared coefficient of variation 1/4. Note that the system is rather asymmetric and the ratios between the per-queue load values are 1:6:6.

To illustrate the fact that the approximation of the distribution is accurate in heavy traffic, Figure 2.4 plots the simulated and approximated density functions of the sojourn-time distributions at Q_1 for FCFS, LCFS and PS service (at all queues) for a heavily loaded system with $\rho = 0.95$. As expected, the approximations closely follow the simulations. Figure 2.4 also illustrates that the differences between the different scheduling disciplines are significant and are well-captured by the asymptotic results.

To proceed, Figure 2.5 shows the simulated and approximated probability density functions for the per-queue sojourn-time distributions for the model with LCFS ser-

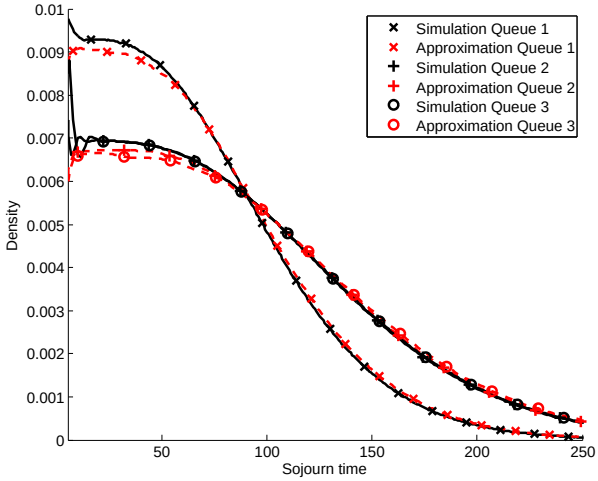


Figure 2.5: Simulated and approximated per-queue sojourn-time distributions for the gated model with $\rho = 0.95$ and LCFS service.

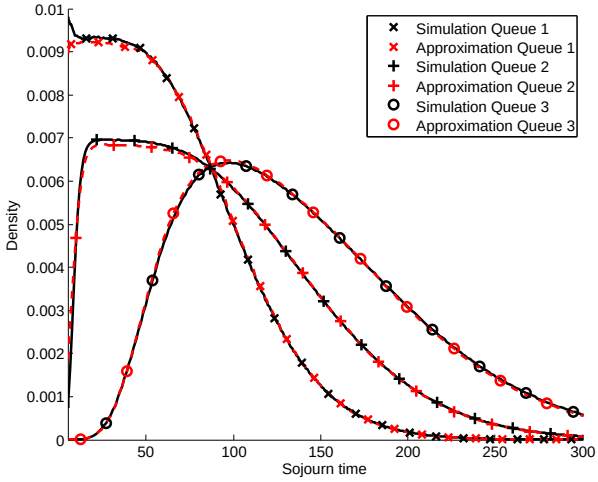


Figure 2.6: Simulated and approximated per-queue sojourn-time distributions for the globally-gated model with $\rho = 0.95$ and LCFS service.

vice at each queue, for a heavily loaded system with $\rho = 0.95$. Figure 2.6 shows the results for the same model but with globally-gated service. The results in Figures 2.5 and 2.6 illustrate the fact that the per-queue sojourn-time distributions are well-captured by the approximations for heavy-traffic scenarios (as they should).

Next, we assess the accuracy of the approximations for the complete range of load

values. To this end, Table 2.2 shows the simulated and approximated values of the mean sojourn times at Q_1 and their relative absolute difference defined as

$$\Delta\% = 100\% \times \frac{|\text{App} - \text{Sim}|}{\text{Sim}}$$

for different values of ρ and for all the scheduling disciplines considered in this chapter. Recall that the mean sojourn-times are the same for FCFS, LCFS and ROS service, but may differ for PS and SJF service. Table 2.3 shows the results for the standard deviations of the sojourn times at Q_1. In Table 2.2 we see that the approximation of

	FCFS/LCFS/ROS			PS			SJF		
ρ	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.10	4.92E00	4.97E00	1.1	4.92E00	4.99E00	1.3	4.92E00	4.97E00	0.9
0.30	5.75E00	6.02E00	4.7	5.75E00	6.07E00	5.6	5.75E00	5.99E00	4.2
0.50	7.29E00	7.87E00	7.9	7.30E00	7.98E00	9.2	7.28E00	7.80E00	7.1
0.70	1.13E01	1.21E01	6.9	1.14E01	1.23E01	8.0	1.13E01	1.19E01	6.1
0.80	1.65E01	1.74E01	5.0	1.68E01	1.78E01	5.8	1.64E01	1.71E01	4.4
0.90	3.22E01	3.31E01	2.6	3.24E01	3.40E01	5.0	3.18E01	3.25E01	2.2
0.95	6.37E01	6.45E01	1.3	6.53E01	6.63E01	1.5	6.26E01	6.33E01	1.1
0.98	1.58E02	1.59E02	0.4	1.62E02	1.63E02	0.6	1.55E02	1.56E02	0.5

Table 2.2: Mean sojourn times for different scheduling disciplines.

the mean sojourn time is most accurate for lightly and heavily loaded systems. This is due to the fact that, by construction, the approximations are asymptotically exact in the limiting cases of $\rho \downarrow 0$ and $\rho \uparrow 1$. For moderately loaded systems, the error is highest, but it still is no more than a few percent. Table 2.3 shows that the results for the standard deviations are accurate for heavily loaded systems, but may become less accurate for low-to-medium loaded systems. This is probably caused by the fact that the approximation for the second (and higher) moments of the sojourn times in (2.34)-(2.35) is asymptotically exact for $\rho \uparrow 1$, but not for $\rho \downarrow 0$ (as opposed to the first moments, which are asymptotically exact for $\rho \downarrow 0$ by construction).

	FCFS			ROS			SJF		
ρ	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.1	3.44E00	2.63E00	23.4	3.44E00	2.78E00	19.1	3.44E00	2.81E00	18.1
0.3	3.93E00	3.30E00	15.9	3.93E00	3.49E00	11.2	3.93E00	3.52E00	10.5
0.5	4.89E00	4.49E00	8.1	4.94E00	4.75E00	3.8	4.93E00	4.77E00	3.1
0.7	7.49E00	7.24E00	3.4	7.71E00	7.66E00	0.6	7.68E00	7.65E00	0.3
0.8	1.08E01	1.07E01	1.8	1.13E01	1.13E01	0.1	1.12E01	1.13E01	0.3
0.9	2.11E01	2.09E01	0.9	2.21E01	2.21E01	0.1	2.19E01	2.19E01	0.0
0.95	4.14E01	4.13E01	0.3	4.37E01	4.37E01	0.1	4.34E01	4.35E01	0.1
0.98	1.03E02	1.03E02	0.3	1.09E02	1.09E02	0.0	1.08E02	1.08E02	0.3

Table 2.3: Standard deviations of the sojourn times for FCFS, ROS and SJF.

In summary, the numerical results (1) illustrate the validity of the asymptotic results, and (2) demonstrate that the sojourn-time approximations nicely capture the impact of the local scheduling policies on the sojourn-time distributions and are accurate over the whole range of load values.

2.6 Concluding remarks

In this chapter, we have studied the impact of scheduling within queues on the waiting-time and sojourn-time distributions in polling systems. We have presented the first HT analysis of polling models where the local scheduling policy is not FCFS, but instead, is varied as LCFS, ROS, PS and SJF. The main contribution is the derivation of asymptotic closed-form expressions for the LST of the scaled waiting-time and sojourn-time distributions under HT conditions. The results raise a number of remarks and challenging open questions for further research, on which we would like to elaborate in the current section.

In this chapter we have assumed that *all* the queues in the polling system follow the (globally) gated service discipline. However, this assumption can easily be relaxed; that is, we only have to assume that the specific queue for which we derive the waiting-time distribution is served according to the gated service discipline (see, also, [40]). For all the other queues, we only have to postulate that the service discipline belongs to the broad class of local branching-type disciplines [121], which includes gated and exhaustive service as special cases.

Furthermore, as [40] argues, the analysis of exhaustive polling systems is more complicated because the waiting times of the customers who are served during a visit are affected by later arrivals which take place during that visit period (which is obviously not the case for gated systems). Extension of the results to a broader class of service disciplines is a challenging topic for further research. In Chapter 3 we will extend the results to the exhaustive service discipline.

Finally, an interesting question is a generic optimization of the system's performance with respect to the choice of the local scheduling disciplines. With respect to mean sojourn times, it follows from [153] that SJF is optimal. For non-anticipating scheduling disciplines, the results in [1] suggest that the optimal discipline for minimizing mean sojourn times belongs to the family of multilevel PS disciplines. Optimization results beyond the mean, e.g. in terms of tails of sojourn times, is still open. In this context, it is worthwhile to note that the sojourn-time distribution at a given queue does not depend on the choice of the local scheduling discipline at any other queue. This implies that the sojourn-time distribution at a queue only depends on the choice of the local service order at that same queue. Therefore, the results presented in this chapter provide a good starting point for tackling this type of optimization problem.

Chapter 3

Exhaustive polling systems with non-FCFS service

3.1 Introduction

In the previous chapter, we derived HT limits of the waiting-time distributions in cyclic polling models (described in Chapter 1) with gated and globally-gated service for the LCFS, ROS, PS and SJF local service orders. In the current chapter, we extend the results to the case of exhaustive service at each of the queues, which is fundamentally more complicated than the gated and globally-gated case (as also stated in [40]). The additional complexity of the exhaustive-service model compared to the (globally-)gated model is that customers that arrive during a visit of the server at a queue may intervene with the customers that were present at the beginning of that visit period.

In this chapter, we study Poisson-driven cyclic polling models with general service-time and switch-over time distributions, and with exhaustive service at all queues (see Section 3.11 for a relaxation of that assumption). For this model, we consider the following seven scheduling policies that determine the local order in which the customers at a given queue are served: FCFS (which is used as a benchmark), LCFS (with and without preemption), ROS, PS, the multi-class priority scheduling (with and without preemption), SJF and SRPT. For these models, we derive new, exact expressions for the waiting-time distributions in terms of the intervisit time distributions for stable systems. Subsequently, we use these expressions to derive the asymptotic waiting-time distributions for each of the local order policies under HT assumptions (i.e., when the load approaches 1). We show that in all cases the asymptotic waiting-time distribution at queue i can be expressed as the product of two independent random variables Γ and Θ_i, where Γ is gamma-distributed with known parameters that are independent of the scheduling policy. Moreover, we derive the distribution of the random variable Θ_i, which expresses the impact of the local service order on the asymptotic waiting-time distribution. The results are exact and give a full characterization of the limiting behavior of the system, and as such provide new

fundamental insight in the influence of the local scheduling policy on the waiting-time performance of polling models. As a by-product, the HT limits suggest simple closed-form approximations for the complete waiting-time distributions for stable systems with arbitrary load values strictly less than 1. The accuracy of the approximations is evaluated by several numerical examples.

The remainder of the chapter is organized as follows. In Section 3.2, the model is described and the notation is introduced. In Section 3.3, we present preliminary results, including the HT asymptotics for FCFS that serve as a benchmark. The waiting-time distributions and HT asymptotics for LCFS, ROS, PS, multi-class priority queues, and SJF and SRPT are derived in Sections 3.4–3.8, respectively. The results are summarized in Section 3.9. Furthermore, Section 3.10 proposes a simple approximation for the waiting-time distributions and present numerical results to evaluate the accuracy of the approximations. Finally, Section 3.11 contains a number of concluding remarks and addresses several topics for further research.

3.2 Model description

In this section we introduce the notation and give a description of the model. Recall that useful notation with respect to a one-dimensional absolutely-continuous random variable X is presented in Table 2.1.

The model is as follows. We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. At each queue, the service discipline is exhaustive; that is, the server proceeds to the next queue when the queue is empty. Customers arrive at Q_i according to a Poisson process $\{N_i(t),\ t \in \mathbb{R}\}$ with rate λ_i. These customers are referred to as type-i customers. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i. The kth moment of the service time of an arbitrary customer is denoted by $\mathbb{E}[B^k] = \sum_{i=1}^{N} \lambda_i \, \mathbb{E}[B_i^k]/\Lambda$, $k = 1, 2, \ldots$. The load offered to Q_i is $\rho_i = \lambda_i \, \mathbb{E}[B_i]$ and the total load offered to the system is equal to $\rho = \sum_{i=1}^{N} \rho_i$. A necessary and sufficient condition for stability of the system is $\rho < 1$. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is a random variable S_i. We let $S = \sum_{i=1}^{N} S_i$ denote the total switch-over time in a cycle. The random variable C_i describes the cycle time of the server, defined as the time between two successive departures of the server from Q_i. The mean cycle time is known to be the same for all queues, and is given by $\mathbb{E}[C_i] = \mathbb{E}[C] = \mathbb{E}[S]/(1-\rho)$. Denote by V_i the visit time at Q_i, defined as the time elapsed between a polling instant at Q_i (i.e., the moment the server arrives at the queue) and the server's successive departure from Q_i. Denote by I_i the intervisit time of Q_i, defined as the time elapsed between a departure of the server from Q_i and the successive polling instant at Q_i. Note that $C_i = I_i + V_i$, for $i = 1, \ldots, N$.

The *local service order policy* of a queue determines the order in which the customers are served during a visit period of the server at that queue. We only consider work-conserving policies. We denote $i \in I_P$ if Q_i receives scheduling policy $P \in$ {FCFS, LCFS, LCFS-PR, ROS, PS, NPRIOR, NPRIOR-PR, SJF, SRPT}; for example, I_{FCFS} is the (index) set of queues that are served on a FCFS basis. We refer to Table 1.1 for a short explanation of the policies

In this chapter we mainly focus on HT limits, i.e., the limiting behavior as ρ approaches 1, see also Chapter 2. Recall that for each variable x that is a function of ρ, we denote its value *evaluated at* $\rho = 1$ by $\hat{x}$.

Let T_i denote the sojourn time of an arbitrary customer at Q_i, defined as the time between the moment of arrival of a customer and the moment at which the customer departs from the system. The waiting time W_i of an arbitrary customer at Q_i is defined as the sojourn time minus the service requirement. When $\rho \uparrow 1$, all queues become unstable, therefore the focus lies on the limiting distribution for $\rho \uparrow 1$ of the random variables $\tilde{W}_i := (1-\rho)W_i$ and $\tilde{T}_i := (1-\rho)T_i$, referred to as the *scaled* waiting times and sojourn times at Q_i, respectively. We denote by $\Gamma(\alpha, \mu)$ a gamma-distributed random variable with shape and rate parameters α and μ, respectively. Moreover, we denote by $U[a, b]$, with $a < b$, a random variable that is uniformly distributed over the interval $[a, b]$. For later reference, recall from (2.3) that the LST of the random variable $U[a, b]\Gamma(\alpha+1, \mu)$, where $U[a, b]$ and $\Gamma(\alpha+1, \mu)$ are independent, is given by, for, $Re(s) > 0$,

$$\mathbb{E}\left[e^{-sU[a,b]\Gamma(\alpha+1,\mu)}\right] = \frac{\mu}{\alpha s(b-a)}\left\{\left(\frac{\mu}{\mu+sa}\right)^{\alpha} - \left(\frac{\mu}{\mu+sb}\right)^{\alpha}\right\}. \tag{3.1}$$

In Sections 3.3 to 3.8 we derive expressions for the LSTs of the waiting-time distributions for the scheduling disciplines shown in Table 1.1.

3.3 Preliminaries and method outline

In this section we formulate a number of known preliminary results that serve as a reference for the remaining sections. In Section 3.1 we give expressions for the asymptotic distributions of the cycle and intervisit times under HT assumptions. In Section 3.2 we use these results to give an expression for the LST of the waiting-time distribution for the case of FCFS service. We refer to [132] for rigorous proofs of these results.

3.3.1 Cycle and intervisit times

To start, let us consider the distribution of the cycle time C_i. Recall that C_i is defined as the time between two successive *departures* of the server from Q_i. A simple

but important observation is that the distribution of C_i does not depend on the local scheduling policy, provided that the policy is work-conserving. This means that we can use the results for the cycle times and also for the intervisit times throughout the rest of this chapter. The following result gives a characterization of the limiting behavior of the scaled cycle-time distributions, stating that the (scaled) cycle times $\tilde{C}_i := (1-\rho)C_i$ converge to a gamma distribution with known parameters in HT.

Property 3.1 (Convergence of cycle times)**.** *For $i = 1, \ldots, N$, as $\rho \uparrow 1$,*

$$\tilde{C}_i \to_d \tilde{\Gamma}, \tag{3.2}$$

where $\tilde{\Gamma}$ has a gamma distribution with parameters

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu := \frac{\delta}{\sigma^2}, \tag{3.3}$$

with

$$\sigma^2 := \frac{\mathbb{E}[B^2]}{\mathbb{E}[B]}, \quad \textit{and} \quad \delta := \sum_{i=1}^{N} \hat{\rho}_i(1-\hat{\rho}_i). \tag{3.4}$$

Note that the distribution of the cycle time C_i is related to the intervisit time I_i in the following way (see e.g. [28]):

$$\mathbb{E}[I_i] = (1-\rho_i)\,\mathbb{E}[C_i], \quad \text{and} \quad \mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))I_i}] = \mathbb{E}[e^{-sC_i}]. \tag{3.5}$$

Here ξ_i is the busy period of a regular M/G/1 queue with arrival rate λ_i and service time B_i. The (scaled) intervisit times $\tilde{I}_i := (1-\rho)I_i$ converge (in distribution) to a gamma distribution with known parameters as stated in the property below.

Property 3.2 (Convergence of intervisit times)**.** *For $i = 1, \ldots, N$, as $\rho \uparrow 1$,*

$$\tilde{I}_i \to_d \tilde{\Gamma}_i, \tag{3.6}$$

where $\tilde{\Gamma}_i$ has a gamma distribution with parameters

$$\alpha := \frac{\mathbb{E}[S]\delta}{\sigma^2}, \qquad \mu_i := \frac{\delta}{(1-\hat{\rho}_i)\sigma^2}, \tag{3.7}$$

where δ and σ^2 are given in Equation (3.4)*.*

In this chapter, we repeatedly use Properties 3.1 and 3.2 to derive expressions for the asymptotic scaled waiting-time distributions associated with each of the service disciplines considered herein. For each policy we use a two-step approach:

(a) we derive an expression for the LST of the limiting distribution of the waiting times in terms of the cycle- and/or intervisit-time distribution;

(b) we combine this expression with Property 3.1 or 3.2 to obtain an expression for the LST of the waiting-time distribution in HT and interpret the resulting LST.

To conclude, we add a remark with intuition for the distribution using the Heavy Traffic Averaging Principle (HTAP).

3.3.2 First-Come-First-Served

Here we illustrate the two-step approach described above for FCFS service. Regarding the first step, the following result gives an expression for the LST of the waiting time W_i in terms of the distribution of the intervisit time I_i (cf. [125]):

Property 3.3 (Waiting times in terms of intervisit times)**.** *For $Re(s) > 0$ and $\rho < 1$,*

$$W_i^*(s) = \frac{(1-\rho_i)s}{s-\lambda_i(1-B_i^*(s))}\frac{1-I_i^*(s)}{s\,\mathbb{E}[I_i]} \quad (i \in I_{FCFS}). \tag{3.8}$$

Next, as step (b), combining Properties 3.2 and 3.3, the expression for $\mathbb{E}[C_i]$, and taking limits we obtain: For $Re(s) > 0$ and $i \in I_{FCFS}$,

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \frac{1}{(1-\hat{\rho}_i)\,\mathbb{E}[S]s}\left\{1-\left(\frac{\mu_i}{\mu_i+s}\right)^{\alpha}\right\}. \tag{3.9}$$

Using (3.1), this leads to the following characterization of the limiting behavior of the scaled waiting-time distribution derived in [136]:

Property 3.4 (Convergence of the waiting times)**.** *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d U_i\tilde{\mathbf{I}}_i \quad (i \in I_{FCFS}), \tag{3.10}$$

where U_i is a uniformly distributed random variable on $[0,1]$, and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ_i, where α and μ_i are given in Equation (3.7).

Note that $\tilde{\mathbf{I}}_i$ is the *length-biased* counterpart of $\tilde{I}_i$, a gamma distributed random variable with parameters α and μ_i as in Equation (3.7). It is well known that if a gamma random variable has parameters α and μ_i, then its length-biased version has parameters $\alpha+1$ and μ_i.

Remark 3.1 (Intuition by the Heavy Traffic Averaging Principle)**.** Property 3.4 states that the limiting behavior of W_i is of the form $U_{FCFS}\Gamma$, where U_{FCFS} is uniformly distributed on the interval $[0,1]$. An intuitive explanation for this follows from the Heavy Traffic Averaging Principle (HTAP) combined with a fluid model ([51; 52; 113]). Loosely speaking, the HTAP principle states that the work in each queue is emptied and refilled at a rate that is much faster than the rate at which the total workload is changing. This implies that the total workload can be considered as a constant during the course of a cycle, while the loads of the individual queues fluctuate like a fluid model.

Figure 3.1 gives a graphical representation of the fluid model. On the horizontal axis, the course of a cycle with fixed length c is plotted. The cycle is divided in two parts, the intervisit time I_i with length $(1-\hat{\rho}_i)c$ and the visit time V_i with length $\hat{\rho}_i c$. On the vertical axis the workload in Q_i is plotted. The cycle starts at the completion of a visit to Q_i. Throughout the cycle, work arrives with intensity 1 and a fraction $\hat{\rho}_i$

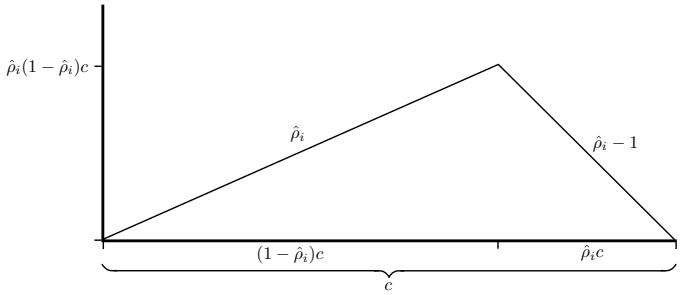


Figure 3.1: Fluid limits in heavy traffic; the amount of fluid in Q_i is plotted over the course of a cycle.

is directed to Q_i. During the visit time V_i work flows out of Q_i with rate 1 until the queue is empty. We refer to [28, p. 34-39] for an intuitive explanation based on this picture.

Here, we opt for a more direct analysis of the fluid model. Let the uniform random variable U on [0,1] denote the fraction of the cycle c that has elapsed at the arrival epoch of this particle. The particle has to wait for the remaining length of the cycle $(1-U)c$ except for the amount of work that arrives at Q_i during the cycle after the arrival of the particle. As work to Q_i arrives at rate $\hat{\rho}_i$, the latter equals $\hat{\rho}_i(1-U)c$. Hence, the waiting time equals $(1-U)c - \hat{\rho}_i(1-U)c = (1-U)(1-\hat{\rho}_i)c$. Using the fact that $U[0,1]$ is in distribution equal to $1-U[0,1]$ and $I_i = (1-\hat{\rho}_i)c$, we conclude that $\tilde{W}_i$ is uniformly distributed on $[0,1]I_i$. This interpretation gives much insight in the heavy-traffic asymptotics.

3.4 Last-Come-First-Served

In this section we consider the LCFS service discipline. In Subsection 3.4.1 we derive the results for LCFS without preemption and in Subsection 3.4.2 we look at queues with LCFS preemptive resume (LCFS-PR) service. In both subsections, we first provide a derivation of the LST of W_i for all $\rho < 1$, giving insight in the terms contributing to the delay. Then we study the behavior of W_i in the HT regime. Since we are interested in deriving the waiting-time distributions of customers that arrive in steady state, it is convenient to define stationary versions of the arrival processes on the entire real line. Hence, each arrival process N_i consists of points $\{T_{i,n}\}_{n\in\mathbb{Z}}$, where $T_{i,0} \leq 0 \leq T_{i,1}$. Associated with each point is the busy period $\xi_{i,n}$ generated by the arriving customer. The points $(T_{i,n}, \xi_{i,n})$ define a marked Poisson process on $\mathbb{R}^2$.

3.4.1 Non-Preemptive LCFS

Now we derive the LST of the waiting time of a tagged customer T that arrives at queue i in steady state. Without loss of generality, we assume that T arrives at time zero. We have to distinguish between the case where T arrives during an intervisit time, and the case where T arrives during a visit time.

Case I: the tagged customer arrives during an intervisit time
In this case, T has to wait for the server to start serving queue i; this is a residual intervisit time. In addition, T has to wait for all customers that arrived after him during the residual intervisit time and for the busy periods they generate. We have, for $i \in I_{LCFS}$,

$$W_i \text{ (\textit{given} } T \text{ arrives during intervisit time)} = I_i^{res} + \sum_{T_{i,k} \in (0, I_i^{res})} \xi_{i,k}. \tag{3.11}$$

Conditioning on I_i^{res} and the number of arrivals during I_i^{res} (as in [40]), we have for $Re(s) > 0$,

$$\begin{aligned}
&\mathbb{E}[e^{-sW_i} | \text{arrival during intervisit time}] \\
&\quad = \int_{t=0}^{\infty} e^{-st} \sum_{n=0}^{\infty} e^{-\lambda_i t} \frac{(\lambda_i t)^n}{n!} \, \mathbb{E}[e^{-s\xi_i}]^n \, \mathrm{d}\,\mathbb{P}(I_i^{res} \le t) \qquad (3.12) \\
&\quad = \int_{t=0}^{\infty} e^{-t(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))} \, \mathrm{d}\,\mathbb{P}(I_i^{res} \le t) \\
&\quad = \frac{1 - \mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))I_i)}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[I_i]} \\
&\quad = \frac{1 - \mathbb{E}[e^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1 - \rho_i)} \qquad (i \in I_{LCFS}), \qquad (3.13)
\end{aligned}$$

where for the final step we use Equation (3.5).

Case II: the tagged customer arrives during a visit time
Note that T now arrives during the service of another customer. Hence, he has to wait for a residual service duration. In addition, he has to wait for the duration of the busy periods generated by the customers that arrived during the residual service time, as they are served before the tagged customer. Hence, we have for $i \in I_{LCFS}$,

$$W_i \text{ (\textit{given} arrival during visit time)} = B_i^{res} + \sum_{T_{i,k} \in (0, B_i^{res})} \xi_{i,k}. \tag{3.14}$$

Using the similarity between (3.11) and (3.14), we immediately see that, for $i \in I_{LCFS}$,

$$\begin{aligned}\mathbb{E}[e^{-sW_i}|\text{arrival during visit time}] &= \frac{1-\mathbb{E}[e^{-(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}])B_i)}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]} \\ &= \frac{1-\mathbb{E}[e^{-s\xi_i}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]},\end{aligned}$$

where the second equality follows from the well known functional equation satisfied by the LST of the busy period of an M/G/1 queue (see e.g., [129, p. 354]). Note that the probability that an arrival occurs during a visit time is equal to ρ_i. This leads to the following proposition.

Proposition 3.1. *For $\rho < 1$, $Re(s) > 0$,*

$$\begin{aligned}W_i^*(s) = \rho_i &\frac{1-\mathbb{E}[e^{-s\xi_i}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[B_i]} \\ &+ (1-\rho_i)\frac{1-\mathbb{E}[e^{-sC_i}]}{(s+\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1-\rho_i)} \quad (i \in I_{LCFS}).\end{aligned} \tag{3.15}$$

Note that the first term appears in the LST of the waiting time in an M/G/1 queue with LCFS service order (see e.g., [129, p. 357]). Also note that Equation (3.15) was found in [122], where intervisit periods are replaced with rest periods.

The following result gives an expression for the asymptotic waiting-time distribution for LCFS service in heavy traffic.

Theorem 3.1. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \begin{cases} 0 & w.p. \quad \hat{\rho}_i \\ U_i\tilde{\mathbf{C}}_i & w.p. \quad 1-\hat{\rho}_i \end{cases} \qquad (i \in I_{LCFS}),$$

where U_i is a uniformly distributed random variable on the interval $[0,1]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (3.3).

Proof. Combining Proposition 3.1 with Property 3.1 gives the following expressions for the LST of the (scaled) waiting-time distribution. For $i \in I_{LCFS}$, $Re(s) > 0$,

$$\begin{aligned}\tilde{W}_i^*(s) &= \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) \\ &= \lim_{\rho\uparrow 1}\left(\rho_i \frac{1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]}{(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))\,\mathbb{E}[B_i]}\right. \\ &\qquad \left. + (1-\rho_i)\frac{1-\mathbb{E}[e^{-s(1-\rho)C_i}]}{(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}\right).\end{aligned} \tag{3.16}$$

Let us first consider the first term on the right-hand side of the final equation:

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \rho_i \frac{1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]}{(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))\,\mathbb{E}[B_i]} \\
&= \lim_{\rho\uparrow 1} \rho_i \frac{(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}])/(1-\rho)}{s\,\mathbb{E}[B_i]+\rho_i((1-\mathbb{E}[e^{-s(1-\rho)\xi_i}])/(1-\rho))} \\
&= \hat{\rho}_i \frac{\mathbb{E}[\xi_i]s}{\mathbb{E}[B_i]s+\hat{\rho}_i\,\mathbb{E}[\xi_i]s} \\
&= \hat{\rho}_i.
\end{aligned}$$

In the second equality, we use l'Hôpital's rule on both the numerator and the denominator, and the fact that the derivative of $\mathbb{E}[e^{-s(1-\rho)\xi_i}]$ at $s(1-\rho)=0$ is equal to $-\mathbb{E}[\xi_i]$. For the third equality we apply the well-known result $\mathbb{E}[\xi_i]=\mathbb{E}[B_i]/(1-\rho_i)$.

Now consider the second term on the right-hand side of (3.16):

$$\begin{aligned}
&\lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-\mathbb{E}[e^{-s(1-\rho)C_i}]}{\mathbb{E}[C](1-\rho_i)(s(1-\rho)+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}]))} \\
&= \lim_{\rho\uparrow 1}(1-\rho_i)\frac{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}}{\mathbb{E}[S](1-\rho_i)(s+\lambda_i(1-\mathbb{E}[e^{-s(1-\rho)\xi_i}])/(1-\rho))} \\
&= (1-\hat{\rho}_i)\frac{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}}{\mathbb{E}[S](1-\hat{\rho}_i)s(1+\lambda_i\,\mathbb{E}[\xi_i])} \\
&= (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\}. \qquad (3.17)
\end{aligned}$$

Combining the above gives

$$\tilde{W}_i^*(s) = \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1-\left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\} \quad (i\in I_{LCFS}), \qquad (3.18)$$

where α and μ are given in (3.3). Note that (3.18) corresponds to the LST of a random variable that is equal to 0 with probability $\hat{\rho}_i$ and to a uniform random variable on $[0,1]$ times a gamma distribution with probability $1-\hat{\rho}_i$. This completes the proof. □

Remark 3.2 (HTAP)**.** The mixed distribution can be intuitively explained with the HTAP and a fluid model, see Figure 3.1. With probability $\hat{\rho}_i$ a particle arrives during V_i. In this case the scaled waiting time is negligible in HT, since the residual service time and the busy periods generated by customers arriving during this time, do not scale with ρ. With probability $(1-\hat{\rho}_i)$ a particle arrives during I_i. Let the uniform random variable U_I denote the fraction of I_i that has elapsed at the arrival epoch of this particle. This arriving particle has to wait for the remaining intervisit time $(1-U_I)I_i$, in addition it has to wait for the busy periods generated by particles that

arrived during that time for duration $\hat{\rho}_i(1-U_I)I_i/(1-\hat{\rho}_i)$, the amount of work built up during the remaining intervisit time divided by the rate at which the queue is emptied. Adding the two terms and noting that $(1-U_I)$ is in distribution equal to U_I we get for the scaled waiting time of a particle arriving during an intervisit time: $W_i^{(I)} = U_I I_i/(1-\hat{\rho}_i) = U_I c$. Now we can use the HTAP and the results from [132] to find the distribution of c and arrive at the result given in Theorem 3.1.

3.4.2 LCFS with Preemptive Resume

The analysis of LCFS-PR service is largely similar to the non-preemptive LCFS case. When an arrival occurs during an intervisit time, the waiting time of the customer consists of the busy periods generated by the customers arriving during the service of the tagged customer, the residual intervisit time and the busy periods generated by the customers arriving during the residual intervisit time. This gives for Case I (see Section 4.1): For $i \in I_{LCFS\text{-}PR}$,

$$W_i \text{ (\textit{given} } T \text{ arrives during intervisit time)} = \sum_{T_{i,k}\in(0,B_i)} \xi_{i,k} + I_i^{res} + \sum_{T_{i,k}\in(0,I_i^{res})} \xi_{i,k}. \tag{3.19}$$

When the arrival occurs during a visit period, the waiting time of T consists of the busy period generated by customers arriving during the service of the tagged customer. We have in Case II: For $i \in I_{LCFS\text{-}PR}$,

$$W_i \text{ (\textit{given} } T \text{ arrives during visit time)} = \sum_{T_{i,k}\in(0,B_i)} \xi_{i,k}. \tag{3.20}$$

Due to the preemptive nature of the discipline, the first term of (3.19) is equal to (3.20), the waiting time in Case II, so we calculate the LST of the waiting time of Case II first. Conditioning on the service time and the number of arrivals therein yields: For $i \in I_{LCFS\text{-}PR}$,

$$\begin{aligned}
\mathbb{E}\left[e^{-sW_i}|T \text{ arrives during visit time}\right] &= \mathbb{E}\left[e^{-s(\sum_{T_{i,k}\in(0,B_i)}\xi_i)}\right] \\
&= \int_{t=0}^{\infty} \sum_{n=0}^{\infty} e^{-\lambda_i t}\frac{(\lambda_i t)^n}{n!}\,\mathbb{E}[e^{-s\xi_i}]^n \,\mathrm{d}\,\mathbb{P}(B_i \le t) \\
&= \int_{t=0}^{\infty} e^{-t(\lambda_i(1-\mathbb{E}[e^{-s\xi_i}]))} \,\mathrm{d}\,\mathbb{P}(B_i \le t) \\
&= B_i^*(\lambda_i(1-\mathbb{E}\left[e^{-s\xi_i}\right])).
\end{aligned}$$

The last two terms of (3.19) are equal to the waiting time of non-preemptive LCFS given in (3.11). We use the corresponding LST given in (3.13) to arrive at (3.21): For

$i \in I_{LCFS\text{-}PR}$, $Re(s) > 0$,

$$\mathbb{E}\left[e^{-sW_i} | T \text{ arrives during intervisit time}\right] = B_i^*(\lambda_i(1 - \mathbb{E}\left[e^{-s\xi_i}\right])) \frac{1 - \mathbb{E}[e^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}. \tag{3.21}$$

Combining the two cases leads to the following expression for the LST of the waiting time at Q_i in terms of the cycle time.

Proposition 3.2. *For $\rho < 1$, $i \in I_{LCFS\text{-}PR}$, $Re(s) > 0$,*

$$W_i^*(s) = B_i^*(\lambda_i(1 - \mathbb{E}[e^{-s\xi_i}])) \times \left(\rho_i + (1-\rho_i)\frac{1 - \mathbb{E}[e^{-sC_i}]}{(s + \lambda_i(1 - \mathbb{E}[e^{-s\xi_i}]))\,\mathbb{E}[C](1-\rho_i)}\right). \tag{3.22}$$

The next result gives the HT limit of the distribution of $\tilde{W}_i$.

Theorem 3.2. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \begin{cases} 0 & w.p. \quad \hat{\rho}_i \\ U_i \tilde{\mathbf{C}}_i & w.p. \quad 1 - \hat{\rho}_i \end{cases} \qquad (i \in I_{LCFS\text{-}PR}),$$

where U_i is a uniformly distributed random variable on the interval $[0,1]$ and $\tilde{\mathbf{C}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ, where α and μ are given in Equation (3.3).

Proof. Using (3.17) and the fact that for $Re(s) > 0$ it holds that $\lim_{\rho\uparrow 1} B_i^*(\lambda_i(1 - \mathbb{E}[e^{-s\xi_i}])) = 1$, we immediately see that the LST of $\tilde{W}_i$ in HT is given by

$$\tilde{W}_i^*(s) := \lim_{\rho\uparrow 1} W_i^*(s(1-\rho)) = \hat{\rho}_i + (1-\hat{\rho}_i)\frac{1}{\mathbb{E}[S]s}\left\{1 - \left(\frac{\mu}{\mu+s}\right)^{\alpha}\right\} \quad (i \in I_{LCFS\text{-}PR}), \tag{3.23}$$

with α and μ given in (3.3). □

Note that the HT scaled waiting-time distribution (3.23) for $i \in I_{LCFS\text{-}PR}$ is equal to the HT scaled waiting-time distribution (3.18) for $i \in I_{LCFS}$. This holds because the busy periods generated by customers arriving during service of the tagged customer do not scale with ρ.

3.5 Random Order of Service

In this section we first derive the LST of the scaled waiting-time distribution for ROS in terms of the intervisit times. Then we use this result to obtain the waiting-time distribution in heavy traffic.

Proposition 3.3. *For $\rho < 1$, $i \in I_{ROS}$, $Re(s) > 0$,*

$$W_i^*(s) =$$
$$\frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\left(\int_{x=\xi_i^*(s)}^1 \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x}\left(B_i^*(\lambda_i(1-x))-B_i^*(s+\lambda_i(1-x))\right)\,\mathrm{d}K(x,s)\right.$$
$$\left.+\int_{x=\xi_i^*(s)}^1 \left(I_i^*(\lambda_i(1-x))-I_i^*(s+\lambda_i(1-x))\right)\,\mathrm{d}K(x,s)\right),$$

with $\xi_i^*(s) = B_i^*(s+\lambda_i(1-\xi_i^*(s)))$, the LST of a busy period at queue i with a dedicated server, and

$$K(x,s) := \exp\left(-\int_{y=x}^1 \frac{1}{y-B_i^*(s+\lambda_i-\lambda_i y)}\,\mathrm{d}y\right). \tag{3.24}$$

Proof. The derivation proceeds along the lines of Kingman [95]. Define the waiting time of a tagged customer T as $w = u + v$. Here u is the time between the arrival instant of T and the time the server begins working on a new type i customer, and v is the time from that moment until T is taken into service. A customer may arrive during an intervisit period of Q_i, in which case $u = I_i^{res}$, or during a visit period, yielding $u = B_i^{res}$.

For v we first consider the transform of the number of customers at moments when the server is able to take a customer from queue i into service, denoted as $Q(z, X)$, with $X \in \{\mathbf{B}_i, \mathbf{I}_i\}$. From Kawasaki et al. [91] we have for an arrival during a visit period:

$$Q(z,\mathbf{B}_i) = \frac{(1-\rho_i)(1-I_i^*(\lambda_i-\lambda_i z))e^{-\lambda_i(1-z)\mathbf{B}_i}}{\lambda_i\,\mathbb{E}[I_i](B_i^*(\lambda_i-\lambda_i z)-z)}.$$

If the customer arrives during an intervisit period we have, for $|z| < 1$, $i \in I_{ROS}$,

$$Q(z,\mathbf{I}_i) = e^{-\lambda_i(1-z)\mathbf{I}_i}.$$

Kingman [95] (Theorem 2) provides the LST of v given the number of customers present. Combining this theorem with the equations above, we obtain the LST of v for an arrival during a visit period while a customer of size $\mathbf{B}_i$ is in service: For $Re(s) > 0$, $i \in I_{ROS}$,

$$\mathbb{E}[e^{-sv}|\mathbf{B}_i \text{ and arrival during visit period}] =$$
$$\int_{\xi_i^*(s)}^1 \frac{(1-\rho_i)(1-I_i^*(\lambda_i-\lambda_i x))e^{-\lambda_i(1-x)\mathbf{B}_i}}{\lambda_i\,\mathbb{E}[I_i](B_i^*(\lambda_i-\lambda_i x)-x)}\,\mathrm{d}K(x,s).$$

Similarly, we have for a customer arriving during an intervisit period of length $\mathbf{I}_i$: For $Re(s) > 0$, $i \in I_{ROS}$,

$$\mathbb{E}[e^{-sv}|\mathbf{I}_i \text{ and arrival during intervisit period}] = \int_{\xi_i^*(s)}^1 e^{-\lambda_i(1-x)\mathbf{I}_i}\,\mathrm{d}K(x,s).$$

Note that given $\mathbf{B}_i$ or $\mathbf{I}_i$, u and v are independent. For an arrival during a visit while a customer of size $\mathbf{B}_i$ is in service, we obtain: For $Re(s) > 0$, $i \in I_{ROS}$,

$$\begin{aligned}\mathbb{E}[e^{-sw}|\mathbf{B}_i] &= \mathbb{E}[e^{-sB_i^{res}}|\mathbf{B}_i]\,\mathbb{E}[e^{-sv}|\mathbf{B}_i] \\ &= \frac{1-e^{-s\mathbf{B}_i}}{s\mathbf{B}_i}\int_{\xi_i^*(s)}^1 \frac{(1-\rho_i)(1-I_i^*(\lambda_i-\lambda_i x))e^{-\lambda_i(1-x)\mathbf{B}_i}}{\lambda_i\,\mathbb{E}[I_i](B_i^*(\lambda_i-\lambda_i x)-x)}\,\mathrm{d}K(x,s) \\ &= \frac{1-\rho_i}{s\lambda_i\,\mathbb{E}[I_i]}\int_{\xi_i^*(s)}^1 \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x}\,\frac{e^{-\lambda_i(1-x)\mathbf{B}_i}-e^{-(s+\lambda_i(1-x))\mathbf{B}_i}}{\mathbf{B}_i}\,\mathrm{d}K(x,s).\end{aligned}$$

Now, using the fact that $\mathbb{E}\left[e^{-\phi\mathbf{B}_i}/\mathbf{B}_i\right] = \frac{B_i^*[\phi]}{\mathbb{E}[B_i]}$ (see [95]), we have for $Re(s) > 0$, $i \in I_{ROS}$,

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[e^{-sw}|\mathbf{B}_i]] = \\ &\quad\frac{1-\rho_i}{s\lambda_i\,\mathbb{E}[I_i]}\int_{\xi_i^*(s)}^1 \frac{1-I_i^*(\lambda_i-\lambda_i x)}{B_i^*(\lambda_i-\lambda_i x)-x}\,\frac{B_i^*(\lambda_i(1-x))-B_i^*(s+\lambda_i(1-x))}{\mathbb{E}[B_i]}\,\mathrm{d}K(x,s).\end{aligned}$$

Again it holds that a customer arrives with probability ρ_i during a visit period. Hence, $W_i^*(s) = \rho_i\,\mathbb{E}[\mathbb{E}[e^{-sw}|\mathbf{B}_i]] + (1-\rho_i)\,\mathbb{E}[\mathbb{E}[e^{-sw}|\mathbf{I}_i]]$. Using similar arguments for the final term in addition to some rewriting, we obtain the result. □

Next, we turn to the heavy-traffic limit. Before we state our result, we define Y as a random variable with pdf and cdf

$$f_Y(y) = \frac{(1-y)^{\frac{\hat\rho_i}{1-\hat\rho_i}}}{(1-\hat\rho_i)}, \qquad F_Y(y) = 1-(1-y)^{\frac{1}{1-\hat\rho_i}}, \qquad y \in [0,1].$$

The r.v. Y is to be interpreted as the fraction of customers, including both present customers and those arriving until the server's departure from the queue, that is served before the arriving customer, see Remarks 3.4 and 3.5.

The next theorem gives the HT limit of the distribution of $\tilde W_i$ in terms of Y.

Theorem 3.3. *For $\rho \uparrow 1$,*

$$\tilde W_i \to_d \begin{cases} U_i^f\tilde{\mathbf{C}} & w.p.\ \hat\rho_i \\ U_i^g\tilde{\mathbf{C}} & w.p.\ 1-\hat\rho_i \end{cases} \qquad (i \in I_{ROS}),$$

where U_i^f has a uniform distribution on the interval $[0, Y\hat\rho_i]$ and U_i^g has a uniform distribution on $[Y\hat\rho_i, 1-\hat\rho_i+Y\hat\rho_i]$.

Proof. First we rewrite the LST of the waiting time given in Proposition 3.3. Noting

that $\frac{\mathrm{d}K(x,s)}{\mathrm{d}x} = \frac{K(x,s)}{x-B_i^*(s+\lambda_i(1-x))}$, we get

$$W_i^*(s) = \frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\Bigg(\int_{x=\xi_i^*(s)}^1 K(x,s)(1-I_i^*(\lambda_i-\lambda_i x)) \\ \times \left(\frac{1}{B_i^*(\lambda_i(1-x))-x} + \frac{1}{x-B_i^*(s+\lambda_i(1-x))} \right) \mathrm{d}x \\ + \int_{x=\xi_i^*(s)}^1 K(x,s)\left(I_i^*(\lambda_i(1-x)) - I_i^*(s+\lambda_i(1-x))\right) \frac{1}{x-B_i^*(s+\lambda_i(1-x))}\,\mathrm{d}x \Bigg).$$

In line with Takagi and Kudoh [127] we take $y = \frac{1-x}{1-\xi_i^*(s)}$; this gives $x = 1-y(1-\xi_i^*(s))$ and $\mathrm{d}x = -(1-\xi_i^*(s))\,\mathrm{d}y$, yielding

$$W_i^*(s) = \frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\Bigg(\int_{y=0}^1 K\big(1-y(1-\xi_i^*(s)),s\big)\big(1-I_i^*(y\lambda_i(1-\xi_i^*(s)))\big) \\ \times \left(\frac{1-\xi_i^*(s)}{B_i^*(y\lambda_i(1-\xi_i^*(s)))-1+y(1-\xi_i^*(s))} \right. \\ \left. + \frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))} \right) \mathrm{d}y \\ + \int_{y=0}^1 K\big(1-y(1-\xi_i^*(s)),s\big)\big(I_i^*(y\lambda_i(1-\xi_i^*(s))) - I_i^*(s+y\lambda_i(1-\xi_i^*(s)))\big) \\ \times \left(\frac{1-\xi_i^*(s)}{1-y(1-\xi_i^*(s))-B_i^*(s+y\lambda_i(1-\xi_i^*(s)))} \right) \mathrm{d}y \Bigg).$$

We now take heavy-traffic limits for the terms separately. We start with the most involved term, $K(x,s)$. Using the substitution $t = \frac{1-y}{1-x}$ in (3.24), we may write

$$K(x,s) = \exp\left(-\int_{t=0}^1 \frac{1-x}{1-t(1-x)-B_i^*(s+\lambda_i t(1-x))}\,\mathrm{d}t \right).$$

Taking the HT limit of $K(1-y(1-\xi_i^*(s)),s)$ we obtain, using l'Hôpital's rule and

some rewriting,

$$\begin{aligned}
&\lim_{\rho\uparrow 1} K(1-y(1-\xi_i^*(s(1-\rho))), s(1-\rho)) = \\
&\quad \exp\left(-\int_{t=0}^{1} \frac{y\,\mathbb{E}[\xi_i]}{-\mathbb{E}[\xi_i]ty + \mathbb{E}[B_i](1+\lambda_i ty\,\mathbb{E}[\xi_i])}\,\mathrm{d}t\right) \\
&\quad = \exp\left(-\frac{y}{1-\hat{\rho}_i}\int_{t=0}^{1}\frac{1}{1-ty}\,\mathrm{d}t\right) \\
&\quad = \exp\left(\frac{1}{1-\hat{\rho}_i}\ln(1-y)\right) \\
&\quad = (1-y)^{\frac{1}{1-\hat{\rho}_i}}.
\end{aligned}$$

In the second step we use the fact that $\mathbb{E}[\xi_i] = \frac{\mathbb{E}[B_i]}{1-\hat{\rho}_i}$. The HT limits for the other terms can be determined using l'Hôpital's rule in addition to some rewriting and the expression for $\mathbb{E}[\xi_i]$ above. In particular, we get

$$\begin{aligned}
&\lim_{\rho\uparrow 1} I_i^*(y\lambda_i(1-\xi_i^*(s(1-\rho)))) = \tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right), \\
&\lim_{\rho\uparrow 1} I_i^*(s(1-\rho)+y\lambda_i(1-\xi_i^*(s(1-\rho)))) = \tilde{I}_i^*\left(\frac{s(1-\hat{\rho}_i+y\hat{\rho}_i)}{1-\hat{\rho}_i}\right), \\
&\lim_{\rho\uparrow 1} \frac{1-\xi_i^*(s(1-\rho))}{B_i^*(y\lambda_i(1-\xi_i^*(s(1-\rho)))) - 1 + y(1-\xi_i^*(s(1-\rho)))} = \frac{1}{y(1-\hat{\rho}_i)}, \\
&\lim_{\rho\uparrow 1} \frac{1-\xi_i^*(s(1-\rho))}{1-y(1-\xi_i^*(s(1-\rho))) - B_i^*(s(1-\rho)+y\lambda_i(1-\xi_i^*(s(1-\rho))))} = \frac{1}{(1-y)(1-\hat{\rho}_i)}.
\end{aligned}$$

Moreover, we have $\tilde{I}_i^*\left(\frac{cs}{1-\hat{\rho}_i}\right) = \tilde{C}_i^*(cs) = \left(\frac{\mu}{\mu+cs}\right)^{\alpha}$ for fixed $c > 0$. Combining the above gives, after some rewriting,

$$\begin{aligned}
\tilde{W}_i^*(s) &= \frac{1-\hat{\rho}_i}{s\,\mathbb{E}[S](1-\hat{\rho}_i)}\left(\int_{y=0}^{1}\left(1-\tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right)\right)\frac{(1-y)^{\frac{1}{1-\hat{\rho}_i}}}{y(1-y)(1-\hat{\rho}_i)}\,\mathrm{d}y\right. \\
&\quad \left.+\int_{y=0}^{1}\left(\tilde{I}_i^*\left(\frac{y\hat{\rho}_i s}{1-\hat{\rho}_i}\right)-\tilde{I}_i^*\left(\frac{s(1-\hat{\rho}_i+y\hat{\rho}_i)}{1-\hat{\rho}_i}\right)\right)\frac{(1-y)^{\frac{1}{1-\hat{\rho}_i}}}{(1-y)(1-\hat{\rho}_i)}\,\mathrm{d}y\right) \\
&= \hat{\rho}_i\int_{y=0}^{1}\frac{1}{s\,\mathbb{E}[S]y\hat{\rho}_i}\left\{1-\left(\frac{\mu}{\mu+y\hat{\rho}_i s}\right)^{\alpha}\right\}\frac{(1-y)^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}}{(1-\hat{\rho}_i)}\,\mathrm{d}y \\
&\quad +(1-\hat{\rho}_i)\int_{y=0}^{1}\frac{1}{s\,\mathbb{E}[S](1-\hat{\rho}_i)}\left\{\left(\frac{\mu}{\mu+y\hat{\rho}_i s}\right)^{\alpha}\right. \\
&\quad \left.-\left(\frac{\mu}{\mu+s(1-\hat{\rho}_i+y\hat{\rho}_i)}\right)^{\alpha}\right\}\frac{(1-y)^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}}{(1-\hat{\rho}_i)}\,\mathrm{d}y.
\end{aligned}$$

This LST corresponds to a mixture of two distributions. With probability $\hat{\rho}_i$ and conditioning on $Y = y$, it is the LST of a uniform $[0, y\hat{\rho}_i]$ times a gamma distribution with parameters $\alpha + 1$ and μ; with probability $1 - \hat{\rho}_i$ and conditioning on $Y = y$, it is the LST of a uniform $[y\hat{\rho}_i, 1 - \hat{\rho}_i + y\hat{\rho}_i]$ times a gamma distribution with the same parameters. This completes the proof. □

Remark 3.3. The expressions for U_i^f and U_i^g in Theorem 3.3 can be rewritten more explicitly, similar to those in Theorem 3.5, see also Remark 3.8.

