Queues with regular variation

Citation for published version (APA):
Deng, Q. (2001). *Queues with regular variation*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Mathematics and Computer Science]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR547197

DOI:
10.6100/IR547197

Document status and date:
Published: 01/01/2001

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Queues with Regular Variation

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Deng, Qing

Queues with Regular Variation/
by Qing Deng. -
Eindhoven : Eindhoven University of Technology, 2001
Proefschrift. - ISBN 90-386-0901-9
NUGI 811
Subject headings : Queueing Theory / Asymptotics
2000 Mathematics Subject Classification : 60K25, 90B22

Printed by Universiteitsdrukkerij Technische Universiteit Eindhoven

Queues with Regular Variation

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. R.A. van Santen, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 5 september 2001 om 16.00 uur

door

Qing Deng

geboren te Jiangxi, China

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. O.J. Boxma
en
prof.dr.ir. S.C. Borst

To my parents

Acknowledgements

This thesis is the result of four years' research carried out in the Department of Econometrics and the CentER for Economic Research at Tilburg University, where I stayed for the first year, and in the Department of Mathematics and Computer Science at Eindhoven University of Technology, where I stayed for the last three years. In various phases of this research, I received much help from a number of people. I would like to take this opportunity to acknowledge their contributions, realizing that such a list can never be complete.

First and foremost, I am greatly indebted to my supervisor prof. Onno Boxma for guiding me through the four years' study. His excellent supervision made it possible for me to accomplish the research in this thesis. His kindness, patience and encouragement have been an important impetus for me to keep going on. Prof. Sem Borst joined the supervision at the last stage of this research. I thank him for carefully reading the manuscript of this thesis and making many valuable comments. I am also grateful to the other members in my defense committee, prof. Michel Mandjes and dr. Hans Blanc for evaluating my work and providing useful comments. In addition, I would like to thank prof. Mufa Chen, who made it possible for me to come to the Netherlands.

I would also like to express my gratitude to Bert Zwart, who created a pleasant work atmosphere in our office. I benefited a lot from the discussions with him. I am grateful to Tjark Vredeveld for sharing the office and helping me in various aspects.

And last, but not least, I would like to thank my family members for their continuous support and encouragement.

Contents

Chapter 1

Introduction

1.1 Background

Queueing theory plays an important role in the design of telecommunication networks. Simple models, like fluid queues or classical single-server queues, can often be used to obtain insightful results, e.g., to predict the global traffic behavior. Traditional queueing models typically assume that the interarrival and service times have finite variance (e.g., exponential or Erlang distribution). As a result, the aggregate traffic that is offered by a collection of sources behaves like white noise.

Recently, it has become clear that delay and buffer content distributions in modern communication networks often do not exhibit such a behavior like white noise. Many studies on traffic measurements from a variety of communication networks, like Ethernet local area networks (LANs) [113], wide area networks (WANs) [89] and variable bit rate (VBR) video over asynchronous transfer mode (ATM) [10], etc., have shown a striking difference between actual network traffic and assumptions in traditional theoretical traffic models. That is, actual network traffic is often *self-similar* or *long-range dependent* in nature. In other words, the traffic looks statistically the same over a wide range of time scales, from milliseconds to minutes and even hours. This conclusion is supported by statistical analysis of numerous high-quality Ethernet and Internet traffic measurements, cf. [74]. In contrast, traditional traffic models focus on a very limited range of time scales and do not have the property of *long-range dependence*.

As pointed out in [114], the long-range dependent nature of Ethernet LAN traffic is caused by the presence of the Noah Effect (high variability) in the traffic generated by the individual source-destination pairs that make up the

aggregate packet stream. Intuitively, the Noah Effect for a fluid queue results in the activity period and/or silence period of an individual traffic source to be very large with nonnegligible probability. More precisely, the activity period and/or silence period (service time and/or interarrival time in an ordinary queue) have infinite variance. A typical example is a power-tailed distribution, like the Pareto distribution: $F(t) = 1 - (\frac{\theta}{\theta+t})^{\nu}$ $(\theta > 0)$, if $1 < \nu < 2$. Many studies have been devoted to this subject and have established a relationship between long-range dependent processes and distributions with infinite variance, see e.g. [93, 94, 106]. It has been pointed out in e.g. [23, 28] that fluid or ordinary queues with input distributions (like the activity period distribution in a fluid queue, or the service time distribution in an ordinary queue) which do *not* have finite variance are useful and tractable models for analyzing the effect of self-similar traffic on system performance.

Tail probabilities are particularly helpful in understanding the performance of queueing systems. There has been a sizable amount of literature which obtains asymptotic results on queueing models with light-tailed input, but not much on queueing models with heavy-tailed input. For the definition of "light-tailed" and "heavy-tailed", see Chapter 2. Note that the class of regularly varying distributions belongs to the class of heavy-tailed distributions and contains the typical example (the Pareto distribution we mentioned above) of distributions with infinite variance.