Remark 3.4 (HTAP)**.** The HT limit states that conditional on $Y = y$, the scaled waiting-time distribution is a uniform times a gamma distribution with probability $\hat{\rho}_i$ and another uniform times a gamma distribution with probability $1 - \hat{\rho}_i$. Here, y is a tag representing the fraction of work from the work present and arriving until the server's departure from the queue that is served before the tagged customer in a fluid model. See Remark 3.5 below for a more intuitive derivation of the tag-distribution $F_Y(\cdot)$.
With probability $1 - \hat{\rho}_i$ a particle arrives during an intervisit time of length $c(1 - \hat{\rho}_i)$. If U_I is the fraction of the intervisit time that has elapsed at the arrival epoch of a tagged particle, it first has to wait $(1-U_I)c(1-\hat{\rho}_i)$ until Q_i is visited. The total work present upon arrival plus the amount of work arriving until the server's departure from Q_i equals the total workload arriving during a cycle and is $\hat{\rho}_i c$. Given the tag $Y = y$, the total scaled waiting time equals $((1-U_I)(1-\hat{\rho}_i)+y\hat{\rho}_i)c$, corresponding to a uniform distribution on $[y\hat{\rho}_i, 1-\hat{\rho}_i+y\hat{\rho}_i]$. With probability $\hat{\rho}_i$ a particle arrives during a visit time of length $\hat{\rho}_i c$. If U_V is the fraction of the intervisit time that remains, the amount work present upon arrival in addition to the remaining amount of work arriving equals $U_V\hat{\rho}_i c$. Given a tag $Y = y$, the scaled waiting time is $yU_V\hat{\rho}_i c$, which is a uniform distribution on $[0, y\hat{\rho}_i]$ times c. Theorem 3.3 thus follows intuitively from HTAP.

Remark 3.5 (Intuition for tag-distribution Y)**.** We provide an intuitive explanation for the distribution of Y using a fluid model for the number of customers or particles. Assume the tagged customer arrives during a visit time, say at time 0, finding x particles present. The queue length is decreasing at rate $1 - \hat{\rho}_i$, i.e. at time t the queue length $L_i(t) = x-(1-\hat{\rho}_i)t$, until the queue is empty at time $x/(1-\hat{\rho}_i)$. Observe that with $L_i(t)$ particles present, the probability for service selection is $1/L_i(t)$. Let $\bar{F}(t)$ be the probability that the tagged customer has not been taken into service at time t. Since there are continuously options for service selection in the fluid model, $\bar{F}(t)$ satisfies the following first-order differential equation (DE), for $0 < t < x/(1-\hat{\rho}_i)$,

$$-\frac{\mathrm{d}}{\mathrm{d}t}\bar{F}(t) = \bar{F}(t) \times \frac{1}{L_i(t)}.$$

Solving the above DE with boundary condition $\bar{F}(0) = 1$ and using the fluid version

of $L_i(t)$, we have, for $0 < t < x/(1-\hat{\rho}_i)$,

$$\bar{F}(t) = \exp\left(\int \frac{1}{x-(1-\hat{\rho}_i)t}\,\mathrm{d}t\right) = \left(1 - t\frac{1-\hat{\rho}_i}{x}\right)^{\frac{1}{1-\hat{\rho}_i}}.$$

Finally, the queue being empty at time $x/(1-\hat{\rho}_i)$ implies that also $x/(1-\hat{\rho}_i)$ particles have been served since time 0. When at least a fraction y of those has been served before the tagged customer is taken into service, then we look for

$$\bar{F}\left(y \times \frac{x}{1-\hat{\rho}_i}\right) = (1-y)^{\frac{1}{1-\hat{\rho}_i}}.$$

This coincides with one minus the cdf of Y.

3.6 Processor Sharing

In a Processor Sharing (PS) queue, all customers present at the queue that is receiving service are served simultaneously and at the same rate. We note that the waiting time W_i (to be interpreted as the delay) is thus defined as the sojourn time minus the service requirement. In this section we will only consider the case of exponentially distributed service time. We extend the work done in [12], where they derive the heavy-traffic limit of the LST of the scaled waiting time conditional on the service requirement. In Subsection 3.6.1, we give the conditional scaled waiting-time distribution. In Subsection 3.6.2 we derive the unconditional scaled waiting-time distribution.

3.6.1 Conditional waiting-time distribution in heavy traffic

Let customers in Q_i have exponentially distributed service requirements with rate b_i. Let x be the required service duration of a tagged customer. Then we have the following theorem for the heavy-traffic limit of the conditional waiting time $W_i|x$:

Theorem 3.4. *For $\rho \uparrow 1$, $x \geq 0$,*

$$\tilde{W}_i|x \to_d \begin{cases} U^f_{i,x}\tilde{\mathbf{I}}_i & w.p.\ \hat{\rho}_i \\ U^g_{i,x}\tilde{\mathbf{I}}_i & w.p.\ 1-\hat{\rho}_i \end{cases} \qquad (i \in I_{PS}),$$

where $U^f_{i,x} = U[0,\omega(x)]$, $U^g_{i,x} = U[\omega(x),\omega(x)+1]$ and $\tilde{\mathbf{I}}_i \sim \Gamma(\alpha+1,\mu_i)$. The parameters α and μ_i can be found in Equation (3.7), and $\omega(x) = \frac{\hat{\rho}_i}{1-\hat{\rho}_i}(1-e^{-b_i x(1-\hat{\rho}_i)})$.

Proof. The authors of [12] derive the LST of the scaled conditional waiting time in heavy traffic: For $\rho \uparrow 1$, $x \geq 0$, $i \in I_{PS}$,

$$\tilde{W}_i^*(s|x) = \frac{\hat{\rho}_i}{s\omega(x)\,\mathbb{E}[S](1-\hat{\rho}_i)} \left\{ 1 - \left(\frac{\mu_i}{\mu_i + s\omega(x)} \right)^{\alpha} \right\} + \frac{1-\hat{\rho}_i}{s\,\mathbb{E}[S](1-\hat{\rho}_i)} \left\{ \left(\frac{\mu_i}{\mu_i + s\omega(x)} \right)^{\alpha} - \left(\frac{\mu_i}{\mu_i + s(\omega(x)+1)} \right)^{\alpha} \right\}. \quad (3.25)$$

From this LST we see that the distribution of the conditional waiting time is a uniform $[0, \omega(x)]$ times a gamma distribution with parameters $\alpha + 1$ and μ_i with probability $\hat{\rho}_i$. With probability $1 - \hat{\rho}_i$, the conditional waiting time has a uniform $[\omega(x), \omega(x) + 1]$ times a gamma distribution with parameters $\alpha + 1$ and μ_i. This completes the proof. □

Remark 3.6 (HTAP)**.** Theorem 3.4 states that the conditional waiting-time distribution is a uniform times a gamma distribution with probability $\hat{\rho}_i$ and another uniform times a gamma distribution with probability $1 - \hat{\rho}_i$. This can be intuitively explained with a fluid model. In the fluid model $\omega(x)c(1 - \hat{\rho}_i)$ is the scaled waiting time of a particle, with service requirement x, arriving at the start of a visit period. With probability $1-\hat{\rho}_i$ a particle arrives during an intervisit period of length $c(1-\hat{\rho}_i)$. If U_I is the fraction of the intervisit time that has elapsed at the arrival epoch of a tagged particle, then the scaled waiting time of this particle is the remaining intervisit time $(1 - U_I)c(1 - \hat{\rho}_i)$ plus $\omega(x)c(1 - \hat{\rho}_i)$. Using the HTAP gives a uniform distribution on $[\omega(x), \omega(x) + 1]$ times a gamma distribution with parameters $\alpha + 1$ and μ_i. A particle arriving during a visit period has to wait an amount of time that is uniformly distributed between 0 (arrive at the end of the visit time) and $\omega(x)c(1 - \hat{\rho}_i)$ (arrive at the start of the visit time). Using the HTAP now gives a uniform distribution on $[0, \omega(x)]$ times a gamma distribution with parameters $\alpha + 1$ and μ_i.

Remark 3.7 (Intuition for $\omega(x)$)**.** The sojourn time of a tagged customer with service time x from the start of the visit time ($\omega(x)I_i$) can be intuitively explained with a fluid model. As long as the tagged customer is present, the amount of service received during $(0, t)$ is $B(t) = \int_0^t 1/L(u)\ \mathrm{d}u$ with $L(u)$ the number of customers at time u. During the visit time, we have in a fluid model $L(t) = L(0) - (1 - \hat{\rho}_i)b_i t$. Hence,

$$B(t) = \int_{u=0}^{t} \frac{1}{L(0) - (1-\hat{\rho}_i)b_i u}\ \mathrm{d}u = -\frac{1}{(1-\hat{\rho}_i)b_i} \left(\ln(L(0) - (1-\hat{\rho}_i)b_i t) - \ln L(0) \right).$$

To obtain the time until service completion, we solve $B(t) = x$ for t. Moreover, using that $L(0) = \hat{\lambda}_i c(1 - \hat{\rho}_i)$ in the fluid model, yields

$$\omega(x) \times I_i = \frac{\hat{\lambda}_i}{(1-\hat{\rho}_i)b_i} \left(1 - e^{-x(1-\hat{\rho}_i)b_i} \right) \times c(1 - \hat{\rho}_i).$$

The result follows from $\hat{\rho}_i = \hat{\lambda}_i / b_i$.

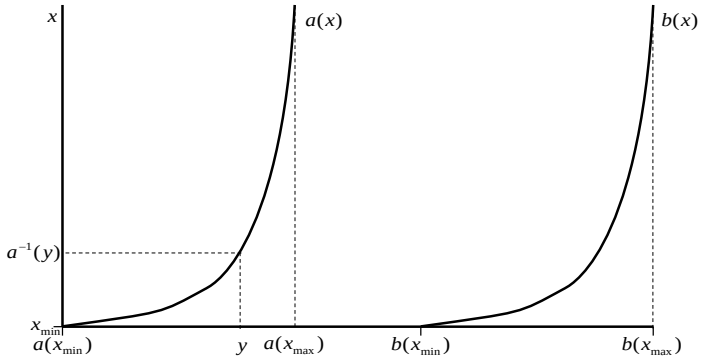


Figure 3.2: Boundaries of the conditional distribution.

3.6.2 Unconditional waiting-time distribution in heavy traffic

In the previous section we derived the heavy-traffic limit of the waiting-time distribution conditional on the service requirement. To obtain the *unconditional* waiting-time distribution, we first consider a more general setting that also covers 'unconditioning' for SJF. Suppose we have a conditional random variable, denoted $T|x$, with pdf $f_{T|x}(y)$, cdf $F_{T|x}(y)$, and $y \in [a(x), b(x)]$, with $a(x) < b(x) \quad \forall x$. We want to find the unconditional distribution $\tilde{T}$. Here, x is a realization of a random variable X with support $x \in [x_{min}, x_{max}]$. We have the following lemma.

Lemma 3.1. *Assume that the conditional random variable $T|x$ has density $f_{T|x}(y)$ and distribution function $F_{T|x}(y)$, with support $y \in [a(x), b(x)]$. Suppose $a(x)$ and $b(x)$ are both increasing in x and $a(x) < b(x) \quad \forall x$. Let $a^{-1}(\cdot)$ be the inverse of $a(\cdot)$ and $b^{-1}(\cdot)$ be the inverse of $b(\cdot)$. Then, the unconditional distribution of $T|x$, denoted by $\tilde{T}$, has probability density function, for $a(x_{max}) \leq b(x_{min})$,*

$$f_{\tilde{T}}(y) = \begin{cases} \int_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [a(x_{min}), a(x_{max})] \\ \int_{x=x_{min}}^{x_{max}} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [a(x_{max}), b(x_{min})] \\ \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [b(x_{min}), b(x_{max})], \end{cases} \tag{3.26}$$

and, for $a(x_{max}) > b(x_{min})$,

$$f_{\tilde{T}}(y) = \begin{cases} \int_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [a(x_{min}), b(x_{min})] \\ \int_{x=b^{-1}(y)}^{a^{-1}(y)} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [b(x_{min}), a(x_{max})] \\ \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x)\, \mathrm{d}x & y \in [a(x_{max}), b(x_{max})]. \end{cases} \tag{3.27}$$

Proof. First consider the case that $a(x_{max}) \leq b(x_{min})$. Figure 3.2 shows an example of the boundaries of the conditional distribution, by plotting $a(x)$ and $b(x)$ with x on the vertical axis. The possible values of $T|x$ then lie between the two lines. To find $f_{\tilde{T}}(y)$, we need to integrate out x with respect to its density function. First, take $y \in [a(x_{min}), a(x_{max})]$, in which case the probability density function $f_{\tilde{T}}(y)$ is obtained from the parts where x is smaller than $a^{-1}(y)$. This gives

$$f_{\tilde{T}}(y) = \int_{x=x_{min}}^{a^{-1}(y)} f_{T|x}(y) f_X(x) \, \mathrm{d}x. \tag{3.28}$$

If $y \in [a(x_{max}), b(x_{min})]$ then y is between the boundaries of the conditional distribution for every $x \in [x_{min}, x_{max}]$. Hence, we get

$$f_{\tilde{T}}(y) = \int_{x=x_{min}}^{x_{max}} f_{T|x}(y) f_X(x) \, \mathrm{d}x. \tag{3.29}$$

Finally, for $y \in [b(x_{min}), b(x_{max})]$, $f_{\tilde{T}}(y)$ can now be obtained from the parts where x is larger than $b^{-1}(y)$. This gives

$$f_{\tilde{T}}(y) = \int_{x=b^{-1}(y)}^{x_{max}} f_{T|x}(y) f_X(x) \, \mathrm{d}x. \tag{3.30}$$

The case $a(x_{max}) > b(x_{min})$ is similar. It may be checked $f_{\tilde{T}}(\cdot)$ is a density function. This completes the proof. □

Note that the distribution in Equation (3.26) is continuous, increasing on the interval $[a(x_{min}), a(x_{max})]$, constant on the interval $[a(x_{max}), b(x_{min})]$ and decreasing on the interval $[b(x_{min}), b(x_{max})]$, which closely resembles the traditional trapezoidal distribution. In line with [61], we refer to (3.26) as a *generalized trapezoidal distribution.*

We now apply Lemma 3.1 to the case $i \in I_{PS}$, in which case we have two conditional distributions, $U_{i,x}^f$ and $U_{i,x}^g$. We need to find the unconditional versions of both uniform distributions.

Theorem 3.5. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \begin{cases} \tilde{U}_i^f \tilde{\mathbf{I}}_i & w.p.\ \hat{\rho}_i \\ \tilde{U}_i^g \tilde{\mathbf{I}}_i & w.p.\ 1-\hat{\rho}_i \end{cases} \qquad (i \in I_{PS}),$$

where $\tilde{U}_i^f$ has a generalized trapezoidal distribution with pdf

$$f_{\tilde{U}_i}(y) = \frac{1}{\hat{\rho}_i} Beta_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i}\left(1 + \frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right) \quad y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i)], \tag{3.31}$$

where $Beta_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$. $\tilde{U}_i^g$ *has a generalized trapezoidal distribution with pdf, for* $\hat{\rho}_i \leq \frac{1}{2}$,

$$g_{\tilde{U}_i}(y) = \begin{cases} 1 - \left(1 - \frac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in \left[0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right) \\ 1 & y \in \left[\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 1\right] \\ \left(1 - \frac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in \left(1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i} + 1\right], \end{cases} \tag{3.32}$$

and, for $\hat{\rho}_i > \frac{1}{2}$,

$$g_{\tilde{U}_i}(y) = \begin{cases} 1 - \left(1 - \frac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in [0,1) \\ \left(1 - \frac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} - \left(1 - \frac{y(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in \left[1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right] \\ \left(1 - \frac{(y-1)(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}} & y \in \left(\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, \frac{\hat{\rho}_i}{1-\hat{\rho}_i} + 1\right], \end{cases}$$

and $\tilde{\mathbf{I}}_i$ *has a gamma distribution with parameters* $\alpha + 1$ *and* μ_i. *The parameters* α *and* μ_i *can be found in Equation* (3.7).

Proof. Let $f_{U_{i,x}}(\cdot)$ and $g_{U_{i,x}}(\cdot)$ be the densities of $U_{i,x}^f$ and $U_{i,x}^g$, respectively. First consider $f_{U_{i,x}}(y) = \frac{1}{\omega(x)}$ for $y \in [0, \omega(x)]$; thus $a(x) = 0$ and $b(x) = \omega(x)$. Here, x is the service requirement, a realization of an exponential distribution, so $x \in [0, \infty)$. Since $\omega(0) = 0$ and $\omega(\infty) = \hat{\rho}_i/(1-\hat{\rho}_i)$ we only have to find the final term of (3.26) and consider the interval $[0, \hat{\rho}_i/(1-\hat{\rho}_i)]$. For a fixed y, the inverse function of ω is $\omega^{-1}(y) = \ln(1 - y(1-\hat{\rho}_i)/\hat{\rho}_i)/(-b_i(1-\hat{\rho}_i))$. By Lemma 3.1, this gives

$$\begin{aligned} f_{\tilde{U}_i}(y) &= \int_{x=\omega^{-1}(y)}^{\infty} f_{B_i}(x) f_{U_{i,x}}(y)\,\mathrm{d}x \\ &= \int_{x=\frac{\ln(1-y(1-\hat{\rho}_i)/\hat{\rho}_i)}{-b_i(1-\hat{\rho}_i)}}^{\infty} b_i e^{-b_i x} \frac{1-\hat{\rho}_i}{\hat{\rho}_i} \left(1 - e^{-b_i x(1-\hat{\rho}_i)}\right)^{-1}\,\mathrm{d}x \\ &= \int_{t=1-y(1-\hat{\rho}_i)/\hat{\rho}_i}^{0} b_i \frac{1-\hat{\rho}_i}{\hat{\rho}_i}(1-t)^{-1} \frac{1}{-b_i(1-\hat{\rho}_i)} t^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}\,\mathrm{d}t \\ &= \int_{t=0}^{1-y(1-\hat{\rho}_i)/\hat{\rho}_i} \frac{1}{\hat{\rho}_i}(1-t)^{-1} t^{\frac{\hat{\rho}_i}{1-\hat{\rho}_i}}\,\mathrm{d}t \\ &= \frac{1}{\hat{\rho}_i} \mathrm{Beta}_{1-y(1-\hat{\rho}_i)/\hat{\rho}_i}\left(1 + \frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0\right). \end{aligned}$$

The third equality is obtained by taking $t = e^{-b_i x(1-\hat{\rho}_i)}$. This leads to an incomplete Beta function.

Now we turn to the second term involving $U_{i,x}^g$. Note that $g_{U_{i,x}}(y) = 1$ for $y \in [\omega(x), \omega(x)+1]$. To apply Lemma 3.1, observe that for $\hat{\rho}_i/(1-\hat{\rho}_i) \leq 1$ it holds that

$a(x_{max}) \leq b(x_{min})$. First assume that $\hat{\rho}_i/(1-\hat{\rho}_i) \leq 1$, implying $\hat{\rho}_i < 1/2$. For a fixed $y \in [0, \hat{\rho}_i/(1-\hat{\rho}_i))$, x needs to be smaller than $\omega^{-1}(y)$, if $y \in [\hat{\rho}_i/(1-\hat{\rho}_i), 1]$, it lies between the boundaries of the uniform distribution for all x and if $y \in (1, \hat{\rho}_i/(1-\hat{\rho}_i)+1]$, then x needs to be larger than $\omega^{-1}(y)$. This gives for the pdf of $\tilde{U}_i^g$

$$g_{\tilde{U}_i}(y) = \begin{cases} F_{B_i}(\omega^{-1}(y)) & y \in \left[0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right) \\ 1 & y \in \left[\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 1\right] \\ 1 - F_{B_i}(\omega^{-1}(y-1)) & y \in \left(1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}+1\right]. \end{cases}$$

Substituting $F_{B_i}(x) = 1 - e^{-b_i x}$ and the inverse of $\omega(\cdot)$ gives Equation (3.32). The case $\hat{\rho}_i > 1/2$ implies $a(x_{max}) > b(x_{min})$ and is similar, completing the proof. □

Remark 3.8 (PS and ROS)**.** For regular GI/M/1 queues, the relation between PS and ROS has been characterized by Borst et al. [34]. It is easily seen that the sample path relations (see Equation (3) of [34]) also hold for the polling models under consideration. More specifically, consider a tagged customer T_i arriving at Q_i when the server visits Q_i. Then, the sojourn-time distribution of T_i for PS, given n_i customers at Q_i upon arrival, is identical to the waiting-time distribution of T_i for ROS, given n_i waiting customers at Q_i upon arrival in addition to the one in service. Under HT scalings, the differences between waiting and sojourn times and the one customer vanish, explaining the equivalence between Theorems 3.3 and 3.5 (see Remark 3.3).

3.7 n-class priority queues

In this section we look at n-class priority queues. Each customer is assigned to a priority index k, $1 \leq k \leq n$, where customers with a low priority index are served before customers with higher priority indices. Within each class the service order is FCFS. In Subsection 3.7.1, the focus lies on the non-preemptive n-class priority regime. We will later use this discipline to find the waiting-time distribution in SJF queues, by letting the number of priority classes go to infinity. In [93], Kella and Yechiali study the M/G/1 queue with single and multiple server vacations under both the preemptive and non-preemptive priority regimes. The M/G/1 queue with multiple vacations is similar to a polling model, since we express the waiting times in cycle times and we can replace vacations by intervisit times. This relation has also been used in [29] to analyze multi-class polling models. We also study the preemptive n-class priority regime in Subsection 3.7.2.

3.7.1 Non-preemptive n-class priority queues

Here, we are interested in the non-preemptive n-class priority regime. We now introduce our notation and terminology based on [93], as this turns out to be useful

and provide intuition for this and the next section. We replace vacation times with intervisit times and add the subscript i to every queue-dependent variable: $\lambda_{i,k}$ is the arrival rate of class-k customers and $B_{i,k}$ is the service duration of class-k customers. Class-a customers are the customers with priority index lower than k, i.e., they are served before class-k customers. They have arrival rate $\lambda_{i,a} = \sum_{j=1}^{k-1} \lambda_{i,j}$ and service duration $B_{i,a}$. Class-b customers are customers with priority index higher than k, their arrival rate is $\lambda_{i,b} = \sum_{j=k+1}^{n} \lambda_{i,j}$ and their service duration is $B_{i,b}$. We have $\rho_{i,a} = \lambda_{i,a}\,\mathbb{E}[B_{i,a}]$ and $\rho_{i,b} = \lambda_{i,b}\,\mathbb{E}[B_{i,b}]$. $\xi_{i,a}$ denotes the length of time from a moment a class-a customer enters service and no other class-a customers are present, until the first moment when there are no class-a customers in the queue. Clearly $\xi_{i,a}$ is the duration of a busy period in a standard M/G/1 queue with arrival rate $\lambda_{i,a}$ and service times $B_{i,a}$. Consequently, the LST of $\xi_{i,a}$ and its mean are given by: For $Re(s) > 0$,

$$\xi^*_{i,a}(s) = B^*_{i,a}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)), \quad \mathbb{E}[\xi_{i,a}] = \mathbb{E}[B_{i,a}]/(1 - \rho_{i,a}). \tag{3.33}$$

For this model, Kella and Yechiali [93] derive the following LST for the waiting-time distribution $W_{i,k}$ of a class-k customer in Q_i: For $Re(s) > 0$, $k = 1, \ldots, n$,

$$\begin{aligned} W^*_{i,k}(s) = {} & \frac{(1-\rho_i)(1 - I^*_i(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \\ & + \frac{\rho_{i,b}(1 - B^*_{i,b}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[B_{i,b}](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \quad (i \in I_{NPRIOR}). \end{aligned} \tag{3.34}$$

The first term of (3.34) corresponds to the waiting time of class-k customers in Q_i that arrive during the time from the start of the intervisit time until the moment a class-b customer at Q_i is taken into service. The second term corresponds to the waiting time of class-k customers that arrive during the time from the moment the first class-b customer is taken into service until the end of the cycle. Note that this expression was also derived in [29]. The following theorem gives the heavy-traffic limit of the distribution of $W_{i,k}$.

Theorem 3.6. *For $\rho \uparrow 1$, $k = 1, \ldots, n$,*

$$\tilde{W}_{i,k} \to_d \begin{cases} 0 & \text{w.p. } \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \\ U_i\tilde{\mathbf{I}}_i & \text{w.p. } \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \end{cases} \quad (i \in I_{NPRIOR}),$$

where U_i is a uniformly distributed random variable that lies between 0 and $\frac{1}{1-\hat{\rho}_{i,a}}$ and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i. The parameters α and μ_i are given in (3.7).

Proof. Combining Equation (3.34) and Property 3.2, we get for the LST of the (scaled)

waiting time of a class-k customer: for $Re(s) > 0$, $k = 1, \ldots, n$, $i \in I_{NPRIOR}$:

$$\begin{aligned}
\tilde{W}^*_{i,k}(s) &= \lim_{\rho\uparrow 1} W^*_{i,k}(s(1-\rho)) \\
&= \lim_{\rho\uparrow 1} \Bigg[\frac{(1-\rho_i)\left(1 - \left(\frac{\mu_i}{\mu_i + s + \lambda_{i,a}(1-\xi^*_{i,a}(s(1-\rho)))/(1-\rho)}\right)^{\alpha}\right)}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s(1-\rho)+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s(1-\rho))) - \lambda_{i,k} + s(1-\rho))} \\
&\qquad + \frac{\rho_{i,b}(1 - B^*_{i,b}(s(1-\rho)+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s(1-\rho))))}{\mathbb{E}[B_{i,b}](\lambda_{i,k}B^*_{i,k}(s(1-\rho)+\lambda_{i,a}-\lambda_{i,a}\xi^*_{i,a}(s(1-\rho))) - \lambda_{i,k} + s(1-\rho))} \Bigg] \\
&= \frac{(1-\hat{\rho}_i)\left(1 - \left(\frac{\mu_i}{\mu_i + s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}\right)^{\alpha}\right)}{\mathbb{E}[S](1-\hat{\rho}_i)s(1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}]))} + \frac{\hat{\rho}_{i,b}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}{1-\hat{\rho}_{i,k}(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])} \\
&= \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \frac{1}{\mathbb{E}[S]s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])(1-\hat{\rho}_i)} \left\{1 - \left(\frac{\mu_i}{\mu_i + s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}\right)^{\alpha}\right\} \\
&\qquad + \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \\
&= \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \frac{1}{\mathbb{E}[S]s(1-\hat{\rho}_i)/(1-\hat{\rho}_{i,a})} \left\{1 - \left(\frac{\mu_i}{\mu_i + s/(1-\hat{\rho}_{i,a})}\right)^{\alpha}\right\} \\
&\qquad + \frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}. \qquad (3.35)
\end{aligned}$$

The third equality was found using l'Hôpital's rule and some basic calculations. After some rewriting we arrive at the fourth equation, and writing out $\mathbb{E}[\xi_{i,a}]$ using (3.33) leads to the final equation. Recognizing this as the LST of a random variable that is equal to zero with probability $\frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ and a uniform times a gamma distribution with probability $\frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ completes the proof. □

Remark 3.9 (HTAP)**.** We can use the fluid model to give some intuition for the asymptotic waiting-time distribution, which corresponds to a uniform times a gamma distribution in addition to a probability mass at zero. In the fluid model, we only consider class a and class k particles, as the impact of class b is negligible in HT. Figure 3.3 gives a graphical representation of the fluid model; the workload of class a and k particles in Q_i is plotted over the course of a cycle of length c. The considered particles arrive at the queue with rate $\hat{\rho}_{i,a} + \hat{\rho}_{i,k}$ and during a visit time they are served with rate 1 until the queue is empty. The cycle is divided in three parts: the first part is the intervisit time I_i with length $(1-\hat{\rho}_i)c$. The second part is the duration between a polling instant and the first time since the start of the cycle for which no class a and k particles are present. This part has length $\frac{(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})(1-\hat{\rho}_i)c}{1-(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})}$. In this part only class a and k particles are served. The last part is the part where class b particles are served, interrupted by classes a and k, having length

$$c - (1-\hat{\rho}_i)c - \frac{(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})(1-\hat{\rho}_i)c}{1-(\hat{\rho}_{i,a}+\hat{\rho}_{i,k})} = \frac{\hat{\rho}_{i,b}c}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}.$$

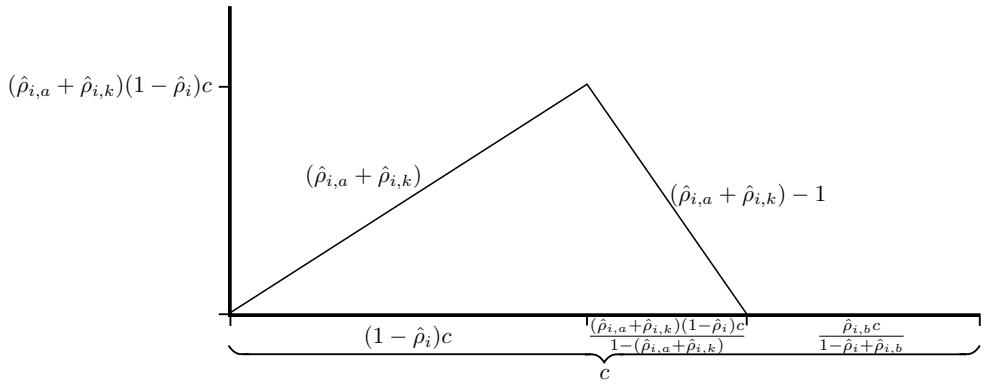


Figure 3.3: Fluid limits in heavy traffic. The workload of class a and k particles in Q_i is plotted over the course of a cycle.

Now, first consider the atom in zero. With probability $\frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ a class-k particle arrives during the last part of the cycle where hardly any class a or k particles are present. In this case the scaled waiting time of the particle is negligible in HT, since the residual service time of the particle in service and the busy periods generated by class-a customers arriving during this remaining service time do not scale with ρ. Second, with probability $\frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ a particle arrives during the first or second part of the cycle. Let the uniform random variable U_i denote the fraction of the length of the first two parts of the cycle together that has elapsed at the arrival epoch of the tagged arriving particle. Similar to FCFS, the scaled waiting time of this particle is the remaining duration $(1-U_i)\frac{(1-\hat{\rho}_i)c}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ minus the time required to serve the class-k work (or extended service time) that arrives during the first two parts of the cycle, but after the tagged particle. Due to class-a interruptions, the extended service time of class k is $\mathbb{E}[B_{i,k}]/(1-\hat{\rho}_{i,a})$. Hence, the scaled waiting time is $(1-U_i)\frac{(1-\hat{\rho}_i)c}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}(1-\frac{\hat{\rho}_{i,k}}{1-\hat{\rho}_{i,a}}) = (1-U_i)\frac{(1-\hat{\rho}_i)c}{1-\hat{\rho}_{i,a}}$. Since $I_i=(1-\hat{\rho}_i)c$, this term corresponds to a uniform distribution on $[0,\frac{1}{1-\hat{\rho}_{i,a}}]I_i$, explaining the result for non-negligible waiting times.

3.7.2 Preemptive n-class priority queues

Similar to the previous section, the results of [93] also allow the derivation of the LST of the time until service in a polling system where different priority classes are served with preemptive priority. Let $W_i^{(q)}$ denote the time until a customer first receives service, or the waiting time in queue. We observe that this is not equal to the waiting time as defined in the current chapter (i.e. sojourn time minus service time) due to service preemptions. For class k, the LST of the time from the start until the end of

service $R_{i,k}$, often referred to as the *residence time*, is

$$R^*_{i,k} = B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)). \tag{3.36}$$

For a class-k customer in Q_i the LST of waiting time in queue is: For $Re(s) > 0$, $k = 1, \ldots, n$ and $i \in I_{NPRIOR\text{-}PR}$,

$$\begin{aligned} W^{(q),*}_{i,k}(s) = &\frac{(1-\rho_i)(1 - I^*_i(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \\ &+ \frac{\rho_{i,b}(\lambda_{i,a}(1 - \xi^*_{i,a}(s)) + s)}{\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s}. \end{aligned} \tag{3.37}$$

For n-class priority queues, the waiting-time distribution in heavy traffic is equal to the case of non-preemptive priority queues. For the scaled waiting time in queue $W^{(q)}_{i,k}$ of a class-k customer in Q_i with preemptive priority service we get using (3.37): For $Re(s) > 0$, $i \in I_{NPRIOR-PR}$, $k = 1 \ldots, n$,

$$\tilde{W}^{(q),*}_{i,k}(s) = \frac{(1-\hat{\rho}_i)\left(1 - \left(\frac{\mu_i}{\mu_i + s(1+\hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}\right)^{\alpha}\right)}{\mathbb{E}[S](1-\hat{\rho}_i)s(1 - \hat{\rho}_{i,k}(1 + \hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}]))} + \frac{\hat{\rho}_{i,b}(1 + \hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])}{1 - \hat{\rho}_{i,k}(1 + \hat{\lambda}_{i,a}\,\mathbb{E}[\xi_{i,a}])},$$

which is equal to (3.35) from the non-preemptive case. As before, α and μ_i are given in (3.7). From (3.36) it follows directly that the residence time can be neglected in heavy traffic.

3.8 SJF and SRPT

The Shortest-Job-First (SJF) service discipline can be thought of as a non-preemptive priority queue with different priority classes. It may be interpreted as the continuous equivalent to having an infinite number of priority classes, where the priority classes correspond to job sizes. Alternatively, in Schrage and Miller [123], for the waiting time conditional on the service requirement x, a three-class priority queue is used where the second class consists of customers of size x. From the heavy-traffic limit derived in the previous section we can immediately derive the heavy-traffic limit of the waiting-time distribution for SJF. In Subsection 3.8.1 we give the scaled waiting-time distribution conditional on the service requirement. In Subsection 3.8.2 we give the unconditional scaled waiting-time distribution. SRPT and preemptive SJF are discussed in Subsection 3.8.3.

3.8.1 Conditional waiting-time distribution in heavy traffic

To go from Equation (3.35) to SJF we let the service time of the customer determine its priority. Note that we can apply Section 3.7.1 if the distribution is discrete. In this

section we assume that the service-time distribution has a density. First we derive the LST of the waiting time conditional on x, the service duration required by a tagged customer. Define $\rho_i(x) = \lambda_i \mathbb{E}[B_i \mathbb{1}_{\{B_i<x\}}]$ which is the continuous equivalent of $\rho_{i,a}$. Because the service-time distribution is continuous, we have $\rho_i - \rho_{i,b} = \rho_{i,a}$. We can now write down the conditional LST using (3.35): For $Re(s) > 0$, $x > 0$,

$$\tilde{W}_i^*(s|x) = \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} \frac{1}{\mathbb{E}[S]s(1-\hat{\rho}_i)/(1-\hat{\rho}_i(x))} \left\{ 1 - \left(\frac{\mu_i}{\mu_i + s/(1-\hat{\rho}_i(x))} \right)^{\alpha} \right\} + \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} \quad (i \in I_{SJF}). \tag{3.38}$$

This result gives rise to the following theorem.

Theorem 3.7. *For $\rho \uparrow 1$,*

$$\tilde{W}_{i,x} \to_d \begin{cases} 0 & \text{w.p. } \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} \\ U_{i,x}\tilde{\mathbf{I}}_i & \text{w.p. } \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} \end{cases} \qquad (i \in I_{SJF}), \tag{3.39}$$

where $U_{i,x}$ is a random variable with a uniform distribution on $[0, \frac{1}{1-\hat{\rho}_i(x)}]$ and $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha + 1$ and μ_i as given in (3.7).

Proof. The results follows directly from (3.38). □

Remark 3.10 (HTAP)**.** The intuition for the asymptotic waiting-time distribution is similar to the n-class priority queue, but slightly simpler. For the fluid model, we only consider particles that are served before a particle with service requirement x, i.e., type-a particles. Figure 3.4 gives a graphical representation of the fluid model; on the horizontal axis the course of a cycle with length c is plotted. On the vertical axis the workload of type-a particles in Q_i is plotted. The cycle is divided in three parts; the first part is the intervisit time I_i with length $(1-\hat{\rho}_i)c$. The second part is the first part of the visit time where type-a particles are being served; it starts at polling instant of Q_i and ends the first moment since the start of the cycle that no type-a particles are present. This part has length $\frac{\hat{\rho}_i(x)(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}$. The last part is the part where the other particles are served and has length

$$c - (1-\hat{\rho}_i)c - \frac{\hat{\rho}_i(x)(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)} = \frac{c(\hat{\rho}_i - \hat{\rho}_i(x))}{1-\hat{\rho}_i(x)}.$$

With probability $\frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)}$ a particle with service requirement x arrives during the last part of the cycle where hardly any type-a particles are present. Again, the scaled waiting time in HT is negligible in this case, since the remaining service duration of the particle in service and the type-a busy periods generated by type-a particles arriving during this remaining duration do not scale with ρ. With probability $\frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)}$ a particle arrives during the duration of the first two parts together. Let the uniform random variable U_i denote the fraction of combined length of the first two parts that

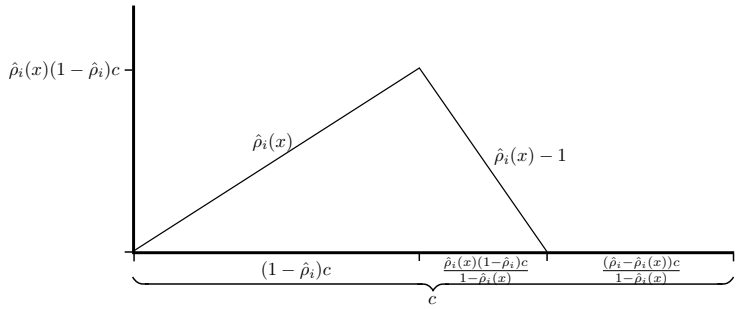


Figure 3.4: Fluid limits in heavy traffic. The amount of type a workload in Q_i is plotted over the course of a cycle.

has elapsed at the arrival epoch of the arriving particle. This particle is served at the start of the third part of the cycle, so the waiting time of this particle is the remaining duration of the first two parts $(1-U_i)\frac{(1-\hat{\rho}_i)c}{1-\hat{\rho}_i(x)}$. Using $I_i = (1-\hat{\rho}_i)c$, it follows that the scaled waiting time is now uniformly distributed on $[0, \frac{1}{1-\hat{\rho}_i(x)}]I_i$.

3.8.2 Unconditional waiting-time distribution in heavy traffic

For the unconditional waiting-time distribution in heavy traffic we have the following theorem. Let $\hat{\rho}_i^{-1}(y)$ denote the inverse function of $\hat{\rho}_i(x)$.

Theorem 3.8. *For $\rho \uparrow 1$,*

$$\tilde{W}_i \to_d \tilde{U}_i \tilde{\mathbf{I}}_i \quad (i \in I_{SJF}), \tag{3.40}$$

where $\tilde{U}_i$ has probability density function

$$f_{\tilde{U}_i}(y) = \begin{cases} 1-\hat{\rho}_i & y \in [0,1] \\ (1-\hat{\rho}_i)\left(1 - F_{B_I}\left(\hat{\rho}_i^{-1}\left(\frac{y-1}{y}\right)\right)\right) & y \in \left(1, \frac{1}{1-\hat{\rho}_i}\right], \end{cases} \tag{3.41}$$

with a point mass at zero of

$$\int_0^\infty \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} f_{B_i}(x)\,\mathrm{d}x, \tag{3.42}$$

and where $\tilde{\mathbf{I}}_i$ has a gamma distribution with parameters $\alpha+1$ and μ_i as given in (3.7).

Proof. Note that the conditional waiting-time distribution in (3.39) can be written as a gamma distribution times a uniform distribution with a point mass at zero; we refer to the latter as "uniform" distribution. To find the unconditional distribution

of the waiting time, we need to find the unconditional "uniform" distribution $\tilde{U}_i$ using Lemma 3.1. The cumulative distribution function of the conditional "uniform" distribution is given by

$$F_{U_{i,x}}(y) = \begin{cases} 0, & y < 0 \\ \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} + \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i(x)} y(1-\hat{\rho}_i(x)), & 0 \le y \le \frac{1}{1-\hat{\rho}_i(x)} \\ 1, & y > \frac{1}{1-\hat{\rho}_i(x)}. \end{cases}$$

The probability density function of $U_{i,x}$ is given by $f_{U_{i,x}}(y) = 1 - \hat{\rho}_i$, for $y \in [0, \frac{1}{1-\hat{\rho}_i(x)}]$, thus we have $a(x) = 0$ and $b(x) = \frac{1}{1-\hat{\rho}_i(x)}$. Note that $\hat{\rho}_i(x_{min}) = 0$ and $\hat{\rho}_i(x_{max}) = \hat{\rho}_i$, by recalling that $\hat{\rho}_i(x) = \hat{\lambda}_i \, \mathbb{E}[B_i \mathbb{1}_{\{B_i < x\}}]$; $b(x)$ thus increases from 1 to $1/(1-\hat{\rho}_i)$. If $y \le 1$, we find

$$f_{\tilde{U}_i}(y) = \int_{x=0}^{\infty} f_{B_i}(x) * f_{U_{i,x}}(y) \, \mathrm{d}x = 1 - \hat{\rho}_i, \quad y \in [0,1].$$

When $y > 1$, $U_{i,x}$ only has probability mass for $x > \hat{\rho}_i^{-1}((y-1)/y)$. We get

$$\begin{aligned} f_{\tilde{U}_i}(y) &= \int_{x=\hat{\rho}_i^{-1}\left(\frac{y-1}{y}\right)}^{\infty} f_{B_i}(x) * f_{U_{i,x}}(y) \, \mathrm{d}x \\ &= (1-\hat{\rho}_i)\left(1 - F_{B_i}\left(\hat{\rho}_i^{-1}\left(\frac{y-1}{y}\right)\right)\right), \quad y \in \left(1, \frac{1}{1-\hat{\rho}_i}\right]. \end{aligned}$$

Combining the results above we see that $\tilde{U}_i$ has probability mass (3.42) in zero, and density (3.41). This completes the proof. □

3.8.3 SRPT and preemptive SJF

In this subsection we consider preemptive size-based scheduling policies. The most common is SRPT, where the customer with the smallest *remaining* service time is preemptively taken into service. A less well-known policy is preemptive SJF, where the customer is preemptively taken into service with the smallest *original* service time. The latter policy also has some desirable properties, see e.g. [15; 80]. Similar to SJF, the waiting-time distribution for preemptive SJF follows directly from the preemptive n-class priority queue of Subsection 3.7.2.

The analysis of SRPT does not follow directly from the results of Kella and Yechiali [93]. Below, we use their framework to derive the LST of the waiting time in queue $W_{i,x}^{(q)}$ for a customer with service time x. We utilize the notation introduced in Section 3.7 and adopt the terminology of [93]. In particular, letting class-a represent customers with service times smaller than x, $\xi_{i,a}^*(s)$ is defined by

$$\xi_{i,a}^*(s) = \frac{1}{F_{B_i}(x)} \int_0^x \exp\left(-t(s + \lambda_{i,a} - \lambda_{i,a}\xi_{i,a}^*(s))\right) f_{B_i}(t) \, \mathrm{d}t, \tag{3.43}$$

with $\lambda_{i,a} = \lambda_i F_{B_i}(x)$, i.e., $\xi^*_{i,a}(s)$ is a type-a busy period. Similarly, let class-b represent customers with service times larger than x and $\lambda_{i,b} = \lambda_i(1 - F_{B_i}(x))$.

Proposition 3.4. *For $\rho < 1$, $i \in I_{SRPT}$, $Re(s) > 0$,*

$$\begin{aligned} W_i^{(q),*}(s) = \frac{1-\rho_i}{s\,\mathbb{E}[I_i]}\left(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s))\right) \\ + \frac{\rho_i - \rho_i(x) - \lambda_{i,b}x}{s}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) \\ + \frac{\lambda_{i,b}}{s}\left(1 - \exp\left(-x(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s))\right)\right). \end{aligned}$$

Proof. We start with the multi-class case, where class-k is the class under consideration having service times in $(x - \epsilon, x]$, for $\epsilon > 0$ small, and classes a and b have priority index lower and higher than k, respectively. That is, the service times of class-a is smaller than $x - \epsilon$ and of class-b is larger than x. Applying the idea of Schrage and Miller [123], customers of size larger than x only affect class-k as soon as their remaining service times become x. Specifically, class-b initiates a delay cycle, as defined in [93], when their remaining service time is x. In the terminology of Kella and Yechiali, we thus have $T_{i,a,k}$ cycles for $T_i = I_i, B_{i,a}, B_{i,k}$, but now also for $T = x$. Since the LST of the waiting time given the cycle during which the customer arrives is known, it remains to specify the probabilities that the system is in a specific delay cycle. In line with [93, p.28], we have the cycle probabilities

$$\begin{aligned} \Pi_{i,0} := \mathbb{P}(\text{no delay}) &= \rho_{i,b} - \lambda_{i,b}x = \rho_i - \rho_{i,a} - \rho_{i,k} - \lambda_{i,b}x, \\ \mathbb{P}(B_{i,a}\text{ cycle}) &= \frac{\Pi_{i,0}\rho_{i,a}}{1-\rho_{i,a}-\rho_{i,k}}, \quad \mathbb{P}(B_{i,k}\text{ cycle}) = \frac{\Pi_{i,0}\rho_{i,k}}{1-\rho_{i,a}-\rho_{i,k}}, \\ \mathbb{P}(I_i\text{ cycle}) &= \frac{1-\rho_i}{1-\rho_{i,a}-\rho_{i,k}}, \quad \mathbb{P}(x\text{ cycle}) = \frac{\lambda_{i,b}x}{1-\rho_{i,a}-\rho_{i,k}}. \end{aligned}$$

Using the probabilities above in Equations (7a) and (8) of [93], we obtain, for $Re(s) > 0$,

$$\begin{aligned} W_{i,k}^{(q),*}(s) = \frac{(1-\rho_i)(1 - I_i^*(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)))}{\mathbb{E}[I_i](\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s)} \\ + \frac{\Pi_{i,0}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) + \lambda_{i,b}\left(1 - \exp\left(-x(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s))\right)\right)}{\lambda_{i,k}B^*_{i,k}(s + \lambda_{i,a} - \lambda_{i,a}\xi^*_{i,a}(s)) - \lambda_{i,k} + s}. \end{aligned} \tag{3.44}$$

Letting $\epsilon \downarrow 0$, and substituting $\Pi_{i,0}$, we obtain the result. □

As in Subsection 3.7.2, $W^{(q)}_{i,x}$ is the waiting time in queue before the customer is first taken into service; this is not the same as the waiting time defined in this chapter.

We note that the residence time is identical to the residence time in a regular SRPT queue, see [123].

For LCFS and multi-class priority queues, the HT limits for the non-preemptive and preemptive policies are identical. The same holds for SJF, preemptive SJF, and SRPT as represented by the following theorem.

Theorem 3.9. *For $\rho \uparrow 1$, the scaled waiting times $\tilde{W}_i$ follow the same probability distribution for SJF, preemptive SJF, and SRPT.*

Proof. Consider the conditional scaled waiting time $\tilde{W}_{i,x}(s)$. For preemptive SJF it can be directly observed from Subsection 3.7.2 that the heavy-traffic limit is identical to the one for SJF. Using Proposition 3.4, it follows that $\lim_{\rho\uparrow 1} W_{i,x}^{(q),*}(s(1-\rho))$ equals the right-hand side of (3.38). Using (3.36) as an upper bound for the residence time, it is evident that the additional delay during the service does not contribute to the HT limit. □

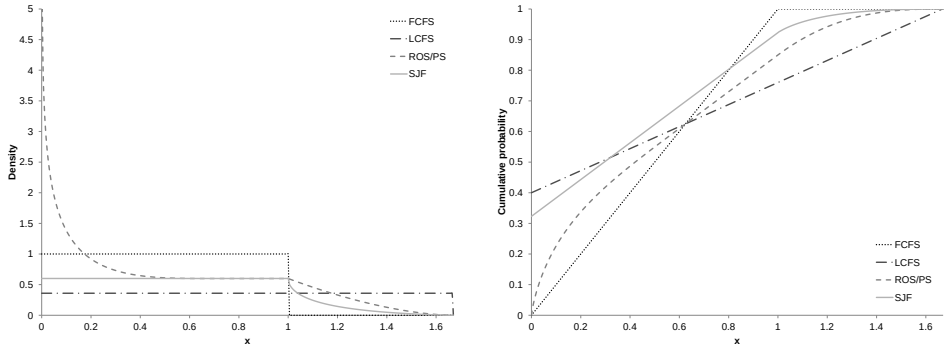
3.9 Summary of the results

In this section we give a summary of the most important results obtained in this chapter. The main result of the chapter is the fact that the scaled waiting-time distribution can always be characterized as a product of two distributions. The first distribution is a service-order specific distribution, the second distribution is a gamma distribution. The gamma distribution is a scaled length-biased intervisit-time distribution or cycle-time distribution; the most intuitive representation for the second depends on the scheduling policy. Due to the fact that for exhaustive service at queue i it holds that $C_i^*(s) = I_i^*(s + \lambda_i(1 - \xi_i^*(s)))$, see also (3.5), we can rewrite the second (gamma) distribution as the scaled length-biased intervisit-time distribution for all scheduling policies.

Let Θ_i denote the service-order specific distribution; the probability density functions for the different service policies are then given in Table 3.1. In Figure 3.5 we plot the pdf $f_{\Theta_i}(x)$ of Θ_i (Figure 3.5a) and also the cumulative distribution functions $F_{\Theta_i}(x)$ (Figure 3.5b). We choose $\hat{\rho}_i = 0.4$. For FCFS, LCFS, ROS, and NPRIOR, the HT limit only depends on the service time distribution through its first moment. This is not the case for PS, SJF, and SRPT. In the figures we took exponential service times for PS and SJF. Figure 3.5a nicely shows how Θ_i behaves; for LCFS and FCFS it is like a uniform distribution, for SJF it is a type of generalized trapezoidal distribution, whereas it slightly deviates from this for ROS and PS. The atoms in zero can be observed from Figure 3.5b. In addition, these cdfs allow us to see the impact of scheduling policy. For instance, SJF is here superior to ROS and PS.

Service order	pdf of Θ_i
FCFS	$f_{\Theta_i}(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$
LCFS/LCFS-PR	$f_{\Theta_i}(x) = (1-\hat{\rho}_i) \begin{cases} 1-\hat{\rho}_i & x \in [0, \frac{1}{1-\hat{\rho}_i}] \\ 0 & \text{otherwise} \end{cases}$ with a point mass of $\hat{\rho}_i$ in zero
ROS/PS	$f_{\Theta_i}(x) = \hat{\rho}_i \begin{cases} \frac{1}{\hat{\rho}_i}\text{Beta}_{1-x(1-\hat{\rho}_i)/\hat{\rho}_i}(1+\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 0) & x \in [0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}] \\ 0 & \text{otherwise} \end{cases}$ $+\mathbb{1}_{\{\hat{\rho}_i \le 1/2\}}(1-\hat{\rho}_i) \begin{cases} 1-g(x) & x \in [0, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}) \\ 1 & x \in [\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, 1] \\ g(x-1) & x \in (1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}+1] \\ 0 & \text{otherwise} \end{cases}$ $+\mathbb{1}_{\{\hat{\rho}_i > 1/2\}}(1-\hat{\rho}_i) \begin{cases} 1-g(x) & x \in [0,1) \\ g(x-1)-g(x) & x \in [1, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}] \\ g(x-1) & x \in (\frac{\hat{\rho}_i}{1-\hat{\rho}_i}, \frac{\hat{\rho}_i}{1-\hat{\rho}_i}+1] \\ 0 & \text{otherwise,} \end{cases}$ where $g(x) = \left(1 - \frac{x(1-\hat{\rho}_i)}{\hat{\rho}_i}\right)^{\frac{1}{1-\hat{\rho}_i}}$
NPRIOR/ NPRIOR-PR	$f_{\Theta_{i,k}}(x) = \frac{1-\hat{\rho}_i}{1-\hat{\rho}_i+\hat{\rho}_{i,b}} \begin{cases} 1-\hat{\rho}_{i,a} & x \in \left[0, \frac{1}{1-\hat{\rho}_{i,a}}\right] \\ 0 & \text{otherwise} \end{cases}$ with a point mass of $\frac{\hat{\rho}_{i,b}}{1-\hat{\rho}_i+\hat{\rho}_{i,b}}$ in zero
SJF/SRPT	$f_{\Theta_i}(x) = \begin{cases} 1-\hat{\rho}_i & x \in [0,1] \\ (1-\hat{\rho}_i)\left(1 - F_{B_i}\left(\hat{\rho}_i^{-1}\left(\frac{x-1}{x}\right)\right)\right) & x \in \left(1, \frac{1}{1-\hat{\rho}_i}\right] \\ 0 & \text{otherwise} \end{cases}$ with a point mass of $\int_0^\infty \frac{\hat{\rho}_i - \hat{\rho}_i(x)}{1-\hat{\rho}_i(x)} f_{B_i}(x)\,\mathrm{d}x$ in zero

Table 3.1: The probability density functions of the service-order specific distributions.



(a) Probability density functions. (b) Cumulative distribution functions.

Figure 3.5: Shapes of the service order specific distributions.

3.10 Closed-form approximations for systems with arbitrary load

In this section we illustrate the results by calculating moments and tail probabilities of the waiting-time distribution for different service disciplines by simulations. Moreover, we use the heavy-traffic limits as the basis for approximations for the waiting-time distributions for stable systems, i.e. with $\rho < 1$. To this end, the asymptotic results suggest the following approximation for the waiting-time distribution for $\rho < 1$: For $i = 1, \ldots, N$,

$$\mathbb{P}(W_i \leq x) \approx \mathbb{P}(\Theta_i \Gamma_i \leq (1-\rho)x). \tag{3.45}$$

The moments of the waiting-time distribution can be approximated using

$$\mathbb{E}[W_i^k] \approx \frac{\mathbb{E}[\Theta_i^k]\,\mathbb{E}[\Gamma_i^k]}{(1-\rho)^k}.$$

See Section 3.11 and references therein for a discussion on convergence of moments.

We consider a polling model with $N = 3$ queues and all queues receive exhaustive service. Service times and switch-over times are exponentially distributed. The mean service durations at queue 1, 2, and 3 equal 2, 3, and 1, respectively. The mean switch-over times are given by $\mathbb{E}[S_1] = \mathbb{E}[S_3] = 1$ and $\mathbb{E}[S_2] = 3$. Arrivals are Poisson and the arrival rates at the different queues are chosen such that the ratios between the arrival rates are 3:2:1, while the total load of the system is varied. Note that the system is rather asymmetric and that the ratios between the loads of the queues are 6:6:1.

We apply the approximation to a system with a load of 0.95 and let the service order be ROS, PS and SJF. We plot the approximated and simulated cumulative distributions

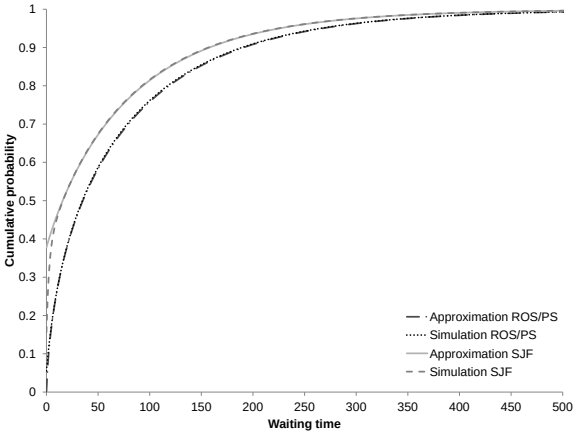


Figure 3.6: Approximated and simulated cumulative distribution functions of the waiting-time distribution in a system with a load of 0.95.

of the waiting time at the first queue. Figure 3.6 shows that the approximation follows the simulation closely. ROS and PS are plotted together, since the distributions are equal. Note that for the SJF service discipline the approximation shows a point mass at zero, this effect does not show up as clearly in the simulation. This is caused by the fact that the point mass at zero only occurs if the load is very close to 1.

To illustrate the differences between the various scheduling policies we plot the approximated cumulative distribution functions of the scaled waiting times at the first queue of the system described above. In Figure 3.7 we clearly see a point mass at zero if the service discipline is LCFS or SJF. The line of SJF always lies above the line of PS; as the service-time distribution is exponential, this indicates that for exponential service times SJF is a better policy than PS. Table 3.2 shows the simulated and approximated values of the mean waiting times at Q_1 and their relative absolute differences defined as

$$\Delta\% = 100\% \times \frac{|\text{App} - \text{Sim}|}{\text{Sim}}$$

for different values of ρ and for different scheduling policies considered in this chapter. The mean waiting times are equal for FCFS, LCFS and ROS and also for PS if the service-time distribution is exponential. In Table 3.3, the results for the standard deviations of the waiting times at Q_1 are given. Both tables show that the relative differences decrease to 0 if ρ increases to 1. It is interesting to note that for lower values of ρ, the error in the standard deviation is quite high, especially if the service order is LCFS. This can be explained by the fact that in HT the waiting time is equal to zero if an arrival occurs during a visit period. For lower loads this effect does not occur, busy periods will influence the waiting time. The numerical approximations

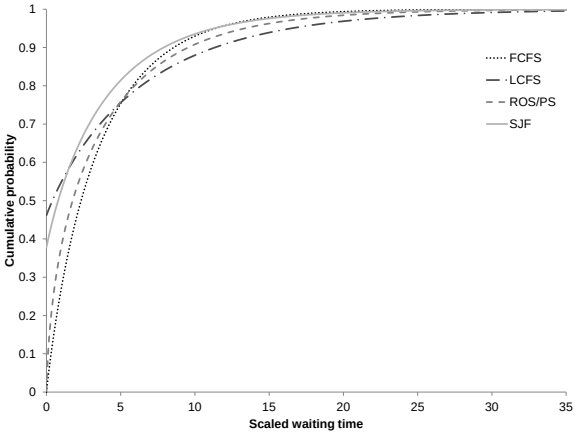


Figure 3.7: Cumulative distribution functions of the scaled waiting times at the first queue for different service orders.

can be improved using an interpolation with light-traffic limits, as carried out in [62] and also in the previous chapter.

	FCFS/LCFS/ROS/PS			SJF		
ρ	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.7	12.25	12.02	1.89	10.11	8.88	12.19
0.8	18.43	18.03	2.15	14.69	13.31	9.34
0.9	36.68	36.07	1.66	28.16	26.63	5.45
0.95	72.79	72.13	0.90	54.86	53.26	2.92
0.98	180.91	180.33	0.32	134.82	133.14	1.25
0.99	361.32	360.66	0.18	267.83	266.29	0.57

Table 3.2: Simulated value, approximated value and delta of the mean waiting time for different service disciplines and loads.

3.11 Discussion and concluding remarks

In this chapter we assume that all queues receive exhaustive service, which is an important extension of the results obtained for similar models but with gated service at all queues in Chapter 2. We emphasize that the exhaustive service case is more complicated than the gated case, despite the fact that both the exhaustive and the gated service disciplines satisfy the well-known branching structure identified in [121]. The complexity lies in the fact that for exhaustive service the local service order of the

	LCFS			ROS/PS			SJF		
ρ	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.7	17.86	20.92	17.10	15.10	16.22	7.37	13.35	14.30	7.12
0.8	28.53	31.38	9.99	23.46	24.33	3.69	20.64	21.44	3.91
0.9	60.02	62.76	4.56	48.02	48.65	1.31	42.16	42.89	1.72
0.95	122.64	125.52	2.35	96.72	97.30	0.60	85.01	85.78	0.90
0.98	310.73	313.80	0.99	242.49	243.26	0.31	213.70	214.44	0.35
0.99	624.20	627.59	0.54	485.78	486.51	0.15	427.88	428.88	0.24

Table 3.3: Simulated value, approximated value and delta of the standard deviation of the waiting time for different service disciplines and loads.

customers during a visit period V_i of the server to a given queue i cannot be determined at the polling instant marking the beginning of V_i; for gated the service order is determined at the beginning of V_i. As a consequence, newly arriving customers at queue i during V_i may change the local service order and the sharing of server capacity among the customers served during V_i, and hence affect the waiting-time and sojourn-time distributions in a complex manner. For example, this complexity manifests itself in the case of PS service and multiple vacations, where analytic results on (conditional) sojourn times, conditioned on the number of customers at the beginning of a service period, are only known under the assumption of exponential service times (see [50]). Even for multiple vacation models, extension of such results to the case of general service times is complicated, because of the complex relation between the number of customers in the system and the remaining amounts of per-customer service times.

The assumption that all queues are served exhaustively can easily be relaxed to the general setting where a subset of the queues receive gated service (or some other branching-type service policy). More specifically, for general mixtures of exhaustive and gated service, let G be the set of indices i for which Q_i receives gated service, and $E := \{1, \ldots, N\} \backslash G$ the subset of queues that receive exhaustive service. Then the results presented above still hold; the only difference is that the parameter δ in (3.4) should be replaced by

$$\delta_{mixture} := 1 - \sum_{i \in E} \hat{\rho}_i^2 + \sum_{i \in G} \hat{\rho}_i^2. \tag{3.46}$$

In the present chapter it is assumed that the arrival processes at the queues are Poisson. This assumption can easily be relaxed to renewal arrivals. Following a well-established line of argumentation (see [51; 52; 114]), one may conjecture that results presented in Section 3.3 to 3.8 are still valid when σ^2 defined in (3.4) is replaced by

$$\sigma^2_{renewal} = \sum_{i=1}^{N} \hat{\lambda}_i \left(\mathrm{Var}[B_i] + c^2_{A_i} \, \mathbb{E}[B_i]^2 \right), \tag{3.47}$$

as was also defined earlier in (2.33).

Finally, we address a number of topics for further research. First, the heavy-traffic results proven in this chapter demonstrate convergence *in distribution* by demonstrating point-wise convergence of the LST's to their limiting regimes, and application of Levy's Continuity Theorem. An interesting question is whether the results can be extended to other types of convergence, and under what assumptions. For example, convergence in distribution does not necessarily imply moment-wise convergence; the latter requires the finiteness of higher moments of the service times and switch-over times. We refer to [136] (Section 3.3) for more detailed discussion about moment-wise convergence. In the case of PS service at queue i we made the additional assumption that the service times are exponentially distributed. Under this assumption, we proved the correctness of Theorem 3.4 by using the results in [12] (Section 5) which, in turn, rely on the classical results by Coffmann et al. [50] for the M/M/1 PS queue (without vacations). It is an open question how the results for PS can be extended to the case of generally distributed service times.

Chapter 4

Transient analysis of cycle times in polling systems

4.1 Introduction

In Chapters 2 and 3, we derived the waiting time distribution in heavy traffic for polling models with various scheduling and service disciplines. In this chapter, we study cyclic polling models with Poisson arrivals, general service and switch-over times and gated or globally gated service. We focus on the transient behavior of the successive cycle times. Our goal is to gain an understanding in the dependency structure between the different cycles. This study is motivated by our interest in systems where disruptions or breakdowns may occur, often leading to excessive cycle lengths. In this context, we are interested in the following questions:

1. If the system encounters an excessively long cycle time (e.g., due to a disruption or a breakdown), then how will that influence the durations of the subsequent cycle times? What is the time needed to recover from excessive cycle times?
2. What is the dependency structure between various residence and cycle times? More specifically, what is the correlation between the successive cycle (and residence) times?

A primary motivation for the second question is that the dependency structure makes polling models challenging to analyze. Insights into the dependency between cycles and residence times might pave the way for approximation methods. For instance, for polling models in tandem, the output of some queues may feed into another queue. The output of a specific queue in a polling system is essentially driven by an on-off source with dependent on and off times ('on' representing visit times and 'off' representing intervisit times). Similar relations have also motivated the study of some vacation models, see e.g. [39; 41; 47]. Finally, we note that waiting-time and queue-length distributions can be expressed in terms of the marginal cycle-time distribution for polling models with (globally) gated and exhaustive service.

In this chapter, we assume that the distribution of the first cycle (in case of globally

gated service) or N residence times (in case of gated service), where N is the number of queues, is known and that the arrivals are Poisson. Using this, we show how the joint LST of all x subsequent cycles (globally gated) can be expressed in terms of the LST of the first cycle. Moreover, for the case of gated service we show how all $x > N$ subsequent residence times can be expressed in terms of the LST of the first cycle. From these joint LST's, we derive the first two moments and correlation coefficients between different cycles. Lastly, we analyze a heavy-tailed first cycle length, due to disruptions or breakdown, or the heavy-traffic regime to provide new fundamental insights into the time-dependent behavior.

The remainder of this chapter is organized as follows. In Section 4.2 the models are described and the method and goals of the chapter are outlined. In Section 4.3 we study the case of globally-gated service, whereas we study the case of gated service in Section 4.4. Both sections contain asymptotic results, such as heavy-tailed initial cycle lengths and heavy traffic, and numerical illustrations.

4.2 Model description

We consider a system of $N \geq 2$ infinite-buffer queues, $Q_1, \ldots, Q_N$, and a single server that visits and serves the queues in cyclic order. Customers arrive at Q_i according to a Poisson process $\{N_i(t),\ t \in \mathbb{R}\}$ with rate λ_i. These customers are referred to as type-i customers. The total arrival rate is denoted by $\Lambda = \sum_{i=1}^{N} \lambda_i$. The service time of a type-i customer is a random variable B_i, with LST $B_i^*(\cdot)$, and kth moment $\mathbb{E}[B_i^k]$, $k = 1, 2, \ldots$, when it is finite. The kth moment of the service time of an arbitrary customer is denoted by $\mathbb{E}[B^k] = \sum_{i=1}^{N} \lambda_i \, \mathbb{E}[B_i^k]/\Lambda$, $k = 1, 2, \ldots$. The load offered to Q_i is $\rho_i = \lambda_i \, \mathbb{E}[B_i]$ and the total load offered to the system is equal to $\rho = \sum_{i=1}^{N} \rho_i$. The switch-over time required by the server to proceed from Q_i to Q_{i+1} is a random variable S_i with mean $\mathbb{E}[S_i]$ and LST $S_i^*(\cdot)$. Let $S = \sum_{i=1}^{N} S_i$, with LST $S^*(\cdot)$, denote the total switch-over time in a cycle. We define $\delta_i(s) := \lambda_i(1 - B_i^*(s))$ and let $\mathbf{e}_i$ be a unit vector with 1 in the ith position and 0 in the other positions.

We consider the gated and globally gated service disciplines. When the service discipline is gated, a gate at Q_i closes when the server arrives at Q_i. Every customer standing in front of the gate is served, while customers arriving at Q_i during service of Q_i must wait for the next cycle, this holds for all $i = 1, \ldots, N$. When the service discipline is globally gated, a gate closes at all queues when the server arrives at Q_1. During the following cycle, every customer standing in front of the gate is served.

Method and goals

Throughout this chapter we assume that the distribution of the length of the first cycle is known. For the gated service discipline, this requires that the joint distribution of the first N residence times is known, where a residence time is a visit time plus the subsequent switch-over time. When the probabilistic behavior of the first cycle is known, the next residence time can be expressed in terms of the first cycle, as it consists of a visit time to serve all the work that arrived at the queue during the first cycle plus the switch-over time. For globally gated, this is true for every queue, as the gate closes at the start of a cycle. For gated, the length of a visit time is always determined by the work that arrived at the corresponding queue during the last N residence times. It can be seen that the second cycle is completely determined in terms of the first cycle. Consequently, the third cycle can be expressed in terms of the second cycle and so also in terms of the first cycle. As a result, every cycle can recursively be expressed in terms of the first cycle. We use this fact to derive the joint LST of x consecutive cycles or residence times in terms of the LST of the first cycle.

Let us first consider the globally gated case. Our goal is to determine the joint LST of x consecutive cycle times, denoted by $\gamma_x(\mathbf{z})$. The vector $\mathbf{z}$ of length x contains the variables $z_1, \ldots, z_x$, corresponding to cycles $1, \ldots, x$, with the LST of the first cycle, $\gamma_1(\mathbf{z})$, assumed to be given. Choosing the z_i in specific ways, enables us to calculate all kinds of useful performance measures. For example, when $z_i = z$ for all $i \in J \subseteq \{1, \ldots, x\}$ and 0 otherwise, we obtain the LST of the sum of cycles of set J. Such a choice is especially convenient to calculate moments, which are then obtained by differentiating with respect to z and taking $z = 0$.

Also, the covariance between cycle 1 and cycle x can be calculated using the following property of the covariance: if X_1 and X_2 are random variables, then $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$, with the variance of a random variable X being $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. To calculate the covariance between cycle x and cycle y ($y \geq x$), we define 3 vectors of length y: $\mathbf{a} = z(\mathbf{e}_x + \mathbf{e}_y)$, $\mathbf{b} = z\mathbf{e}_x$ and $\mathbf{c} = z\mathbf{e}_y$. Let C_i denote the ith cycle. The first vector is then used to calculate the variance of $C_x + C_y$, the other two vectors are used for the variance of C_x and C_y, respectively. The covariance is then given by

$$\mathrm{Cov}(C_x, C_y) = \frac{1}{2}\left(\mathrm{Var}(C_x + C_y) - \mathrm{Var}(C_x) - \mathrm{Var}(C_y)\right). \tag{4.1}$$

If, due to some external event or disaster, the first cycle is very long, we can easily derive the duration of the effect by calculating the duration of subsequent cycles, until it converges to the expected duration of $\mathbb{E}[S]/(1-\rho)$, if $\rho < 1$. In addition, the recursive relations allow us to derive exact asymptotic transient results for some limiting regimes, such as heavy traffic and heavy-tailed initial cycle times.

For gated, the same methods can be applied, for residence times instead of cycle times. By choosing N consecutive z_i equal to z, it is also possible to look at cycles. Note

that the state description for gated in terms of residence times can also be applied to the globally gated case to obtain information about the residence times.

4.3 Analysis of globally gated service

In this section we consider the joint LST of x cycles for a globally gated service discipline. This result is used to derive various moments and asymptotic properties. We suppose that the distribution of the first cycle, C_1, is known, and that its LST is given by $\gamma_1(z) := \mathbb{E}[e^{-zC_1}]$. Because of the globally gated service discipline, the length of a cycle determines the number of customers that are served during the next cycle. The number of customers that are served during a cycle plus the switch-over times together determine the length of that cycle. As such every cycle length can be expressed in the length of the previous cycle, and it is possible to express every cycle length recursively in terms of the first cycle. If we want to give the joint LST of C_1 and C_2, in terms of the LST of C_1, we can first condition on the value of C_1 and then integrate over the density of C_1, as follows:

$$\begin{aligned}\mathbb{E}[e^{-z_1C_1-z_2C_2}] &= \int_{c_1=0}^{\infty} \mathbb{E}[e^{-z_1c_1-z_2C_2}]\,\mathrm{d}\,\mathbb{P}(C_1 \le c_1) \\ &= \int_{c_1=0}^{\infty} \mathbb{E}[e^{-z_1c_1-\sum_{k=1}^{N}\lambda_k(1-B_k^*(z_2))c_1}]\,\mathbb{E}[e^{-z_2S}]\,\mathrm{d}\,\mathbb{P}(C_1 \le c_1) \\ &= C_1^*\left(z_1+\sum_{k=1}^{N}\delta_k(z_2)\right)S^*(z_2), \end{aligned} \tag{4.2}$$

where $C_1^*(\cdot)$ is the LST of C_1, which we defined as $\gamma_1(\cdot)$. More generally, we can write the joint LST of the first x cycles, denoted by $\gamma_x(\cdot)$, in terms of the LST of the first cycle. Here, $\mathbf{z}$ is a vector of length x, with z_i corresponding to cycle i.

Theorem 4.1. *The joint transform of the x cycle lengths is given by*

$$\gamma_x(\mathbf{z}) = \mathbb{E}[e^{-\sum_{i=1}^{x} z_iC_i}] = \gamma_1(z_1+\zeta_1^{(x)}(\mathbf{z}))\prod_{j=1}^{x-1} S^*(z_{j+1}+\zeta_{j+1}^{(x-j)}(\mathbf{z})), \tag{4.3}$$

with

$$\begin{aligned}&\zeta_n^{(1)}(\mathbf{z}) = 0 \\ &\zeta_n^{(i)}(\mathbf{z}) = \textstyle\sum_{k=1}^{N}\delta_k(z_{n+1}+\zeta_{n+1}^{(i-1)}(\mathbf{z})).\end{aligned} \tag{4.4}$$

Note that $\zeta_n^{(i)}$ can be interpreted as the amount of work that arrived during the nth cycle and recursively contains the amount of work that arrives during the next $(i-1)$ cycles, i.e., in some sense it defines a descendant set (see also Remark 4.1 below). In the proof of Theorem 4.1, we use the following lemma (with the proof deferred to Section 4.5.1).

Lemma 4.1. *For $i = 2, 3, \ldots$ and $n = 1, 2, \ldots$, we have*

$$\zeta_n^{(i)}(\mathbf{z}') = \zeta_n^{(i+1)}(\mathbf{z}), \tag{4.5}$$

with $\mathbf{z}' = \mathbf{z} + \zeta_{n+i-1}^{(2)}(\mathbf{z}) \cdot \mathbf{e}_{n+i-1}$.

Proof of Theorem 4.1. We use induction to prove the theorem. It evidently holds for $x = 1$. Now assume that Equation (4.3) holds for $k \leq x$. Taking $k = x + 1$ gives

$$\begin{aligned}
\gamma_{(x+1)}(\mathbf{z}) &= \gamma_x(z_1, z_2, \ldots, z_x + \sum_{k=1}^{N} \delta_k(z_{x+1}))S^*(z_{x+1}) \\
&= \gamma_x(z_1, z_2, \ldots, z_x + \zeta_x^{(2)}(\mathbf{z}))S^*(z_{x+1} + \zeta_{x+1}^{(1)}(\mathbf{z})) \\
&= \gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}')) \prod_{j=1}^{x-1} S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}'))S^*(z_{x+1} + \zeta_{x+1}^{(1)}(\mathbf{z})) \\
&= \gamma_1(z_1 + \zeta_1^{(x+1)}(\mathbf{z})) \prod_{j=1}^{x-1} S^*(z_{j+1} + \zeta_{j+1}^{(x-j+1)}(\mathbf{z}))S^*(z_{x+1} + \zeta_{x+1}^{(1)}(\mathbf{z})) \\
&= \gamma_1(z_1 + \zeta_1^{(x+1)}(\mathbf{z})) \prod_{j=1}^{x} S^*(z_{j+1} + \zeta_{j+1}^{(x-j+1)}(\mathbf{z})).
\end{aligned}$$

For the first equality, we use a reasoning similar to Equation (4.2), then we rewrite the result using (4.4). For the third equality, the induction hypotheses is used. For the fourth equality, we use Lemma 4.1, completing the proof. □

Remark 4.1 (Link with the Descendant Set Approach)**.** The expression in Theorem 4.1 can be explained along the lines of the Descendant Set Approach (DSA) [97]. The DSA considers original customers (originators) and non-original customers, where an original customer arrives during a switch-over period, and a non-original customer arrives during the service of another customer; The children of a customer T arrive during the service of T. The descendant set of T is recursively defined to consist of T, its children and the descendants of its children. In our case, the first cycle and the $x - 1$ switch-over times can be interpreted as originators and the recursive definition of $\zeta_n^{(i)}(\mathbf{z})$ represents the descendant sets. We note that in this case all quantities are expressed in time (or amount of work), while in the DSA this is typically in terms of number of customers.

By choosing specific values for z_i, $i = 1, \ldots, x$, many performance measures can be determined, see also Section 4.2. Below, we focus on the first two moments and the correlation coefficient. Here we assume that z_i is either z or 0 (depending on whether cycle i is included). Let $J \subseteq \{1, 2, \ldots, x\}$ be the set of cycles that is included, i.e., $z_i = z$ if and only if $i \in J$. Taking the derivative of γ_x, with respect to z_x, taking

$z = 0$ and multiplying by -1, gives the expected length of cycle x. The derivative of γ_x, with respect z, then becomes

$$\begin{aligned}\frac{d}{dz}\gamma_x(\mathbf{z}) = \frac{d}{dz}\gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}))\left(\mathbb{1}_{\{1\in J\}} + \frac{d}{dz}\zeta_1^{(x)}(\mathbf{z})\right)\prod_{j=1}^{x-1} S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z})) \\ + \gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}))\sum_{j=1}^{x-1}\frac{d}{dz}S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}))\left(\mathbb{1}_{\{j+1\in J\}} + \frac{d}{dz}\zeta_{j+1}^{(x-j)}(\mathbf{z})\right) \\ \times \prod_{k\neq j}^{x-1} S^*(z_{k+1} + \zeta_{k+1}^{(x-k)}(\mathbf{z})),\end{aligned} \tag{4.6}$$

with $\frac{d}{dz}\zeta_n^{(i)}(\mathbf{z})$ recursively defined as

$$\frac{d}{dz}\zeta_n^{(i)}(\mathbf{z}) = \sum_{k=1}^{N}\frac{d}{dz}\delta_k(z_{n+1} + \zeta_{n+1}^{(i-1)}(\mathbf{z}))\left(\mathbb{1}_{\{n+1\in J\}} + \frac{d}{dz}\zeta_{n+1}^{(i-1)}(\mathbf{z})\right),\ i \geq 2, \tag{4.7}$$

and $\frac{d}{dz}\zeta_n^{(i)}(\mathbf{z}) = 0$ if $i = 1$. Here $\frac{d}{dz}\delta_k(z) = -\lambda_k B_k^*(z)$. Taking $z = 0$ gives $\frac{d}{dz}\delta_k(z)|_{z=0} = \rho_k$, and filling in $z = 0$ in Equation (4.7) gives the next property:

Property 4.1. *We have, for $i = 2, 3, \ldots$ and $n = 1, 2, \ldots$,*

$$\frac{d}{dz}\zeta_n^{(i)}(\mathbf{z})|_{z=0} = \sum_{k=1}^{i-1}\rho^k\,\mathbb{1}_{\{n+k\in J\}}. \tag{4.8}$$

Proof. Equation (4.8) can be proved by induction. First, take $i = 2$, yielding

$$\frac{d}{dz}\zeta_n^{(2)}(\mathbf{z}) = \sum_{k=1}^{N}\frac{d}{dz}\delta_k(z_{n+1})\mathbb{1}_{\{n+1\in J\}}.$$

If $z = 0$, this equals $\sum_{k=1}^{N}\rho_k\mathbb{1}_{\{n+1\in J\}} = \rho\mathbb{1}_{\{n+1\in J\}}$, which agrees with (4.8). Now assume (4.8) is true for all $k \leq i$. Taking $k = i + 1$ gives

$$\frac{d}{dz}\zeta_n^{(i+1)}(\mathbf{z}) = \sum_{k=1}^{N}\frac{d}{dz}\delta_k(z_{n+1} + \zeta_{n+1}^{(i)}(\mathbf{z}))\left(\mathbb{1}_{\{n+1\in J\}} + \frac{d}{dz}\zeta_{n+1}^{(i)}(\mathbf{z})\right)$$

according to (4.7), and for $z = 0$, we obtain

$$\begin{aligned}\frac{d}{dz}\zeta_n^{(i+1)}(\mathbf{z})|_{z=0} &= \sum_{k=1}^{N}\rho_k\left(\mathbb{1}_{\{n+1\in J\}} + \sum_{l=1}^{i-1}\rho^l\mathbb{1}_{\{n+1+l\in J\}}\right)\\ &= \rho\mathbb{1}_{\{n+1\in J\}} + \sum_{l=1}^{i-1}\rho^{l+1}\mathbb{1}_{\{n+1+l\in J\}}\\ &= \rho\mathbb{1}_{\{n+1\in J\}} + \sum_{l=2}^{i}\rho^{l}\mathbb{1}_{\{n+l\in J\}} \quad = \quad \sum_{k=1}^{i}\rho^l\mathbb{1}_{\{n+l\in J\}},\end{aligned}$$

where the first equality holds due to the induction hypotheses, completing the proof. □

Combining the above yields the following proposition, for the expectation of the summation of the lengths of cycles $i \in J$.

Proposition 4.1 (First moments)**.** *The expected total length of all cycles in set J is*

$$\mathbb{E}\left[\sum_{j\in J}C_j\right] = \mathbb{E}[\gamma_1]\sum_{k=0}^{x-1}\rho^k\mathbb{1}_{\{k+1\in J\}} + \mathbb{E}[S]\sum_{j=1}^{x-1}\sum_{l=0}^{x-j-1}\rho^l\mathbb{1}_{\{j+l+1\in J\}}, \tag{4.9}$$

where $\mathbb{E}[\gamma_1]$ is the expected length of the initial cycle.

To calculate the variances of specific cycles, or covariances between two cycles, we also need the second moments. To this end, we take the derivative of Equation (4.6), resulting in Equation (4.19) in Section 4.5.2. The derivative of (4.7) is given by Equation (4.20) in Section 4.5.2. Taking $z = 0$ leads to the following property:

Property 4.2. *We have, for $i = 2, 3, \ldots$ and $n = 1, 2, \ldots$,*

$$\frac{d^2}{dz^2}\zeta_n^{(i)}(\mathbf{z})|_{z=0} = -\Lambda\,\mathbb{E}[B^2]\sum_{k=0}^{i-2}\rho^k\left(\sum_{j=0}^{i-k-2}\rho^j\mathbb{1}_{\{n+j+k+1\in J\}}\right)^2.$$

The proof is by induction and is similar to the proof of Property 4.1.

Proposition 4.2 (Second moments)**.** *Let $\mathbb{E}[\gamma_1^2]$ be the second moment of the initial*

cycle. The second moment of the total cycle length of all cycles in J is

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j\in J} C_j\right)^2\right] &= \mathbb{E}[\gamma_1^2]\left(\sum_{k=0}^{x-1}\rho^k \mathbb{1}_{\{k+1\in J\}}\right)^2 \\ &+ \mathbb{E}[\gamma_1]\left(2\,\mathbb{E}[S]\left(\sum_{k=0}^{x-1}\rho^k \mathbb{1}_{\{k+1\in J\}}\right)\left(\sum_{j=1}^{x-1}\sum_{k=0}^{x-j-1}\rho^k \mathbb{1}_{\{j+1+k\in J\}}\right)\right. \\ &\qquad \left. + \Lambda\,\mathbb{E}[B^2]\sum_{k=0}^{x-2}\rho^k\left(\sum_{j=0}^{x-k-2}\rho^j \mathbb{1}_{\{j+k+2\in J\}}\right)^2\right) \\ &+ \mathbb{E}[S]\Lambda\,\mathbb{E}[B^2]\sum_{j=1}^{x-1}\sum_{k=0}^{x-j-2}\rho^k\left(\sum_{l=0}^{x-j-k-2}\rho^l \mathbb{1}_{\{j+l+k+2\in J\}}\right)^2 \\ &+ \mathbb{E}[S^2]\sum_{j=1}^{x-1}\left(\sum_{k=0}^{x-j-1}\rho^k \mathbb{1}_{\{j+k+1\in J\}}\right)^2 \\ &+ \mathbb{E}[S]^2\left(\left(\sum_{j=1}^{x-1}\sum_{k=0}^{x-j-1}\rho^k \mathbb{1}_{\{j+k+1\in J\}}\right)^2 - \sum_{j=1}^{x-1}\left(\sum_{k=0}^{x-j-1}\rho^k \mathbb{1}_{\{j+k+1\in J\}}\right)^2\right).\end{aligned} \tag{4.10}$$

Using Equation (4.1) and the moments derived above, we can calculate the covariance between cycles 1 and x. This gives

$$\begin{aligned}\mathrm{Cov}(C_1, C_x) &= \frac{1}{2}\big(\mathbb{E}[\gamma_x(\mathbf{a})^2] - \mathbb{E}[\gamma_x(\mathbf{a})]^2 \\ &\qquad - (\mathbb{E}[\gamma_x(\mathbf{b})^2] - \mathbb{E}[\gamma_x(\mathbf{b})]^2) - (\mathbb{E}[\gamma_x(\mathbf{c})^2] - \mathbb{E}[\gamma_x(\mathbf{c})]^2)\big) \\ &= \big(\mathbb{E}[\gamma_1^2] - \mathbb{E}[\gamma_1]^2\big)\,\rho^{x-1},\end{aligned} \tag{4.11}$$

where the second equality is obtained by filling in Equations (4.9) and (4.10). Interestingly, we see that the covariance between the first cycle and cycle x only depends on the variance of the first cycle and the load of the complete system.

Remark 4.2 (Steady-state distribution)**.** In polling models, it is common to relate the length of a cycle to the length of the previous cycle. Assuming the system to be in steady state, this provides an equation for the LST of the steady-state cycle length (see e.g. [158]). In fact Equation (27) of [158] is equivalent to (4.2) if we take $z_1 = 0$. A similar recursive scheme can also be found in [38]. The distinguishing feature is that [38; 158] focus on steady-state results, whereas our aim is to derive transient performance.

Alternatively, putting $z_x = z$ and z_i equal to zero for $i < x$, gives information about cycle x. Taking $x \to \infty$, also gives information about a cycle in steady state. This

limit in Equation (4.8) gives exactly $\mathbb{E}[S]/(1-\rho)$, if $\rho < 1$. Doing the same with Equation (4.10), gives exactly the second moment of a steady state cycle. This shows that our expressions are in line with known results (see again [38; 158]).

4.3.1 Asymptotic properties

From Theorem 4.1 and Equation (4.11) we may derive more explicit results for the different asymptotic regimes, such as light and heavy traffic and large switch-over times. Moreover, we consider the case in which the first cycle has a heavy-tailed distribution, representing the situation that this cycle is affected by an external event.

Light and heavy traffic In heavy traffic, when $\rho \uparrow 1$, we directly obtain that the coefficient of correlation between cycles 1 and x tends to 1, see (4.11). Hence, the distribution of C_x is identical to the distribution of C_1 in heavy traffic.

For light traffic, we observe from Equation (4.11) that when the load tends to zero, the covariance between the first cycle and any other cycle is equal to zero, which also holds for the correlation coefficient. For approximations based on light and heavy traffic interpolations, the derivative of the light-traffic regime with respect to the load is useful. Remarkable is that the derivative of the covariance is zero for cycles $x = 3, 4, \ldots$, but it is strictly positive for the second cycle.

Large switch-over times Another common asymptotic regime is the situation of large switch-over times, see [131]. Let the switch-over times be deterministic with length r, i.e. $S^*(z) = e^{-rz}$. We scale the cycle times by dividing by r and let $r \to \infty$. This gives

$$\begin{aligned}\lim_{r\to\infty} \gamma_x(\mathbf{z}/r) &= \lim_{r\to\infty} \gamma_1(z_1/r + \zeta_1^{(x)}(\mathbf{z}/r)) \prod_{j=1}^{x-1} e^{-(z_{j+1}/r + \zeta_{j+1}^{(x-j)}(\mathbf{z}/r))r} \\ &= \prod_{j=1}^{x-1} e^{-(\sum_{k=0}^{x-j-1} \rho^k z_{j+k+1})}.\end{aligned}$$

implying that the scaled cycle lengths become deterministic and are simple expressions in terms of the system load ρ.

Heavy-tailed initial cycle length Let us assume that the first cycle is regularly varying of index $-\nu$ (denoted as $C_1 \in \mathcal{R}_{-\nu}$; and with $m < \nu < m+1$), whereas the service and switch-over times have a lighter tail. This situation is of interest in cases where a disaster or external events leads to a long cycle. Of primary interest is how this initial cycle affects future cycle lengths.

More specifically, we assume that $\mathbb{P}(C_1 > y) \sim L(y)y^{-\nu}$, $\nu > 1$, with $L(y)$ some slowly varying function. We use the notation $f(y) \sim g(y)$ to indicate that $f(y)/g(y) \to 1$ as

$y \to \infty$. A function $L(\cdot)$ is called slowly varying if $L(\eta y) \sim L(y)$, for all $\eta > 1$. For the service and switch-over times, we assume the following:

Assumption 4.1. $\mathbb{P}(B_i > y) = o(\mathbb{P}(C_1 > y))$ and $\mathbb{P}(S_i > y) = o(\mathbb{P}(C_1 > y))$ $(i = 1, \ldots, N)$, where $f(y) = 0(g(y))$ is equivalent to $\lim_{y\to\infty} f(y)/g(y) = 0$.

Proposition 4.3. *Suppose that $C_1 \in \mathcal{R}_{-\nu}$ and Assumption 4.1 is satisfied. Then, for $n = 1, 2, \ldots,$*

$$\mathbb{P}(C_n > y) \sim \mathbb{P}\left(C_1 > \frac{y}{\rho^{n-1}}\right), \quad \text{as } y \to \infty$$

The result of Proposition 4.3 can be intuitively explained and is related to the fact that rare events tend to occur due to a single most probable cause in systems with heavy-tailed characteristics. More precisely, the most likely way for a large n-th cycle time is a large initial cycle time, whereas the system shows average behavior otherwise. As traffic arrives at rate ρ, the amount of traffic arriving during the first cycle is about ρC_1; this means cycle 1 needs to be larger than y/ρ for cycle 2 to be at least of length y.

The proof of the above proposition relies on the relation between the asymptotic behavior of regularly varying distributions and the behavior of the LST near the origin. For clearness of exposition, we restate that result (see [26; 27; 36]).

Theorem 4.2. *Let X be a non-negative random variable, $L(\cdot)$ a slowly varying function, $\nu \in (m, m+1)$, and $D \geq 0$. Then the following statements are equivalent:*

$$\begin{aligned} \mathbb{P}(X > y) &= (D + o(1))y^{-\nu}L(y), \quad \text{as } y \to \infty, \\ \mathbb{E}\left[e^{-zX}\right] - \sum_{j=0}^{m} \mathbb{E}[X^j]\frac{(-z)^j}{j!} &= (-1)^m\Gamma(1-\nu)(D + o(1))z^{\nu}L(1/z), \quad \text{as } z \downarrow 0. \end{aligned}$$

To derive Proposition 4.3, we use the series expansion of $\mathbb{E}\left[e^{-zC_n}\right]$ and rely on Lemma 3 of [58] stating the series expansion of iterated functions that are regularly varying.

Proof of Proposition 4.3. We first show that $\mathbb{P}(C_2 > y) \sim \mathbb{P}(C_1 > y/\rho)$. The result for C_n then follows by induction. Now, first assume that the switch-over time at the end of the second cycle can be neglected, and let the corresponding cycle length be $\tilde{C}_n$. Let $t(z) = \sum_k \delta_k(z)$. Then $\mathbb{E}\, e^{-z\tilde{C}_2} = \gamma_1(t(z))$, which is precisely an iterated function of the type considered in [58, Lemma 3]. Using the definition of $\delta_k(\cdot)$, we obtain $t(z) = \rho z + o(z)$, when $y \to \infty$. In the notation of [58], we have $\psi_\nu = 0$ (due to Assumption 4.1), $\phi_\nu = (-1)^m\Gamma(1-\nu)$ (due to Theorem 4.2), and $\psi_1 = \rho$ (due to

the series expansion of $t(z)$ above). Using [58, Lemma 3], yields

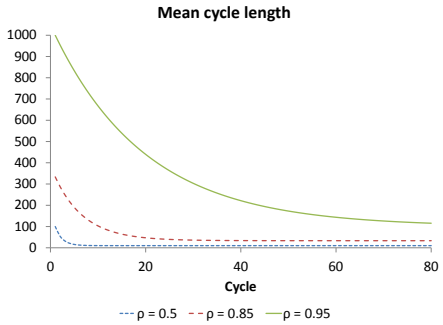
$$\mathbb{E}\left[e^{-z\tilde{C}_2}\right] = \gamma_1(t(z)) = \sum_{i=0}^{m} \theta_i z^i + ((-1)^m \Gamma(1-\nu) + o(1))\rho^\nu z^\nu L(1/z).$$

Another application of Theorem 4.2 provides the desired asymptotic tail probability of C_2 in the case of zero switch-over times. From Feller [67], p. 271, and Assumption 4.1 it follows that $\mathbb{P}(\tilde{C}_2 + \sum_i S_i > y) \sim \mathbb{P}(\tilde{C}_2 > y)$. Hence, the switch-over times are negligible and the proof is completed. □

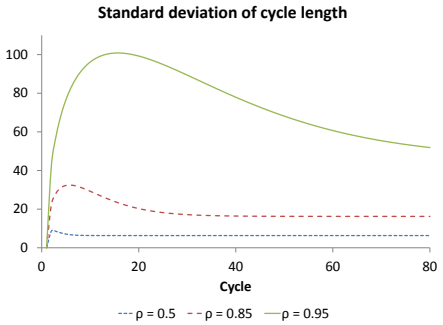
4.3.2 Numerical results

In this section we show the impact of the first cycle on the succeeding cycles by plotting mean cycle lengths, standard deviations and correlations between different cycles. The parameters that are needed, are the first two moments of the total switch-over times, the total arrival rate and the overall load, the second moment of the service distribution of an arbitrary customer and the first two moments of the first cycle length. Let us consider the following system; switch-over times are exponentially distributed, with parameters $\mathbb{E}[S] = 5$ and $\mathbb{E}[S^2] = 50$. The total arrival rate to the system is $\Lambda = 3$, the service times have an exponential distribution with $\mathbb{E}[B^2] = \frac{2}{3}\rho^2$. The first cycle is deterministic, with length $\frac{10\,\mathbb{E}[S]}{1-\rho}$, so it is 10 times longer than an average cycle. The means and standard deviations of the cycle lengths for different values of ρ are plotted in Figure 4.1. Figure 4.1a shows that the mean cycle length decreases to the length of an average cycle, for lower loads this decrease is faster than for higher loads. Figure 4.1b illustrates that the standard deviation first increases and then decreases, until it converges to the standard deviation of an average cycle. The low standard deviation in the beginning is explained by the fact that the first cycle is deterministic.