The main goal of this thesis is to investigate the effect of regularly varying (sometimes heavy-tailed) service times on the tail of the waiting time or workload distributions for several queueing models. More specifically, we study in detail the following four queueing models: (i) the $M/G/1$ queue with priority classes, (ii) the tandem queueing system with Poisson input processes and identical service times at both queues, (iii) the cyclic polling system with Poisson input processes, and (iv) the $M/G/2$ queue with heterogeneous servers. For these models, we assume that at least one of the service times has a regularly varying (sometimes heavy-tailed) distribution. We find, for models (i), (ii) and (iii), that the service time with the largest tail probability governs the tail behavior of the waiting time and workload distributions. For the multiserver queue (the $M/G/2$ queue with heterogeneous servers), the waiting time tail behavior depends not only on the service time tail behavior, but also on the total traffic load. We also develop intuitive arguments which give insight into the most likely way in which large waiting times or workloads may occur.

The remainder of this introductory chapter is organized as follows. Section 1.2 is devoted to some basic knowledge of queueing theory and the performance analysis of computer-communication networks. In Section 1.3, we briefly in-

troduce the concepts of self-similarity and long-range dependence and the corresponding stochastic modeling. We describe some results for queueing models with heavy-tailed distributions in Section 1.4, where we focus on results which are relevant to this thesis. An overview of this thesis is given in Section 1.5.

1.2 Queueing theory and the performance analysis of computer-communication systems

Queueing theory

In general, a queueing model describes a system in which customers arrive who require a certain amount of work to be done by the servers. An important characteristic of queueing models is the randomness of the interarrival and service times. We refer to Cooper [38] for an excellent survey paper which covers many of the most important results in queueing theory till the end of the eighties.

In the following we describe the most basic queueing model, the $G/G/1$ queue. Here we use the notational convention introduced by Kendall [66]. To define the arrival process, it is assumed that customers always arrive individually. Let A_n $(n = 2, 3, ...)$ denote the interarrival time between the nth customer and the $(n-1)$th customer, B_n $(n = 1, 2, ...)$ the amount of service that the nth customer needs, and W_n $(n = 1, 2, ...)$ the waiting time of the nth customer. Here interarrival time is defined as the time interval between two consecutive arrivals, and waiting time is defined as the time interval between the arrival epoch of a customer and the epoch that the server starts to serve that customer. In many queueing studies, the waiting time distribution is one of the important quantities to focus on. If the service policy is First-Come-First-Served (FCFS), i.e., customers are served in order of arrival, then it is well-known that the waiting time W_n can be represented as

$$W_n = \max(W_{n-1} + B_{n-1} - A_n, 0), \quad n \geq 2.$$

The interarrival time sequence $\{A_n : n = 2, 3, ...\}$ and service time sequence $\{B_n : n = 1, 2, ...\}$ are usually assumed to be sequences of i.i.d. (independent and identically distributed) random variables. Furthermore, the arrival process is assumed to be independent of the service process. In order to emphasize that the interarrival time sequence and service time sequence are i.i.d., this model is also called $GI/GI/1$ queue.

The service discipline plays an important role in determining the waiting time distribution. Various service disciplines have been proposed and studied

in the queueing literature. The FCFS discipline mentioned above is probably the most common among them. In multi-class (or multi-queue) systems, however, customers may be served according to a non-FCFS discipline. We now mention two such disciplines which are particularly relevant for this thesis: (i) priority strategy; and (ii) polling strategy. In priority models, there are several classes of customers. Each class is assigned a fixed ranking. Waiting customers from higher ranking classes always have priority for service over those from lower ranking classes, while customers from the same class are served in order of arrival. The service discipline is either nonpreemptive or preemptive resume. In the nonpreemptive case, the service of customers of lower rank cannot be interrupted by the arrival of customers of higher rank. In the preemptive resume case, interruptions occur during the service of customers of lower rank when customers of higher rank arrive, and the server starts serving the higher ranking customers first. When the server finishes serving all higher ranking customers, it resumes the interrupted service of the lower ranking customers. The second discipline that has a prominent place in this thesis is the polling strategy. In polling systems, there is one server attending to a number of queues. Customers arrive at each queue independently. The server visits the queues in a certain order, e.g., cyclicly. Each queue is served according to some polling discipline, e.g., the exhaustive service discipline, which means that the server continues serving a queue until it becomes empty.