The correlation between cycles 1 and x is given by the covariance between cycles 1 and x, Equation (4.11), divided by the standard deviation of cycle 1 times the standard deviation of cycle x. Suppose that the first cycle is an average cycle, then both standard deviations are equal, so the covariance is divided by the variance of an average cycle. It is immediately clear that the correlation between cycles 1 and x is then given by ρ^{x-1}. This is plotted in Figure 4.2. Figure 4.2a shows that the correlation between cycles 1 and 2 is equal to ρ; for cycles that are further apart the correlation decreases rapidly unless the systems load is considerable. For instance, the correlation between cycles 1 and 50 is only significant if the load well exceeds 0.9. Figure 4.2b illustrates how the correlation decreases for fixed ρ by letting the distance between cycles increase. Note that when the load is equal to 1, the correlation does not decrease, which would give a horizontal line.

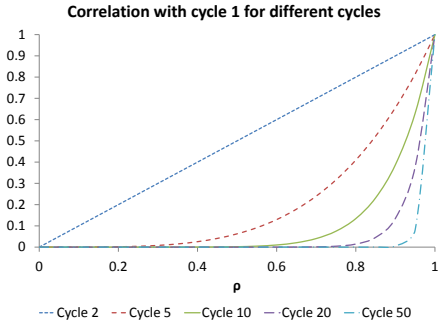


(a) Mean length per cycle.

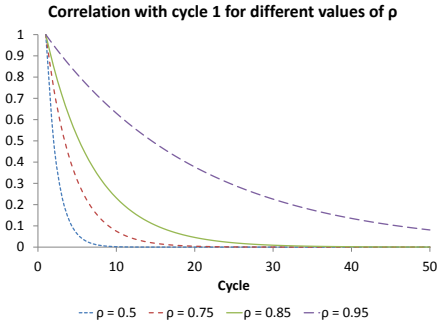


(b) Standard deviation per cycle.

Figure 4.1: Mean and standard deviation per cycle, for different values of ρ, if the first cycle is ten times the average length.



(a) Correlation as function of ρ.



(b) Correlation as function of x.

Figure 4.2: The correlation with cycle one, as function of ρ for different values of x and as function of x for different values of ρ.

4.4 Analysis of gated service

First, we note that the structure and methods of proof in this section are similar to those for the globally gated case of Section 4.3. Recall that with the gated service policy, the server only serves the customers that were present at the queue, the moment the server arrived. Analogously to the globally gated case, we again assume that the probabilistic behavior of the first cycle is known. But for gated it is not sufficient to have information about the length of the cycle, we need information about the lengths of the first N visits and switch-overs. The visit time plus the following switch-over time, i.e. $V_i + S_i$, is called the residence time. Assuming the first N residence times are probabilistically known, we can write down the joint LST of the first x residence times, for $x \geq N$. This LST is denoted by $\hat{\gamma}_x(\mathbf{z})$, and is given in the following theorem:

Theorem 4.3. *For $x \geq N$, we have*

$$\begin{aligned}\hat{\gamma}_x(\mathbf{z}) &= \mathbb{E}[e^{-\sum_{i=1}^{x} z_i(V_i+S_i)}] \\ &= \hat{\gamma}_N(z_1 + d_1^{(x-1)}(\mathbf{z}), \ldots, z_N + d_N^{(x-N)}(\mathbf{z})) \prod_{k=1}^{x-N} S_{\tilde{k}}^*(z_{N+k} + d_{N+k}^{(x-N-k)}(\mathbf{z})),\end{aligned}$$

with $\tilde{x} = x \mod N$ and

$$d_n^{(i)}(\mathbf{z}) = \sum_{k=n+1}^{N+n} \delta_{\tilde{k}}(z_k + d_k^{(i+n-k)}(\mathbf{z}))\mathbb{1}_{\{N<k\leq n+i\}}.$$

Note that the vector $\mathbf{z}$ has length x, and every element z_i of the vector now corresponds to the ith residence time for $i = 1, \ldots, x$. The recursive term $d_n^{(i)}(\mathbf{z})$ is similar to the ζ of the globally gated case. In order to prove Theorem 4.3, we first establish the following lemma (with the proof deferred to Section 4.5.1).

Lemma 4.2. *For $n = 1, 2, \ldots$ fixed and $x \geq n$ and $x \geq N$, we have*

$$(z'_n + d_n^{(x-n)}(\mathbf{z}')) = z_n + d_n^{(x-n+1)}(\mathbf{z}), \tag{4.12}$$

with $\mathbf{z}' = \mathbf{z} + d_x^{(1)}(\mathbf{z}) \sum_{i=x-N+1}^{x} \mathbf{e}_i$

Proof of Theorem 4.3. For the proof of Theorem 4.3, we use induction. First we take the case $x = N$, giving

$$\begin{aligned}\hat{\gamma}_N(\mathbf{z}) &= \hat{\gamma}_N(z_1 + d_1^{(N-1)}(\mathbf{z}), \ldots, z_N + d_N^{(N-N)}(\mathbf{z})) \prod_{k=1}^{0} S_{\tilde{k}}^*(z_{N+k} + d_{N+k}^{(-k)}(\mathbf{z})) \\ &= \hat{\gamma}_N(z_1, \ldots, z_N),\end{aligned}$$

because $d_n^{(N-n)}(\mathbf{z}) = 0$ for all n, and the product of switch-overs is empty and thus equals 1. Next we assume that Theorem 4.3 holds for all $k \leq x$, then for $k = x+1$, we have

$$\begin{aligned}
\hat{\gamma}_{x+1}(\mathbf{z}) &= \hat{\gamma}_x(z_1, \ldots, z_{x-N}, z_{x-N+1} + d_x^{(1)}, \ldots, z_x + d_x^{(1)}) S^*_{\widetilde{x+1}}(z_{x+1}) \\
&= \left(\hat{\gamma}_N(z'_1 + d_1^{(x-1)}(\mathbf{z}'), \ldots, z'_N + d_N^{(x-N)}(\mathbf{z}')) \prod_{k=1}^{x-N} S^*_{\tilde{k}}(z'_{N+k} + d_{N+k}^{(x-N-k)}(\mathbf{z}')) \right) \\
&\qquad \times S^*_{\widetilde{x+1}}(z_{x+1}) \\
&= \hat{\gamma}_N(z_1 + d_1^{(x)}(\mathbf{z}), \ldots, z_N + d_N^{(x-N+1)}(\mathbf{z})) \prod_{k=1}^{x-N} S^*_{\tilde{k}}(z_{N+k} + d_{N+k}^{(x-N-k+1)}(\mathbf{z})) \\
&\qquad \times S^*_{\widetilde{x+1}}(z_{x+1}) \\
&= \hat{\gamma}_N(z_1 + d_1^{(x)}(\mathbf{z}), \ldots, z_N + d_N^{(x-N+1)}(\mathbf{z})) \prod_{k=1}^{x-N+1} S^*_{\tilde{k}}(z_{N+k} + d_{N+k}^{(x-N-k+1)}(\mathbf{z})).
\end{aligned}$$

For the first equality, we use the same reasoning that was used for Equation (4.2). For the second equality, the induction hypothesis is used. The third equality uses Lemma 4.2. For the last equality, note that $\widetilde{x+1} = x+1 \mod N$, which equals $\widetilde{x-N+1} = x-N+1 \mod N$. This completes the proof. □

As in Section 4.3, we define $J \subseteq \{0, 1, 2, \ldots, x\}$ as the set of residence times we wish to include. Specifically, we take z_j equal to z if $j \in J$, and 0 otherwise. For example, if we want to calculate the distribution of the length of the first residence time in the second cycle we set $z_{N+1} = z$ and all other z_i equal to 0. Moments can now again be derived by differentiating with respect to z. The derivative of $\hat{\gamma}_x$ with respect to z (with z_i either z or 0) is then given by:

$$\begin{aligned}
\frac{d}{dz}\hat{\gamma}_x(\mathbf{z}) = &\sum_{j=1}^{N} \hat{\gamma}_N^{(j)}(z_1 + d_1^{(x-1)}(\mathbf{z}), \ldots, z_N + d_1^{(x-N)}(\mathbf{z})) \left(\mathbb{1}_{\{j \in J\}} + \frac{d}{dz} d_j^{(x-j)}(\mathbf{z}) \right) \\
&\times \prod_{k=1}^{x-N} S^*_{\tilde{k}}(z_{N+k} + d_{N+k}^{(x-N-k)}(\mathbf{z})) \\
&+ \hat{\gamma}_N(z_1 + d_1^{(x-1)}(\mathbf{z}), \ldots, z_N + d_N^{(x-N)}(\mathbf{z})) \\
&\times \sum_{j=1}^{x-N} \frac{d}{dz} S^*_{\tilde{j}}(z_{N+j} + d_{N+j}^{(x-N-j)}(\mathbf{z})) \left(\mathbb{1}_{\{N+j \in J\}} + \frac{d}{dz} d_{N+j}^{(x-N-j)}(\mathbf{z}) \right) \\
&\times \prod_{k \neq j} S^*_{\tilde{k}}(z_{N+k} + d_{N+k}^{(x-N-k)}(\mathbf{z})),
\end{aligned} \tag{4.13}$$

where $\hat{\gamma}_N^{(j)}(\cdot)$ is the partial derivative of $\hat{\gamma}_N(\cdot)$, with respect to the jth parameter. The

derivative of $d_j^{(x-j)}(\mathbf{z})$ is recursively defined as

$$\frac{d}{dz}d_j^{(x-j)}(\mathbf{z}) = \sum_{k=j+1}^{N+j} \frac{d}{dz}\delta_{\tilde{k}}(z_k + d_k^{(x-k)}(\mathbf{z}))\left(\mathbb{1}_{\{k\in J\}} + \frac{d}{dz}d_k^{(x-k)}(\mathbf{z})\right)\mathbb{1}_{\{N<k\leq x\}}. \tag{4.14}$$

For convenience, we define $\alpha_j = \frac{d}{dz}d_j^{(x-j)}(\mathbf{z})|_{\{z=0\}}$. The expected length of residence times under consideration (that is, in set J) is given by the following proposition.

Proposition 4.4 (First moments)**.** *The expected total residence time for set J is*

$$\mathbb{E}\left[\sum_{j\in J} V_j + S_j\right] = \sum_{j=1}^{N}\left(\mathbb{E}[\hat{V}_j] + \mathbb{E}[\hat{S}_j]\right)\left(\mathbb{1}_{\{j\in J\}} + \alpha_j\right) + \sum_{j=1}^{x-N}\mathbb{E}[S_{\tilde{j}}]\left(\mathbb{1}_{\{N+j\in J\}} + \alpha_{N+j}\right),$$

with

$$\alpha_j = \sum_{k=j+1}^{N+j}\rho_{\tilde{k}}\left(\mathbb{1}_{\{k\in J\}} + \alpha_k\right)\mathbb{1}_{\{N<k\leq x\}}. \tag{4.15}$$

Example 4.1. An interesting special case is where we are only interested in two subsequent cycles, i.e. $x = 2N$. Take $J = \{N+1, \ldots, x\}$. From (4.15) it may be easily verified that

$$\alpha_i = \sum_{j=N+1}^{N+i}\rho_{\tilde{j}} + \rho_{\tilde{j}}\sum_{k=j+1}^{x}\rho_{\tilde{k}}\prod_{l=j+1}^{k-1}(1+\rho_{\tilde{l}}). \tag{4.16}$$

For the second moment we need to differentiate Equation (4.13), the result can be found in Section 4.5.2, Equation (4.21). Taking $z = 0$ yields the second moments.

Proposition 4.5 (Second moments)**.** *The second moment of the total residence time*

for set J is

$$\mathbb{E}\left[\left(\sum_{j\in J}(V_j+S_j)\right)^2\right] = \sum_{j=1}^{N}\sum_{l=1}^{N}\mathbb{E}[(\hat{V}_j+\hat{S}_j)(\hat{V}_l+\hat{S}_l)]\left(\mathbb{1}_{\{l\in J\}}+\alpha_l\right)\left(\mathbb{1}_{\{j\in J\}}+\alpha_j\right)$$
$$-\sum_{j=1}^{N}\mathbb{E}[(\hat{V}_j+\hat{S}_j)]\frac{d^2}{dz^2}d_j^{(x-j)}(\mathbf{z})|_{\{z=0\}}$$
$$+2\sum_{j=1}^{N}\mathbb{E}[(\hat{V}_j+\hat{S}_j)]\left(\mathbb{1}_{\{j\in J\}}+\alpha_j\right)\sum_{k=1}^{x-N}\mathbb{E}[S_{\tilde{k}}]\left(\mathbb{1}_{\{N+k\in J\}}+\alpha_{N+k}\right)$$
$$+\sum_{j=1}^{x-N}\Bigg[\mathbb{E}[S_{\tilde{j}}^2]\left(\mathbb{1}_{\{N+j\in J\}}+\alpha_{N+j}\right)^2-\mathbb{E}[S_{\tilde{j}}]\frac{d^2}{dz^2}d_{N+j}^{(x-N-j)}(\mathbf{z})|_{\{z=0\}}$$
$$+\mathbb{E}[S_{\tilde{j}}]\left(\mathbb{1}_{\{N+j\in J\}}+\alpha_{N+j}\right)\sum_{k\neq j}\mathbb{E}[S_{\tilde{k}}]\left(\mathbb{1}_{\{N+k\in J\}}+\alpha_{N+k}\right)\Bigg],$$

with α_j given by (4.15) *and $\frac{d^2}{dz^2}d_j^{(x-j)}(\mathbf{z})|_{\{z=0\}}$ recursively given by*

$$\frac{d^2}{dz^2}d_j^{(x-j)}(\mathbf{z})|_{\{z=0\}} =$$
$$\sum_{k=j+1}^{N+j}\left[-\lambda_{\tilde{k}}\,\mathbb{E}[B_{\tilde{k}}^2]\left(\mathbb{1}_{\{k\in J\}}+\alpha_k\right)^2+\rho_{\tilde{k}}\frac{d^2}{dz^2}d_k^{(x-k)}(\mathbf{z})|_{\{z=0\}}\right]\mathbb{1}_{\{N<k\leq x\}}.$$

Remark 4.3 (Comparison with globally gated)**.** For globally gated we are able to find closed-form expressions for both the mean and the second moment. For gated this turned out to be involved, so the recursive terms are left intact. Assuming the initial cycle to be in steady-state, the relations between the residence times of the second cycle expressed in terms of the first cycle give rise to a system of equations. We refer to e.g. Takagi [125] for such an approach for the analysis of the number of customers at polling instants.

By adding more parameters to the joint LST of the residence times, the visit times and switch-over times can be tracked separately. Analogously to the gated case, the globally gated case can be extended to also record the visit and switch-over times per cycle. This can be useful, for example, for determining the output process.

4.4.1 Asymptotic properties

The asymptotic results are similar to those for the globally gated case. However, the expressions are more involved and concern iteratively defined functions. Below, we focus on the two most interesting cases: heavy traffic and heavy-tailed initial cycle length.

Heavy traffic For the heavy-traffic regime, we consider the usual scaling of residence times $\tilde{R}_i := \tilde{V}_i + \tilde{S}_i = (1-\rho)(V_i + S_i)$. For each variable x that is a function of ρ, we denote its value evaluated at $\rho = 1$ by $\hat{x}$.

Proposition 4.6. *For the joint LST of scaled residence times, we have*

$$\lim_{\rho\uparrow 1} \hat{\gamma}_x(\mathbf{z}(1-\rho)) = \hat{\gamma}_N^{\mathrm{HT}}(z_1 + \hat{\alpha}_1, \ldots, z_N + \hat{\alpha}_N),$$

with $\hat{\gamma}_N^{\mathrm{HT}}(\cdot)$ *the joint transform of scaled initial residence times, and* $\hat{\alpha}_j$ *given by the recursion*

$$\hat{\alpha}_j = \sum_{k=j+1}^{N+j} \hat{\rho}_{\tilde{k}}(z_k + \hat{\alpha}_k)\mathbb{1}_{\{N<k\le x\}}. \tag{4.17}$$

Proof. The heavy-traffic limit follows from

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \hat{\gamma}_x(\mathbf{z}(1-\rho)) = \\
&\quad \lim_{\rho\uparrow 1} \hat{\gamma}_N(z_1(1-\rho) + d_1^{(x-1)}(\mathbf{z}(1-\rho)), \ldots, z_N(1-\rho) + d_N^{(x-N)}(\mathbf{z}(1-\rho))) \\
&\qquad \times \prod_{k=1}^{x-N} S_{\tilde{k}}^*(z_{N+k}(1-\rho) + d_{N+k}^{(x-N-k)}(\mathbf{z}(1-\rho))).
\end{aligned}$$

Using Theorem 4.3 and by applying l'Hôpital's rule, we obtain

$$\begin{aligned}
\lim_{\rho\uparrow 1} \frac{d_n^{(i)}(\mathbf{z}(1-\rho))}{1-\rho} = \lim_{\rho\uparrow 1} \sum_{k=n+1}^{N} &\delta_{\tilde{k}}'(z_k(1-\rho) + d_k^{(i+n-k)}(\mathbf{z}(1-\rho))) \\
&\times \left(z_k + \frac{\mathrm{d}}{\mathrm{d}(1-\rho)} d_k^{(i+n-k)}(\mathbf{z}(1-\rho))\right) \mathbb{1}_{\{N<k\le n+i\}}.
\end{aligned}$$

The recursion (4.17) then follows by letting $\hat{\alpha}_j = \lim_{\rho\uparrow 1} d_n^{(i)}(\mathbf{z}(1-\rho))/(1-\rho)$. Combining the above provides the result. □

Proposition 4.6 shows that in heavy traffic, the joint LST of residence times is fully characterized by the LST of the initial cycle. As such, we see that an 'averaging principle' applies to our transient results, as the future evolution of residence times are specified by average input values. However, we note that, depending on the composition of the first cycle, the scaled cycle lengths may either increase or decrease; see Example 4.2 for an illustration.

Example 4.2. For simplicity, take $N = 2$ and $x = 4$. Similar to (4.16), we may obtain that $\hat{\alpha}_1 = \hat{\rho}_1 z_3 + \hat{\rho}_1\hat{\rho}_2 z_4$ and $\hat{\alpha}_2 = \hat{\rho}_1 z_3 + (1+\hat{\rho}_1)\hat{\rho}_2 z_4$, yielding

$$\begin{aligned}
&\lim_{\rho\uparrow 1} \hat{\gamma}_x(\mathbf{z}(1-\rho)) = \\
&\quad \mathbb{E}\left[\exp\{-z_1\tilde{R}_1 - z_2\tilde{R}_2 - z_3\hat{\rho}_1(\tilde{R}_1 + \tilde{R}_2) - z_4\hat{\rho}_2(\hat{\rho}_1\tilde{R}_1 + (1+\hat{\rho}_1)\tilde{R}_2)\}\right].
\end{aligned}$$

Addressing expected values, it holds in equilibrium that $(1-\hat{\rho}_1)\,\mathbb{E}[\tilde{R}_1^*] = \hat{\rho}_1\,\mathbb{E}[\tilde{R}_2^*]$. If we assume that the initial residence time of queue 1 is relatively short, i.e., that $(1-\hat{\rho}_1)\,\mathbb{E}[\tilde{R}_1^*] \leq \hat{\rho}_1\,\mathbb{E}[\tilde{R}_2^*]$, then it may be verified that $\mathbb{E}[\tilde{R}_3] \geq \mathbb{E}[\tilde{R}_1]$ and $\mathbb{E}[\tilde{R}_4] \leq \mathbb{E}[\tilde{R}_2]$. Moreover, for the total cycle length we have $\mathbb{E}[\tilde{R}_3 + \tilde{R}_4] \geq \mathbb{E}[\tilde{R}_1 + \tilde{R}_2]$. This means that the residence times converge to their equilibrium value, whereas the overall cycle length is increasing. For a relatively long initial residence time of queue 1, i.e., $(1-\hat{\rho}_1)\,\mathbb{E}[\tilde{R}_1^*] \geq \hat{\rho}_1\,\mathbb{E}[\tilde{R}_2^*]$, all inequalities are reversed.

Heavy-tailed initial cycle length Clearly, an excessively long residence time affects subsequent residence times. For convenience, we assume here that queue N (which may be arbitrary) has a heavy-tailed residence time. When the residence time of queue $i < N$ would have a heavy tail, this would affect the tail behavior of queues $i+1,\ldots,N$ as well, which is precisely the effect we aim to study. More precisely, we assume that the residence time of queue N is regular varying of index $-\nu$ (with $m < \nu < m+1$) and asymptotically dominates the tail of the residence times at queues $1,\ldots,N-1$, that is

$$\mathbb{P}(V_1+S_1 > \eta_1 y,\ldots,V_{N-1}+S_{N-1} > \eta_{N-1} y, V_N+S_N > y) \sim \mathbb{P}(V_N+S_N > y).$$

This leads to the following assumption.

Assumption 4.2. The initial cycle length is asymptotically dominated by the residence time of queue N, giving

$$\hat{\gamma}_N(f_1(z),\ldots,f_{N-1}(z),z) = \sum_{i=0}^{m} a_i z^i + (-1)^m \Gamma(1-\nu)(1+o(1)) z^\nu L(1/z).$$

As in the globally gated case, it holds that an excessively long n-th residence time is most likely due to an excessively long initial visit of queue N, while the system shows average behavior otherwise. The average behavior is specified by a similar recursive scheme as for the mean residence times and heavy-traffic asymptotics.

Proposition 4.7. *Suppose that $V_N \in \mathcal{R}_{-\nu}$ and Assumptions 4.1 and 4.2 are satisfied. Then,*

$$\mathbb{P}(V_x+S_x > y) \sim \mathbb{P}\left(V_N > y/\tilde{\alpha}_N\right),$$

with $\tilde{\alpha}_x = 1$ and $\tilde{\alpha}_N$ given by the recursion

$$\tilde{\alpha}_j = \sum_{k=j+1}^{N+j} \rho_{\tilde{k}} \tilde{\alpha}_k \mathbb{1}_{\{N<k\leq x\}}. \tag{4.18}$$

Proof. The proof is along the same line as the proof of Proposition 4.3. Consider the visit of queue x, i.e., take $z_x = z$ and $z_i = 0$ for $i = 1,\ldots,x-1$. First, we assume that all switch-over times equal zero, represented by $\tilde{V}_j$ in the notation for the visit time

of the j-th queue. From Theorem 4.3, we obtain $\mathbb{E}\, e^{-z\tilde{V}_x} = \hat{\gamma}_N(d_1^{(x-1)}, \ldots, d_N^{(x-N)})$. We thus need to consider the series expansion of the terms $d_j^{(x-j)}$, which are in fact iterated functions. In view of [58, Lemma 3] and Assumption 4.1, we have $\psi_\nu = 0$ (in the notation of [58]; see also the proof of Proposition 4.3) and we only need to consider the first term of the series expansion of $d_j^{(x-j)}$. Using Theorem 4.3, it easily follows that $d_j^{(x-j)} = \tilde{\alpha}_j z + o(z)$ where $\tilde{\alpha}_j$ follows from the recursion (4.18). Applying the same arguments as in [58, Lemma 3] combined with Assumption 4.2, we obtain

$$\mathbb{E}\left[e^{-z\tilde{V}_x}\right] = \hat{\gamma}_N(d_1^{(x-1)}, \ldots, d_N^{(x-N)}) = \sum_{i=0}^{m} \theta_i z^i + ((-1)^m \Gamma(1-\nu) + o(1))\tilde{\alpha}_N^\nu z^\nu L(1/z).$$

Using Theorem 4.2 yields $\mathbb{P}(\tilde{V}_x > y) \sim \mathbb{P}(V_N > y/\tilde{\alpha}_N)$. Again, under Assumption 4.1, it is straightforward to show that the contribution of switch-over times and the contribution of work arriving during these switch-over times are negligible. This completes the proof. □

4.4.2 Numerical results

In this section we use a numerical example to illustrate the properties of a gated system. Consider a symmetric system with $N = 4$ queues, $\lambda_i = 1$, $\mathbb{E}[B_i] = \rho_i$, $\mathbb{E}[B_i^2] = 2\rho_i^2$, $\rho_i = \rho/N$, $\mathbb{E}[S_i] = \frac{5}{4}$, and $\mathbb{E}[S_i^2] = \frac{25}{8}$, for all i. Because of a disaster, all visit times take ten times longer than average, but switch-over times are not affected, so $\mathbb{E}[(\hat{V}_i + \hat{S}_i)] = 10\frac{\rho\,\mathbb{E}[S_i]}{1-\rho} + \mathbb{E}[S_i]$. We take the residence times deterministic, giving $\mathbb{E}[(\hat{V}_i + \hat{S}_i)(\hat{V}_j + \hat{S}_j)] = \mathbb{E}[(\hat{V}_i + \hat{S}_i)]^2$, for $i, j = 1, 2, 3, 4$.

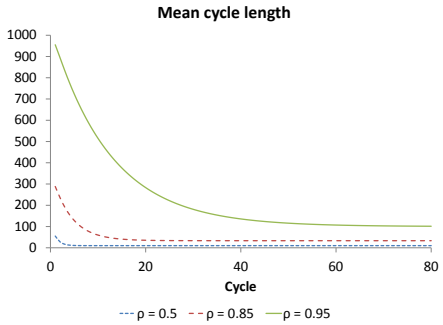
The mean and standard deviation of the cycle lengths of this system are plotted in Figure 4.3. We see that the figures look very similar to the globally gated case in Figure 4.1. Figure 4.4 shows the correlation between cycle 1 and various other cycles, where the first cycle is distributed as an average cycle. These figures also look similar to the globally gated case in Figure 4.2. Recall that the correlation between cycle 1 and 2 was equal to ρ in the globally gated system, Figure 4.4a shows that this is not the case for the gated system, however, the difference is small.

4.5 Appendix

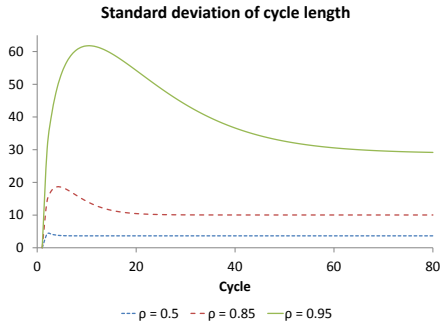
4.5.1 Proofs

Proof of Lemma 4.1. We prove Lemma 4.1 by induction. First we show that the lemma holds for $i = 2$:

$$\zeta_n^{(2)}(\mathbf{z}') = \sum_{k=1}^{N} \delta_k(z_{n+1} + \zeta_{n+1}^{(2)}(\mathbf{z}) + \zeta_{n+1}^{(1)}(\mathbf{z})) = \zeta_n^{(3)}(\mathbf{z}).$$

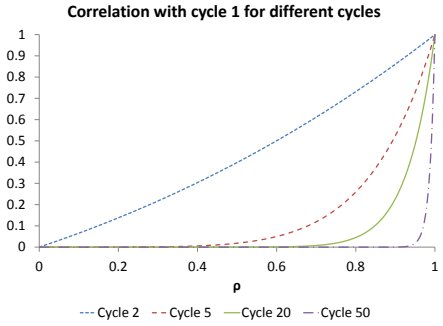


(a) Mean length per cycle.

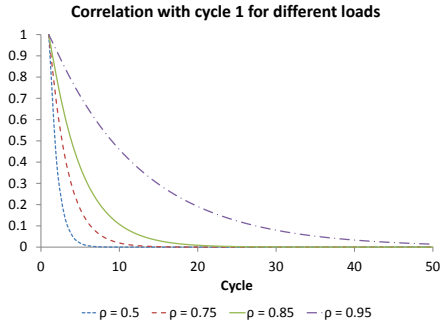


(b) Standard deviation per cycle.

Figure 4.3: Mean and standard deviation per cycle, for different values of ρ, if the first cycle is ten times normal length and deterministic.



(a) Correlation as function of ρ.



(b) Correlation as function of x.

Figure 4.4: The correlation with cycle one, as function of ρ for different values of x and as function of x for different values of ρ.

Both equalities follow from the definition of ζ, given in Equation (4.4). Now assume that (4.5) is true for all $k \leq i$. We have for $k = i + 1$:

$$\begin{aligned}\zeta_n^{(i+1)}(\mathbf{z}') &= \sum_{k=1}^{N} \delta_k(z_{n+1} + \zeta_{n+1}^{(i)}(\mathbf{z}')) \\ &= \sum_{k=1}^{N} \delta_k(z_{n+1} + \zeta_{n+1}^{(i+1)}(\mathbf{z})) \quad = \quad \zeta_n^{(i+2)}(\mathbf{z}).\end{aligned}$$

The first and the last equality use the definition of ζ, and for the second equality, the induction hypotheses is used. This completes the proof. □

Proof of Lemma 4.2. We prove Lemma 4.2 using induction. First we show that the lemma holds for $n = x$, because then $x - n = 0$, providing

$$(z'_x + d_x^{(0)}(\mathbf{z}')) = z_x + d_x^{(1)}(\mathbf{z}) + 0.$$

This follows from the fact that z_x is the last element of $\mathbf{z}$ and $d_x^{(0)} = 0$. Now assume that Equation (4.12) holds for all $k \leq x - n$. For $k + 1 = x - n$, we need to consider two cases: (i) the case where z_{x-k-1} is not one of the last N elements of $\mathbf{z}$ and (ii) the case where it is. For case (i) we have $x - k - 1 < x - N + 1$, thus $k > N - 2$, giving

$$\begin{aligned}(z'_{x-k-1} + d_{x-k-1}^{(k+1)}(\mathbf{z}')) &= z_{x-k-1} + \sum_{l=x-k}^{N+x-k-1} \delta_{\tilde{l}}(z'_l + d_l^{(x-l)}(\mathbf{z}')) \\ &= z_{x-k-1} + \sum_{l=x-k}^{N+x-k-1} \delta_{\tilde{l}}(z_l + d_l^{(x-l+1)}(\mathbf{z}))\mathbb{1}_{\{N<l\leq x\}} \\ &= z_{x-k-1} + d_{x-k-1}^{(k+2)}(\mathbf{z}).\end{aligned}$$

For the first equality, the definition of $d_n^{(i)}(\mathbf{z})$ is used, the second equality uses the induction hypotheses and the final equality uses the definition again. Note that $N + x - k - 1 \leq x$ and thus also $N + x - k - 1 \leq x + 1$, so the last equality is indeed true. For the second case, we have

$$\begin{aligned}(z'_{x-k-1} + d_{x-k-1}^{(k+1)}(\mathbf{z}')) &= z_{x-k-1} + d_x^{(1)}(\mathbf{z}) + \sum_{l=x-k}^{N+x-k-1} \delta_{\tilde{l}}(z'_l + d_l^{(x-l)}(\mathbf{z}'))\mathbb{1}_{\{N<l\leq x\}} \\ &= z_{x-k-1} + \delta_{\widetilde{x+1}}(z_{x+1} + d_{x+1}^{(0)}(\mathbf{z})) + \sum_{l=x-k}^{N+x-k-1} \delta_{\tilde{l}}(z_l + d_l^{(x-l+1)}(\mathbf{z}))\mathbb{1}_{\{N<l\leq x\}} \\ &= z_{x-k-1} + \sum_{l=x-k}^{N+x-k-1} \delta_{\tilde{l}}(z_l + d_l^{(x-l+1)}(\mathbf{z}))\mathbb{1}_{\{N<l\leq x+1\}} = z_{x-k-1} + d_{x-k-1}^{(k+2)}(\mathbf{z}).\end{aligned}$$

The first equality uses the definition of $d_n^{(i)}(\mathbf{z})$, whereas the second equality also uses the induction hypothesis. For the third equality we use the fact that the indicator function determines the end of the summation at x and the extra term contains the case were $k = x + 1$. The final equality follows again from the definition, completing the proof. □

4.5.2 Second-order derivatives

For globally gated, the second derivative of $\gamma_x(\mathbf{z})$ given in Equation (4.3), is given by

$$\begin{aligned}
\frac{d^2}{dz^2}\gamma_x(\mathbf{z}) &= \frac{d^2}{dz^2}\gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}))\left(\mathbb{1}_{\{1\in J\}} + \frac{d}{dz}\zeta_1^{(x)}(\mathbf{z})\right)^2 \prod_{j=1}^{x-1} S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z})) \\
&+ 2\frac{d}{dz}\gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}))\left(\mathbb{1}_{\{1\in J\}} + \frac{d}{dz}\zeta_1^{(x)}(\mathbf{z})\right) \\
&\quad \times \sum_{j=1}^{x-1} \frac{d}{dz}S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}))\left(\mathbb{1}_{\{j+1\in J\}} + \frac{d}{dz}\zeta_{j+1}^{(x-j)}(\mathbf{z})\right)\prod_{k\neq j}^{x-1} S^*(z_{k+1} + \zeta_{k+1}^{(x-k)}(\mathbf{z})) \\
&+ \frac{d}{dz}\gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z}))\frac{d^2}{dz^2}\zeta_1^{(x)}(\mathbf{z})\prod_{j=1}^{x-1} S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z})) \\
&+ \gamma_1(z_1 + \zeta_1^{(x)}(\mathbf{z})) \qquad (4.19) \\
&\quad \times \sum_{j=1}^{x-1}\left[\frac{d^2}{dz^2}S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}))\left(\mathbb{1}_{\{j+1\in J\}} + \frac{d}{dz}\zeta_{j+1}^{(x-j)}(\mathbf{z})\right)^2\prod_{k\neq j}^{x-1} S^*(z_{k+1} + \zeta_{k+1}^{(x-k)}(\mathbf{z}))\right. \\
&+ \frac{d}{dz}S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}))\frac{d^2}{dz^2}\zeta_{j+1}^{(x-j)}(\mathbf{z})\prod_{k\neq j}^{x-1} S^*(z_{k+1} + \zeta_{k+1}^{(x-k)}(\mathbf{z})) \\
&+ \frac{d}{dz}S^*(z_{j+1} + \zeta_{j+1}^{(x-j)}(\mathbf{z}))\left(\mathbb{1}_{\{j+1\in J\}} + \frac{d}{dz}\zeta_{j+1}^{(x-j)}(\mathbf{z})\right) \\
&\quad \left.\times \sum_{k\neq j}^{x-1}\frac{d}{dz}S^*(z_{k+1} + \zeta_{k+1}^{(x-k)}(\mathbf{z}))\left(\mathbb{1}_{\{k+1\in J\}} + \frac{d}{dz}\zeta_{k+1}^{(x-k)}(\mathbf{z})\right)\prod_{l\neq k}^{x-1} S^*(z_{l+1} + \zeta_{l+1}^{(x-l)}(\mathbf{z}))\right].
\end{aligned}$$

The second derivative of $\zeta_n^{(i)}(\mathbf{z})$, given in (4.4), is equal to

$$\begin{aligned}
\frac{d^2}{dz^2}\zeta_n^{(i)}(\mathbf{z}) = \sum_{k=1}^{N}\Bigg[&\frac{d^2}{dz^2}\delta_k(z_{n+1} + \zeta_{n+1}^{(i-1)}(\mathbf{z}))\left(\mathbb{1}_{\{n+1\in J\}} + \frac{d}{dz}\zeta_{n+1}^{(i-1)}(\mathbf{z})\right)^2 \\
&+ \frac{d}{dz}\delta_k(z_{n+1} + \zeta_{n+1}^{(i-1)}(\mathbf{z}))\frac{d^2}{dz^2}\zeta_{n+1}^{(i-1)}(\mathbf{z})\Bigg]. \qquad (4.20)
\end{aligned}$$

For gated, the second derivative of $\hat{\gamma}_x(\mathbf{z})$, given in Theorem 4.3, is given by

$$
\begin{aligned}
&\frac{d^2}{dz^2}\hat{\gamma}_x(\mathbf{z}) = \\
&\sum_{j=1}^{N}\sum_{l=1}^{N}\hat{\gamma}_N^{(j,l)}(z_1+d_1^{(x-1)}(\mathbf{z}),\ldots,z_N+d_1^{(x-N)}(\mathbf{z}))\left(\mathbb{1}_{\{l\in J\}}+\frac{d}{dz}d_l^{(x-l)}(\mathbf{z})\right)\\
&\qquad\times\left(\mathbb{1}_{\{j\in J\}}+\frac{d}{dz}d_j^{(x-j)}(\mathbf{z})\right)\prod_{k=1}^{x-N}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\\
&+\sum_{j=1}^{N}\hat{\gamma}_N^{(j)}(z_1+d_1^{(x-1)}(\mathbf{z}),\ldots,z_N+d_1^{(x-N)}(\mathbf{z}))\frac{d^2}{dz^2}d_j^{(x-j)}(\mathbf{z})\prod_{k=1}^{x-N}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\\
&+2\sum_{j=1}^{N}\hat{\gamma}_N^{(j)}(z_1+d_1^{(x-1)}(\mathbf{z}),\ldots,z_N+d_1^{(x-N)}(\mathbf{z}))\left(\mathbb{1}_{\{j\in J\}}+\frac{d}{dz}d_j^{(x-j)}(\mathbf{z})\right)\\
&\qquad\times\sum_{k=1}^{x-N}\frac{d}{dz}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\left(\mathbb{1}_{\{N+k\in J\}}+\frac{d}{dz}d_{N+k}^{(x-N-k)}(\mathbf{z})\right)\\
&\qquad\qquad\times\prod_{l\neq k}S_{\tilde{l}}^*(z_{N+l}+d_{N+l}^{(x-N-l)}(\mathbf{z}))\\
&+\hat{\gamma}_N(z_1+d_1^{(x-1)}(\mathbf{z}),\ldots,z_N+d_N^{(x-N)}(\mathbf{z}))\qquad(4.21)\\
&\qquad\times\sum_{j=1}^{x-N}\left[\frac{d^2}{dz^2}S_{\tilde{j}}^*(z_{N+j}+d_{N+j}^{(x-N-j)}(\mathbf{z}))\left(\mathbb{1}_{\{N+j\in J\}}+\frac{d}{dz}d_{N+j}^{(x-N-j)}(\mathbf{z})\right)^2\right.\\
&\qquad\qquad\times\prod_{k\neq j}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\\
&\qquad+\frac{d}{dz}S_{\tilde{j}}^*(z_{N+j}+d_{N+j}^{(x-N-j)}(\mathbf{z}))\frac{d^2}{dz^2}d_{N+j}^{(x-N-j)}(\mathbf{z})\prod_{k\neq j}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\\
&\qquad+\frac{d}{dz}S_{\tilde{j}}^*(z_{N+j}+d_{N+j}^{(x-N-j)}(\mathbf{z}))\left(\mathbb{1}_{\{N+j\in J\}}+\frac{d}{dz}d_{N+j}^{(x-N-j)}(\mathbf{z})\right)\\
&\qquad\qquad\times\sum_{k\neq j}\frac{d}{dz}S_{\tilde{k}}^*(z_{N+k}+d_{N+k}^{(x-N-k)}(\mathbf{z}))\left(\mathbb{1}_{\{N+k\in J\}}+\frac{d}{dz}d_{N+k}^{(x-N-k)}(\mathbf{z})\right)\\
&\qquad\qquad\qquad\left.\times\prod_{l\neq k}S_{\tilde{l}}^*(z_{N+l}+d_{N+l}^{(x-N-l)}(\mathbf{z}))\right],
\end{aligned}
$$

with $\frac{d^2}{dz^2} d_j^{(x-j)}(\mathbf{z})$, being the derivative of (4.14), recursively given by

$$\frac{d^2}{dz^2} d_j^{(x-j)}(\mathbf{z}) = \sum_{k=j+1}^{N+j} \left[\frac{d^2}{dz^2} \delta_{\tilde{k}}(z_k + d_k^{(x-k)}(\mathbf{z})) \left(\mathbb{1}_{\{k \in J\}} + \frac{d}{dz} d_k^{(x-k)}(\mathbf{z}) \right)^2 + \frac{d}{dz} \delta_{\tilde{k}}(z_k + d_k^{(x-k)}(\mathbf{z})) \frac{d^2}{dz^2} d_k^{(x-k)}(\mathbf{z}) \right] \mathbb{1}_{\{N<k\leq x\}}.$$

Note that $\frac{d}{dz} d_j^{(x-j)}(\mathbf{z})$ is also recursively defined and given in (4.14).

Chapter 5

Queue-length distributions in DPS queues with batch arrivals

5.1 Introduction

In this chapter, we study a queueing model with the Discriminatory Processor Sharing (DPS) service policy. In contrast to the polling models from the previous chapters (Chapters 2, 3 and 4), where the server only serves one job type, the server now serves all job types simultaneously regardless of their type. The customer classes are assigned with different weights indicating their relative priority (see Chapter 1 for a more elaborate description). The majority of papers on DPS only considers single arrivals, i.e., whenever an arrival occurs, only one new customer joins the system. In some applications it is possible that multiple customers arrive at the same time, where these customers could possibly belong to different classes. Such an arrival pattern is captured by allowing multi-class batch arrivals. Our result gives insight in the relation between the simultaneity of arrivals and the joint queue length in the DPS model.

The possibility of simultaneous arrivals of batches of different types strongly enhances the modeling capabilities of DPS models. Examples are found in communication networks. For instance, consider a Web server that needs to respond to numerous document-retrieval requests initiated by the end users. A Web document generally consists of a number of files (e.g., pieces of text, in-line images, audio or video files), each of which generates a separate file-retrieval request to the Web server. Each of these requests generates one or more data flows to be transferred via a multitude of connections (typically TCP-based connections with different characteristics) that compete for access to a shared medium. This way, a document request can be seen as a batch of flows that arrive simultaneously to a DPS node. Other examples can be found in computer systems where threads compete for access to shared processors in a processor sharing fashion. Efficient thread-spawning algorithms create batches of additional threads to reduce congestion during temporary overload situations, and

vice versa, can terminate threads when no longer needed. At the operating system level, different thread types may have different priorities. This way, the creation of threads can be seen as a batch of jobs arriving to a DPS node.

DPS models have received much attention in the literature; we refer to Altman et al. [7] for a survey on DPS queues. The (conditional) moments of the response times and number of customers in a DPS queue and their finiteness are studied in [11; 66; 82]. Analysis of overloaded regimes in PS and DPS models can be found in [6; 21; 64; 88]. DPS models in the heavy-traffic (HT) regime have been studied in [120; 141]. These papers assume single arrivals and the approach they follow differs from our approach. The analysis we follow builds on the study of Verloop et al. [145], who analyze a DPS queue with phase-type service time distributions by considering a Markovian framework. Their main result is the joint distribution of the scaled number of customers in the system in the HT regime. Grishechkin [75] allows for batch arrivals and explored the relationship between Processor Sharing models and Crump-Mode-Jagers branching processes. Recently, Izagirre et al. [87] proposed an approximation for the mean sojourn time in a DPS queue with Poisson arrivals and general service times by interpolating between heavy traffic and light traffic.

The motivation for this chapter is two-fold. First, it is of a fundamental interest to explicitly quantify the impact of correlations between the arrival processes of the different customers classes on the number of each type in the system. Using a specific class of correlation structures, we take a significant step in that direction. In fact, our results *explicitly quantify the impact of the simultaneity of the arrivals* on the joint distribution of the number of jobs in DPS systems in HT. Moreover, the result also leads to sharp approximations of the impact of batched arrivals for stable systems (i.e., with load strictly less than 1), providing new insight in the performance of DPS systems, a class of models that is notoriously hard to analyze in an exact manner. Second, in several applications of DPS models the arrival processes of the different job types are correlated (see the examples above). In view of those applications it is important to be able to predict the queueing behavior accurately, in particular when the system load is significant. The effectiveness of the existing numerical techniques (like simulations) tends to degrade strongly when the system is heavily loaded. This raises the need for the development of simple and fast approximations for the delay incurred at each of the queues, explicitly capturing the impact of correlated arrivals.

We study a DPS queue with batch arrivals that occur according to a Poisson process and exponential service times. Each arriving batch may contain customers of multiple types and the number of customers per type can be larger than one. The size of a batch is according to a general joint batch-size distribution. We are interested when the system is in HT. To obtain this, we scale the arrival rate and let the total load of the system go to 1. We analyze the scaled joint queue-length distribution and show that a state-space collapse occurs in the HT limit. More specifically, the joint distribution of the scaled number of customers is given by a vector of constants multiplied by an exponential distribution. This result is similar to the result of Verloop et al. [145] for

Poisson arrivals of single customers, where the authors find the same constant vector times an exponential distribution. The difference with [145] is in the parameter of the exponential distribution, which now contains the second moments and correlation structure of the batches. In the HT regime, the batch arrivals only affect the mean of the scaled joint queue-length distribution. For polling models a similar phenomenon is observed in, e.g., [135].

The remainder of this chapter is organized as follows. In Section 5.2 the model is described in detail and we introduce the notation. In Section 5.3, we state the main result and include some intuition. The proof of the main result is given in Section 5.4. Finally, in Section 5.5, we discuss numerical results.

5.2 Model description

We consider a system with N classes. Arrivals occur according to a Poisson process with rate λ. Each arrival consists of a batch of size $\mathbf{K} = (K_1, \ldots, K_N)$, where K_i stands for the number of class-i customers. Denote the joint batch-size distribution by $p(k_1, \ldots, k_N) = p(\mathbf{k}) := \mathbb{P}(K_1 = k_1, \ldots, K_N = k_N)$ and let the corresponding probability generating function (PGF) of $\mathbf{K}$ be $K(\mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_N)$ and $|z_i| < 1$, for $i = 1, \ldots, N$. The arrival rate of class-i customers is denoted by $\lambda_i := \lambda\,\mathbb{E}[K_i]$. Customers of type i have an exponentially distributed service requirement with mean $1/\mu_i$. The N customer types share a common resource of capacity 1. Associated with every class i, there is a strictly positive weight w_i, $i = 1, \ldots, N$. When there are $\mathbf{q} := (q_1, \ldots, q_N)$ customers present in the system, with q_i the number of type-i customers, each type-i customer is served at rate

$$\frac{w_i}{\sum_{j=1}^{N} w_j q_j}, \quad i = 1, \ldots, N.$$

We denote the random variable of the number of type-i customers in the system by Q_i and denote its joint steady-state distribution by $\pi(\mathbf{q}) := \mathbb{P}(\mathbf{Q} = \mathbf{q})$, with $\mathbf{Q} = (Q_1, \ldots, Q_N)$. The load of type-$i$ is given by

$$\rho_i := \frac{\lambda_i}{\mu_i},$$

and the load of the system is

$$\rho := \lambda \sum_{j=1}^{N} \frac{\mathbb{E}[K_j]}{\mu_j} = \sum_{j=1}^{N} \rho_j.$$

We analyze the system when it is near saturation, i.e., for $\rho \uparrow 1$. To obtain this regime we scale the arrival rate by letting

$$\lambda \uparrow \hat{\lambda} := \left(\sum_{i=1}^{N} \frac{\mathbb{E}[K_i]}{\mu_i} \right)^{-1}, \tag{5.1}$$

while keeping μ_i, $i = 1, \ldots, N$, and the batch-size distribution $p(\mathbf{k})$ fixed. In HT, the load per customer type is given by

$$\hat{\rho}_i = \frac{\hat{\lambda}_i}{\mu_i}, \quad i = 1, \ldots, N,$$

with $\hat{\lambda}_i = \hat{\lambda}\,\mathbb{E}[K_i]$. The total load is equal to $\hat{\rho} = \sum_{i=1}^{N} \hat{\rho}_i = 1$. We let $\mathbf{e}_i$ denote the ith unit vector.

5.3 Main result

In this section we state the main result. The proof of this result can be found in Section 5.4.

Theorem 5.1. *As $\rho \uparrow 1$, the joint distribution of the scaled queue lengths is given by*

$$(1-\rho)(Q_1, Q_2, \ldots, Q_N) \to_d (\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_N) =_d \left(\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_N}{w_N}\right) X, \qquad (5.2)$$

where X is exponentially distributed with mean

$$\mathbb{E}[X] = \frac{\sum_{j=1}^{N} \hat{\rho}_j \frac{1}{\mu_j} + \hat{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}[K_i K_j] \frac{1}{\mu_i} \frac{1}{\mu_j}}{2 \sum_{j=1}^{N} (\hat{\rho}_j / w_j) \frac{1}{\mu_j}}.$$

The intuition behind the result is as follows. Observe that the total amount of work is the same for any work-conserving M/G/1 queue and is exponentially distributed in HT. Hence, the total amount of work in a DPS queue in HT is also an exponential random variable X. From the theorem above, we see that there is balance in how the total amount of work is distributed among the N classes. This is reflected by the fact that the exponential random variable is multiplied by a constant vector; the different customer types all have their own portion of the exponential random variable equal to $\hat{\rho}_i / w_i$, $i = 1, \ldots, N$. The number of type-i customers grows with rate $\hat{\lambda}_i$ and is depleted with rate $w_i q_i \mu_i \left(\sum_{j=1}^{N} w_j q_j\right)^{-1}$. Due to the balance of type-i customers for a certain realization of X, these two rates should be equal. We can solve this equation for q_i to get: $q_i = (\hat{\rho}_i / w_i) \sum_{j=1}^{N} w_j q_j$. We see that $\sum_{j=1}^{N} w_j q_j$ is a constant common to all q_i, $i = 1, \ldots, N$, giving relative portions according to the vector $(\hat{\rho}_1 / w_1, \ldots, \hat{\rho}_N / w_N)$.

Observe that the joint batch arrivals only influence the deterministic vector in Theorem 5.1 through the load. Also, the random variable X remains exponential; the effect of the joint batch arrivals only appears in the mean of X. In case of single-class

arrival processes, it holds that $\mathbb{E}[K_i K_j] = 0$ if $i \neq j$. Rewriting $\mathbb{E}[X]$ for single and single-class arrivals, respectively, yields

$$\mathbb{E}[X] = \frac{\sum_{j=1}^{N} \hat{\rho}_j \frac{1}{\mu_j}}{\sum_{j=1}^{N} (\hat{\rho}_j / w_j) \frac{1}{\mu_j}} \qquad \text{(for single arrivals)}$$

$$\mathbb{E}[X] = \frac{\sum_{j=1}^{N} \hat{\rho}_j \frac{1}{\mu_j} \left(1 + \mathbb{E}[K_j^2] / \mathbb{E}[K_j]\right)}{2 \sum_{j=1}^{N} (\hat{\rho}_j / w_j) \frac{1}{\mu_j}} \qquad \text{(for single-type batch arrivals).}$$

Analogous to the derivation of Verloop et al. [145], we can extend our main result to phase type service-time distributions and even to a more general Markovian framework. In this framework, after service completion a customer of type i becomes a customer of type j with probability $\hat{p}_{ij}$, or the customer leaves the system with probability $\hat{p}_{i0}$. The service duration of a type-i customer is still exponential with rate μ_i, but now a customer has to complete multiple services with different service rates (because the customer changes type after a service completion).

Conjecture 5.1. *Consider the general Markovian framework described in [145]. The joint distribution of the scaled queue length $(\hat{Q}_1, \ldots, \hat{Q}_N)$ is given by*

$$(\hat{Q}_1, \ldots, \hat{Q}_N) =_d \left(\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_N}{w_N} \right) X,$$

with $\hat{\rho}_i$, $i = 1, \ldots, N$ the load corresponding to type-i customers and X exponentially distributed with mean

$$\mathbb{E}[X] = \frac{\sum_{j=1}^{N} \hat{\rho}_j \, \mathbb{E}[R_j] + \hat{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}[K_i K_j] \, \mathbb{E}[R_i] \, \mathbb{E}[R_j]}{2 \sum_{j=1}^{N} (\hat{\rho}_j / w_j) \, \mathbb{E}[R_j]},$$

where R_i is the remaining service durations of customers of type i, $i = 1, \ldots, N$.

5.4 Analysis

In this section we derive the limiting distribution of the number of customers in the queue in HT (i.e., when $\rho \uparrow 1$). To this end, we start by formulating the balance equations for the limiting distribution $\pi(\mathbf{q})$; these balance equations will be used to derive the functional equation. When we have the functional equations, we scale it with a factor $(1 - \rho)$ and take the limit $\rho \uparrow 1$. This leads to a partial differential equation. The solution to this equation gives the desired distribution up to a single random variable. The final step is finding this random variable.

5.4.1 Balance equations and functional equation

In this subsection, we derive the functional equation. First, we introduce a transformation that leads to more convenient expressions. Define

$$r(\mathbf{0}) = 0, \quad \text{and} \quad r(\mathbf{q}) = \frac{\pi(\mathbf{q})}{\sum_{j=1}^{N} w_j q_j}, \quad \text{for } \mathbf{q} \neq \mathbf{0}. \tag{5.3}$$

Let $P(\mathbf{z})$ and $R(\mathbf{z})$ denote the generating functions of $\pi(\mathbf{q})$ and $r(\mathbf{q})$, respectively. That is

$$P(\mathbf{z}) = \mathbb{E}\left[z_1^{Q_1} \dots z_N^{Q_N}\right] = \sum_{q_1=0}^{\infty} \cdots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \dots \cdot z_N^{q_N} \pi(\mathbf{q}),$$

and

$$R(\mathbf{z}) = \mathbb{E}\left[\frac{z_1^{Q_1} \cdot \dots \cdot z_N^{Q_N}}{\sum_{i=1}^{N} w_i Q_i \mathbb{1}_{\{\sum_{j=1}^{N} Q_j > 0\}}}\right] = \sum_{q_1=0}^{\infty} \dots \sum_{q_N=0}^{\infty} z_1^{q_1} \cdot \dots \cdot z_N^{q_N} r(\mathbf{q}).$$

The following lemma formulates a functional equation for $R(\mathbf{z})$:

Lemma 5.1. *For $\rho < 1$, a functional equation for $R(\mathbf{z})$ is given by*

$$\lambda(1-\rho)(1-K(\mathbf{z})) = \sum_{i=1}^{N} \left(\lambda z_i (K(\mathbf{z}) - 1) + \mu_i (1 - z_i)\right) w_i \frac{\partial}{\partial z_i} R(\mathbf{z}). \tag{5.4}$$

Proof. Assuming $\rho < 1$, the equilibrium distribution $\pi(\mathbf{q})$ satisfies the following balance equations

$$\lambda \pi(\mathbf{0}) = \sum_{i=1}^{N} \mu_i \pi(\mathbf{e}_i), \tag{5.5}$$

and, for $\mathbf{q} \neq \mathbf{0}$,

$$\left(\lambda + \frac{\sum_{i=1}^{N} w_i \mu_i q_i}{\sum_{i=1}^{N} w_i q_i}\right) \pi(\mathbf{q}) = \lambda \sum_{k_1=0}^{q_1} \cdots \sum_{k_N=0}^{q_N} p(\mathbf{k}) \pi(\mathbf{q} - \mathbf{k}) + \sum_{i=1}^{N} \mu_i \frac{w_i (q_i + 1)}{\sum_{j=1}^{N} w_j q_j + w_i} \pi(\mathbf{q} + \mathbf{e}_i).$$

Now we take the generating function, yielding

$$\begin{aligned}
&\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}\mathbb{1}_{\{\sum_{j=1}^N q_j>0\}}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\left(\lambda+\frac{\sum_{i=1}^N w_i\mu_i q_i}{\sum_{i=1}^N w_i q_i}\right)\pi(\mathbf{q})\\
&\quad=\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\lambda\sum_{k_1=0}^{q_1}\cdots\sum_{k_N=0}^{q_N}p(\mathbf{k})\pi(\mathbf{q}-\mathbf{k})\\
&\qquad+\sum_{i=1}^N\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}\mathbb{1}_{\{\sum_{j=1}^N q_j>0\}}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\mu_i\frac{w_i(q_i+1)}{\sum_{j=1}^N w_j q_j+w_i}\pi(\mathbf{q}+\mathbf{e}_i).
\end{aligned}$$

To get rid of the indicator functions, we add Equation (5.5), change the order of summation in the second line and start the corresponding summations at 0 by a change of variable. This leads to

$$\begin{aligned}
&\lambda\pi(\mathbf{0})+\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}\mathbb{1}_{\{\sum_{j=1}^N q_j>0\}}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\left(\lambda+\frac{\sum_{i=1}^N w_i\mu_i q_i}{\sum_{i=1}^N w_i q_i}\right)\pi(\mathbf{q})\\
&\quad=\lambda\sum_{k_1=0}^{\infty}\cdots\sum_{k_N=0}^{\infty}z_1^{k_1}\cdot\dots\cdot z_N^{k_N}p(\mathbf{k})\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\pi(\mathbf{q})\\
&\qquad+\sum_{i=1}^N\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\mu_i\frac{w_i(q_i+1)}{\sum_{j=1}^N w_j q_j+w_i}\pi(\mathbf{q}+\mathbf{e}_i).
\end{aligned}$$

We now apply the transformation from (5.3). Note that for the first term on the right-hand side, we have to take into account that $r(\mathbf{0})=0$, but $\pi(\mathbf{0})\neq 0$. Hence, we obtain

$$\begin{aligned}
&\lambda\pi(\mathbf{0})+\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\left(\lambda\sum_{i=1}^N w_i q_i+\sum_{i=1}^N w_i\mu_i q_i\right)r(\mathbf{q})\\
&\quad=\lambda K(\mathbf{z})\left(\pi(\mathbf{0})+\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}r(\mathbf{q})\sum_{i=1}^N w_i q_i\right)\\
&\qquad+\sum_{i=1}^N\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}\mu_i w_i(q_i+1)r(\mathbf{q}+\mathbf{e}_i).
\end{aligned}$$

Taking partial derivatives of $R(\mathbf{z})$ with respect to z_i, we get

$$\frac{\partial}{\partial z_i}R(\mathbf{z})=z_i^{-1}\sum_{q_1=0}^{\infty}\cdots\sum_{q_N=0}^{\infty}z_1^{q_1}\cdot\dots\cdot z_N^{q_N}q_i r(\mathbf{q}).$$

Using this, we can rewrite the functional equation as

$$\lambda\pi(\mathbf{0}) + \sum_{i=1}^{N}\left(\lambda w_i z_i \frac{\partial}{\partial z_i}R(\mathbf{q}) + \mu_i w_i z_i \frac{\partial}{\partial z_i}R(\mathbf{q})\right)$$
$$= \lambda K(\mathbf{z})\left(\pi(\mathbf{0}) + \sum_{i=1}^{N} w_i z_i \frac{\partial}{\partial z_i}R(\mathbf{q})\right) + \sum_{i=1}^{N}\mu_i w_i \frac{\partial}{\partial z_i}R(\mathbf{q}).$$

Rearranging the terms and using $\pi(\mathbf{0}) = 1-\rho$ completes the proof. □

5.4.2 Heavy-traffic limit

For convenience we use the change of variables $z_i = e^{-s_i}$, with $s_i > 0$, $i = 1,\ldots,N$. We use the notation $\mathbf{s} = (s_1,\ldots,s_N)$ and $e^{-(1-\rho)\mathbf{s}} = (e^{-(1-\rho)s_1},\ldots,e^{-(1-\rho)s_N})$. For the heavy-traffic limit, we define

$$\hat{R}(\mathbf{s}) = \mathbb{E}\left[\frac{1 - e^{-s_1\hat{Q}_1}\cdot\ldots\cdot e^{-s_N\hat{Q}_N}}{\sum_{j=1}^{N}\hat{Q}_j w_j}\mathbb{1}_{\{\sum_{j=1}^{N}\hat{Q}_j>0\}}\right]. \tag{5.6}$$

Now we can formulate the following lemma.

Lemma 5.2. *If* $\lim_{\rho\uparrow 1} P(e^{-(1-\rho)\mathbf{s}})$ *exists, then the function* $\hat{R}(\mathbf{s})$ *satisfies the following partial differential equation:*

$$0 = \sum_{i=1}^{N} F_i(\mathbf{s})\frac{\partial\hat{R}(\mathbf{s})}{\partial s_i} = \mathbf{F}(\mathbf{s})\nabla\hat{R}(\mathbf{s}), \quad \forall\ \mathbf{s}\geq\mathbf{0},$$

where $\mathbf{F}(\mathbf{s}) = (F_1(\mathbf{s}),\ldots,F_N(\mathbf{s}))$, *and*

$$F_i(\mathbf{s}) = w_i\left(\mu_i s_i - \hat{\lambda}\sum_{j=1}^{N} s_j\,\mathbb{E}[K_j]\right), \quad i = 1,\ldots,N,$$

with $\hat{\lambda}$ *as defined in* (5.1).

Proof. We divide both sides of (5.4) by $(1-\rho)$ and apply the change of variables. Note that $\lim_{\rho\uparrow 1}(1-K(e^{-(1-\rho)\mathbf{s}}))/(1-\rho) = \sum_{j=1}^{N} s_j\,\mathbb{E}[K_j]$. Taking the limit $\rho\uparrow 1$ gives the partial differential equation

$$0 = \sum_{i=1}^{N}\left(\mu_i s_i - \hat{\lambda}\sum_{j=1}^{N} s_j\,\mathbb{E}[K_j]\right) w_i\frac{\partial}{\partial s_i}\hat{R}(\mathbf{s}). \tag{5.7}$$

This completes the proof. □

The lemma above is similar to Lemma 2 in [145]; in our case $p_{ij} = 0$ if $j > 0$ and it turns out that $p_{0j} = \mathbb{E}[K_j]$ due to batch arrivals. The next step is to establish the state-space collapse. Due to the similarity between our functional equation in Lemma 5.2 and the functional equation in [145, Lemma 2], we can rely on Lemma 3 of [145]. Specifically, [145, Lemma 3] gives that $\hat{R}(\mathbf{s})$ is constant on an $(N-1)$-dimensional hyperplane, see [145] for a geometric interpretation. Essentially, this provides that the N-dimensional random vector of queue lengths reduces to a deterministic vector times a single random variable in heavy traffic. Applying [145, Lemma 3] and the subsequent analysis, we get

$$(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_N) =_d \left(\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_N}{w_N} \right) \frac{w_1}{\hat{\rho}_1} \hat{Q}_1, \tag{5.8}$$

Note that [145, Lemma 3] holds in our case, as its proof does not depend on the fact that the p_{0j} add up to 1, and we take p_{0j} equal to $\mathbb{E}[K_j]$, $j = 1, \ldots, N$. Equation (5.8) is now equivalent to (5.2), with X distributed as $\frac{w_1}{\hat{\rho}_1}\hat{Q}_1$. It remains to find the distribution of X.

5.4.3 Specifying the common distribution

The distribution of X is given in the following lemma.

Lemma 5.3. *X is exponentially distributed with mean*

$$\mathbb{E}[X] = \frac{\sum_{j=1}^{N} \hat{\rho}_j \frac{1}{\mu_j} + \hat{\lambda} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}[K_i K_j] \frac{1}{\mu_i} \frac{1}{\mu_j}}{2 \sum_{j=1}^{N} (\hat{\rho}_j / w_j) \frac{1}{\mu_j}}.$$

Proof. Denote by B the total amount of work that an arbitrary arriving batch brings into the system. From Kingman [94] we have that the total amount of work in the system W in the GI/GI/1 queue, when scaled by $(1-\rho)$, has a proper distribution as $\rho \uparrow 1$. In particular,

$$(1-\rho)W \to_d \hat{W},$$

where $\hat{W}$ is exponentially distributed with mean

$$\mathbb{E}[\hat{W}] = \frac{\mathbb{E}[B^2]}{2\,\mathbb{E}[B]}.$$

For our DPS model, we can represent the total workload as

$$W = \sum_{j=1}^{N} \sum_{h=1}^{Q_j} R_{j,h},$$

where $R_{j,h}$ is the remaining service requirement of the hth type-j customer. Since we have exponential service requirements, the remaining service requirements are

in distribution equal to the original service requirements: $R_{j,h} =_d B_{j,h}$, with $B_{j,h}$ exponentially distributed with mean $\mathbb{E}[B_j] = 1/\mu_j$. Using the representation of the total workload, we may write

$$(1-\rho)W = \sum_{j=1}^{N}(1-\rho)Q_j \times \frac{1}{Q_j}\sum_{h=1}^{Q_j} B_{j,h}.$$

Observe that $(1-\rho)Q_j \to X\hat{\rho}_j/w_j$ according to Theorem 5.1 and $\frac{1}{Q_j}\sum_{h=1}^{Q_j} B_{j,h} \to \mathbb{E}[B_j]$ due to the strong law of large numbers. This suggests that

$$\hat{W} = X\sum_{j=1}^{N}\frac{\hat{\rho}_j}{w_j}\,\mathbb{E}[B_j].$$

This equation is formally shown in [145, Equation (17)].

Combining the two expressions for $\mathbb{E}[\hat{W}]$ above gives

$$\frac{\mathbb{E}[B^2]}{2\,\mathbb{E}[B]} = \mathbb{E}[X]\sum_{j=1}^{N}\frac{\hat{\rho}_j}{w_j}\,\mathbb{E}[B_j],$$

and thus

$$\mathbb{E}[X] = \frac{\mathbb{E}[B^2]/(2\,\mathbb{E}[B])}{\sum_{j=1}^{N}(\hat{\rho}_j/w_j)\,\mathbb{E}[B_j]}.$$

Note that B can be rewritten as

$$B = \sum_{j=1}^{N}\sum_{i=1}^{K_j} B_{j,i}.$$

Using the law of total expectation, we derive the moments of B:

$$\mathbb{E}[B] = \mathbb{E}[\mathbb{E}[B|\mathbf{K}]] = \mathbb{E}\left[\sum_{j=1}^{N}\sum_{i=1}^{K_j}\mathbb{E}[B_{j,i}]\right] = \sum_{j=1}^{N}\mathbb{E}[K_j]\frac{1}{\mu_j} = \frac{1}{\hat{\lambda}},$$

and

$$\begin{aligned}
\mathbb{E}[B^2] &= \mathbb{E}\left[\mathbb{E}[B^2|\mathbf{K}]\right] = \mathbb{E}\left[\mathrm{Var}\left[B|\mathbf{K}\right] + \left(\mathbb{E}[B|\mathbf{K}]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{N}\sum_{i=1}^{K_j}\mathrm{Var}[B_{j,i}] + \sum_{i=1}^{N}\sum_{j=1}^{N}K_iK_j\,\mathbb{E}[B_i]\,\mathbb{E}[B_j]\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{N}K_j\frac{1}{\mu_j^2} + \sum_{i=1}^{N}\sum_{j=1}^{N}K_iK_j\frac{1}{\mu_i}\frac{1}{\mu_j}\right] \\
&= \sum_{j=1}^{N}\mathbb{E}[K_j]\frac{1}{\mu_j^2} + \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[K_iK_j]\frac{1}{\mu_i}\frac{1}{\mu_j}.
\end{aligned}$$

Substituting the above in the equation for $\mathbb{E}[X]$ completes the proof. □

Proof of Theorem 5.1. The proof of Theorem 5.1 follows directly from combining Lemma 5.2 and [145, Lemma 3], leading to Equation (5.8), and Lemma 5.3 for the distribution of the remaining random variable. □

5.5 Numerical results

In this section we perform some numerical experiments and compare simulation results with the closed-form expressions from the HT limit. In Subsection 5.5.1, we plot the queue-length distribution obtained from simulation, to demonstrate the state-space collapse. In Subsection 5.5.2, we present the scaled mean queue lengths for different loads and show that the mean queue lengths indeed converge to their HT limit. We will use the heavy-traffic result as an approximation for smaller loads and show the errors in a table. Finally, in Subsection 5.5.3, we compare the mean queue lengths in the system with joint batch arrivals to the mean queue lengths in a system with batch arrivals of one customer class and a system with single arrivals.

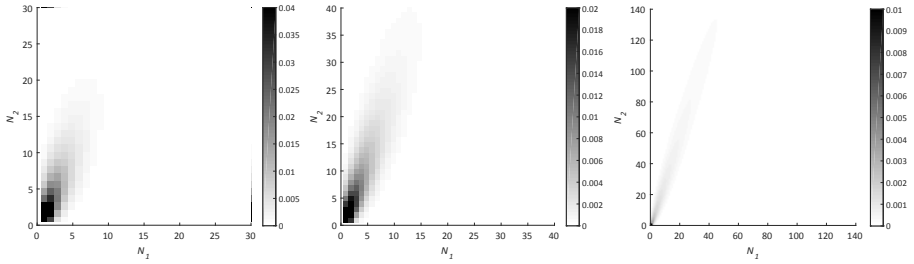
5.5.1 State-space collapse

The basic DPS queue that we use for our experiments is a system with two customer classes and batches of at most 2 arrivals per class. We use the batch-size distribution $p(0,1) = p(1,0) = p(1,1) = p(1,2) = p(0,2) = 1/5$, i.e., there are five possible batches that have the same probability of occurrence. We take $w_1 = 2$, $w_2 = 1$, $\mu_1 = 0.75$ and $\mu_2 = 1$. The arrival rate λ is varied to allow for different loads. In Figure 5.1, we plot the joint queue-length distribution obtained by simulation for three different loads: $\rho = 0.8$ (5.1a), $\rho = 0.9$ (5.1b) and $\rho = 0.99$ (5.1c). For every point (Q_1, Q_2), the color of the point represents the density. We see that for higher loads, the density is more concentrated on a single line, demonstrating the state-space collapse.

5.5.2 Convergence and approximation of moments

We use the same model instances as in Subsection 5.5.1. In Figure 5.2, the scaled mean queue lengths $(1-\rho)\,\mathbb{E}[Q_i]$ are plotted for loads of $\rho = 0.7$ and larger. The dashed lines are the simulation results, the solid lines correspond to the HT limit without any further modification. Observe that the scaled simulated queue lengths converge to the HT limit.

The HT result also provides the following for the scaled marginal queue length distribution: $(1-\rho)Q_i \to_d (\hat{\rho}_i/w_i)X$, with X an exponential random variable. This



(a) $\rho = 0.8$. (b) $\rho = 0.9$. (c) $\rho = 0.99$.

Figure 5.1: Joint queue-length distribution for different values of ρ.

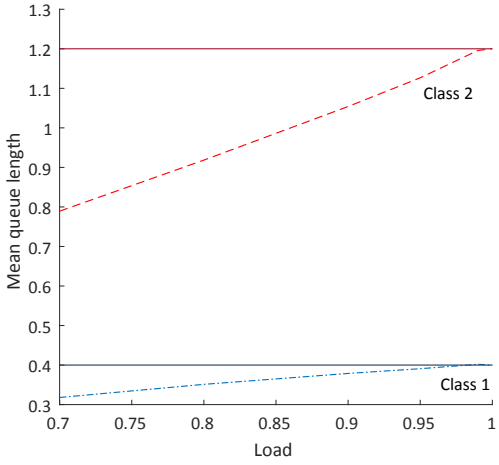


Figure 5.2: Scaled mean queue length for both customer classes and for different loads.

	$\mathbb{E}[Q_1]$			$\mathbb{E}[Q_1^2]$		
ρ	Sim	App	$\Delta\%$	Sim	App	$\Delta\%$
0.70	1.06	1.33	25.79	3.30	3.56	7.80
0.80	1.76	2.00	13.90	7.88	8.00	1.58
0.90	3.79	4.00	5.60	32.27	32.00	0.85
0.95	7.82	8.00	2.35	129.48	128.00	1.15
0.99	40.19	40.00	0.48	3212.30	3200.00	0.38

Table 5.1: Comparison between simulation and heavy-traffic approximation for type-1 customers.

provides the basis for an approximation of the number of type-i customers in a system with $\rho < 1$. Specifically, Q_i is then approximately exponentially distributed with mean

$$\mathbb{E}[Q_i] = \frac{(\hat{\rho}_i/w_i)\,\mathbb{E}[X]}{1-\rho}$$

and with second moment

$$\mathbb{E}[Q_i^2] = \frac{2((\hat{\rho}_i/w_i)\,\mathbb{E}[X])^2}{(1-\rho)^2}, \quad i = i, \ldots, N.$$

We compare the approximations above with simulation results for different values of ρ (by changing λ) using the absolute percentual error, given by

$$\Delta\% = 100\% \times \frac{|\text{App} - \text{Sim}|}{\text{Sim}}.$$

From Table 5.1 we see that the approximation works better for higher loads. This is to be expected, since the approximation is exact in HT. For loads around 0.9, the approximation is reasonable, for lower loads the error increases substantially. Note that we only studied one specific setting, but we expect similar results in other settings.

5.5.3 The impact of batch arrivals

Finally, we experiment with the impact of batch arrivals on the (scaled) mean queue length. To do so, we consider a system with joint batch arrivals to similar systems with single-class batch arrivals and systems with single arrivals only. The arrivals process is modified such that the systems have as many features in common as possible, like the load per class. In the system with joint batch arrivals we again take: $p(0,1) = p(1,0) = p(1,1) = p(1,2) = p(0,2) = 1/5$. In the system with batch arrivals of a single type we have: $p(1,0) = 3/7$, $p(0,1) = 2/7$ and $p(0,2) = 2/7$, and in the system with single arrivals: $p(1,0) = 1/3$ and $p(0,1) = 2/3$. We simulate the mean queue lengths of type-1 customers for different loads. The results are scaled by a factor

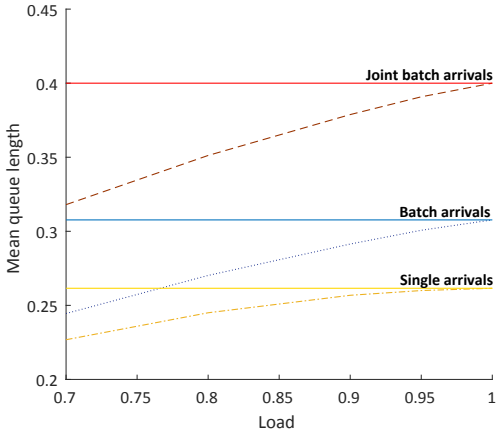


Figure 5.3: Scaled mean queue lengths of type-1 customers in systems with different arrival types.

$(1-\rho)$ and plotted in Figure 5.3 (dashed lines). We see that the scaled mean queue length is smaller if the batches are of a single type and smallest if there are only single arrivals. This can be explained by the variability in the arrival process, where an arrival process consisting of only single customers has the smallest variation. This also explains why the convergence to the HT limits (solid lines) is faster in case of single arrivals.

We conclude that the influence on the queue-length distributions of the joint batch arrivals is significant and that the HT approximation works reasonably well for systems with high loads.

Chapter 6

Access times in appointment-driven systems and level-dependent MAP/G/1 queues

6.1 Introduction

This chapter is motivated by a real health care application. We consider a single-server queue with two types of customers. The number of type-2 customers in the system is infinite, so with queue length we refer to the queue length of type-1 customers. Type-2 customers are only taken into service when the queue length is short enough, taking away resources from type-1. As soon as the queue length of type-1 customers becomes too long, type-2 customers are preempted. We are interested in the waiting time of type-1 customers and the fraction of time that type 2 cannot receive service. In the remainder of this chapter we omit any reference to type 2 or multi-class queues and we only consider type 1. The model is motivated by the actual dynamics of a health care appointment system, therefore, this section contains relevant information and literature on health care and appointment systems.

Specifically, this chapter finds its motivation from problems of long access to ambulatory care. In [42], the authors describe various aspects and issues with respect to waiting in the US health care domain, including the imbalance between supply and demand; see also [111], [117], and references therein. Using programs as 'Advanced Access' [111] and the Dutch program 'Sneller Beter' [14], health organizations aim to improve the access to elective care. The impact of such programs is often temporary. It is much more common that access times are stable through time and are in the order of several weeks; see e.g., [42] and the Dutch article [130].

When designing waiting time standards, a natural question is the amount of capacity required to meet such standards, viz. the desired utilization. This question is at the realm of queueing theory, due to the randomness in the number of care requests and available capacity [110]. In practice, we observe quite lengthy, but rather stable waiting lists. The classical models suffer from their capability to display such behavior.

For instance, using the Pollaczek-Khinchine formula for the M/D/1 queue, we may see that the load has to be close to the critical region to give expected waiting times in the order of weeks for an appointment system with daily operations (see also Subs. 6.5.1). It is quite unlikely that so many health organizations operate just in this small region of the load. Moreover, a consistent (but small) backlog is efficient for planning of follow-up health activities. For example, scheduling of surgeries can be performed efficiently when there is a manageable pool of surgeries to schedule from.

In our experience with Dutch hospitals, we have observed that the capacity is often adapted to the current status of the waiting list. When the waiting list tends to be (too) long, patients are scheduled in overtime or are squeezed in between other appointments; we refer to this as *overbooking*, as patients are booked in excess of the available time. Of course, this is at the cost of the tardiness and waiting time on the day itself. Essentially, the capacity can be interpreted to be flexible and to be adapted based on the level of congestion. It is even not uncommon that the utilization of the outpatient capacity is over 100% for a prolonged period, resulting from a structural deficit in capacity with respect to demand [22]. Overbookings have been addressed from the domain of appointment scheduling as well [102; 105; 160], where overbookings are used to counter no-shows.

The goal of this chapter is to develop queueing models that produce waiting-time results that may be recognized in practice, i.e. provide considerable waiting for a wider range of the load. We accomplish this by considering queues with queue-length dependent features, e.g., the capacity, leading to level-dependent M/G/1 or MAP/G/1 queues. Although the motivation of our queueing model stems from health systems with flexible capacity, the model also captures level-dependent arrivals reflecting the decision to join the waiting list or not depending on its length. To the best of our knowledge, waiting-list dependent overbookings in the context of access times have not been considered before.

The aim of the current chapter is three-fold. First, we provide macro-scopic models that describe access times in appointment systems. In these models, we take many relevant features into account, including randomness in available capacity (e.g., due to partial absence of the medical staff) and overbookings. Second, we derive exact and intuitively appealing results for the level-dependent M/G/1 and MAP/G/1 queue. We exploit that the queueing dynamics are level independent above some threshold. And third, using some numerical experiments, we identify that these level-dependent queues yield waiting times that may be observed in practice. In fact, these experiments indicate that it is efficient for appointment systems to operate close to or just above their level of stability, provided that some extra flexible capacity can be used.

The literature on appointment systems and ambulatory care is primarily focused on appointment scheduling rather than controlling access times; see e.g. [85; 117] for an overview of advanced scheduling problems, and [77] for an excellent treatment of appointment systems in health care. In [77] the authors distinguish between *indirect*

and *direct* waiting; indirect waiting refers to the time between making an appointment and the time of appointment, whereas direct waiting is the time between arrival and the moment the actual appointment starts. In [111] the authors note that direct waiting is an inconvenience for the patient, whereas indirect waiting involves patient safety. This chapter focuses on indirect waiting.

We first discuss some related literature on access times (or indirect waiting) for ambulatory care and appointment systems. One of the pioneering papers on queueing models for hospital waiting lists is [156]. The model in [156] assumes an M/G/s queue with a linearly decreasing state-dependent arrival rate; the latter reflects a feedback mechanism where patients are discouraged when they encounter a longer waiting list. An interesting approach is the provision of a waiting time guarantee for elective treatment, inspired by its introduction in Denmark; see [101]. The analysis in [101] is based on CTMC in combination with discrete-event simulation. In our numerical experiments, we see that the combination of appropriate capacity levels and some flexible capacity can be effective in providing waiting-time guarantees.

An intuitive approach to model access times is to use discrete-time bulk service queues, as in [86; 99]. A problem occurs when deriving access times in case of variable capacity and overbookings. In [99], the available capacity is deterministic; Izady [86] considers a model with random available capacity, but does not provide the access (or waiting) time. Applying overbookings further complicates this issue. In [54; 55], advanced queueing models are considered for the access time in appointment-driven systems. Due to its structure of arrival and service sessions, the size of the waiting list is analyzed using vacation models. These models do also not include variability in available capacity and overbookings.

The authors in [65] study the reduction in access time by temporarily allowing extra capacity to (partly) clear waiting lists. Adaptive allocation of capacity is studied in [146] based on an extensive case analysis. Both papers rely on simulation. Also, in [74; 89; 105; 160], access times (or indirect waiting) has been evaluated using queueing models, typically of the M/D/1 type. Common in [74; 105; 160] is the role of no shows. In [74; 105], the focus is on the panel size and in [160] the focus is on the relation between the no-show probability and the corresponding access time. These papers do not consider the control of the access time.

Second, state-dependent M/G/1 queues are studied in [2; 37; 152]. The paper [152] considers a switch-over policy with two threshold levels, whereas the authors in [2] focus on the general $M_n/G_n/1$ queue. Although the M/G/1 variant of our model is a special case of [2], the representation of the results are considerably different. We find the results in [37] most related to our M/G/1 case; the set-up in [37] is a one-dimensional random walk having a similar structure as our level-dependent M/G/1 queue.

Moreover, there are some studies related to our level-dependent MAP/G/1 model. In fact our model is a special case of [84], as we restrict ourselves to the case that

the queueing dynamics are level-independent above some finite threshold. As a consequence, we are able to obtain more explicit and intuitive results than those in [84]. The model in [124] is also related to ours; the main difference is in the interpretation of the model and intuitive presentation of the results. The author in [140] considers a multi-type MMAP/PH/1 queue with adaptive arrivals, yielding level-dependent dynamics as well. Finally, we mention [43] as a key paper on level-dependent QBD's.

The organization of the chapter is now as follows. In Section 6.2 we propose two models for studying the access time for appointment systems, yielding level-dependent M/G/1 and MAP/G/1 queues. The former is analyzed in Section 6.3, whereas the latter is analyzed in Section 6.4. The results for both queues follow a similar structure. Some experiments in Section 6.5 give insight in the behavior of the system as a function of the load.

6.2 Model description

We relate the access time for appointment systems to the dynamics of state-dependent M/G/1 and MAP/G/1 queues. In Subsection 6.2.1 we describe the characteristics of appointment systems and the elements that should be captured by the model. Measuring the access time in slots, we propose a first model (model I) based on M/D/1 queues in Subsection 6.2.2. In Subsection 6.2.3 we allow for a richer, but more involved, model (model II) based on MAP/D/1 queues.

6.2.1 Characteristics of appointment systems

Most appointment systems have daily operations, meaning that a natural time unit is days. In this chapter, we adhere to this situation. However, the modeling and results carry over to other settings; e.g., by defining all quantities at the time scale of weeks.

Arrivals Patients arrive randomly in time to make an appointment. In particular, we assume that the requests for appointments arrive according to a Poisson process with rate λ per day. The detailed scheduling of appointment requests in practice is complicated by many (often personal) factors, such as preferences for certain times or days. In the model, we abstract from these small scale details and assume that arriving requests are scheduled on a FCFS basis; i.e. patients are assigned to the first place available. Such a scheduling mechanism also represents the backlog in slots.

Slots In the basic setting, patients are served in slots of deterministic length. Although the actual duration of an appointment is random, the scheduled time for an

appointment is known and deterministic. The length of the waiting list is not affected by the randomness in service time, as possible overtime does not carry over to the next day. Allowing for differences in scheduled service time can be accomplished by assigning multiple slots to a single patient.

Capacity For outpatient departments, the available capacity is typically recorded in a so-called blueprint (or basic capacity). The blueprint describes the number of slots available for each day and is often drawn on a yearly basis. As such, unavailability of staff is generally not taken into account, which usually adds up to 20% - 25% of the capacity of the blueprint. Cancellation of the basic capacity is thus quite common; in fact, we often encounter that the available capacity per day fluctuates, leading to a random number of available slots. We define m_1 as the number of slots per day according to the blueprint, and assume that independently on each day i slots are available with probability $s_i^{(1)}$, $i = 1, \ldots, m_1$, with $\sum_{i=1}^{m_1} s_i^{(1)} = 1$. Let $s^{(1)} = \sum_{i=1}^{m_1} i s_i^{(1)}$, such that $s^{(1)}$ represents the average offered basic capacity. In periods of excessive congestion, overbooking occurs; this implies that additional capacity is made available. Let $\tilde{m}$ be the potential extra capacity, yielding capacity $m_2 = m_1 + \tilde{m}$ during such periods. The availability of capacity in case of overbooking is $s_i^{(2)}$, $i = 1, \ldots, m_2$, with $s^{(2)}$ defined accordingly.

Access time The time between the appointment request and the day the actual appointment takes place is called the *backlog* or *access time* and is measured in days to weeks. In [77] this is also referred to as indirect waiting time, as opposed to direct waiting time which refers to physical waiting in the waiting room, and is measured in minutes to hours. To control the access time, we assume a simple threshold type of control; when the access time is at or below L, the system operates in the usual mode, whereas overbooking is used when the access time exceeds L.

We assume that the system with extra capacity is stable, i.e. $\lambda < s^{(2)}$. The system operating with basic capacity is not necessarily stable; we allow for $\lambda \geq s^{(1)}$. In practice, we encounter both doctors and specialisms that achieve utilizations of the basic capacity exceeding 100%, implying that overbooking happens on a more structural basis.

6.2.2 Model I: backlog in slots

To model the appointment system with a queue in continuous time, the intervals during which the system is closed are cut out and the hours of operation are glued together. Customers then arrive according to a continuous time Poisson process and require a single slot as service time. Access time in appointment systems have been

modeled using the $M/D/1$ queue, see e.g., [74; 105]. The state is

$$X(t) = \text{backlog in slots for patients arriving at time } t.$$

Equivalently, $X(t)$ is the access time or number of slots at the waiting list for a patient arriving at time t. Due this interpretation, the arrival rate is measured in terms of slots. Hence, in basic capacity mode, the arrival rate is $\lambda_1 = \lambda/m_1$. With extra capacity, the slots should be considered to be squeezed such that m_2 slots now represent a single day instead of m_1. The arrival rate when the backlog exceeds L is thus $\lambda_2 = \lambda/m_2$.

With this interpretation we see that the backlog in the appointment system $\{X(t), t \geq 0\}$ can be modeled with an $M/D/1$ (or $M/G/1$) queue with state-dependent arrival rates that continuously depend on the backlog. This model is analyzed in Section 6.3, leading to intuitive results. When the backlog is small, appointment requests are scheduled on a short notice leading to same day appointments. For outpatient departments, this is not always realistic; for ambulatory services with a more urgent character this is more common. We should note that (almost) empty waiting lists are rather uncommon for outpatient departments in the Netherlands, and the impact of this assumption seems limited.

One of the main disadvantages of the $M/D/1$ interpretation is that it remains unclear how to incorporate the fluctuation in capacity. Moreover, the backlog is now measured in slots whereas actual access times are measured in days. Although there evidently is one-to-one relation, one should be careful with the interpretation due to the difference in time scale of the process below and above level L.

6.2.3 Model II: backlog in days

An alternative way of modeling the access time is by considering the backlog in days directly. As we also need the available number of slots at the first available day to maintain the Markov property, we consider the two-dimensional Markov process $\{(X(t), J(t)), t \geq 0\}$, with

$$\begin{aligned} X(t) &= \text{backlog in days for patients arriving at time } t, \\ J(t) &= \text{number of patients scheduled on the first available day when arriving at} \\ &\qquad \text{time } t. \end{aligned}$$

We have $X(t) \in \{0, 1, \ldots\}$ and $J(t) \in \{1, \ldots, m\}$ with m the maximum available capacity per day. For this process, we have the following dynamics. A service time corresponds to the elapsing of a day, and is thus deterministic. When a new patient arrives and the state before an arrival is (x, j), then there are two possible transitions of the Markov process. Let $i = 1 + \mathbb{1}(x > L)$ denote whether the system uses the basic capacity ($i = 1$) or the extra capacity ($i = 2$). First, with probability $p_{j+1}^{(i)}$ the

patient took the last available slot on that day and the next patient will be placed on the next day; the Markov process moves to state $(x+1,0)$. Second, with probability $\bar{p}_{j+1}^{(i)} = 1 - p_{j+1}^{(i)}$ there are more available slots left and the Markov process moves to state $(x, j+1)$. The probabilities $p_j^{(i)}$ are linked to the availability of slots $s_j^{(i)}$ via the relation

$$\bar{p}_j^{(i)} = \frac{s_j^{(i)}}{\sum_{k \geq j} s_k^{(i)}} \quad \text{for } i = 1, 2, \text{ and } j = 1, \ldots, m-1.$$

This implies that the access time can be modeled with a level-dependent MAP/D/1 (or MAP/G/1) queue. The arrival process is then defined by the $m \times m$ matrices $D_k^{(i)}$, $k = 0, 1$ and $i = 1, 2$, where $D_0^{(i)}$ gives the phase transitions and $D_1^{(i)}$ represents the arrival rate leading to an additional day of backlog. For the model as described above, we have

$$D_0^{(i)} = \begin{bmatrix} -\lambda_i & \lambda_i \bar{p}_1^{(i)} & & & \\ & -\lambda_i & \lambda_i \bar{p}_2^{(i)} & & \\ & & \ddots & \ddots & \\ & & & -\lambda_i & \lambda_i \bar{p}_{m-1}^{(i)} \\ & & & & -\lambda_i \end{bmatrix}, \qquad D_1^{(i)} = \begin{bmatrix} \lambda_i p_1^{(i)} & 0 & \ldots \\ \lambda_i p_2^{(i)} & 0 & \ldots \\ \vdots & \vdots & \ddots \\ \lambda_i p_{m-1}^{(i)} & & \\ \lambda_i & & \end{bmatrix}.$$

The matrices above imply that the interarrival times for an additional day of backlog follow a mixed Erlang distribution. We allow for a more flexible setup, where these matrices are just special cases. In particular, we allow for general level-dependent matrices $D_0^{(i)}$ and $D_1^{(i)}$, with $i = 0, 1, \ldots, L$, which become level-independent D_0 and D_1 above level L. Such a more general level-dependent MAP/G/1 queue is analyzed in Section 6.4, where the results lead to similar intuitive findings as for the M/G/1 case.

Finally, we note that this model allows for many other features to be incorporated. For example, we can model heterogeneous customer classes requiring multiple slots by assuming that the phase can increase by more than one. Also, the arrival rate λ_i may be level-dependent reflecting that customers balk when access times tend to be longer. Such features will typically be reflected in the matrices $D_0^{(i)}$ and $D_1^{(i)}$.

6.3 State-dependent M/G/1 queue

6.3.1 Model and method outline

We have an M/G/1 queue, where the arrival rate is λ_1 when the number of customers in the system is at most L, and the arrival rate is λ_2 when there are more than L

customers present. Note that this notation differs from the other chapters, where λ_i typically refers to the arrival rate of class i. The arrival rate is continuously adjusted, i.e. may change upon arrival and departure instants. We denote the service time distribution by $H(\cdot)$ and its LST by $H^*(\cdot)$. We assume that $\lambda_2 \mathbb{E}[H] = \rho_2 < 1$.

The analysis now proceeds along similar lines as in [17] for the workload process in Lévy-driven queues. Let x_n be the steady-state probability that n customers are left behind by a departing customer. We then set up a set of equations using balancing principles, which are utilized in the following procedure:

Step 1 Determine the distribution x_n for $n < L$ up to a constant, which is independent of x_n for $n \geq L$.

Step 2 Using x_n for $n < L$, the generating function (GF) is completely determined upto a constant. Rewriting the GF and applying inversion, we determine x_n for $n \geq L$.

Step 3 Find the remaining constant x_0; this step can be included at the end of Step 1 (MAP/G/1) or Step 2 (M/G/1).

The method above typically leads to intuitively appealing results, where x_n has a clear interpretation both for $n < L$ as well as $n \geq L$. The derivation in Section 6.4 for the level-dependent MAP/G/1 queue is much along the same lines. We note that in the subsequent analysis we do not treat Step 3 separately.

6.3.2 Performance analysis

We analyze the queueing model, embedded at departure instants; we refer to Theorem 6.3 for the relation between queue lengths at arbitrary and departure instants. Then,

$$x_n = \sum_{i=1}^{n+1} x_i \alpha_{n+1-i}^{(i)} + x_0 \alpha_n^{(1)}, \tag{6.1}$$

where $\alpha_i^{(n)}$ is the probability of i arrivals during a service time when the number of customers just after the previous service completion is n. These probabilities are given by

$$\alpha_i^{(n)} = \begin{cases} \int_{t=0}^{\infty} \frac{(\lambda_1 t)^i}{i!} e^{-\lambda_1 t} \, \mathrm{d}H(t) =: \alpha_i & \text{if } n+i \leq L \\ \int_{t=0}^{\infty} \frac{(\lambda_2 t)^i}{i!} e^{-\lambda_2 t} \, \mathrm{d}H(t) =: \hat{\alpha}_i & \text{if } n > L \\ \int_{t=0}^{\infty} \int_{u=0}^{t} \frac{\lambda_1^{L-n+1} u^{L-n}}{(L-n)!} e^{-\lambda_1 u} & \\ \quad \times \frac{(\lambda_2 (t-u))^{n+i-(L+1)}}{(n+i-(L+1))!} e^{-\lambda_2 (t-u)} \, \mathrm{d}u \, \mathrm{d}H(t) & \text{if } n \leq L, n+i > L. \end{cases}$$

As the first two cases do not depend on n, we drop the n from the notation there. The third case follows from conditioning on the moment that level $L+1$ is hit, when

we start at level n, which follows an Erlang distribution consisting of $L-n+1$ phases with rate λ_1.

Step 1. We first determine x_n for $n < L$. Observe that Equation (6.1) reads $x_n = \sum_{i=1}^{n+1} x_i \alpha_{n+1-i} + x_0 \alpha_n$ corresponding to the dynamics of an M/G/1/L queue with arrival rate λ_1. Intuitively, this naturally follows from the fact that at downcrossings of level L the time until the next arrival is again exponential due to the lack-of-memory property of the arrival process. Hence, restricting to the periods that there are no more than L customers present, the sample path is indistinguishable from the sample path in an isolated M/G/1/L queue. In particular, let X^{Q1} be the steady-state number of customers in the M/G/1/L queue with arrival rate λ_1 and service time H, such that $x_n \propto \mathbb{P}(X^{Q1} = n)$. This gives the following lemma.

Lemma 6.1. *The number of customers left behind by a departing customer is, for* $n = 0, 1, \ldots, L-1$,

$$x_n = x_0 \frac{\mathbb{P}(X^{Q1} = n)}{\mathbb{P}(X^{Q1} = 0)}.$$

Remark 6.1. The analysis can be adapted to the case that the arrival rate at or below L would be state-dependent as well. In that case, x_n for $n = 0, 1, \ldots, L-1$ would follow from a set of equations similar to the MAP/G/1 case analyzed in Section 6.4. Moreover, from an up- and downcrossing argument, we may see that x_L can be obtained similar to x_i, $i = 0, \ldots, L-1$ using the M/G/1/$L+1$ queue.

Step 2. Using the above and taking GF in (6.1), we may directly obtain the GF of the number of customers left behind at service completions: $X^*(z) = \sum_{n=0}^{\infty} x_n z^n$. To obtain intuitively appealing results for x_n, we first consider the time between the moment that the number of customers hits $L+1$ until the first service completion. Let T_i be the time between hitting $L+1$ from below until the next service completion, given that i customers were present at the moment that this service started; this is depicted in Figure 6.1. T_i is equal to zero if this service ended before the number of customers reached $L+1$. Moreover, let $\hat{T}_i = T_i \mid T_i > 0$ be the random variable T_i conditioned that it is positive and denote $\hat{T}_i^*(z) = \mathbb{E}[e^{-z\hat{T}_i}]$ as its LST. Let $P(s,x)$ be the regularized gamma function, i.e.,

$$P(s,x) = \frac{1}{(s-1)!} \int_{t=0}^{x} t^{s-1} e^{-t} \, \mathrm{d}t. \tag{6.2}$$

Lemma 6.2. *The conditional first time above* L *starting from* $i = 0, 1, \ldots, L$ *until the first service completion* $\hat{T}_i$ *has LST*

$$\hat{T}_i^*(z) = \frac{1}{1 - \mathbb{P}(T_i = 0)} \int_{t=0}^{\infty} e^{-zt} \left(\frac{\lambda_1}{\lambda_1 - z} \right)^{L-i+1} P(L-i+1, t(\lambda_1 - z)) \, \mathrm{d}H(t),$$

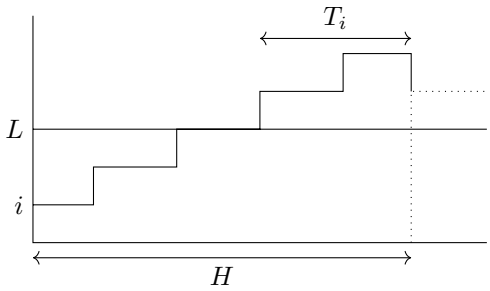


Figure 6.1: Graphical representation of T_i.

where

$$\mathbb{P}(T_i = 0) = \int_{t=0}^{\infty} \sum_{n=0}^{L-i} \frac{(\lambda_1 t)^n}{n!} e^{-\lambda_1 t} \, \mathrm{d}H(t). \tag{6.3}$$

Proof. Let $i = 0, 1, \ldots, L$ be the starting level. Note that the probability that T_i is 0 is equal to the probability that at most $L - i$ customers arrived during the service time. Conditioning on the service time yields (6.3).

Now consider the LST of T_i. The conditional LST of $\hat{T}_i$ then follows straightforwardly. Let the service time be fixed at t; at the end we integrate over the service time. Conditioning on the moment when level $L + 1$ is hit, the density of T_i, for $u \in (0, t)$, reads $f_{T_i}(u) = \frac{\lambda_1(\lambda_1(t-u))^{L-i}}{(L-i)!} e^{-\lambda_1(t-u)}$. This follows from the fact that the corresponding hitting time has an Erlang distribution consisting of $L - i + 1$ phases with rate λ_1. The LST of T_i for fixed service time t then equals

$$\begin{aligned}
\mathbb{E}[e^{-zT_i}] &= \mathbb{P}(T_i = 0) + \int_{u=0}^{t} e^{-zu} e^{-\lambda_1(t-u)} \frac{\lambda_1(\lambda_1(t-u))^{L-i}}{(L-i)!} \, \mathrm{d}u \\
&= \mathbb{P}(T_i = 0) + e^{-zt} \int_{u=0}^{t} e^{-u(\lambda_1 - z)} u^{L-i} \, \mathrm{d}u \frac{\lambda_1^{L-i+1}}{(L-i)!} \\
&= \mathbb{P}(T_i = 0) + e^{-zt} \int_{v=0}^{t(\lambda_1 - z)} e^{-v} \left(\frac{v}{\lambda_1 - z} \right)^{L-i} \frac{1}{\lambda_1 - z} \, \mathrm{d}v \frac{\lambda_1^{L-i+1}}{(L-i)!} \\
&= \mathbb{P}(T_i = 0) + e^{-zt} \left(\frac{\lambda_1}{\lambda_1 - z} \right)^{L-i+1} P(L - i + 1, t(\lambda_1 - z)).
\end{aligned} \tag{6.4}$$

For the third equality we use the substitution $v = u(\lambda_1 - z)$ and in the fourth equality we use (6.2). Finally, integrating over the service time and using the conditional expectation completes the proof. □

Remark 6.2. In special cases, the expression for the LST of $\hat{T}_i$ can be considerably simplified depending on the service time distribution. When H is a mixture of expo-

nentials, it is possible to work out all integrals and summations. If H is deterministic, then the outer integral vanishes.

Since we already determined x_n for $n < L$, we present the GF of the number of customers as follows.

Theorem 6.1. *The GF of the number of customers at departure moments is given by*

$$X^*(z) = \sum_{n=0}^{L-1} x_n z^n + z^L x_0 \,\mathbb{P}(T_1 > 0) \frac{H^*(\lambda_2(1-z)) - z\hat{T}_1^*(\lambda_2(1-z))}{(H^*(\lambda_2(1-z)) - z)} + z^L \sum_{i=1}^{L} x_i \,\mathbb{P}(T_i > 0) \frac{H^*(\lambda_2(1-z)) - z\hat{T}_i^*(\lambda_2(1-z))}{(H^*(\lambda_2(1-z)) - z)}, \tag{6.5}$$

with

$$x_0 = \mathbb{P}(X^{Q1} = 0)(1-\rho_2) \times \left((1-\rho_2) \sum_{n=0}^{L-1} \mathbb{P}(X^{Q1} = n) + \mathbb{P}(X^{Q1} = 0)\,\mathbb{P}(T_1 > 0)(\lambda_2\,\mathbb{E}[\hat{T}_1] + 1 - \rho_2) + \sum_{i=1}^{L} \mathbb{P}(X^{Q1} = i)\,\mathbb{P}(T_i > 0)(\lambda_2\,\mathbb{E}[\hat{T}_i] + 1 - \rho_2) \right)^{-1}.$$

The proof can be found in Subsection 6.6.1 and follows by taking the GF in (6.1) and some tedious rewriting. Here, we interpret the specific form of $X^*(z)$. Specifically, the first term at the rhs of (6.5) corresponds to the GF on the set $\{0, 1, \ldots, L-1\}$ which we already obtained. The second and third term correspond to the GF on $\{L, L+1, \ldots\}$, as those terms can be interpreted as the convolution of L (corresponding to z^L) with a non-negative random variable; hence, the convolution only has probability mass at or above L. The quantity

$$\frac{H^*(\lambda_2(1-z)) - z\hat{T}_i^*(\lambda_2(1-z))}{(H^*(\lambda_2(1-z)) - z)}$$

corresponds to an M/G/1 queue with arrival rate λ_2, service time H, and exceptional first service time $\hat{T}_i$ in a busy period, see Takagi [126, p. 129]. Let $X_{(i)}^{Q2}$ be the steady-state number of customers in such a queue.

Now, the second term at the right-hand side of (6.5) represents the probability that the set $\{L+1, L+2, \ldots\}$ is entered from level 0 with probability $x_0\,\mathbb{P}(T_1 > 0)$, in which case the process above L behaves as $X_{(1)}^{Q2}$, i.e., an M/G/1 queue with exceptional first service time $\hat{T}_1$ in a busy period. Similarly, the third term provides the possibilities of entering the set $\{L+1, L+2, \ldots\}$ starting a service time from level i. Combining the above and inverting the GF in Theorem 6.1, yields the following corollary.

Corollary 6.1. *For the distribution of the number of customers at departure epochs, we have, for* $n = 0, 1, \ldots, L-1$,

$$x_n = x_0 \frac{\mathbb{P}(X^{Q1} = n)}{\mathbb{P}(X^{Q1} = 0)}$$

and, for $n = 0, 1, \ldots$, *with* $c_i = (1 - \rho_2 + \lambda_2 \mathbb{E}\,\hat{T}_i)/(1 - \rho_2)$,

$$x_{L+n} = x_0 c_1 \,\mathbb{P}(T_1 > 0)\,\mathbb{P}(X^{Q2}_{(1)} = n) + \sum_{i=1}^{L} x_i c_i \,\mathbb{P}(T_i > 0)\,\mathbb{P}(X^{Q2}_{(i)} = n).$$

6.4 Level-dependent MAP/G/1 queue

We consider the LD-MAP/G/1 queue. While analyzing this model, we will use the standard queueing terminology. For our application, recall that an arriving patient causes a change in the phase of the arrival process of the queueing model. An arrival in the queueing model means that a new day is filled with appointments. The service duration represents one day.

6.4.1 The arrival process

As described in Subsection 6.2.3, we have $m \times m$ matrices $D_n^{(k)}$, $n = 0, 1$ and $k \in \mathbb{N}_0$. The matrix generating function is given by $D^{(k)}(z) = D_0^{(k)} + D_1^{(k)} z$, for $|z| \leq 1$. Since the matrices $D_n^{(k)}$ are level-independent for $k > L$, we drop the superscript form the notation when the level is above L. Let $\boldsymbol{\pi}$ be the stationary probability vector of $D_0 + D_1$ such that the arrival rate above L is given by $\lambda_2 = \boldsymbol{\pi} D_1 \mathbf{e}$, with $\mathbf{e}$ a column vector of ones. Again, denote $\rho_2 = \lambda_2 \,\mathbb{E}[H] < 1$.

The generator matrix Q is defined as follows, with empty entries representing zeros:

$$Q = \begin{pmatrix} D_0^{(0)} & D_1^{(0)} & & & & & & \\ & D_0^{(1)} & D_1^{(1)} & & & & & \\ & & \ddots & \ddots & & & & \\ & & & D_0^{(L)} & D_1^{(L)} & & & \\ & & & & D_0 & D_1 & & \\ & & & & & D_0 & \ddots & \\ & & & & & & \ddots & \end{pmatrix}.$$

Let $P_n^{(k)}(t)$ be the matrix whose entries $P_{n,ij}^{(k)}(t)$ denote the conditional probability that n arrivals occur in the interval $(0, t]$ and the arrival phase at time t is j, given that

the arrival phase at time 0 was i (and the number of customers present at time 0 was k). According to Hofmann [84, Theorem 2.1], we have $P_n^{(k)}(t) = \left(e^{Qt}\right)_{k,n+k}$, where $(\cdot)_{i,j}$ denotes the (i,j)th block of the matrix. Note that the matrix Q has infinite size, so we cannot compute the matrix exponential. To work around this, define, for $k = 0, 1, \ldots, L$, the finite matrix

$$\hat{Q}^{(k)} = \begin{pmatrix} D_0^{(k)} & D_1^{(k)} & & & \\ & D_0^{(k+1)} & D_1^{(k+1)} & & \\ & & \ddots & \ddots & \\ & & & D_0^{(L-1)} & D_1^{(L-1)} \\ & & & & D_0^{(L)} \end{pmatrix}.$$

We have $P_n^{(k)}(t) = \left(e^{\hat{Q}^{(k)}t}\right)_{0,n} = \left(e^{\hat{Q}^{(0)}t}\right)_{k,n+k}$, for $n \leq L-k$. For $n > L-k$ we condition on the moment that there are $L+1$ customers in the system, starting from level k:

$$P_n^{(k)}(t) = \int_{v=0}^{t} P_{L-k}^{(k)}(v) D_1^{(L)} P_{n-(L-k+1)}(t-v)\,\mathrm{d}v, \quad n > L-k. \tag{6.6}$$

The matrix $P_{n-(L-k+1)}(t)$ can be calculated using the standard method as described by, e.g., Lucantoni [106], since the arrival process is level independent above L.

The matrix $A_n^{(k)}$, with entries $A_{n,ij}^{(k)}$, contains the probabilities of having n arrivals during a service, and at the end of the service duration the arrival process is in phase j, given that the arrival process started in phase i and the number of customers was k at the beginning of the service. Using the definition of $P_n^{(k)}(t)$, we see that

$$A_n^{(k)} = \int_{t=0}^{\infty} P_n^{(k)}(t)\,\mathrm{d}H(t), \quad k > 0. \tag{6.7}$$

Note that in case of deterministic service durations of length b we have $A_n^{(k)} = P_n^{(k)}(b)$. The matrix transform of the matrix $A_n^{(k)}$ is given by $A^{(k)}(z) = \sum_{n=0}^{\infty} z^n A_n^{(k)}$.

Using Equation (6.6) we get after some rewriting, for $0 < k \leq L$,

$$A^{(k)}(z) = \int_{t=0}^{\infty} \left(\sum_{n=0}^{L-k} z^n P_n^{(k)}(t) + \int_{v=0}^{t} z^{L-k+1} P_{L-k}^{(k)}(v) D_1^{(L)} e^{(t-v)D(z)}\,\mathrm{d}v \right) \mathrm{d}H(t).$$

When $k = 0$, we have $A^{(0)}(z) = -D_0^{(0)-1} D_1^{(0)} A^{(1)}(z)$, because there has to be an arrival before there can be a service. On the other hand, when $k > L$, we can use the expression given in Lucantoni [106]: $A(z) = \int_{t=0}^{\infty} e^{D(z)t}\,\mathrm{d}H(t)$. The matrix $A^{(k)}$ is equal to $A^{(k)}(1)$.

We will need the vector $\mathbf{a}^{(k)}$, the mean number of arrivals during a service period, given that the number of customers at the start of the service is $k \geq 1$; we drop the superscript again when $k > L$. For $\mathbf{a}^{(k)}$, we take the derivative of $A^{(k)}(z)$ with respect to z and evaluate at $z = 1$, then multiply by $\mathbf{e}$:

$$\mathbf{a}^{(k)} = \begin{cases} -D_0^{(0)-1} D_1^{(0)} \mathbf{a}^{(1)} & k = 0 \\ \frac{\mathrm{d}}{\mathrm{d}z} A^{(k)}(z) \Big|_{z=1} \mathbf{e} & 0 < k \leq L \\ \frac{\mathrm{d}}{\mathrm{d}z} \int_{t=0}^{\infty} e^{D(z)t} \, \mathrm{d}H(t) \Big|_{z=1} \mathbf{e} & k > L. \end{cases}$$

6.4.2 Performance analysis

We derive the stationary distribution of the number of customers at departure instants $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots)$. The derivation of $\mathbf{x}_k$ at or below the threshold L is similar to Shin and Pearce [124].

Starting point is the balance equation (note the similarity with (6.1))

$$\mathbf{x}_k = \mathbf{x}_0 A_k^{(0)} + \sum_{i=1}^{k+1} \mathbf{x}_i A_{k+1-i}^{(i)}. \tag{6.8}$$

Defining the vector generating function $X^*(z) = \sum_{n=0}^{\infty} \mathbf{x}_n z^n$, we obtain with standard techniques that the GF satisfies

$$X^*(z) \left[zI - A(z) \right] = \mathbf{x}_0 A^{(0)}(z)(z-1) + \sum_{i=0}^{L} \mathbf{x}_i z^i \left[A^{(i)}(z) - A(z) \right], \tag{6.9}$$

see also Equation (3.1) in [124].

Step 1. The matrix G is defined as in Lucantoni [106]; the entries $G_{i,j}$ are the probabilities to enter level $n-1$ in phase j, given that the process started in level n and phase i for $n > L$. The matrix can be computed iteratively using

$$G = \sum_{\nu=0}^{\infty} A_\nu G^\nu = \int_{t=0}^{\infty} e^{D[G]t} \, \mathrm{d}H(t),$$

with

$$D[G] = \sum_{j=0}^{\infty} D_j G^j = D_0 + D_1 G.$$

The probabilities $\mathbf{x}_n$ for $n = 0, 1, \ldots, L-1$ are given in the following lemma, based on [124, Lemma 1]; the difference with [124] are the matrices at the boundary $\bar{A}_L^{(0)}$ and $\bar{A}_{L+1-i}^{(i)}$, $i = 1, \ldots, L$. The first moment of X is given in Subsection 6.6.3.

Lemma 6.3. *Let* $(\mathbf{X}^*, \mathbf{J}^*) = \{(X_n^*, J_n^*),\ n \geq 0\}$ *denote the censored Markov chain obtained by embedding* $\{(X_n^*, J_n^*),\ n \geq 0\}$ *at the epochs when it visits the set of states* $\{(i,j) :\ 0 \leq i \leq L,\ 1 \leq j \leq m\}$. *Then the transition probability matrix* Q^* *of* $(\mathbf{X}^*, \mathbf{J}^*)$ *is given by*

$$Q^* = \begin{pmatrix} A_0^{(0)} & A_1^{(0)} & A_2^{(0)} & \cdots & A_{L-1}^{(0)} & \bar{A}_L^{(0)} \\ A_0^{(1)} & A_1^{(1)} & A_2^{(1)} & \cdots & A_{L-1}^{(1)} & \bar{A}_L^{(1)} \\ 0 & A_0^{(2)} & A_1^{(2)} & \cdots & A_{L-2}^{(2)} & \bar{A}_{L-1}^{(2)} \\ 0 & 0 & A_0^{(3)} & \cdots & A_{L-3}^{(3)} & \bar{A}_{L-2}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_0^{(L)} & \bar{A}_1^{(L)} \end{pmatrix},$$

with

$$\bar{A}_{L-k+1}^{(k)} = \int_{t=0}^{\infty} \int_{v=0}^{t} P_{L-k}^{(k)}(v) D_1^{(L)} e^{D[G](t-v)}\, \mathrm{d}v\, \mathrm{d}H(t), \quad 0 < k \leq L, \tag{6.10}$$

and

$$\bar{A}_L^{(0)} = -D_0^{(0)-1} D_1^{(0)} \bar{A}_L^{(1)}.$$

If the invariant probability vector of Q^* *is, in partitioned form,* $\mathbf{p} = [\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_L]$, *where* $\mathbf{p}_i$ *is an* m*-vector, then the vectors* $\mathbf{x}_i$ *take the form*

$$\mathbf{x}_i = c\mathbf{p}_i, \quad 0 \leq i \leq L,$$

where

$$c = (1-\rho_2) / \left(U_p(1)(I - A + \mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{a} + \mathbf{p}_0\mathbf{e} + \sum_{i=0}^{L} \mathbf{p}_i(\mathbf{a}^{(i)} - \mathbf{a}) \right)$$

and $U_p(1) = \sum_{i=0}^{L} \mathbf{p}_i(A^{(i)} - A)$.

Proof. The matrix $\bar{A}_n^{(k)}$ is given by $\bar{A}_n^{(k)} = \sum_{\nu=n}^{\infty} A_\nu^{(k)} \prod_{j=0}^{\nu-n-1} G^{(k+\nu-1-j)}$, with $G^{(l)}$ the level-dependent version of G, see [84]. This matrix contains the probabilities to go from level k to $k+n-1$ during a service with n or more arrivals, with respect to the arrival phases. The specific matrices we need are $\bar{A}_{L-k+1}^{(k)}$, $k \geq 1$; using the definition we have

$$\begin{aligned} \bar{A}_{L-k+1}^{(k)} &= \sum_{\nu=0}^{\infty} A_{\nu+L-k+1}^{(k)} \prod_{j=0}^{\nu-1} G^{(\nu+L-j)} \\ &= \sum_{\nu=0}^{\infty} A_{\nu+L-k+1}^{(k)} G^{\nu} \\ &= \sum_{\nu=0}^{\infty} \int_{t=0}^{\infty} \int_{v=0}^{t} P_{L-k}^{(k)}(v) D_1^{(L)} P_\nu(t-v)\, \mathrm{d}v\, \mathrm{d}H(t) G^{\nu}. \end{aligned}$$

For the second equality, we use the fact that $\nu + L - j \geq L + 1$. Rearranging terms and using the equality $\sum_{\nu=0}^{\infty} P_\nu(t-v)G^\nu = e^{D[G](t-v)}$ gives the final result. □

Step 2. The generating function $X^*(z)$ for the distribution of $\mathbf{x}$, as derived by Shin and Pearce [124], is given in Equation (6.9). Our goal is to obtain an intuitively more appealing form analogous to the M/G/1 case, for the distribution of $\mathbf{x}$. To do so, we use the following representation of its GF (see Subsection 6.6.2 for the derivation using tedious calculus):

Theorem 6.2. *The GF of the number of customers at departure instants is*

$$\begin{aligned} X^*(z) &= \sum_{n=0}^{L-1} \mathbf{x}_n z^n + z^L \mathbf{x}_0 \bar{A}_L^{(0)} \left[zB^{(0)}(z) - A(z) \right] [zI - A(z)]^{-1} \\ &\quad + z^L \sum_{i=1}^{L} \mathbf{x}_i \bar{A}_{L+1-i}^{(i)} \left[zB^{(i)}(z) - A(z) \right] [zI - A(z)]^{-1}, \end{aligned}$$

with $\bar{A}_{L+1-i}^{(i)}$ *given in Equation* (6.10) *and*

$$B^{(i)}(z) = (\bar{A}_{L+1-i}^{(i)})^{-1} \int_{t=0}^{\infty} \int_{v=0}^{t} P_{L-i}^{(i)}(v) D_1^{(L)} e^{D(z)(t-v)} \, \mathrm{d}v \, \mathrm{d}H(t).$$

$\bar{A}_L^{(0)}$ *and* $B^{(0)}(z)$ *are defined similarly.*

The expression in Theorem 6.2 can be interpreted similar to the M/G/1 case. The corresponding GF can be decomposed into two parts. The first term $\sum_{n=0}^{L-1} \mathbf{x}_n z^n$ contains all stationary probabilities with less than L customers. The other terms at the rhs correspond to situations with at least L customers. This follows from the term z^L times the GF of a specific $MAP/G/1$ queue, similar as in Equation (20) from Lucantoni [106], which only has probability mass on the non-negative real line. Using this interpretation we obtain the following corollary.

Corollary 6.2. *Let* $\mathbf{x}^{(i)}$ *be the stationary probability vector of the number of customers at departure instants in an* $MAP/G/1$ *queue with* $\tilde{B}(z) = B^{(i)}(z)$*, for* $i = 0, 1, \ldots, L$*,* $\mathbf{x}_0^{(0)} = x_0 \bar{A}_L^{(0)}$*, and* $\mathbf{x}_0^{(i)} = \mathbf{x}_i \bar{A}_{L+1-i}^{(i)}$*, for* $i = 1, \ldots, L$*. Then, for* $n = 1, 2, \ldots,$

$$\mathbf{x}_{L+n} = \sum_{i=0}^{L} \mathbf{x}_n^{(i)}.$$

Hence, the stationary probability vector above L *is a mixture of stationary probabilities of* $MAP/G/1$ *queues with exceptional first service times.*

6.4.3 Queue length at arbitrary moments

The queue length at arbitrary moments is typically determined using the key renewal theorem, followed by lengthy calculations, see e.g., [84] for the level-dependent MAP/G/1 queue. Here, we use a more direct derivation based on Palm theory to relate functionals of the queue length at arbitrary moments to that of the queue length upon service completions. A similar approach has been used for the state-dependent G/G/1 queue [16] to relate the virtual waiting time at arbitrary moments to those at arrival epochs. Let Y denote the queue length at an arbitrary moment and let $f(\cdot)$ be some positive function such that $\mathbb{E}[f(Y)]$ is well defined; for convenience we here assume that $f(0) = 0$. Moreover, let H^r be the stationary excess random variable of H.

Theorem 6.3. *Suppose that $\mathbb{E}[f(Y)]$ is well defined. Then,*

$$\mathbb{E}[f(Y)] = \frac{\mathbb{E}[H]}{\mathbb{E}[H] - \mathbf{x}_0 D_0^{(0)-1}\mathbf{e}} \left(\sum_{l=1}^{\infty} \mathbf{x}_l \sum_{k=0}^{\infty} P_k^{(l)}(H^r)\mathbf{e} f(l+k) - \mathbf{x}_0 D_0^{(0)-1} D_1^{(0)} \sum_{k=0}^{\infty} P_k^{(1)}(H^r)\mathbf{e} f(k+1) \right).$$

The choice of $f(\cdot)$ depends on the performance measure of interest. For instance, for the marginal distribution of the number of customers, take $f(x) = \mathbb{1}(x = k)$ with $k = 1, 2, \ldots$. Using this function, we may readily rederive the distribution of Y according to Hofmann [84]. We note that in [84], the author considers both the level and the phase, whereas Theorem 6.3 only provides the level. To determine the nth moment, take $f(x) = x^n$. In particular, let the vector of the phase-dependent mean number of arrivals in time t, starting from level k, be defined by (see [84])

$$\mathbf{n}^{(k)}(t) = \sum_{n=1}^{\infty} n P_n^{(k)}(t)\mathbf{e}.$$

The first moment of Y in terms of $\mathbb{E}[X]$ and the mean number of arrivals is presented in the following corollary.

Corollary 6.3. *The expected number of customers at arbitrary moments is given by*

$$\mathbb{E}[Y] = \frac{\mathbb{E}[H]}{\mathbb{E}[H] - \mathbf{x}_0 D_0^{(0)-1}\mathbf{e}} \left(\mathbb{E}[X] + \sum_{l=1}^{\infty} \mathbf{x}_l \mathbf{n}^{(l)}(H^r) - \mathbf{x}_0 D_0^{(0)-1} D_1^{(0)} (\mathbf{n}^{(1)}(H^r) + 1) \right).$$

Proof of Theorem 6.3. The proof is based on Palm theory, relating time averages to event averages. Let the events be the departures of customers from the queue and let I denote a generic interdeparture time. Note that interdeparture times correspond to

service times in case the level is positive, whereas an additional arrival is required in case the level is zero. Hence, we have $\mathbb{E}[I] = \mathbb{E}[H] - \mathbf{x}_0 D_0^{(0)-1}\mathbf{e}$.

Now, the key identity that we exploit is (see, e.g., [13, Section 1.3]),

$$\mathbb{E}[f(Y)] = \frac{1}{\mathbb{E}[I]}\,\mathbb{E}\left[\int_0^I f(X_t)\,\mathrm{d}t\right],$$

where X_t is the number of customers at time t and time 0 is a departure epoch. In case $X_0 > 0$, then I corresponds to a service time. Conditioning on the service time, we have

$$\begin{aligned}\mathbb{E}\left[\int_0^H f(X_t)\,\mathrm{d}t\right] &= \mathbb{E}\left[\int_{x=0}^{\infty}\int_{t=0}^{x} f(X_t)\,\mathrm{d}t\,\mathrm{d}H(x)\right]\\ &= \mathbb{E}\left[\int_{t=0}^{\infty}\mathbb{P}(H \geq t) f(X_t)\,\mathrm{d}t\right] = \mathbb{E}[H]\,\mathbb{E}[f(X_{H^r})],\end{aligned}$$

where the second equality follows from interchanging integrals. Noting that X_t only increases between $(0, I)$ by arrivals, we have

$$\mathbb{E}[f(X_t); X_0 > 0] = \sum_{l=1}^{\infty}\sum_{k=0}^{\infty}\mathbf{x}_l P_k^{(l)}(t)\mathbf{e} f(l+k).$$

For the case that $X_0 = 0$, note that the (i,j)th element of $-\mathbf{x}_0 D_0^{(0)-1} D_1^{(0)}$ corresponds to the probability that starting at level 0 in phase i, the first arrival occurs when the process moves to state j. Thus,

$$\mathbb{E}[f(X_t); X_0 = 0] = -\mathbf{x}_0 D_0^{(0)-1} D_1^{(0)} \sum_{k=0}^{\infty} P_k^{(1)}(t)\mathbf{e} f(k+1).$$

Combining the above completes the proof. □

6.5 Numerical experiments

We consider an outpatient department and study the impact of varying the load of the system for various scenarios. Our interest is in the expected waiting time and the probability that extra capacity is used. For all experiments, we fix the service time at exactly 1 (representing one slot or one day) and vary λ yielding different traffic loads expressed in terms of the basic capacity $s^{(1)}$; this means we express the load as $\lambda/s^{(1)} \times 100\%$. The performance measures are at the beginning of the day, i.e., at departure instants in the corresponding queueing model.

6.5.1 Level-independent case

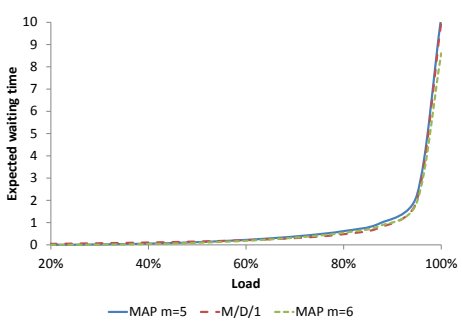
First assume that an outpatient department does not use overbooking of capacity, i.e. $m \equiv m_1 = m_2$. In the first set of experiments, the capacity *exactly* equals five (and six, see Subs 6.5.2, respectively). For this case, the M/D/1 queue (model I) should give a good approximation. For model II, this means that the arrival process actually follows an Erlang distribution with m phases. Since, the service time are deterministic, the variability in the offered traffic is relatively small. This implies that the system can operate at relatively large values of the load with acceptable waiting times. This can also be seen in Figure 6.2a, where the expected waiting time is plotted against the load $\lambda/s^{(1)} \times 100\%$ by varying λ. We see that the expected waiting time is small for low loads and increases slowly with the load until a load of about 90%. When the load gets close to 100%, the expected waiting time increases sharply. Once the load reaches 100% (or more), the available capacity is not sufficient to handle all patients in the long run and the waiting time will explode. We see that the difference between having a capacity of five or six is small, the waiting time in the case of $m = 5$ is only slightly higher than for the case $m = 6$. Also, the M/D/1 approximation fits nicely.

The second experiment is the situation in which the available capacity is random, e.g., due to unavailability of staff. We assume a maximum capacity of nine, whereas the mean number of slots available per day, $\mathbb{E}[m]$, is equal to five (and six, respectively). The arrival process for model II (the time for filling a day) is then based on a Coxian (or mixed-Erlang) distribution. In Figure 6.2b, we see that the shape of the expected waiting time as a function of the load is similar as in Figure 6.2a (and for the Pollaczek-Khinchine formula); however, the expected waiting time is considerably larger for loads above 90% compared to the situation without randomness in the capacity. We also see that the M/D/1 approximation is not able to capture the variability in the capacity available. Note that the effect of having one extra slot available ($\mathbb{E}[m] = 6$) decreases the expected waiting time somewhat.

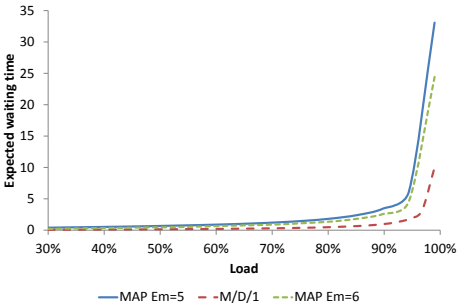
From these two experiments, we observe that M/D/1 queues severely underestimate the waiting time when the capacity is random. Moreover, the range of load values where the expected waiting time is in the order of several weeks (say, between 10 and 30 days) is small, showing that level-dependent features are required to model access times in actual appointment systems.

6.5.2 Level-dependent case

We analyze the level-dependent case using similar experiments as in Subsection 6.5.1. The basic capacity is again equal to five; when the waiting time is more than ten, the capacity is increased to either six or seven. This is denoted by $(5, 6)$ and $(5, 7)$ in the legends of Figures 6.3 and 6.4. First, we consider the case that the basic capcity is always used and is thus deterministic. This situation can be well approximated with



(a) Without variation in available capacity.



(b) With some variation in available capacity.

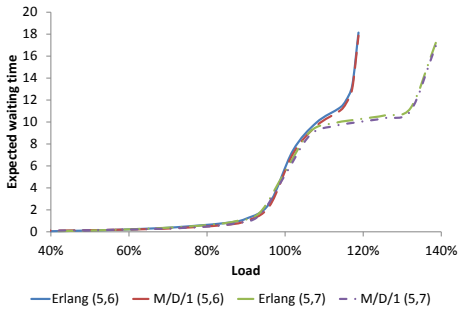
Figure 6.2: Expected waiting times for different loads in the level-independent case.

model I (M/D/1 type of system), as shown in Figures 6.3a and 6.4a. The second experiment is for the case with variability in the basic capacity, i.e. where some slots might be unavailable. The results for this case are depicted in Figures 6.3b and 6.4b.

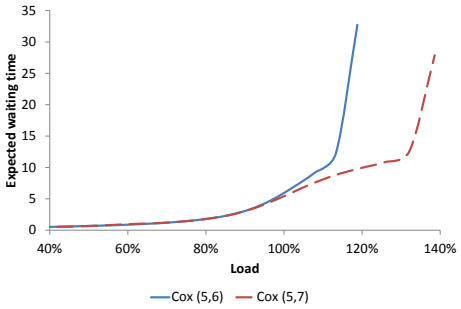
Figure 6.3 illustrates the effect of the load on the expected waiting time. When the load is roughly between 90% and 100%, the waiting time increases sharply. However, the expected waiting time then stabilizes for loads above 100%, in contrast with the level-independent situation. Hence, there is a much larger range of loads leading to expected waiting times in the order of several weeks, as we also observe in practice. Of course, when the load approaches $\lambda/s^{(2)} \times 100\%$ the system becomes unstable and the waiting time tends to explode. With more extra capacity (m_2 is 7 compared to 6), this will happen for higher load values.

Comparing the situations with and without randomness in available capacity, i.e. Figures 6.3a and 6.3b, then we see that the expected waiting time is (somewhat) larger in case of random capacity. Also, the behavior of the expected waiting time around 100% evolves more gradually for random capacity compared to fixed capacity. For random capacity, we omitted the comparison with model I due to the poor fit.

Another interesting performance measure is the probability that extra capacity is used; these probabilities can be found in Figure 6.4. We see that the extra capacity is only used if the load is close to 100%, or higher. When the available capacity is deterministic, the extra capacity is typically always used or not used at all (depending on the load); there is only a small range of loads where the probability of using extra capacity is between 0 and 1 (Figure 6.4a). With more variability in the capacity, the probability of using extra capacity is less steep, especially if there is more extra capacity available (Figure 6.4b).

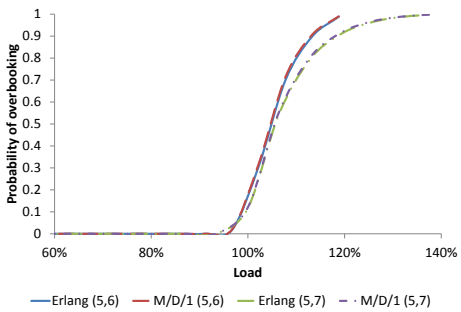


(a) Without variation in available capacity.

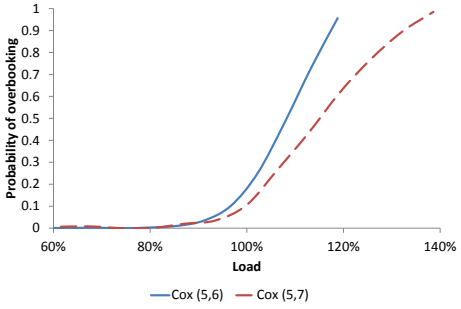


(b) With some variation in available capacity.

Figure 6.3: Expected waiting times for different loads in the level-dependent case.



(a) Without variation in available capacity.



(b) With some variation in available capacity.

Figure 6.4: Probability of overbooking for different loads in the level-dependent case.

							Costs	
L	m_1	m_2	SCV1	SCV2	$\mathbb{E}[X]$	$\mathbb{P}(X \geq L)$	$c_2 = 1$	$c_2 = 1.5$
5	5	7	0.45	0.25	5.62	0.89	6.79	7.68
		8	0.45	0.21	5.03	0.84	7.53	8.79
		9	0.45	0.17	4.88	0.82	8.30	9.95
		10	0.45	0.15	4.82	0.82	9.08	11.12
	6	7	0.40	0.30	5.40	0.69	**6.69**	7.03
		8	0.40	0.24	4.28	0.53	7.06	7.59
		9	0.40	0.20	4.08	0.48	7.44	8.17
		10	0.40	0.17	4.01	0.46	7.86	8.78
	7	7	0.37	0.37	4.13	0.34	7.00	**7.00**
		8	0.37	0.29	2.40	0.14	7.14	7.21
		9	0.37	0.23	2.27	0.11	7.22	7.34
		10	0.37	0.19	2.22	0.10	7.31	7.46
10	5	7	0.45	0.25	10.62	0.89	6.79	7.68
		8	0.45	0.21	10.03	0.84	7.53	8.79
		9	0.45	0.17	9.88	0.82	8.30	9.95
		10	0.45	0.15	9.82	0.82	9.08	11.12
	6	7	0.40	0.30	10.23	0.67	**6.67**	7.01
		8	0.40	0.24	9.07	0.51	7.02	7.52
		9	0.40	0.20	8.85	0.46	7.38	8.07
		10	0.40	0.17	8.78	0.44	7.77	8.66
	7	7	0.37	0.37	4.13	0.10	7.00	**7.00**
		8	0.37	0.29	3.38	0.03	7.03	7.05
		9	0.37	0.23	3.31	0.03	7.05	7.08
		10	0.37	0.19	3.29	0.02	7.07	7.10

Table 6.1: Performance measures and costs for different scenarios.

6.5.3 Optimization

The experiments above indicate that outpatient departments can operate at high load while maintaining acceptable waiting times. Now, we set up an experiment for finding an optimal capacity configuration. To capture some variability in capacity, we assume here that 75% of the base capacity m_1 will be available on average. For example, if the base capacity is 5, there are on average 3.75 slots available per day. The extra capacity, $(m_2 - m_1)$, is always used, so the expected number of available slots above L is given by $0.75m_1 + (m_2 - m_1) = m_2 - 0.25m_1$. We are interested in the expected waiting time $\mathbb{E}[X]$ and the probability that extra capacity is used $\mathbb{P}(X \geq L)$. Suppose that every unit of basic capacity has cost $c_1 = 1$, and extra capacity costs c_2 if it is used. The total costs are then equal to $m_1 + c_2(m_2 - m_1)\,\mathbb{P}(X \geq L)$. These costs are calculated for different values of L, m_1, m_2 and c_2. The results can be found in Table 6.1.

From Table 6.1 it can be seen that $\mathbb{E}[X]$ lies around L, so if we have a constraint $\mathbb{E}[X] \leq 6$, we can set $L = 5$. If $c_2 = 1$, the cheapest option is to have a basic capacity of six and add one extra slot if the waiting time exceeds L. If L is chosen higher, extra capacity is used less often, reducing the costs, but increasing the expected waiting time. If $c_2 = 1.5$, we see that the cheapest solution is not to use overbookings. With a basic capacity of seven, it holds that $\lambda/s^{(1)} = 5/(0.75 \times 7) = 20/21 \approx 95.2\%$. In this case, roughly 5% of the slots will remain unfilled. Using a lower basic capacity and adding extra capacity if needed reduces this idle time and could be a reasonable choice.

6.6 Appendix

6.6.1 Proof of Theorem 6.1

Using the balance equations (6.1) and taking GF's, we may write

$$\begin{aligned}
X^*(z) &= \sum_{n=0}^{L} x_n z^n + \sum_{n=L+1}^{\infty} \left(\sum_{i=1}^{n+1} x_i \alpha_{n+1-i}^{(i)} + x_0 \alpha_n^{(1)} \right) z^n \\
&= \sum_{n=0}^{L} x_n z^n + \sum_{n=L+1}^{\infty} \sum_{i=1}^{L} x_i \alpha_{n+1-i}^{(i)} z^n + \sum_{n=L+1}^{\infty} \sum_{i=L+1}^{n+1} x_i \alpha_{n+1-i}^{(i)} z^n \\
&\quad + x_0 \sum_{n=L+1}^{\infty} \alpha_n^{(1)} z^n \\
&= \sum_{n=0}^{L} x_n z^n + \frac{1}{z} \sum_{i=L+1}^{\infty} x_i z^i \sum_{l=0}^{\infty} z^l \hat{\alpha}_l - z^L x_{L+1} \hat{\alpha}_0 \\
&\quad + x_0 \sum_{n=L+1}^{\infty} z^n \alpha_n^{(1)} + \sum_{n=L+1}^{\infty} z^n \sum_{i=1}^{L} x_i \alpha_{n+1-i}^{(i)} \\
&= \sum_{n=0}^{L} x_n z^n + \frac{1}{z} X^*(z) H^*(\lambda_2(1-z)) - \frac{1}{z} \sum_{n=0}^{L} x_i z^i H^*(\lambda_2(1-z)) \\
&\quad - z^L x_{L+1} \hat{\alpha}_0 + x_0 \sum_{n=L+1}^{\infty} z^n \alpha_n^{(1)} + \sum_{i=1}^{L} x_i \sum_{n=L+1}^{\infty} z^n \alpha_{n+1-i}^{(i)}.
\end{aligned}$$

Rearranging terms gives

$$\begin{aligned}
X^*(z) \left(1 - \frac{1}{z} H^*(\lambda_2(1-z)) \right) &= \sum_{n=0}^{L} x_n z^n \left(1 - \frac{1}{z} H^*(\lambda_2(1-z)) \right) - z^L x_{L+1} \hat{\alpha}_0 \\
&\quad + x_0 \sum_{n=L+1}^{\infty} z^n \alpha_n^{(1)} + \sum_{i=1}^{L} x_i \sum_{n=L+1}^{\infty} z^n \alpha_{n+1-i}^{(i)}. \quad (6.11)
\end{aligned}$$

We now rewrite the final term of (6.11) using probabilistic arguments. Note that the $\alpha_{n+1-i}^{(i)}$ in the final term of (6.11) is the probability of going from $i \leq L$ to $n+1 > L+1$ customers during a service time. This is equal to the probability of having $n-L$ arrivals during T_i, given that T_i is positive, i.e.,

$$\begin{aligned}
\alpha_{n+1-i}^{(i)} &= \mathbb{P}(A(0,H) = n+1-i | X(0) = i) \\
&= \mathbb{P}(A(0,T_i) = n-L | X(0) = i, T_i > 0)\, \mathbb{P}(T_i > 0),
\end{aligned}$$

where $A(0,t)$ denotes the number of arrivals during $(0,t)$. Using the above, we may write

$$\begin{aligned}\sum_{n=L+1}^{\infty} z^n \alpha_{n+1-i}^{(i)} &= \sum_{n=L+1}^{\infty} z^n \,\mathbb{P}(A(0,T_i)=n-L|T_i>0)\,\mathbb{P}(T_i>0)\\ &= z^{L+1}\sum_{n=0}^{\infty} z^n \,\mathbb{P}(A(0,T_i)=n+1|T_i>0)\,\mathbb{P}(T_i>0)\\ &= z^{L}\,\mathbb{P}(T_i>0)\left(\sum_{n=0}^{\infty} z^n \,\mathbb{P}(A(0,T_i)=n|T_i>0)-\mathbb{P}(A(0,T_i)=0|T_i>0)\right).\end{aligned}$$

For the term with the summation, we obtain

$$\begin{aligned}\sum_{n=0}^{\infty} z^n \,\mathbb{P}(A(0,T_i)=n|T_i>0) &= \int_{t=0}^{\infty}\sum_{n=0}^{\infty} z^n e^{-\lambda_2 t}\frac{(\lambda_2 t)^n}{n!}\,\mathrm{d}\,\mathbb{P}(T_i\le t|T_i>0)\\ &= \int_{t=0}^{\infty} e^{-\lambda_2(1-z)t}\,\mathrm{d}\,\mathbb{P}(T_i\le t|T_i>0)\\ &= \hat{T}_i^*(\lambda_2(1-z)).\end{aligned}$$

Based on these probabilistic arguments, we can rewrite (6.11) as

$$\begin{aligned}X^*(z)\left(1-\frac{1}{z}H^*(\lambda_2(1-z))\right) &= \sum_{n=0}^{L} x_n z^n\left(1-\frac{1}{z}H^*(\lambda_2(1-z))\right)-z^L x_{L+1}\hat{\alpha}_0\\ &+ z^L x_0\,\mathbb{P}(T_1>0)\left(\hat{T}_1^*(\lambda_2(1-z))-\mathbb{P}(A(0,T_1)=0|T_1>0)\right)\\ &+ z^L\sum_{i=1}^{L} x_i\,\mathbb{P}(T_i>0)\left(\hat{T}_i^*(\lambda_2(1-z))-\mathbb{P}(A(0,T_i)=0|T_i>0)\right).\end{aligned}\tag{6.12}$$

Rearranging terms in (6.1) yields the relation

$$\begin{aligned}x_L = x_{L+1}\hat{\alpha}_0 &+ x_0\,\mathbb{P}(T_1>0)\,\mathbb{P}(A(0,T_1)=0|T_1>0)\\ &+\sum_{i=1}^{L} x_i\,\mathbb{P}(T_i>0)\,\mathbb{P}(A(0,T_i)=0|T_i>0).\end{aligned}$$

Using this in Equation (6.12) and multiplying by $-z$, gives

$$\begin{aligned}X^*(z)\left(H^*(\lambda_2(1-z))-z\right) = \sum_{n=0}^{L} x_n z^n\left(H^*(\lambda_2(1-z))-z\right)+z^{L+1}x_L\\ - z^{L+1}x_0\,\mathbb{P}(T_1>0)\hat{T}_1^*(\lambda_2(1-z))\\ - z^{L+1}\sum_{i=1}^{L} x_i\,\mathbb{P}(T_i>0)\hat{T}_i^*(\lambda_2(1-z)).\end{aligned}\tag{6.13}$$

Finally, we use another type of level crossing argument; the rate of moving from $L+1$ to L (that is x_L) should be equal to the rate of moving from L to $L+1$,

$$x_L = x_0\,\mathbb{P}(T_1 > 0) + \sum_{i=1}^{L} x_i\,\mathbb{P}(T_i > 0). \tag{6.14}$$

We add $z^L H^*(\lambda_2(1-z))(x_0\,\mathbb{P}(T_1 > 0) + \sum_{i=1}^{L} x_i\,\mathbb{P}(T_i > 0))$ to Equation (6.13) and subtract $z^L H^*(\lambda_2(1-z))x_L$, which is equal by (6.14). Rearranging terms gives Equation (6.5).

Now it remains to determine x_0, because the generating function in Equation (6.5) is completely determined in terms of x_0, due to Corollary 6.1. Using the fact that $X^*(1) = 1$, we see that

$$\begin{aligned}
1 = &\sum_{n=0}^{L-1} x_0 \frac{\mathbb{P}(X^{Q1}=n)}{\mathbb{P}(X^{Q1}=0)} \\
&+ x_0 \frac{\mathbb{P}(X^{Q1}=0)}{\mathbb{P}(X^{Q1}=0)}\,\mathbb{P}(T_1>0)\lim_{z\to 1}\frac{H^*(\lambda_2(1-z)) - z\hat{T}_1^*(\lambda_2(1-z))}{(H^*(\lambda_2(1-z))-z)} \\
&+ \sum_{i=1}^{L} x_0 \frac{\mathbb{P}(X^{Q1}=i)}{\mathbb{P}(X^{Q1}=0)}\,\mathbb{P}(T_i>0)\lim_{z\to 1}\frac{H^*(\lambda_2(1-z)) - z\hat{T}_i^*(\lambda_2(1-z))}{(H^*(\lambda_2(1-z))-z)} \\
= &\frac{x_0}{\mathbb{P}(X^{Q1}=0)}\left(\sum_{n=0}^{L-1}\mathbb{P}(X^{Q1}=n) + \mathbb{P}(X^{Q1}=0)\,\mathbb{P}(T_1>0)\frac{-\rho_2+\lambda_2\,\mathbb{E}[\hat{T}_1]+1}{1-\rho_2}\right. \\
&\left.+ \sum_{i=1}^{L}\mathbb{P}(X^{Q1}=i)\,\mathbb{P}(T_i>0)\frac{-\rho_2+\lambda_2\,\mathbb{E}[\hat{T}_1]+1}{1-\rho_2}\right).
\end{aligned}$$

For the limit we use l'Hôpital's rule . Some rewriting gives the result for x_0, completing the proof.

6.6.2 Proof of Theorem 6.2

Using (6.8) and similar steps as for the M/G/1 case, we obtain

$$\begin{aligned}
X^*(z)\,[zI - A(z)] \;=\; &\sum_{n=0}^{L}\mathbf{x}_n z^n\,[zI - A(z)] - z^{L+1}\mathbf{x}_{L+1}A_0 \\
&+ z\mathbf{x}_0\sum_{n=L+1}^{\infty} A_n^{(0)} z^n + z\sum_{i=1}^{L}\mathbf{x}_i\sum_{n=L+1}^{\infty} A_{n+1-i}^{(i)} z^n.
\end{aligned} \tag{6.15}$$

For an intuitively appealing form, it is now crucial to rewrite $\sum_{n=L+1}^{\infty} z^n A_{n+1-i}^{(i)}$ for $i = 0, 1, \ldots, L$. To do this, we focus on the final three terms at the right-hand side of

(6.15). First, rearranging terms in (6.8), yields

$$\mathbf{x}_{L+1}A_0 = \mathbf{x}_L - \mathbf{x}_0 A_L^{(0)} - \sum_{i=1}^{L} \mathbf{x}_i A_{L+1-i}^{(i)}.$$

Using the above to rewrite the final three terms at the rhs of (6.15), we have

$$\begin{aligned}
&-z^{L+1}\mathbf{x}_{L+1}A_0 + z\mathbf{x}_0 \sum_{n=L+1}^{\infty} A_n^{(0)} z^n + z\sum_{i=1}^{L}\mathbf{x}_i \sum_{n=L+1}^{\infty} A_{n+1-i}^{(i)} z^n \\
&= -z^{L+1}\mathbf{x}_L + z^{L+1}\mathbf{x}_0 \left(A_L^{(0)} + \sum_{n=L+1}^{\infty} A_n^{(0)} z^{n-L} \right) \\
&\quad + z^{L+1} \sum_{i=1}^{L} \mathbf{x}_i \left(A_{L+1-i}^{(i)} + \sum_{n=L+1}^{\infty} A_{n+1-i}^{(i)} z^{n-L} \right) \\
&= -z^{L+1}\mathbf{x}_L + z^{L+1}\mathbf{x}_0 \sum_{n=0}^{\infty} A_{L+n}^{(0)} z^n + z^{L+1} \sum_{i=1}^{L} \mathbf{x}_i \sum_{n=0}^{\infty} A_{L-i+1+n}^{(i)} z^n.
\end{aligned}$$

Now, we use the level-dependent version of the equation at the bottom of page 186 of Ramaswami [119]:

$$\mathbf{x}_L = \mathbf{x}_0 \bar{A}_L^{(0)} + \sum_{i=1}^{L} \mathbf{x}_i \bar{A}_{L+1-i}^{(i)}, \tag{6.16}$$

where $\bar{A}_{L+1-i}^{(i)} = \sum_{\nu=L+1-i}^{\infty} A_\nu^{(i)} G^{\nu-(k+1-i)}$, see also [84, Theorem 5.1]. Right multiply Equation (6.16) with $A(z)$ and with z^L, and combining the above, we obtain

$$\begin{aligned}
X^*(z)\left[zI - A(z)\right] &= \sum_{n=0}^{L} \mathbf{x}_n z^n \left[zI - A(z)\right] - z^L \mathbf{x}_L \left[zI - A(z)\right] \\
&+ z^L \mathbf{x}_0 \bar{A}_L^{(0)} \left[zB^{(0)}(z) - A(z) \right] + z^L \sum_{i=1}^{L} \mathbf{x}_i \bar{A}_{L+1-i}^{(i)} \left[zB^{(i)} - A(z) \right],
\end{aligned} \tag{6.17}$$

with $B^{(0)}(z) = (\bar{A}_L^{(0)})^{-1} \sum_{n=0}^{\infty} A_{L+n}^{(0)} z^n$ and $B^{(i)}(z) = (\bar{A}_{L+1-i}^{(i)})^{-1} \sum_{n=0}^{\infty} A_{L+1-i+n}^{(i)} z^n$.

Now, rewriting (6.17) yields the result. Conditioning on the moment that the process hits level $L+1$ twice yields more explicit expressions for $B^{(i)}(z)$.

6.6.3 Mean queue length for LD-MAP/G/1

The stationary mean queue length at departure epochs is (cf. [124]),

$$\mathbb{E}[X] = \frac{1}{2(1-\rho_2)} \left[\mathbf{p}(1)A''\mathbf{e} + U''(1)\mathbf{e} + 2\left[U'(1) - \mathbf{p}(1)(I - A')\right](I - A' + \mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{a} \right],$$

where we define $A^{'(k)} = \frac{\mathrm{d}}{\mathrm{d}z}A^{(k)}(z)\Big|_{z=1}$, $A^{''(k)} = \frac{\mathrm{d}^2}{\mathrm{d}z^2}A^{(k)}(z)\Big|_{z=1}$, and

$$\mathbf{p}(1) = \left(\boldsymbol{\pi} + \sum\nolimits_{i=0}^{L} \mathbf{x}_i(A^{(i)} - A)\right)(I - A + \mathbf{e}\boldsymbol{\pi})^{-1}$$

$$U'(1) = \mathbf{x}_0 A^{(0)} + \sum_{i=1}^{L} i\mathbf{x}_i(A^{(i)} - A) + \sum_{i=0}^{L} \mathbf{x}_i(A^{'(i)} - A'),$$

$$U''(1) = 2\mathbf{x}_0 A^{'(0)} + \sum_{i=2}^{L} i(i-1)\mathbf{x}_i(A^{(i)} - A) + 2\sum_{i=1}^{L} i\mathbf{x}_i(A^{'(i)} - A') + \sum_{i=0}^{L} \mathbf{x}_i(A^{''(i)} - A'').$$

Chapter 7

Waiting-time distributions in call blending models with abandonments

7.1 Introduction

In this chapter we consider a multi-server queue with two types of customers. Just like in the previous chapter, the server can handle both types of customers and there is an infinite supply of the second type. Here type-2 customers are only taken into service if there are no type-1 customers waiting and there are enough free servers to handle incoming type-1 jobs immediately. This chapter is motivated from a call center perspective. We refer to the system as a blended system, where homogeneous servers handle two classes of customers: *urgent* and *best effort*. The question is to determine the performance of such a blended system. Such a performance analysis is key to determine appropriate staffing levels as well as the control of assigning servers to customers. Motivated by practice and theoretical optimality results, we use a threshold structure for the number of servers that should always be kept available for the 'urgent' class. We derive the probability to abandon and the waiting-time distribution by considering the waiting time process of the first customer in line. Herewith, we extend the method in [18] by incorporating elements of the so-called system-point method due to Brill and Posner [45; 44]. As such, we avoid to analyze quantities involving the number of customers, which are typically less relevant for the management of the system.

The idea of studying the waiting time of the first customer in line directly was addressed in [18], but is now more involved as the total service rate depends on the types of customers in service and due to abandonments. The method in this chapter is thus an extension of the method followed in [18]. We see a dichotomy in results depending on the service rates of the urgent and best effort class. If these are the same, the performance measures have a similar structure as for the M/M/s+G model. If the service rates are different, the waiting time distribution of the first customer in line can iteratively be found as solutions of second-order differential equations. In the case

of infinite patience, the waiting times can be expressed as a mixture of exponential terms. The required constants follow from a set of linear equations.

We envisage two streams of related literature: (1) queueing models for blended call centers, and (2) multi-server queues with setup and/or vacation times. Regarding (1), the literature related to queueing models for multi-class call centers is extensive; we will only highlight the literature that is most closely connected to the current chapter. The performance of a blended call center has been analyzed in [23; 60; 118]; [23; 118] use simulation, whereas [60] developed approximative models using Markov chains. Blending policies have been considered in [25; 72; 104]; the models in those papers are related to ours. In their analysis, both [25; 72] use a Markov decision process framework to determine the structure of effective routing policies. They both conclude that threshold control is optimal when the expected service times are identical for the two classes, but this is no longer necessarily true if the expected service times are different. In [104], the authors assume a threshold policy. They first obtain the steady-state number of customers in the system and exploit that result to give involved expressions for the waiting time. Note that these papers all assume that customers do not abandon.

In the call center literature, many papers haven been devoted to an asymptotic analysis of multi-server queues, often in the many-server heavy-traffic regime. Here, we only refer to settings with homogeneous servers. In that case, the models in [8; 9] consider a system with call backs, i.e., the customer is given the opportunity to be called back, which is advantageous for customers that do not need a swift response (and are essentially best effort customers). In addition to the asymptotic analysis, the callback option also leads to a modified model with state-dependent arrival rates. The papers [78; 109] consider service-level differentiation in queues with fully flexible (or homogeneous) servers; in [78] the authors derive asymptotically optimal staffing and scheduling schemes, whereas [109] considers minimal staffing subject to SL constraints. The authors in [116] consider a system that is closely related to ours; the model only differs with respect to the patience distribution, as [116] requires exponential patience. The key difference with [116], however, is that they consider a many-server heavy traffic scaling, whereas we focus on exact analysis of the waiting time.

Using the analysis of the waiting time of the first customer in line does not occur frequently in the literature. The method in this chapter builds on [18], where a considerably different model is studied. Another example is the working paper [103], where the authors use a similar idea to study a multi-server queue with generally distributed abandonment times. The latter paper however is geared towards optimization by considering a discrete state space. Moreover, there is only a single class of customers, whereas they do not obtain any closed-form results for performance measures.

For (2), we note that the dynamics of the urgent class show similarities with M/M/s (or actually M/M/s+G) queues with setup times or vacations; if an urgent customer arrives and the server has to switch between classes, this could be interpreted as

a setup time or vacation. The literature on multi-server queues with vacations is limited; see [92] for a short survey. Moreover, these systems do typically assume that customers do not abandon. The authors of [161] study a vacation queue, where the server that completes a service will take a vacation, but only if there are less than d servers already on vacation. This model is closely related to our blended system, albeit without abandonments. The authors find the waiting time in terms of its Laplace-Stieltjes transform by analyzing the number of customers first. Models with different vacation scenarios, less relevant to our model, can be found in [128; 24; 157]. A study where abandoments do occur is [5], which considers an M/M/s model with server vacations and abandonments; a key assumption in [5] is though that abandonments only occur when the server is absent.

Finally, there is also little literature on multi-server queues with set-up times. In [10] the number of customers and the waiting-time distributions are derived, but their model assumptions are somewhat restrictive. In [70; 71], M/M/s queues with set-up times are analyzed, but the performance is either approximate or not closed-form. We refer to [71] for additional references, also related to multi-server queues with vacations. Moreover, all these papers assume that customers have infinite patience.

The remainder of the chapter is organized as follows. The model for blending of traffic is introduced in Section 7.2. For the analysis we distinguish two cases depending on the average service times of both classes; when they are equal the analysis is much simpler and can be found in Section 7.3, the analysis in case of unequal mean service times can be found in Section 7.4.

7.2 Model description

We consider a multi-server queue with two types of traffic, type 1 and type 2, see Figure 1.4. We refer to traffic as jobs, but they may equivalently be interpreted as customers, calls, or patients, depending on the specific application. Jobs of type i have independent exponentially distributed service requirements with rate μ_i. Type-1 jobs arrive according to a Poisson process with rate λ, and there is an infinite supply of type-2 jobs. There are s identical servers and there is an infinite waiting room for type 1. Let $\rho = \lambda/(s\mu_1)$. Type-1 jobs have general patience of length G, which is a random variable with cumulative distribution function $G(\cdot)$ and hazard rate $h_G(\cdot)$. The hazard rate function is assumed to be continuous. If the waiting time in the queue of a type-1 job exceeds its patience, the job abandons the queue and will not receive service. In case customers have infinite patience, we let $G(\cdot) \equiv 0$ and $h_G(\cdot) \equiv 0$ and assume that $\rho < 1$ for stability of type 1.

Type 1 represents the class of 'urgent' jobs. They typically have to meet a traditional service level (SL) target in terms of the waiting time distribution. Let W denote the stationary waiting time for type-1 jobs. Then the objective is to evaluate $\mathbb{P}(W > t)$ for some constant t indicating whether the SL is met. The probability to abandon is

another performance measure of interest. For type 2 the performance measure is its throughput, i.e., the long-term average number of type-2 jobs that are served. This may be equivalently interpreted as the long-term average amount of time available for type 2.

To benefit from blending the two traffic classes, we consider a simple threshold type of control. Let $N \in \{0, 1, \ldots, s\}$ denote the threshold specifying the number of servers that we keep available for type 1. This is in contrast to Chapters 2–5, where N denotes the number of queues. Since N servers are always kept available for type 1, at most $s - N$ type-2 jobs can be simultaneously in service. The threshold policy takes the following actions:

- Type-1 jobs are taken into service as soon as a server is available.
- When the type-1 queue is empty and there are no more than $s - N$ type-1 jobs in service, as many type-2 jobs are taken into service such that N servers remain idle (to handle future type-1 traffic); type-2 traffic only starts to be served after the $(N + 1)$-th server becomes available. The number of type-1 jobs in service does not play a role (as long as this does not exceed $s-N$). If $l = 0, 1, \ldots, s-N$ type 1 jobs are in service, $s - N - l$ type-2 jobs are taken into service.
- Preemption of service is not allowed; if preemption would be allowed it would always be preferred to serve a type-2 job over being idle.

Type 1 is protected from type 2 by $N > 0$, whereas the utilization of the service facility is increased compared to a single type-1 system by letting $N < s$. Finally, we note that optimal control for this system has been investigated in [25] in case the target for type 1 is in terms of the average waiting time instead of the tail distribution (the latter being considerably more involved). For $\mu_1 = \mu_2$, the authors show that threshold control is optimal in that setting, whereas numerical experiments indicate that threshold control is near optimal for unequal service requirements; the authors do not consider abandonments in their model.

7.3 Analysis for equal service requirements

In this section we assume that $\mu := \mu_1 = \mu_2$, which considerably simplifies the analysis. In this case, the analysis could follow the lines of the M/M/s+G model. However, we propose a more direct approach by directly considering the waiting time, or more specifically, the waiting time of the first type-1 job in line (FIL), see, e.g., [18]. This approach turns out to be extremely valuable in the next section as well, where $\mu_1 \neq \mu_2$. Now, we first give a description of the process in Subsection 7.3.1. Then we give the analysis of the FIL distribution in Subsection 7.3.2 and we conclude the section with our performance measures in Subsection 7.3.3.

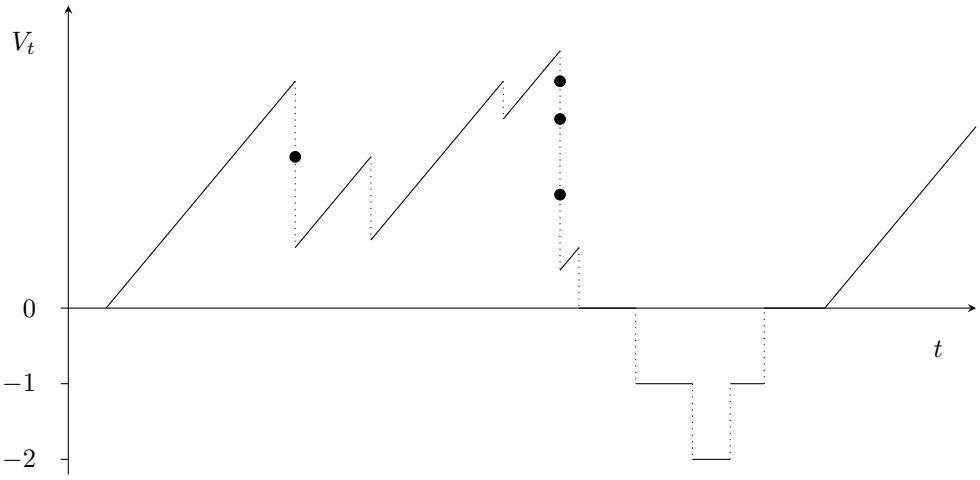


Figure 7.1: Sample path of V_t.

7.3.1 Process description

Let V_t denote the waiting time of the first job in line at time t (e.g. with $V_t = 0$ if the queue is empty) and let N_t denote the number of free servers. The process $(V_t, N_t)_{t\geq 0}$ is then a piecewise deterministic Markov process as considered in [56]. Note that $V_t > 0$ if and only if $N_t = 0$, i.e. free servers only occur when there are no type-1 jobs waiting. For convenience, we omit N_t and extend V_t to the non-positive integers, where $V_t \leq 0$ denotes that the queue is empty and $-V_t$ servers are free. The state space then consists of all negative integers not smaller than $-N$ and all positive real numbers

$$\{-N, -N+1, ..., -1, 0\} \cup (0, \infty).$$

A sample path of the process $(V_t)_{t\geq 0}$ is depicted in Figure 7.1 for the case $N = 2$; the black dots represent jobs that abandoned before they became the first job in line. For $V_t > 0$, the process increases linearly at rate 1 until the first job in line leaves the queue; this can either be due to a service completion, occurring with rate $s\mu$, or due to this job becoming impatient. If the first job in line leaves the queue, the next job that did not abandon (if any) moves to the first position.

For this situation, suppose that at time t, the n-th arrival leaves the first position in the queue and thus the $(n+1)$-th arrival is the next candidate to become first in line. Note that the interarrival time between jobs n and $n+1$ is exponentially distributed and let A_λ denote an exponentially distributed random variable with rate λ. Now, different situations may occur. First, with probability $\mathbb{P}(A_\lambda > V_{t^-})$ the $(n+1)$-th job has not arrived at time t, meaning that at time t^+ the queue is empty whereas all servers are occupied, i.e. $V(t^+) = 0$. Second, if job $n+1$ has arrived his accumulated

waiting time would be $V_{t^-} - A_\lambda$. With probability $1 - G(V_{t^-} - A_\lambda)$, this time does not exceed his patience and job $n+1$ becomes first in line. Finally, with probability $G(V_{t^-} - A_\lambda)$ customer $n+1$ abandoned and customer $n+2$ is the next candidate to become first in line. Now, the scenarios above apply for customer $n+2$ and this procedure repeats.

Consequently, when the first job in line leaves the queue at time t, it holds that

$$V_{t^+} = \max\{V_{t^-} - \tilde{A}_\lambda, 0\}, \tag{7.1}$$

with $\tilde{A}_\lambda$ the jump size, a sum of one or more interarrival times depending on the number of customers that abandoned. The distribution of $\tilde{A}_\lambda$ is considered in Lemma 7.1 below. As indicated, if there are no arrivals before t that did not abandon, the process moves to state 0; the fact that at time t^- a customer is waiting implies that all servers are occupied. In the boundary case that $V_t = 0$ an arrival may occur with rate λ, in which case V_t starts to increase linearly again. The other boundary transitions are standard; in state $-l$, $l = 1, \ldots, N$, jobs arrive with rate λ moving the process to state $-l+1$, and in state $-l$, $l = 0, \ldots, N-1$, jobs depart at rate $(s-l)\mu$ moving the process to state $-l-1$.

Since interarrival times and service times are exponentially distributed, the process $(V_t)_{t\geq 0}$ has the Markov property. In addition, it is a regenerative process and thus has a steady-state distribution. We will use the abbreviation 'FIL' for 'first in line'. Let $W^{FIL}(x) = \lim_{t\to\infty} \mathbb{P}(V_t \leq x)$ be the steady-state distribution with density $w^{FIL}(x)$. For the boundary states we write: $w^{FIL}(-l) = \lim_{t\to\infty} \mathbb{P}(V_t = -l)$, for $l = 0, \ldots, N$.

In Lemma 7.1 we present the distribution function of the jump size $\tilde{A}_\lambda$. Specifically, let $F_x(y)$ be the probability that the FIL process is smaller than $y \in [0, x]$, when right before the jump the process was at x and a jump occurred.

Lemma 7.1. *The distribution function $F_x(y)$, $0 \leq y \leq x$, of the jump size with starting position x is given by*

$$F_x(y) = \exp\left\{-\lambda \int_y^x \bar{G}(u)\, \mathrm{d}u\right\}.$$

Proof. Define W_x^+ as the FIL process just after a jump, when the FIL process right before the jump is equal to x. We are interested in

$$F_x(y) = \mathbb{P}(W_x^+ \leq y).$$

Conditioning gives

$$\begin{aligned} F_x(y) &= e^{-\lambda(x-y)} + \int_y^x \lambda e^{-\lambda(x-u)} G(u) F_u(y)\, \mathrm{d}u \\ &= e^{-\lambda(x-y)} + e^{-\lambda x}\lambda \int_y^x e^{\lambda u} G(u) F_u(y)\, \mathrm{d}u. \end{aligned}$$

Now we take the derivative with respect to x, leading to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}F_x(y) &= -\lambda e^{-\lambda(x-y)} - \lambda e^{-\lambda x}\lambda\int_y^x e^{\lambda u}G(u)F_u(y)\,\mathrm{d}u + \lambda e^{-\lambda x}e^{\lambda x}G(x)F_x(y)\\ &= -\lambda F_x(y) + \lambda G(x)F_x(y)\\ &= -\lambda\bar{G}(x)F_x(y),\end{aligned}$$

where $\bar{G}(x) = 1 - G(x)$. We now have a first-order ordinary differential equation (DE), with solution

$$F_x(y) = C\exp\left\{-\int_0^x \lambda\bar{G}(u)\,\mathrm{d}u\right\}.$$

Using the boundary condition $F_y(y) = 1$, we can determine the constant C. This gives $C = \exp\left\{\int_0^y \lambda\bar{G}(u)\,\mathrm{d}u\right\}$, completing the proof. □

7.3.2 Analysis of FIL distribution

For convenience, define $\rho_i = \lambda/((s-i)\mu)$, $i = 0, 1, \ldots, N$, and let an empty product be equal to 1. The next theorem provides the steady-state distribution of the FIL process.

Proposition 7.1. *The density of the FIL process is, for $x > 0$,*

$$w^{FIL}(x) = \lambda w^{FIL}(0)\bar{G}(x)\exp\left\{\int_0^x \lambda\bar{G}(u)\,\mathrm{d}u - s\mu x\right\},$$

where

$$w^{FIL}(0) = \left(\int_0^\infty \lambda\bar{G}(x)\exp\left\{\int_0^x \lambda\bar{G}(u)\,\mathrm{d}u - s\mu x\right\}\,\mathrm{d}x + \sum_{k=0}^{N}\prod_{j=0}^{k-1}\frac{1}{\rho_j}\right)^{-1},$$

For the boundary states, we have, $k = 1, \ldots, N$,

$$w^{FIL}(-k) = w^{FIL}(0)\prod_{j=0}^{k-1}\frac{1}{\rho_j}. \tag{7.2}$$

Proof. For $x > 0$, it follows from level crossings that

$$w^{FIL}(x) = \int_x^\infty F_y(x)(s\mu + h_G(y))w^{FIL}(y)\,\mathrm{d}y. \tag{7.3}$$

The left-hand side corresponds to upcrossings of level x and the right-hand side corresponds to the long-run average number of downcrossings through level x. Observe that we have continuous upcrossings of waiting-time levels and downcrossings by jumps,

where the jump sizes correspond to (multiple) interarrival times between successive customers (in contrast to workloads in single-server queues). Filling in $F_y(x)$ as given in Lemma 7.1 and some rewriting, we get

$$w^{FIL}(x) = \exp\left\{\lambda \int_0^x \bar{G}(u)\,\mathrm{d}u\right\} \int_x^\infty \exp\left\{-\lambda \int_0^y \bar{G}(u)\,\mathrm{d}u\right\} (s\mu + h_G(y)) w^{FIL}(y)\,\mathrm{d}y.$$

Taking derivatives with respect to x yields

$$\begin{aligned}
\frac{\mathrm{d}w^{FIL}(x)}{\mathrm{d}x} &= \lambda \bar{G}(x) \exp\left\{\lambda \int_0^x \bar{G}(u)\,\mathrm{d}u\right\} \\
&\quad \times \int_x^\infty \exp\left\{-\lambda \int_0^y \bar{G}(u)\,\mathrm{d}u\right\} (s\mu + h_G(y)) w^{FIL}(y)\,\mathrm{d}y \\
&\quad - \exp\left\{\lambda \int_0^x \bar{G}(u)\,\mathrm{d}u\right\} \exp\left\{-\lambda \int_0^x \bar{G}(u)\,\mathrm{d}u\right\} (s\mu + h_G(x)) w^{FIL}(x) \\
&= \lambda \bar{G}(x) w^{FIL}(x) - (s\mu + h_G(x)) w^{FIL}(x) \\
&= (\lambda \bar{G}(x) - s\mu - h_G(x)) w^{FIL}(x).
\end{aligned}$$

This is a first-order differential equation, which is easily solved as

$$w^{FIL}(x) = D \exp\left\{\int_0^x \lambda \bar{G}(u) - s\mu - h_G(u)\,\mathrm{d}u\right\}. \tag{7.4}$$

Now we find the constant D. Balancing the transitions in and out of $(0, \infty)$ results in

$$\lambda w^{FIL}(0) = \int_0^\infty F_y(0)(s\mu + h_G(y)) w^{FIL}(y)\,\mathrm{d}y. \tag{7.5}$$

From (7.3) and (7.5) we get $\lim_{x \downarrow 0} w^{FIL}(x) = \lambda w^{FIL}(0)$ and from we get (7.4) $\lim_{x \downarrow 0} w^{FIL}(x) = D$. Hence, $D = \lambda w^{FIL}(0)$, implying

$$w^{FIL}(x) = \lambda w^{FIL}(0) \exp\left\{\int_0^x \lambda \bar{G}(u) - s\mu - h_G(u)\,\mathrm{d}u\right\}.$$

Due to properties of the hazard rate, this can be rewritten using $\int_0^x h_G(u)\,\mathrm{d}u = -\log \bar{G}(x)$.

Moreover, for the boundary states (with free servers), we have the well-known balance equations, for $i = 1, \ldots, N$,

$$\lambda w^{FIL}(-i) = (s - i + 1)\mu w^{FIL}(-i + 1).$$

Expressing the constants in terms of $w^{FIL}(0)$ yields Equation (7.2). The result follows by normalization. □

7.3.3 Performance measures

From the FIL process we may obtain the actual performance measures of interest: the waiting time W of served customers, the probability to abandon, and the throughput of the best-effort class. The derivation of the waiting-time distribution from the FIL process follows from a generalization of PASTA; the actual waiting time of a served customer corresponds to the value of the FIL process embedded at moments when a new customer is taken into service. In particular, the waiting time is zero when there are idle agents available upon arrival such that a customer can be directly taken into service. Using the PASTA property, we have

$$\mathbb{P}(W=0) = \sum_{k=1}^{N} w^{FIL}(-k) = w^{FIL}(0) \sum_{k=1}^{N} \prod_{j=0}^{k-1} \frac{1}{\rho_j}.$$

We define the empty sum to be equal to 0 for the case $N = 0$.

Given that a served customer has to wait, the waiting time is the value of the FIL process at epochs just before a downward jump due to a service completion. Before addressing these positive waiting times, we first consider the probability to receive service (and thus also the probability to abandon).

Corollary 7.1 (Probability of service)**.** *The probability that a customer receives service is given by*

$$\mathbb{P}(\textit{service}) = w^{FIL}(0) \left(\sum_{k=1}^{N} \prod_{j=0}^{k-1} \frac{1}{\rho_j} + s\mu \int_0^\infty \bar{G}(x) \exp\left\{ \int_0^x \lambda \bar{G}(u)\, \mathrm{d}u - s\mu x \right\} \mathrm{d}x \right),$$

with $w^{FIL}(0)$ given in Proposition 7.1.

Proof. If we divide the rate at which customers are taken into service by the total arrival rate, we can calculate the probability that a customer receives service. The rate at which customers are taken into service is given by:

$$\begin{aligned} \lambda^* &= \lambda\, \mathbb{P}(W=0) + \int_0^\infty s\mu w^{FIL}(x)\, \mathrm{d}x \\ &= \lambda w^{FIL}(0) \left(\sum_{k=1}^{N} \prod_{j=0}^{k-1} \frac{1}{\rho_j} + s\mu \int_0^\infty \bar{G}(x) \exp\left\{ \int_0^x \lambda \bar{G}(u)\, \mathrm{d}u - s\mu x \right\} \mathrm{d}x \right). \quad (7.6) \end{aligned}$$

Dividing by λ completes the proof. □

The stationary waiting time distribution is given in the following theoreom.

Theorem 7.1 (Waiting time type 1)**.** *The tail of the steady-state waiting time distribution of a job that received service is*

$$\mathbb{P}(W > \alpha | \textit{service}) = \frac{s\mu \int_\alpha^\infty \bar{G}(v) \exp\left\{\int_0^v \lambda \bar{G}(u)\, \mathrm{d}u - s\mu v\right\}\, \mathrm{d}v}{\sum_{k=1}^N \prod_{j=0}^{k-1} \frac{1}{\rho_j} + s\mu \int_0^\infty \bar{G}(x) \exp\left\{\int_0^x \lambda \bar{G}(u)\, \mathrm{d}u - s\mu x\right\}\, \mathrm{d}x}.$$

Proof. We denote with $N(a,b)$ the number of customers taken into service between moments a and b. Let us consider the infinitesimal interval $(t, t+h]$, where $h > 0$. Note that

$$\mathbb{P}(V_t > v, N(t,t+h) = 1) = \int_v^\infty s\mu h w^{FIL}(x)\, \mathrm{d}x + o(h).$$

In addition, $\lim_{h\to 0} \mathbb{P}(N(t,t+h) = 1)/h = \lambda^*$, in stationarity, since the rate at which customers are taken into service is equal to the arrival rate of customers that are served for a stable system. Then, we may compute

$$\begin{aligned}\mathbb{P}(W > \alpha | \text{service}) &= \lim_{h\to 0} \mathbb{P}(V_t > \alpha | N(t,t+h) = 1) = \lim_{h\to 0} \frac{\mathbb{P}(V_t > \alpha, N(t,t+h) = 1)}{\mathbb{P}(N(t,t+h) = 1)} \\ &= \frac{1}{\lambda^*} \int_\alpha^\infty s\mu w^{FIL}(v)\, \mathrm{d}v = \frac{1}{\rho \mathbb{P}(\text{service})} \int_\alpha^\infty w^{FIL}(v)\, \mathrm{d}v.\end{aligned}$$

Using Proposition 7.1, we obtain the result. □

Remark 7.1. If $N = s$, our model represents a regular M/M/s+G queue with no blending, as studied by Zeltyn and Mandelbaum [159]. In this case, we can rewrite $\mathbb{P}$(service) to equal their expression for the same probability. In [90], more performance measures for this model are derived, including $\mathbb{P}(W > \alpha|\text{service})$, which coincides with our expression in the case $N = s$.

The throughput of type 2 may also be obtained from Proposition 7.1 after some straightforward calculations; see [25] for the case of infinite patience.

Corollary 7.2 (Throughput type 2)**.** *The throughput for type-2 traffic is*

$$TH = w^{FIL}(0) \left(\sum_{k=1}^{N} ((s-k)\mu - \lambda) \prod_{j=0}^{k-1} \frac{1}{\rho_j} + s\mu \right),$$

with $w^{FIL}(0)$ given in Proposition 7.1.

Proof. First, observe that the fraction of time that all servers are busy is

$$w^{FIL}(0) + \int_0^\infty w^{FIL}(y) \mathrm{d}y.$$

Now, the throughput of type 2 is the total throughput minus the type-1 throughput. The long-run average service rate is

$$\sum_{k=1}^{N}(s-k)\mu w^{FIL}(-k) + s\mu w^{FIL}(0) + s\mu \int_0^\infty w^{FIL}(y)\,\mathrm{d}y.$$

Using (7.2), the fact that the type-1 throughput is λ^*, cf. (7.6), and some rewriting, we obtain the result. □

7.4 Analysis for different service requirements

In this section, we consider the much more involved case $\mu_1 \neq \mu_2$. We describe the model and derive the level crossing equations in Subsection 7.4.1. Our aim is to derive the type-1 waiting-time distribution (Theorem 7.2) and the throughput of type 2 (Corollary 7.4); both results can be found in Subsection 7.4.5. We do this by directly considering the waiting time of the first job in line, as in the previous section. In Subsections 7.4.2 and 7.4.3, we derive the steady state distribution of the FIL process, for general and infinite impatience, respectively. Situations with idle servers are addressed in Subsection 7.4.4 providing a linear set of equations for obtaining the remaining constants. With this approach, we avoid the analysis of the number of customers in the system. Note that it is a non-trivial step to obtain the waiting-time distribution from the number of customers present.

7.4.1 Level crossings

Let us define the stochastic process $(X_t)_{t\geq 0}$ as the vector of two stochastic processes $(V_t, Y_t)_{t\geq 0}$; for a fixed t, $X_t = (V_t, Y_t)$ where $(V_t)_{t\geq 0}$ is the waiting time of the first in line customer or the number of free servers when the queue is empty, as in Subsection 7.3.1, and $(Y_t)_{t\geq 0}$ is the number of type-2 jobs in service. Again, for $(V_t)_{t\geq 0}$, we denote the number of empty servers as non-positive integers: $0, -1, -2, ..., -N$. The two-dimensional state space is given by

$$(\{-N, -N+1, ..., -1, 0\} \cup (0, \infty)) \times \{0, ..., s-N\}.$$

The process $(X_t)_{t\geq 0}$ is again a piecewise-deterministic Markov process as in [56]. We will refer to the state of Y_t as "page". If $Y_t = k$ during some interval, the sample path of the FIL process evolves as in Section 7.3 (see Figure 7.1 for an example). As soon as Y_t changes, the FIL process jumps to a different page, implying that the overall service rate changes. This is in line with the 'System Point' method of Brill and Posner [45; 44].

We introduce the steady-state version of the FIL process with k type-2 jobs in service as $W^{FIL}(v,k) = \lim_{t\to\infty} \mathbb{P}(V_t \leq v, Y_t = k)$. The joint steady-state density is

$w^{FIL}(v,k)$, for $v > 0$. For cases when there are no customers waiting, $w^{FIL}(-l,k)$ denotes the steady-state probability that there are l idle servers and k type-2 jobs in service, $l \in \{0,1,\ldots,N\}$ and $k \in \{0,1,\ldots,s-N\}$. Also, let $\xi_k = \lim_{t\to\infty} \mathbb{P}(Y_t = k)$. The total service rate is determined by the number of type-2 jobs in service. For convenience, define

$$r_k = (s-k)\mu_1 + k\mu_2$$

as the total service rate at page k if all servers are occupied.

The analysis proceeds as follows. First, we derive level-crossing equations for the FIL process. Using these equations we recursively determine the FIL distribution in Subsections 7.4.2 and 7.4.3.

Lemma 7.2. *We consider the level-crossing equations at page k with upcrossings of level v on the left-hand side and downcrossings on the right-hand for two different cases.*

(i) For $k \in \{0,1,\ldots,s-N-1\}$ and $v > 0$

$$\begin{aligned} w^{FIL}(v,k) + (k+1)\mu_2 \int_v^\infty (1-F_u(v))\, w^{FIL}(u,k+1)\mathrm{d}u \\ = k\mu_2 \left(\xi_k - W^{FIL}(v,k)\right) + \int_v^\infty w^{FIL}(u,k)((s-k)\mu_1 + h_G(u))F_u(v)\mathrm{d}u. \end{aligned}$$

(ii) For $k = s-N$ and $v > 0$

$$\begin{aligned} w^{FIL}(v,s-N) =& (s-N)\mu_2 \left(\xi_{s-N} - W^{FIL}(v,s-N)\right) \\ &+ \int_v^\infty w^{FIL}(u,s-N)(N\mu_1 + h_G(u))F_u(v)\mathrm{d}u. \end{aligned}$$

Proof. Consider case (i) and fix page $k \in \{0,1,\ldots,s-N-1\}$. Using standard infinitesimal arguments and (7.1) gives us, for $h > 0$ small,

$$\begin{aligned} &\mathbb{P}(V_{t+h} > v+h, Y_{t+h} = k) = \\ &\quad \int_v^\infty (1 - hk\mu_2 - h(s-k)\mu_1 - hh_G(u))\, \mathrm{d}\mathbb{P}(V_t \le u, Y_t = k) \\ &\qquad + \int_v^\infty h((s-k)\mu_1 + h_G(u))\, \mathbb{P}(\tilde{A}_\lambda \le u - v)\, \mathrm{d}\mathbb{P}(V_t \le u, Y_t = k) \\ &\qquad + h(k+1)\mu_2 \mathbb{P}(V_t - \tilde{A}_\lambda > v, Y_t = k+1) + o(h). \end{aligned}$$

Subtracting $\mathbb{P}(V_t > v+h, Y_t = k)$ from both sides, dividing by h, and letting $h \downarrow 0$,

we obtain:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(V_t > v, Y_t = k) = &-\frac{\mathrm{d}}{\mathrm{d}v}\mathbb{P}(V_t > v, Y_t = k)\\ &-\int_v^\infty (k\mu_2 + (s-k)\mu_1 + h_G(u))\,\mathrm{d}\mathbb{P}(V_t \le u, Y_t = k)\\ &+\int_v^\infty ((s-k)\mu_1 + h_G(u))(1-F_u(v))\,\mathrm{d}\mathbb{P}(V_t \le u, Y_t = k)\\ &+(k+1)\mu_2\mathbb{P}(V_t - \tilde{A}_\lambda > v, Y_t = k+1).\end{aligned}$$

Now, let $t \to \infty$ and note that, for $l = k+1$,

$$\mathbb{P}(V - \tilde{A}_\lambda \le v, Y = l) = W^{FIL}(v,l) + \int_v^\infty F_u(v) w^{FIL}(u,l)\mathrm{d}u.$$

The above yields

$$\begin{aligned}0 = &\, w^{FIL}(v,k) - \int_v^\infty (k\mu_2 + (s-k)\mu_1 + h_G(u))w(u,k)\,\mathrm{d}u\\ &+\int_v^\infty ((s-k)\mu_1 + h_G(u))(1-F_u(v))w(u,k)\,\mathrm{d}u\\ &+(k+1)\mu_2\left(\xi_{k+1} - W^{FIL}(v,k+1) - \int_v^\infty F_u(v)w^{FIL}(u,k+1)\,\mathrm{d}u\right).\end{aligned}$$

Using the fact that

$$\int_v^\infty w(u,k)\,\mathrm{d}u = \xi_k - W^{FIL}(v,k),$$

and some rewriting, gives

$$\begin{aligned}w^{FIL}(v,k) = &\, k\mu_2\left(\xi_k - W^{FIL}(v,k)\right) + \int_v^\infty ((s-k)\mu_1 + h_G(u))w^{FIL}(u,k)F_u(v)\mathrm{d}u\\ &-(k+1)\mu_2\left(\xi_{k+1} - W^{FIL}(v,k+1)\right)\\ &+(k+1)\mu_2\int_v^\infty w^{FIL}(u,k+1)F_u(v)\mathrm{d}u.\end{aligned}$$

Rewriting $\xi_{k+1} - W^{FIL}(v,k+1)$ in terms of its density provides the level crossing equation for case (i). For case (ii), the term involving page $k+1$ disappears. □

7.4.2 General impatience

In this subsection, we show that the steady-state distribution of the FIL process can be iteratively solved and written as the solution of linear second-order differential equations; this is next presented in Proposition 7.2. We exclude the case that $N = s$,

as no type-2 traffic is taken into service in that case. In fact, this case can be directly obtained from the results in Section 7.3.

Before we present the result, we define, for $k \in \{1, \ldots, s-N\}$,

$$\begin{aligned} a_k(v) &= r_k + h_G(v) - \lambda\bar{G}(v), \\ b_k(v) &= -\lambda\bar{G}(v)\mu_2 k. \end{aligned}$$

For the FIL distribution, we need solutions of linear second-order DE's. Let $\tilde{w}_k(v)$ be the solution of the homogeneous second-order DE

$$w''(v) + a_k(v)w'(v) + b_k(v)w(v) = 0, \qquad \text{with } w(0) = 1, w(\infty) = 0. \tag{7.7}$$

Let, for $j = k+1, \ldots, s-N$, $\tilde{w}^{\text{part}}_{k,j}(v)$ be a particular solution of

$$w''(v) + a_k(v)w'(v) + b_k(v)w(v) = -\lambda\bar{G}(v)(k+1)\mu_2\tilde{w}^{\text{part}}_{k+1,j}(v), \tag{7.8}$$

where we define $\tilde{w}^{\text{part}}_{j,j}(v) = \tilde{w}_j(v)$.

In specific cases, the second-order DE's may be solved analytically. A primary example is the case of infinite patience, see Subsection 7.4.3. In general, there are no closed-form solutions of linear second-order DE's, but packages are available for numerical solutions. An alternative representation can be obtained by rewriting a linear second-order DE with boundary condition as a Fredholm integral equation of the second kind, see e.g. [100]. Such a Fredholm integral equation may be solved by successive substitutions leading to solutions of an infinite sum of iterated kernels. As such a solution does not provide any additional insight, we present our results in terms of solutions of DE's.

Proposition 7.2. *The tail distribution of the FIL process at page $k \in \{1, \ldots, s-N\}$ is*

$$\bar{W}^{FIL}(v,k) = C_k\tilde{w}_k(v) + \sum_{j=k+1}^{s-N} C_j\tilde{w}^{part}_{k,j}(v),$$

with $\tilde{w}_k(\cdot)$ and $\tilde{w}^{part}_{k,j}(\cdot)$ as defined according to (7.7) and (7.8), respectively. At page 0, the FIL density is

$$\begin{aligned} w^{FIL}(v,0) &= C_0\bar{G}(v)\exp\left\{\int_0^v \lambda\bar{G}(u)\,\mathrm{d}u - s\mu_1 v\right\} \\ &\quad + \int_{t=0}^{v} \bar{G}(v)\exp\left\{\int_t^v \lambda\bar{G}(u)\,\mathrm{d}u - s\mu_1(v-t)\right\}\lambda\mu_2\bar{W}^{FIL}(t,1)\,\mathrm{d}t. \end{aligned}$$

Here C_j, $j = 0, \ldots, s-N$, are constants that are determined in Subsection 7.4.4 below.

Observe that the proposition provides an iterative scheme to determine the FIL distribution. Starting at page $s-N$, the tail of the FIL distribution can be iteratively solved at page k using the analysis at pages $k+1, \ldots, s-N$.

Proof. To derive the FIL distribution, we start at page $s-N$ and then iteratively determine the FIL distribution at page k given its tail distribution at page $k+1$.

Page $s-N$: Let us start with page $s-N$, for which we need the solution of case (ii) in Lemma 7.2. Using Lemma 7.1 for the distribution of the jumps, taking the derivative with respect to v in (ii) of Lemma 7.2 and applying similar arguments as in the proof of Proposition 7.1, yields

$$\frac{\mathrm{d}}{\mathrm{d}v}w^{FIL}(v,s-N) = -(s-N)\mu_2 w^{FIL}(v,s-N) - (N\mu_1 + h_G(v))w^{FIL}(v,s-N) \\ + \lambda\bar{G}(v)\left[\int_v^\infty (N\mu_1 + h_G(u))w^{FIL}(u,s-N)\exp\left\{-\lambda\int_v^u \bar{G}(z)\,\mathrm{d}z\right\}\mathrm{d}u\right].$$

Observe that the terms in square brackets may be rewritten using the equation in (ii) of Lemma 7.2, from which we obtain

$$\frac{\mathrm{d}}{\mathrm{d}v}w^{FIL}(v,s-N) + (r_{s-N} + h_G(v) - \lambda\bar{G}(v))w^{FIL}(v,s-N) \\ = \lambda\bar{G}(v)(s-N)\mu_2 W^{FIL}(v,s-N) - \lambda\bar{G}(v)(s-N)\mu_2\xi_{s-N}. \tag{7.9}$$

As $w^{FIL}(v,s-N)$ is the derivative of $W^{FIL}(v,s-N)$, we have a linear second-order differential equation. For convenience, we consider the tail of the FIL distribution $\bar{W}^{FIL}(v,k) = \xi_k - W^{FIL}(v,k)$. Using the fact that $\mathrm{d}/\mathrm{d}v(\bar{W}^{FIL}(v,k)) = -w^{FIL}(v,k)$, Equation (7.9) may be written as the following linear second-order differential equation:

$$\frac{\mathrm{d}^2}{\mathrm{d}v^2}\bar{W}^{FIL}(v,s-N) + a_{s-N}(v)\frac{\mathrm{d}}{\mathrm{d}v}\bar{W}^{FIL}(v,s-N) + b_{s-N}(v)\bar{W}^{FIL}(v,s-N) = 0.$$

If customers have infinite patience, the coefficients are constant and there is a direct way to solve these equations; this case will be treated separately in the next subsection. For the boundary conditions, it should hold that $\lim_{v\to\infty}\bar{W}^{FIL}(v,k) = 0$. The second boundary condition (at $v=0$) is determined later in Subsection 7.4.4 as the solution of a linear system of equations. Hence, the general solution of the above DE is

$$\bar{W}^{FIL}(v,k) = C_{s-N}\tilde{w}_{s-N}(v),$$

where C_{s-N} is the remaining unknown constant. This completes the FIL distribution at page $s-N$.

Page $k \in \{1, \ldots, s-N-1\}$: The approach is similar to the approach for page $s-N$. First we rewrite (i) of Lemma 7.2, yielding

$$\begin{aligned}
&w^{FIL}(v,k) + (k+1)\mu_2 \int_v^\infty w^{FIL}(u,k+1)\,\mathrm{d}u \\
&\quad - (k+1)\mu_2 \exp\left\{\lambda \int_0^v \bar{G}(z)\,\mathrm{d}z\right\} \int_v^\infty \exp\left\{-\lambda \int_0^u \bar{G}(z)\,\mathrm{d}z\right\} w^{FIL}(u,k+1)\,\mathrm{d}u \\
&= k\mu_2(\xi_k - W^{FIL}(v,k)) + \exp\left\{\lambda \int_0^v \bar{G}(z)\,\mathrm{d}z\right\} \\
&\quad \times \int_v^\infty \exp\left\{-\lambda \int_0^u \bar{G}(z)\,\mathrm{d}z\right\} ((s-k)\mu_1 + h_G(u)) w^{FIL}(u,k)\,\mathrm{d}u.
\end{aligned}$$

Again, we take the derivative with respect to v and use the level crossing equation of Lemma 7.2 to rewrite the remaining integrals, providing (after some tedious rewriting)

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}v} w^{FIL}(v,k) + (r_k + h_G(v) - \lambda\bar{G}(v)) w^{FIL}(v,k) - \lambda\bar{G}(v) k\mu_2 W^{FIL}(v,k) \\
&\qquad = \quad -\lambda\bar{G}(v) k\mu_2 \xi_k + \lambda\bar{G}(v)\mu_2(k+1)\left(\xi_{k+1} - W^{FIL}(v,k+1)\right). \qquad (7.10)
\end{aligned}$$

Note that the above corresponds to a linear second-order differential equation again, but now involving an inhomogeneous term due to page $k+1$. As at page $s-N$, the coefficients are constant for the appealing special case in which customers have infinite patience, see Subsection 7.4.3. For the case of impatient customers, we follow the procedure at page $s-N$. Considering the tail of the FIL distribution $\bar{W}^{FIL}(v,k)$ again, Equation (7.10) may be written as the following linear second-order differential equation:

$$\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}v^2} \bar{W}^{FIL}(v,k) + a_k(v) \frac{\mathrm{d}}{\mathrm{d}v} \bar{W}^{FIL}(v,k) + b_k(v) \bar{W}^{FIL}(v,k) \\
&\qquad = -\lambda\bar{G}(v)\mu_2(k+1)\bar{W}^{FIL}(v,k+1), \qquad (7.11)
\end{aligned}$$

We now show by induction that the tail of the FIL distribution can be iteratively determined by

$$\bar{W}^{FIL}(v,k) = C_k \tilde{w}_k(v) + \sum_{j=k+1}^{s-N} C_j \tilde{w}_{k,j}^{\mathrm{part}}(v), \qquad (7.12)$$

with $\tilde{w}_k(\cdot)$ and $\tilde{w}_{k,j}^{\mathrm{part}}(\cdot)$ as defined in (7.7) and (7.8), respectively. For $k = s-N$ the result follows directly from the analysis at page $s-N$ above. Assuming that (7.12) is valid for $k+1$, we show that $\bar{W}^{FIL}(v,k)$ also satisfies (7.12).

Observe that $\bar{W}^{FIL}(v,k)$ satisfies an inhomogeneous second-order DE, cf. (7.11). Using (7.7), the general solution of the complementary homogeneous DE is $C_k\tilde{w}_k(x)$. For the particular solution we rely on the iterative scheme of calculating the tail of

the FIL distribution. Due to the induction hypothesis, we need particular solutions of

$$w''(v) + a_k(v)w'(v) + b_k(v)w(v) = \\ -\lambda\bar{G}(v)\mu_2(k+1)\left(C_{k+1}\tilde{w}_{k+1}(v) + \sum_{j=k+2}^{s-N} C_j\tilde{w}^{\text{part}}_{k+1,j}(v)\right).$$

Now, using that $\tilde{w}_{k+1}(v) = \tilde{w}^{\text{part}}_{k+1,k+1}(v)$, $\tilde{w}^{\text{part}}_{k,k+1}(\cdot)$ is a particular solution of

$$w''(v) + a_k(v)w'(v) + b_k(v)w(v) = -\lambda\bar{G}(v)\mu_2(k+1)\tilde{w}^{\text{part}}_{k+1,k+1}(v).$$

Also, for $j = k+2, \ldots, s-N$, $\tilde{w}^{\text{part}}_{k,j}(\cdot)$ is a particular solution of

$$w''(v) + a_k(v)w'(v) + b_k(v)w(v) = -\lambda\bar{G}(v)\mu_2(k+1)\tilde{w}^{\text{part}}_{k+1,j}(v).$$

Using standard arguments about DE's, it follows that a particular solution of (7.11) is given by $\sum_{j=k+1}^{s-N} C_j\tilde{w}^{\text{part}}_{k,j}(v)$. Combining the above, the induction step is shown.

Page 0**:** We use the same approach as for page k, for $k \in \{1, \ldots, s-N-1\}$, yielding Equation (7.10). As $k = 0$ in this case, we now obtain a first-order inhomogeneous DE for the FIL-density (instead of the FIL-distribution),

$$\frac{\mathrm{d}}{\mathrm{d}v}w^{FIL}(v,k) + (s\mu_1 + h_G(v) - \lambda\bar{G}(v))w^{FIL}(v,k) = \\ \lambda\bar{G}(v)\mu_2\left(\xi_1 - W^{FIL}(v,1)\right). \tag{7.13}$$

The solution to the corresponding homogeneous equation is equivalent to the situation of Section 7.3. Specifically, the solution to the homogeneous equation is, cf. (7.4),

$$C_0 \exp\left\{\int_0^v \lambda\bar{G}(u) - s\mu_1 - h_G(u)\,\mathrm{d}u\right\},$$

corresponding to the FIL waiting time in a corresponding M/M/s+G system. Using standard analysis, the general solution to (7.13) is

$$w^{FIL}(v,0) = C_0 \exp\left\{\int_0^v \lambda\bar{G}(u) - s\mu_1 - h_G(u)\,\mathrm{d}u\right\} \\ + \int_{t=0}^v \exp\left\{\int_t^v \lambda\bar{G}(u) - s\mu_1 - h_G(u)\,\mathrm{d}u\right\}\lambda\bar{G}(t)\mu_2\bar{W}^{FIL}(t,1)\,\mathrm{d}t.$$

Using that $-\int_t^v h_G(u)\,\mathrm{d}u = \log(\bar{G}(v)/\bar{G}(t))$ and some rewriting completes the proof. □

7.4.3 Infinite patience

The steady-state distribution of the FIL process can be found explicitly in terms of the mixture of exponential terms in case customers have infinite patience, as presented in Proposition 7.3. Define ϕ_k as the negative root of the following quadratic equation

$$\phi^2 + (r_k - \lambda)\,\phi - \lambda k \mu_2 = 0, \tag{7.14}$$

that is,

$$\phi_k = \frac{\lambda - r_k - \sqrt{(r_k - \lambda)^2 + 4\lambda k \mu_2}}{2}. \tag{7.15}$$

For convenience, we assume that $\lambda \neq s(\mu_1 - \mu_2)$, such that all ϕ_k are different, for different k's, see also Remark 7.2.

Proposition 7.3. *The distribution of the FIL process at page k is*

$$W^{FIL}(v,k) = \begin{cases} \xi_k + \sum_{j=k}^{s-N} C_{j,k} e^{\phi_j v}, & \text{for } k \in \{1,\ldots,s-N\}; \\ \xi_0 - \frac{\tilde{C}_{0,0}}{s\mu_1(1-\rho)} e^{-s\mu_1(1-\rho)v} + \sum_{j=1}^{s-N} C_{j,0} e^{\phi_j v}, & \text{for } k = 0, \end{cases}$$

where, for $j \in \{k+1,\ldots,s-N\}$,

$$C_{j,k} = (-1)^{j-k} C_{j,j} \prod_{l=k}^{j-1} \frac{\lambda(l+1)\mu_2}{\phi_j^2 + (r_l - \lambda)\phi_j - \lambda l \mu_2} \tag{7.16}$$

and $\tilde{C}_{0,0} = C_{0,0}\phi_0$.

We note that $\phi_0 = \lambda - s\mu_1 = -s\mu_1(1-\rho)$. For convenience, we may write $C_{0,0} = \tilde{C}_{0,0}/\phi_0$.

Proof. Again, for the derivation we start at page $s-N$ and then iteratively determine the FIL distribution at page k given its distribution at page $k+1$.

Page $s-N$: Using the infinite patience assumption, it holds that $a_k(v) = r_k - \lambda$ and $b_k = -\lambda k \mu_2$. Thus (7.7) is a second-order differential equation with constant coefficients. Using standard calculus, the general solution for the FIL distribution is

$$\xi_{s-N} + C^{(1)}_{s-N} e^{\phi^{(1)}_{s-N} v} + C^{(2)}_{s-N} e^{\phi^{(2)}_{s-N} v},$$

where $\phi^{(1)}_{s-N}$ and $\phi^{(2)}_{s-N}$ are the negative and positive root of Equation (7.14) with $k = s-N$. Since $e^{\phi^{(2)}_{s-N} v} \to \infty$, as $v \to \infty$ (due to $\phi^{(2)}_{s-N}$ being positive), and $W^{FIL}(v, s-N)$ is a distribution function, it should hold that $C^{(2)}_{s-N} = 0$. This completes the FIL distribution at page $s-N$.

Page $k \in \{1, \ldots, s-N-1\}$: For infinite patience, the second-order differential equation of (7.10) can be written as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}v}w^{FIL}(v,k) + (r_k - \lambda)w^{FIL}(v,k) - \lambda k\mu_2 W^{FIL}(v,k) \\ = \quad -\lambda k\mu_2\xi_k + \lambda(k+1)\mu_2\left(\xi_{k+1} - W^{FIL}(v,k+1)\right). \end{aligned} \tag{7.17}$$

We now show by induction that the FIL distribution is a mixture of exponentials:

$$W^{FIL}(v,k) = \xi_k + \sum_{j=k}^{s-N} C_{j,k}e^{\phi_j v},$$

with $C_{j,k}$ given by (7.16). Note that the result holds for $k = s-N$ from the analysis of page $s-N$ above. Assuming the above FIL-distribution for page $k+1$, we show that this provides the FIL-distribution for page k.

Note that (7.17) is again an inhomogeneous second-order DE with constant coefficients, where the homogeneous part is similar to that of page $s-N$. Hence, the solution of the complementary homogeneous DE is

$$C_k^{(1)}e^{\phi_k^{(1)}v} + C_k^{(2)}e^{\phi_k^{(2)}v},$$

with $\phi_k^{(1)}$ and $\phi_k^{(2)}$ the negative and positive root of Equation (7.14). As for page $s-N$, it should hold that $C_k^{(2)} = 0$, as $W^{FIL}(v,k)$ is a distribution function, leaving $C_{k,k}\exp(\phi_k v)$ as solution to the homogeneous DE.

For the particular solution, we rely on the iterative scheme of calculating the FIL distribution, i.e., the induction assumption. Trying particular solutions of the type $W_p^{FIL}(v,k) = \xi_k + \sum_{j=k+1}^{s-N} A_j e^{\phi_j v}$, gives expressions for A_j (representing $C_{j,k}$) in terms of $C_{j,j}$:

$$\begin{aligned}A_j &= -C_{j,k+1}\frac{\lambda(k+1)\mu_2}{\phi_j^2 + (r_k - \lambda)\phi_j - \lambda k\mu_2} \\ &= (-1)\times(-1)^{j-k-1}C_{j,j}\prod_{l=k+1}^{j-1}\frac{\lambda(l+1)\mu_2}{\phi_j^2 + (r_l-\lambda)\phi_j - \lambda l\mu_2}\times\frac{\lambda(k+1)\mu_2}{\phi_j^2 + (r_k-\lambda)\phi_j - \lambda k\mu_2},\end{aligned}$$

where the second equality follows from the induction hypothesis. The general solution of (7.17) is then

$$W^{FIL}(v,k) = \xi_k + C_{k,k}e^{\phi_k v} + \sum_{j=k+1}^{s-N} C_{j,k}e^{\phi_j v},$$

where, for $j \in \{k+1, \ldots, s-N\}$, $C_{j,k}$ is given by Equation (7.16). This shows the results for $W^{FIL}(v,k)$.

Page 0: For infinite patience, Equation (7.13) reads

$$\frac{\mathrm{d}}{\mathrm{d}v}w^{FIL}(v,0) + (s\mu_1 - \lambda)w^{FIL}(v,k) = \lambda\mu_2\left(\xi_1 - W^{FIL}(v,1)\right).$$

The solution to the corresponding homogeneous equation is directly seen to be

$$\tilde{C}_{0,0}e^{(\lambda - s\mu_1)v},$$

corresponding to the waiting time density in a corresponding Erlang delay model. For the particular solution, we use the FIL-distribution at page 1. Hence, we look for a solution of the type $w_p^{FIL}(v,0) = \sum_{j=1}^{s-N} A_j e^{\phi_j v}$, showing that

$$A_j = -C_{j,1}\frac{\lambda\mu_2}{\phi_j + s\mu_1 - \lambda}.$$

The general solution of the FIL-density then is

$$w^{FIL}(v,0) = \tilde{C}_{0,0}e^{-s\mu_1(1-\rho)v} + \sum_{j=1}^{s-N} A_j e^{\phi_j v}.$$

We now obtain the distribution function from the density:

$$\begin{aligned} W^{FIL}(v,0) &= \xi_0 - \int_v^\infty w^{FIL}(u,0)\mathrm{d}u \\ &= \xi_0 - \frac{\tilde{C}_{0,0}}{s\mu_1(1-\rho)}e^{-s\mu_1(1-\rho)v} + \sum_{j=1}^{s-N}\frac{A_j}{\phi_j}e^{\phi_j v}. \end{aligned}$$

Finally, note that the constant $C_{j,0} := A_j/\phi_j$, for $j \in \{1,\ldots,s-N\}$, also follows from the recursion (7.16). □

Remark 7.2. We note that for $\lambda = s(\mu_1 - \mu_2)$, it holds that all ϕ_k are identical to ϕ_0. In that case, the FIL distribution $W^{FIL}(v,k)$ is of the form $\sum_{j=k}^{s-N}\hat{C}_{j,k}v^{s-N-j}e^{\phi_0 v}$ instead of a mixture of exponential terms.

7.4.4 Constants and boundary conditions

To complete the FIL distribution, we consider the 'boundary' case in which there are no type 1 customers waiting. In particular, we present the boundary equations in the lemma below, which are essentially flow balance equations. These equations are used to construct a set of linear equations to find the full FIL distribution.

Lemma 7.3. *We consider the boundary equations for state $(-l, k)$, with l the number of idle servers.*

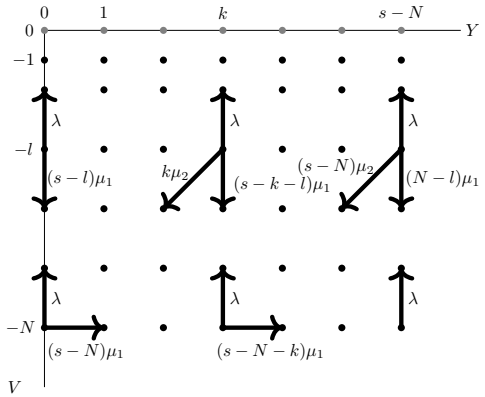


Figure 7.2: System dynamics in case of idle servers.

(*i*) *For* $l = 0$ *and* $k \in \{0, 1, \ldots, s-N\}$

$$\begin{aligned}(\lambda + k\mu_2 + (s-k)\mu_1)\, w^{FIL}(0,k) \\ = \lambda w^{FIL}(-1,k) + \int_0^\infty ((s-k)\mu_1 + h_G(u)) w^{FIL}(u,k) F_u(0)\mathrm{d}u \\ + \mathbb{1}_{\{k<s-N\}}(k+1)\mu_2 \int_0^\infty w^{FIL}(u,k+1) F_u(0)\mathrm{d}u.\end{aligned} \tag{7.18}$$

(*ii*) *For* $l \in \{1, \ldots, N-1\}$ *and* $k \in \{0, 1, \ldots, s-N\}$

$$\begin{aligned}(\lambda + k\mu_2 + (s-k-l)\mu_1)\, w^{FIL}(-l,k) \\ = \lambda w^{FIL}(-l-1,k) + (s-k-l+1)\mu_1 w^{FIL}(-l+1,k) \\ + \mathbb{1}_{\{k<s-N\}}(k+1)\mu_2 w^{FIL}(-l+1,k+1).\end{aligned} \tag{7.19}$$

(*iii*) *For* $l = N$ *and* $k \in \{0, 1, \ldots, s-N\}$

$$\begin{aligned}(\lambda + (s-N-k)\mu_1)\, w^{FIL}(-N,k) \\ = (s-N-k+1)\mu_1 \left[\mathbb{1}_{\{k>0\}} w^{FIL}(-N,k-1) + w^{FIL}(-N+1,k)\right] \\ + \mathbb{1}_{\{k<s-N\}}(k+1)\mu_2 w^{FIL}(-N+1,k+1).\end{aligned} \tag{7.20}$$

Proof. The flow balance equations for the states $(-l,k) \in \{1,\ldots,N\} \times \{0,\ldots,s-N\}$ follow directly from Figure 7.2. Now consider states $(0,k)$ with $k \in \{0,\ldots,s-N-1\}$.

Using infinitesimal arguments, we obtain

$$\begin{aligned}\mathbb{P}(V_{t+h}=0,Y_{t+h}=k) &= (1-\lambda h - k\mu_2 h - (s-k)\mu_1 h)\mathbb{P}(V_t=0,Y_t=k)\\ &\quad + \lambda h \mathbb{P}(V_t=-1,Y_t=k)\\ &\quad + \int_0^\infty h((s-k)\mu_1 + h_G(u))\mathbb{P}(\tilde{A}_\lambda > u)\,\mathrm{d}\mathbb{P}(V_t\le u, Y_t=k)\\ &\quad + (k+1)\mu_2 h\mathbb{P}(V_t\le \tilde{A}_\lambda, Y_t=k+1)+o(h).\end{aligned}$$

Note that $\mathbb{P}(\tilde{A}_\lambda > u) = F_u(0)$. Subtracting $\mathbb{P}(V_t = 0, Y_t = k)$, dividing by h and taking the limits $h \to 0$ and $t \to \infty$, yields the result. For state $(0, s-N)$, the final term vanishes. □

Note that the terms $w^{FIL}(u,k)$ appear in the integrals; this can be obtained as the derivative of the solutions to the differential equations. Now, it remains to determine the $(s-N+1)(N+3)$ unknowns: C_k, ξ_k and $w^{FIL}(-l,k)$, with $k=0,1,\ldots,s-N$ and $l=0,1,\ldots,N$. The required number of $(s-N+1)(N+3)$ equations to find the unknowns are then as follows:

1. $(s-N+1)(N+1)$ boundary equations in Lemma 7.3;
2. Letting $v \downarrow 0$ in Proposition 7.2 and using $\tilde{w}_k(0) = 1$, yields the following $(s-N+1)$ equations:

$$\xi_k - \sum_{l=0}^{N} w^{FIL}(-l,k) = C_k + \sum_{j=k+1}^{s-N} C_j \tilde{w}^{\text{part}}_{k,j}(0), \quad k=1,2,\ldots,s-N, \tag{7.21}$$

   and, with $w^{FIL}(u,0)$ a linear function of C_j, $j=0,\ldots,s-N$,

$$\xi_0 - \sum_{l=0}^{N} w^{FIL}(-l,0) = \int_0^\infty w^{FIL}(u,0)\,\mathrm{d}u;$$

3. $(s-N)$ set balance equations for page $k \in \{0,1,\ldots,s-N-1\}$:

$$\begin{aligned} &w^{FIL}(-N,k)(s-N-k)\mu_1 + \mathbb{1}_{\{k>0\}}\left(\xi_k - w^{FIL}(-N,k)\right)k\mu_2 =\\ &\quad \left(\xi_{k+1} - w^{FIL}(-N,k+1)\right)(k+1)\mu_2\\ &\qquad + \mathbb{1}_{\{k>0\}} w^{FIL}(-N,k-1)(s-N-k+1)\mu_1;\end{aligned} \tag{7.22}$$

4. Normalization equation

$$\sum_{k=0}^{s-N} \xi_k = 1. \tag{7.23}$$

Remark 7.3 (Constants and equations for infinite patience)**.** When the patience of the jobs is infinite, we have unknown constants $C_{k,k}$, instead of C_k, $k=0,1,\ldots,s-N$. Equation (7.18) can be more explicit, by writing out the integrals using the form of $W^{FIL}(v,k)$ given in Proposition 7.3.

7.4.5 Performance measures

The waiting-time distribution follows from the FIL process by considering specific epochs, i.e., moments of service completions. Our main result is presented in Theorem 7.2 and gives the tail of the steady-state waiting-time distribution, for customers that do not abandon. The constants follow from a system of linear equations representing the boundary conditions.

Theorem 7.2 (Waiting time type 1)**.** *The tail of the steady-state waiting-time distribution is*

$$\mathbb{P}(W > \alpha | service) = \frac{\sum_{k=0}^{s-N} r_k \bar{W}^{FIL}(\alpha, k)}{\lambda \mathbb{P}(W = 0) + \sum_{k=0}^{s-N} r_k \bar{W}^{FIL}(0, k)},$$

where $\bar{W}^{FIL}(\alpha, k)$ follows from Proposition 7.2, and an atom in 0,

$$\mathbb{P}(W = 0) = \sum_{k=0}^{s-N} \sum_{l=1}^{N} w^{FIL}(-l, k). \tag{7.24}$$

Here, $\bar{W}^{FIL}(0, k)$ and $w^{FIL}(-l, k)$ follow from Proposition 7.2 and the system of linear equations (7.18)–(7.23).

Proof. We use the same method as in the proof of Theorem 7.1. Specifically, the waiting time is zero when there are idle servers upon arrival. The PASTA property yields (7.24). The arrival rate of customers that are taken into service is given by

$$\lambda^* = \lambda \mathbb{P}(W = 0) + \sum_{k=0}^{s-N} r_k \bar{W}^{FIL}(0, k). \tag{7.25}$$

For positive waiting times, we consider the moments at which jobs are taken into service:

$$\begin{aligned} \mathbb{P}(V_t > \alpha, N(t, t+h) = 1) &= \sum_{k=0}^{s-N} \mathbb{P}(V_t > \alpha, Y_t = k, N(t, t+h) = 1) \\ &= \sum_{k=0}^{s-N} r_k h \bar{W}^{FIL}(\alpha, k) + o(h). \end{aligned}$$

Now,

$$\begin{aligned} \mathbb{P}(W > \alpha | \text{service}) &= \lim_{h \to 0} \mathbb{P}(V_t > \alpha | N(t, t+h) = 1) = \lim_{h \to 0} \frac{\mathbb{P}(V_t > \alpha, N(t, t+h) = 1)}{\mathbb{P}(N(t, t+h) = 1)} \\ &= \frac{1}{\lambda^*} \sum_{k=0}^{s-N} r_k \bar{W}^{FIL}(\alpha, k). \end{aligned}$$

Filling in λ^* (into Equation (7.25)) completes the proof. □

We also directly have another relevant performance measure, the probability that a customer receives service.

Corollary 7.3 (Probability of service)**.** *The probability that a customer receives service is given by*

$$\mathbb{P}(\textit{service}) = \lambda^*/\lambda,$$

with λ^* given in (7.25).

The throughput of type 2 follows directly from the analysis of the FIL process.

Corollary 7.4 (Throughput type 2)**.** *The throughput for type-2 traffic is*

$$TH = \sum_{k=0}^{s-N} \xi_k k \mu_2.$$

This formula for the throughput also holds if the patience of type-1 jobs is infinite. The following corollary gives the waiting-time distribution for type-1 jobs in the case of infinite patience, leading to more explicit expressions.

Corollary 7.5 (Waiting time for infinite patience)**.** *The tail of the steady-state waiting-time distribution of type 1 jobs is*

$$\mathbb{P}(W > \alpha) = e^{-s\mu_1(1-\rho)\alpha} \times \frac{-C_{0,0}}{\rho} + \sum_{j=1}^{s-N} e^{\phi_j \alpha} \times \frac{-\sum_{k=0}^{j} r_k C_{j,k}}{\lambda},$$

with $C_{j,k}$ and ϕ_k given in (7.16) and (7.15), respectively, and an atom in 0 given in Equation (7.24)*. See also Remark 7.3.*

Remark 7.4 (Numerical results)**.** We did some numerical experiments for the model with infinite patience. Figure 7.3 shows the tail probability of the waiting time and the throughput of type-2 jobs for a specific set of parameters. On the left axis, the probability of waiting less than 1/3, the required service level (SL) can be found, where the dotted line gives the required SL. The throughput of type-2 jobs is plotted on the right axis. We see that the SL increases with N and the throughput of type-2 jobs decreases. As N increases, the ascent in SL decreases, i.e., the SL increases in a concave manner. The decrease in throughput is for intermediate values of N roughly linear. If $\mathbb{P}(W < \alpha)$ is above the dotted line, the 80-20 service level is met, 80% of the jobs is served within 20 seconds (1/3 minutes). For $N = 3$, the SL requirement is met and the throughput of type-2 jobs is equal to 1.42.

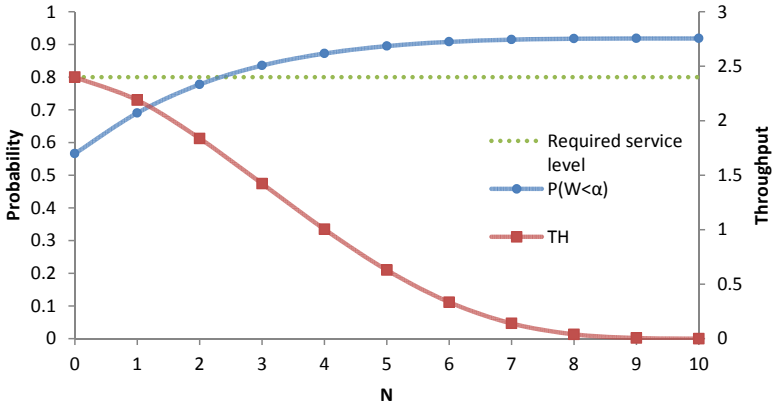


Figure 7.3: Blended system with $\lambda = 7$, $\mu_1 = 1$, $\mu_2 = 0.8$, $s = 10$, and $\alpha = 1/3$.

Bibliography

[1] S. Aalto, U. Ayesta, and R. Righter. Properties of the gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25(3):269–288, 2011.

[2] H. Abouee-Mehrizi and O. Baron. State-dependent M/G/1 queueing systems. *Queueing Systems*, 82(1-2):121–148, 2016.

[3] O. Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production & Operations Management*, 16(6):665–688, 2007.

[4] E. Altman and D. Fiems. Expected waiting time in symmetric polling systems with correlated walking times. *Queueing Systems*, 56(3):241–253, 2007.

[5] E. Altman and U. Yechiali. Analysis of customers impatience in queues with server vacations. *Queueing Systems*, 52(4):261–279, 2006.

[6] E. Altman, T. Jimenez, and D. Kofman. DPS queues with stationary ergodic service times and the performance of TCP in overload. In *Proceedings IEEE INFOCOM 2004*, volume 2, pages 975–983. IEEE, 2004.

[7] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53(1-2):53–63, 2006.

[8] 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.

[9] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. *Operations Research*, 52(4):527–545, 2004.

[10] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Analysis of a multiserver queue with setup times. *Queueing Systems*, 51(1):53–76, 2005.

[11] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings IEEE 24th Annual Joint Conference*

of the IEEE Computer and Communications Societies., volume 2, pages 784–795. IEEE, 2005.

[12] U. Ayesta, O. J. Boxma, and I. M. Verloop. Sojourn times in a processor sharing queue with multiple vacations. *Queueing Systems*, 71(1-2):53–78, 2012.

[13] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, volume 26. Springer Science & Business Media, 2013.

[14] P. J. M. Bakker. *It can be done: Better care for less money (in Dutch: Het kan echt: betere zorg voor minder geld).* TPG, Amsterdam, 2004.

[15] N. Bansal and D. Gamarnik. Handling load with less stress. *Queueing Systems*, 54(1):45–54, 2006.

[16] R. Bekker, S. C. Borst, O. J. Boxma, and O. Kella. Queues with workload-dependent arrival and service rates. *Queueing Systems*, 46(3-4):537–556, 2004.

[17] R. Bekker, O. J. Boxma, and J. A. C. Resing. Lévy processes with adaptable exponent. *Advances in Applied Probability*, 41(1):117–205, 2009.

[18] R. Bekker, G. M. Koole, B. F. Nielsen, and T. B. Nielsen. Queues with waiting time dependent service. *Queueing Systems*, 68(1):61–78, 2011.

[19] R. Bekker, P. Vis, J. L. Dorsman, R. D. van der Mei, and E. M. M. Winands. The impact of scheduling policies on the waiting-time distributions in polling systems. *Queueing Systems*, 79(2):145–172, 2015.

[20] R. Bekker, G. M. Koole, P. Vis, and B. Zaber. Waiting-time distributions in blended multi-server queues with impatient customers. *Submitted*, 2017.

[21] A. Ben Tahar and A. Jean-Marie. The fluid limit of the multiclass processor sharing queue. *Queueing Systems*, 71(4):347–404, 2012.

[22] J. C. Bennett and D. J. Worthington. An example of a good but partially successful OR engagement: Improving outpatient clinic operations. *Interfaces*, 28(5):56–69, 1998.

[23] H. G. Bernett, M. J. Fischer, and D. M. B. Masi. Blended call center performance analysis. *IT Professional*, 4(2):33–38, 2002.

[24] C. Bhargava and M. Jain. Unreliable multiserver queueing system with modified vacation policy. *Opsearch*, 51(2):159–182, 2014.

[25] S. Bhulai and G. M. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48(8):1434–1438, 2003.

[26] N. H. Bingham and R. A. Doney. Asymptotic properties of supercritical branching processes I: The Galton-Watson process. *Advances in Applied Probability*, 6(4):711–731, 1974.

[27] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27. Cambridge university press, 1989.

[28] M. A. A. Boon. *Polling Models, From Theory to Traffic Intersections*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2011.

[29] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67(6):468–484, 2010.

[30] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A two-queue polling model with two priority levels in the first queue. *Discrete Event Dynamic Systems*, 20 (4):511–536, 2010.

[31] M. A. A. Boon, E. M. M. Winands, I. J. B. F. Adan, and A. C. C. van Wijk. Closed-form waiting time approximations for polling systems. *Performance Evaluation*, 68:290–306, 2010.

[32] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science*, 16 (2):67–82, 2011.

[33] S. C. Borst and R. D. van der Mei. Waiting-time approximations for multiple-server polling systems. *Performance Evaluation*, 31(3-4):163–182, 1998.

[34] S. C. Borst, O. J. Boxma, J. A. Morrison, and R. Núñez Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31(4):254–262, 2003.

[35] S. C. Borst, R. Núñez-Queija, and A. P. Zwart. Sojourn time asymptotics in processor-sharing queues. *Queueing Systems*, 53(1):31–51, 2006.

[36] O. J. Boxma and V. Dumas. The busy period in the fluid queue. In *ACM SIGMETRICS Performance Evaluation Review*, volume 26, pages 100–110. ACM, 1998.

[37] O. J. Boxma and V. I. Lotov. On a class of one-dimensional random walks. *Markov Processes and Related Fields*, 2(2):349–362, 1996.

[38] O. J. Boxma, J. A. Weststrate, and U. Yechiali. A globally gated polling system with server interruptions, and applications to the repairman problem. *Probability in the Engineering and Informational Sciences*, 7(2):187–208, 1993.

[39] O. J. Boxma, M. R. H. Mandjes, and O. Kella. On a queuing model with service interruptions. *Probability in the Engineering and Informational Sciences*, 22 (04):537–555, 2008.

[40] O. J. Boxma, J. Bruin, and B. Fralix. Sojourn times in polling systems with various service disciplines. *Performance Evaluation*, 66(11):621–639, 2009.

[41] O. J. Boxma, O. Kella, and M. R. H. Mandjes. On a generic class of Lévy-driven vacation models. *Probability in the Engineering and Informational Sciences*, 24 (01):1–12, 2010.

[42] L. Brandenburg, P. Gabow, G. Steele, J. Toussaint, and B. J. Tyson. Innovation and best practices in health care scheduling. *Technical report*, 2015.

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

[44] P. H. Brill and M. J. M. Posner. The system point method in exponential queues: A level crossing approach. *Mathematics of Operations Research*, 6(1): 31–49, 1981.

[45] P. H. Brill and M. J. M. Posner. A two server queue with nonwaiting customers receiving specialized service. *Management Science*, 27(8):914–925, 1981.

[46] T. Bu and D. Towsley. Fixed point approximations for TCP behavior in an AQM network. In *ACM SIGMETRICS Performance Evaluation Review*, volume 29, pages 216–225. ACM, 2001.

[47] W. Bux and H. L. Truong. Mean-delay approximation for cyclic-service queueing systems. *Performance Evaluation*, 3(3):187–196, 1983.

[48] S. K. Cheung, J. L. van den Berg, R. J. Boucherie, R. Litjens, and F. Roijers. An analytical packet/flow-level modelling approach for wireless LANs with quality-of-service support. In *Proceedings 19th International Teletraffic Congress*, pages 1651–1662. Beijing University Post and Telecommunications Press, 2005.

[49] A. Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[50] E. G. Coffman, R. R. Muntz, and H. Trotter. Waiting-time distributions for processor-sharing systems. *Journal of the ACM*, 17:123–130, 1970.

[51] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. Polling systems with zero switchover times: A heavy-traffic averaging principle. *The Annals of Applied Probability*, 5(3):681–719, 1995.

[52] E. G. Coffman, A. A. Puhalskii, and M. I. Reiman. Polling systems in heavy traffic: A bessel process limit. *Mathematics of Operations Research*, 23(2): 257–304, 1998.

[53] J. W. Cohen. *The single server queue*. North-Holland, 1982.

[54] S. Creemers and M. Lambrecht. An advanced queueing model to analyze appointment-driven service systems. *Computers & Operations Research*, 36 (10):2773–2785, 2009.

[55] S. Creemers and M. Lambrecht. Queueing models for appointment-driven systems. *Annals of Operations Research*, 178(1):155–172, 2010.

[56] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46:353–388, 1984.

[57] F. de Véricourt and O. B. Jennings. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[58] Q. Deng. A two-queue polling model with regularly varying service and/or switchover times. *Stochastic Models*, 19(4):507–526, 2003.

[59] B. T. Denton. *Handbook of Healthcare Operations Management*. Springer, 2013.

[60] A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. N. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

[61] J. R. Dorp and S. Kotz. Generalized trapezoidal distributions. *Metrika*, 58(1): 85–97, 2003.

[62] J. L. Dorsman, R. D. van der Mei, and E. M. M. Winands. A new method for deriving waiting-time approximations in polling systems with renewal arrivals. *Stochastic Models*, 27(2):318–332, 2011.

[63] R. R. Egorova. *Sojourn Time Tails in Processor-Sharing Systems*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2009.

[64] R. R. Egorova, S. C. Borst, and A. P. Zwart. Bandwidth-sharing networks in overload. *Performance Evaluation*, 64(9):978–993, 2007.

[65] S. G. Elkhuizen, S. F. Das, P. J. M. Bakker, and J. A. M. Hontelez. Using computer simulation to reduce access time for outpatient departments. *Quality and Safety in Health Care*, 16(5):382–386, 2007.

[66] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27(3):519–532, 1980.

[67] W. Feller. *An Introduction to Probability Theory and its Applications, Volume II*. Wiley, 1971.

[68] I. Frigui and A. S. Alfa. Analysis of a time-limited polling system. *Computer Communications*, 21(6):558–571, 1998.

[69] S. W. Fuhrmann. Performance analysis of a class of cyclic schedules. Bell laboratories technical memorandom, 1981.

[70] A. Gandhi, M. Harchol-Balter, and I. J. B. F. Adan. Server farms with setup costs. *Performance Evaluation*, 67(11):1123–1138, 2010.

[71] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems*, 77(2):177–209, 2014.

[72] N. Gans and Y. P. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271, 2003.

[73] N. Gans, G. M. Koole, and A. Mandelbaum. Telephone call centers: A tutorial and literature review. *Manufacturing & Service Operations Management*, 5(2): 79–141, 2002.

[74] L. V. Green and S. I. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[75] S. A. Grishechkin. On a relationship between processor-sharing queues and Crump–Mode–Jagers branching processes. *Advances in Applied Probability*, 24 (3):653–698, 1992.

[76] L. Guo and I. Matta. Scheduling flows with unknown sizes: Approximate analysis. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30, pages 276–277. ACM, 2002.

[77] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819, 2008.

[78] I. Gurvich, M. Armony, and A. Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[79] R. W. Hall (ed.). *Patient Flow: Reducing Delay in Healthcare Delivery.* Springer Science & Business Media, 2013.

[80] M. Harchol-Balter. *Queueing Disciplines.* John Wiley & Sons, Inc., 2009.

[81] R. Hariharan, W. K. Ehrlich, P. K. Reeser, and R. D. van der Mei. Performance of web servers in a distributed computing environment. *Teletraffic Engineering in the Internet Era*, 4:137–148, 2001.

[82] M. Haviv and J. van der Wal. Mean sojourn times for phase-type discriminatory processor sharing systems. *European Journal of Operational Research*, 189(2): 375–386, 2008.

[83] Y. Hayel and B. Tuffin. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *International Conference on Research in Networking*, pages 816–827. Springer, 2005.

[84] J. Hofmann. The BMAP/G/1 queue with level-dependent arrivals – an overview. *Telecommunication Systems*, 16(3-4):347–359, 2001.

[85] P. J. H. Hulshof, N. Kortbeek, R. J. Boucherie, E. W. Hans, and P. J. M. Bakker. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.

[86] N. Izady. Appointment capacity planning in specialty clinics: A queueing approach. *Operations Research*, 63(4):916–930, 2015.

[87] A. Izagirre, U. Ayesta, and I. M. Verloop. Sojourn time approximations for a discriminatory processor sharing queue. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 1(1):5, 2016.

[88] A. Jean-Marie and P. Robert. On the transient behavior of the processor sharing queue. *Queueing Systems*, 17(1):129–136, 1994.

[89] H. Jiang, Z. Pang, and S. I. Savin. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management*, 14(4):654–669, 2012.

[90] O. Jouini, G. M. Koole, and A. Roubos. Performance indicators for call centers with impatient customers. *IIE Transactions*, 45(3):341–354, 2013.

[91] N. Kawasaki, H. Takagi, Y. Takahashi, S.-J. Hong, and T. Hasegawa. Waiting time analysis of $M^X/G/1$ queues with/without vacations under random order of service discipline. *Journal of the Operations Research Society of Japan*, 43 (4):455–468, 2000.

[92] J. C. Ke, C. H. Wu, and Z. G. Zhang. Recent developments in vacation queueing models: A short survey. *International Journal of Operations Research*, 7(4):3–8, 2010.

[93] O. Kella and U. Yechiali. Priorities in M/G/1 queue with server vacations. *Naval Research Logistics*, 35:23–34, 1988.

[94] J. F. C. Kingman. The single server queue in heavy traffic. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 902–904. Cambridge University Press, 1961.

[95] J. F. C. Kingman. On queues in which customers are served in random order. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 58, pages 79–91. Cambridge University Press, 1962.

[96] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14(2):242–261, 1967.

[97] A. G. Konheim, H. Levy, and M. M. Srinivasan. Descendant set: An efficient approach for the analysis of polling systems. *IEEE Transactions on Communications*, 42(234):1245–1253, 1994.

[98] G. M. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1):41–59, 2002.

[99] N. Kortbeek, M. E. Zonderland, A. Braaksma, I. M. H. Vliegen, R. J. Boucherie, N. Litvak, and E. W. Hans. Designing cyclic appointment schedules for

outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation*, 80:5–26, 2014.

[100] I. Kotsireas. A survey on solution methods for integral equations. *The Ontario Research Centre for Computer Algebra*, 47, 2008.

[101] D. Kozlowski and D. J. Worthington. Use of queue modelling in the analysis of elective patient treatment governed by a maximum waiting time policy. *European Journal of Operational Research*, 244(1):331–338, 2015.

[102] L. R. LaGanga and S. R. Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276, 2007.

[103] B. Legros and G. M. Koole. A uniformization approach for the dynamic control of multi-server queueing systems with abandonments. *Working paper*, 2016.

[104] B. Legros, O. Jouini, and G. M. Koole. Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430, 2015.

[105] N. Liu and S. Ziya. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production & Operations Management*, 23(12): 2209–2223, 2014.

[106] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models*, 7(1):1–46, 1991.

[107] C. Mack. The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19:173–178, 1957.

[108] C. Mack, T. Murphy, and N. L. Webb. The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19:166–172, 1957.

[109] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5): 1189–1205, 2009.

[110] M. F. Murray. Improving access to specialty care. *The Joint Commission Journal on Quality and Patient Safety*, 33(3):125–135, 2007.

[111] M. F. Murray and D. M. Berwick. Advanced access: Reducing waiting and delays in primary care. *Jama*, 289(8):1035–1040, 2003.

[112] R. Núñez-Queija. *Procesor-Sharing Models for Integrated-Service Networks*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2000.

[113] T. L. Olsen and R. D. van der Mei. Polling systems with periodic server routeing in heavy traffic: Distribution of the delay. *Journal of Applied Probability*, 40 (2):305–326, 2003.

[114] T. L. Olsen and R. D. van der Mei. Polling systems with periodic server routing in heavy traffic: Renewal arrivals. *Operations Research Letters*, 33(1):17–25, 2005.

[115] M. Olvera-Cravioto, J. Blanchet, and P. Glynn. On the transition from heavy traffic to heavy tails for the M/G/1 queue: The regularly varying case. *The Annals of Applied Probability*, 21(2):645–668, 2011.

[116] G. Pang and O. Perry. A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91, 2014.

[117] J. Patrick and A. Aubin. Models and methods for improving patient access. In *Handbook of Healthcare Operations Management*, pages 403–420. Springer, 2013.

[118] J. Pichitlamken, A. Deslauriers, P. L'Ecuyer, and A. N. Avramidis. Modelling and simulation of a telephone call center. In *Proceedings of the 35th conference on Winter simulation: Driving innovation*, pages 1805–1812. Winter Simulation Conference, 2003.

[119] V. Ramaswami. A stable recursion for the steady state vector in Markov chains of M/G/1 type. *Stochastic Models*, 4(1):183–188, 1988.

[120] K. M. Rege and B. Sengupta. Queue-length distribution for the discriminatory processor-sharing queue. *Operations Research*, 44(4):653–657, 1996.

[121] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(4):409–426, 1993.

[122] M. Scholl and L. Kleinrock. On the M/G/1 queue with rest periods and certain service-independent queueing disciplines. *Operations Research*, 31(4):705–719, 1983.

[123] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14(4):670–684, 1966.

[124] Y. W. Shin and C. E. M. Pearce. The BMAP/G/1 vacation queue with queue-length dependent vacation schedule. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 40(2):207–221, 1998.

[125] H. Takagi. *Analysis of Polling Systems*. MIT press, 1986.

[126] H. Takagi. *Queueing analysis: A foundation of performance evaluation, vol. 1: vacation and priority systems*. North-Holland, 1991.

[127] H. Takagi and S. Kudoh. Symbolic higher-order moments of the waiting time in an M/G/1 queue with random order of service. *Stochastic Models*, 13(1): 167–179, 1997.

[128] N. Tian and Z. G. Zhang. A two threshold vacation policy in multiserver queueing systems. *European Journal of Operational Research*, 168(1):153–163, 2006.

[129] H. C. Tijms. *A First Course in Stochastic Models*. Wiley, 2003.

[130] C. van Aartsen. Wachttijden ziekenhuis stijgen verder. *Zorgvisie*, 2017. URL `https://www.zorgvisie.nl/Kwaliteit/Nieuws/2017/1/Wachttijden-ziekenhuis-stijgen-verder/`.

[131] R. D. van der Mei. Delay in polling systems with large switch-over times. *Journal of Applied Probability*, 36(1):232–243, 1999.

[132] R. D. van der Mei. Distribution of the delay in polling systems in heavy traffic. *Performance Evaluation*, 38(2):133–148, 1999.

[133] R. D. van der Mei. Polling systems in heavy traffic: Higher moments of the delay. *Queueing Systems*, 31(3):265–294, 1999.

[134] R. D. van der Mei. Polling systems with switch-over times under heavy load: Moments of the delay. *Queueing Systems*, 36(4):381–404, 2000.

[135] R. D. van der Mei. Waiting-time distributions in polling systems with simultaneous batch arrivals. *Annals of Operations Research*, 113(1-4):155–173, 2002.

[136] R. D. van der Mei. Towards a unifying theory on branching-type polling systems in heavy traffic. *Queueing Systems*, 57(1):29–46, 2007.

[137] R. D. van der Mei and S. C. Borst. Analysis of multiple-server polling systems by means of the power-series algorithm. *Stochastic Models*, 13(2):339–369, 1997.

[138] R. D. van der Mei, R. Hariharan, and P. K. Reeser. Web server performance modeling. *Telecommunication Systems*, 16(3-4):361–378, 2001.

[139] K. van Eeden, D. Moeke, and R. Bekker. Care on demand in nursing homes: A queueing theoretic approach. *Health Care Management Science*, 19(3):227–240, 2016.

[140] B. van Houdt. Analysis of the adaptive MMAP[K]/PH[K]/1 queue: A multitype queue with adaptive arrivals and general impatience. *European Journal of Operational Research*, 220(3):695–704, 2012.

[141] G. van Kessel, R. Núñez-Queija, and S. C. Borst. Asymptotic regimes and approximations for discriminatory processor sharing. *SIGMETRICS Performance Evaluation Review*, 32(2):44–46, 2004.

[142] M. J. G. van Uitert. *Generalized processor sharing queues*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2003.

[143] M. van Vuuren and E. M. M. Winands. Iterative approximation of k-limited polling systems. *Queueing Systems*, 55(3):161–178, 2007.

[144] A. C. C. van Wijk. *Pooling and Polling: Creation of Pooling in Inventory and Queueing Models*. Ph.D. thesis, Eindhoven University of Technology, The Netherlands, 2012.

[145] I. M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Operations Research*, 59(3):648–660, 2011.

[146] I. B. Vermeulen, S. M. Bohte, S. G. Elkhuizen, H. Lameris, P. J. M. Bakker, and H. La Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46(1):67–80, 2009.

[147] P. Vis and R. Bekker. Access times in appointment-driven systems and level-dependent MAP/G/1 queues. *Submitted*, 2017.

[148] P. Vis, R. Bekker, and R. D. van der Mei. Heavy-traffic limits for polling models with exhaustive service and non-FCFS service order policies. *Advances in Applied Probability*, 47(4):989–1014, 2015.

[149] P. Vis, R. Bekker, and R. D. van der Mei. Transient analysis of cycle lengths in cyclic polling systems. *Performance Evaluation*, 91:303–317, 2015.

[150] P. Vis, R. Bekker, R. D. van der Mei, and R. Núñez-Queija. Influence of batch arrivals on the queue-length distribution in DPS queues. *Submitted*, 2017.

[151] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67(2):173–220, 2006.

[152] P. P. Wang. Queueing models with delayed state-dependent service times. *European Journal of Operational Research*, 88(3):614–621, 1996.

[153] A. Wierman, E. M. M. Winands, and O. J. Boxma. Scheduling in polling systems. *Performance Evaluation*, 64(9):1009–1028, 2007.

[154] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[155] E. M. M. Winands, I. J. B. F. Adan, and G. J. van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54(1):35–44, 2006.

[156] D. J. Worthington. Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, 38(5):413–422, 1987.

[157] X. Xu and Z. G. Zhang. Analysis of multi-server queue with a single vacation (e, d)-policy. *Performance Evaluation*, 63(8):825–838, 2006.

[158] U. Yechiali. Analysis and control of polling systems. In *Performance Evaluation of Computer and Communication Systems*, pages 630–650. Springer, 1993.

[159] 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.

[160] B. Zeng, H. Zhao, and M. Lawley. The impact of overbooking on primary care patient no-show. *IIE Transactions on Healthcare Systems Engineering*, 3(3): 147–170, 2013.

[161] Z. G. Zhang and N. Tian. An analysis of queueing systems with multi-task servers. *European Journal of Operational Research*, 156(2):375–389, 2004.

Summary

Queueing models are typically used to analyze stochastic systems where congestion occurs. Prominent examples are grocery stores, amusement parks and road networks (visible queues), and call centers, communication networks, manufacturing and computer systems (at a more abstract level). In this thesis, we study multi-class queues, or more specifically, we consider a single queueing node that is used by multiple customer classes. Three common types of multi-class queues are: *priority* queues, *polling models*, and *Processor Sharing* queues. In priority queues, the customer classes are subject to a priority structure in which high priority classes have preferential treatment over lower priority classes. In polling models, customers of different classes arrive in different queues. There is a single server that can serve only one queue at a time and then switches to a different queue. Finally, Processor Sharing queues are characterized by the fact that all customers receive service simultaneously. Nonetheless, high priority customers may receive a larger share of the server. Following this hierarchy, priority queues are found in Chapters 6 and 7, polling models in Chapters 2–4, and Process Sharing queues in Chapter 5.

In Chapter 2 we analyze polling models with gated and globally gated service disciplines. We consider the following five local scheduling policies: FCFS, LCFS, ROS, PS and SJF (see Table 1.1 for a description). For each configuration, we derive the distribution of the waiting time in the heavy-traffic (HT) regime, i.e., when the load tends to 1. We show that the waiting-time distribution in HT is the product of two random variables. The first random variable captures the impact of the local scheduling policy, whereas the second random variable has a gamma distribution with known parameters and is the same for all local scheduling policies. These asymptotic results, combined with low-traffic results, are used to derive closed-form approximations for the waiting-time distributions in polling models with arbitrary load. The performance of these approximations is evaluated with simulations. The numerical results show that the approximations are accurate for all possible load values.

The model and type of results of Chapter 3 are similar to those in Chapter 2, but now the service discipline is exhaustive. This policy is more challenging to analyze than the gated policies, since we now have to deal with customers arriving during the service of the queue. We derive new closed-form expressions for the asymptotic

waiting-time distribution under exhaustive service. The waiting-time distribution in HT is the product of two random variables, where the first random variable captures the impact of the local scheduling policy. The second random variable has a gamma distribution and is the same for all local scheduling policies. The difference with the gated case is the fact that the first random variables are generally more complicated. The results lead again to closed-form approximations for the waiting-time distributions in polling models with arbitrary loads, which are evaluated using simulations. The approximations are accurate for all systems with reasonable loads.

In Chapter 4 we study polling systems with globally gated or gated service disciplines and FCFS as local scheduling policy. We are interested in the transient behavior of the cycle lengths. By deriving the joint LST of x cycles in terms of the first cycle, we are able to analyze the dependency structure between the different cycles. This is useful in, e.g., systems where breakdowns or other disruptions might occur, leading to long cycle lengths. The time to recover from such events is a primary performance measure. From the joint LST, we derive first and second moments and correlation coefficients between different cycles. Numerical results show the influence of cycle lengths on subsequent cycle lengths.

In Chapter 5 we analyze a Discriminatory Processor Sharing (DPS) queue. This is a queue, where all jobs that are present are served simultaneously. The different job types are assigned different weights and, depending on those weights, each job receives a share of the server's capacity. Jobs with higher weights receive more server capacity than jobs with lower weights and thus are served relatively fast. We assume that the service times are exponential and batches of jobs of various types arrive according to a Poisson process. We are interested in the joint queue-length distribution. We show that, in the HT regime, the scaled distribution is given by a vector of known constants multiplied by a single exponentially distributed random variable (with known parameter), also referred to as a state-space collapse. This simple result can be used to approximate the joint queue-length distribution in stable DPS systems. Numerical results show the usefulness of the asymptotic results for stable systems.

In Chapter 6 we study a specific single-server priority queue with two types of jobs. This chapter is motivated by a health-care application, more specifically, by access times for an appointment at a hospital's outpatient department. The type-1 jobs are patients arriving according to a Poisson process and the type-2 jobs are other tasks (e.g., administration tasks). We assume that there is an infinite number of type-2 jobs. If the queue length of type-1 jobs is above a certain threshold level, then more type-1 jobs are taken into service, by doing less type-2 jobs. This causes type-1 to be served faster. If the queue length of type-1 jobs drops below the threshold, more type-2 jobs will be taken into service again. We are interested in the waiting-time distribution of type-1 jobs and the fraction of time that less type-2 jobs can be done. To this end, we develop two different models, where the second model also allows for randomness in the number of type-2 jobs that can be done. Based on numerical experiments, we see that such systems may efficiently operate at high loads of type 1.

In Chapter 7 we study a specific multi-server priority queue with two types of jobs. This chapter is motivated by a call center application. Type-1 jobs (e.g., inbound calls) arrive according to a Poisson process and have non-preemptive priority over type-2 jobs (e.g., emails, outbound calls). We assume again an infinite number of type-2 jobs and the service-time distribution is exponential, with different means for the different job types. Type-1 jobs have a general patience distribution, they abandon the queue if their patience is smaller than their waiting time. If there is no queue of type-1 jobs, some of the servers will be kept idle, so that they are able to immediately handle arriving type-1 jobs. For the type-1 jobs, we derive the waiting-time distribution and the probability to abandon. The waiting-time distribution is given by the solution of second-order differential equations. When customers have infinite patience, the waiting-time distribution can be written as a mixture of exponentials. For the type-2 jobs, we determine the throughput.