There are many variants of the basic $G/G/1$ queue which have been studied in the literature. For example, we may consider the $G/G/c/c+d$ queue, in which there are c servers instead of 1, and d positions for customers waiting for service. In such a system new customers are blocked and lost from the system if all waiting positions are occupied. When $d = \infty$, we omit the last symbol in the notation. An important class of queueing models is the $M/G/c$ queue, in which the arrival process is assumed to be a Poisson process. The symbol M stands for the Markovian (or Memoryless) property of the Poisson process. Poisson arrival processes occur quite naturally, see [39]. Accordingly, we restrict ourself to Poisson arrival processes in this thesis.

The performance analysis of computer-communication networks

Queueing theory has always been inspired by new questions occurring in the performance modeling and analysis of manufacturing systems, computer systems, and in particular communication networks. The first queueing problems were formulated and studied by the famous Danish scientist A.K. Erlang during the years 1908-1922 in the context of telephone networks. Later it turned out that problems arising in telephone networks are also relevant for vari-

ous other fields of research: engineering, economics, manufacturing, computer communications, etc. In communication networks, there is typically a service resource (e.g. a transmission link) and tasks (e.g. calls or data packets) requiring service from that resource. This raises the issue of how one should organize the resource and its buffer to guarantee a particular level of performance. The basic queueing model described above provides a suitable starting point for studying that issue.

In general, queueing theoretic data flow analysis in computer systems and communication networks is presently known as *performance evaluation.* Performance evaluation methods in data communications were strongly stimulated by Kleinrock's pioneering work [70], in which he made a connection between message switching and packet switching networks and (Jackson) queueing networks. For general references in the vast area of performance evaluation of data communication networks, we refer to the classical text books of Kleinrock [71] and the methodology of Heidelberger and Lavenberg [62].

The book by Walrand and Varaiya [110] is a good source which provides a firm background on modern networking technology and explains how questions in computer-communication networks are related to queueing theory. A recent landmark in performance modeling of communication networks and the Internet is the discovery of self-similar traffic by Willinger et al. [113]. This discovery has stimulated several works that provide mathematical models of long-range dependent traffic with a view towards facilitating performance analysis in a queueing theoretic sense, e.g., [47, 75, 83, 88, 108]. These works establish basic performance limits by investigating queueing models with long-range dependent input, which exhibit fundamentally different performance characteristics from corresponding systems with Markovian input. The analysis of such non-Markovian queueing systems is highly nontrivial and provides fundamental insight into the performance impact of long-range dependent traffic.

1.3 Long-range dependence phenomena in teletraffic

Concepts of self-similarity and long-range dependence

Self-similarity and fractals are concepts pioneered by Mandelbrot [77]. They describe the phenomenon where a certain property of an object is preserved under scaling in space and/or time. Self-similar stochastic processes were introduced by Cox [40]. Let $X = \{X_n : n \geq 1\}$ denote a stochastic process with finite mean $\mu = \mathrm{E}[X_n]$ and finite variance $\sigma^2 = \mathrm{E}[(X_n - \mu)^2]$. Furthermore,

we assume X to be "stationary" in the sense that its behavior or structure is invariant with respect to shifts in time. We call X *strictly stationary* if $(X_{t_1}, ..., X_{t_i})$ and $(X_{t_1+k}, ..., X_{t_i+k})$ have the same joint distribution for all $i, t_1, ..., t_i, k \geq 1$. Since strict stationarity is too restrictive in many cases, we introduce a weaker form of stationarity — *second-order stationarity* — which requires that the autocorrelation function $r_X(k) := \frac{1}{\sigma^2}\mathrm{Cov}(X_n, X_{n+k})$ $(k \geq 0)$ does not depend on n. Consider the corresponding aggregated process $X^{(m)} = \{X_n^{(m)} : n \geq 1\}$ which is derived from X by setting $X_n^{(m)} = n^{-1}(X_{(m-1)n+1} + ... + X_{mn})$ for $m, n \geq 1$. Following Cox [40] and Park and Willinger [86] we give the following definitions of self-similarity.

Definition 1.3.1 *A stationary stochastic process X is called* self-similar *with Hurst parameter $0 < H < 1$ if X and $m^{1-H}X^{(m)}$ have the same finite-dimensional distributions for all $m \geq 1$.*

Definition 1.3.2 *A second-order stationary stochastic process X is said to be* second-order self-similar *with Hurst parameter $0 < H < 1$ if X and $m^{1-H}X^{(m)}$ have identical correlation function for all $m \geq 1$. X is said to be* asymptotically second-order self-similar *if the limiting process $m^{1-H}X^{(m)}$ for $m \to \infty$ is second-order self-similar.*

The above definitions can easily be extended to continuous-time stochastic processes. A continuous-time stochastic process $Y = \{Y_t : 0 \leq t < \infty\}$ is called *self-similar* with Hurst parameter $0 < H < 1$ if $\{Y_{at} : 0 \leq t < \infty\}$ and $\{a^H Y_t : 0 \leq t < \infty\}$ have identical finite-dimensional distributions for all $a > 0$. The definitions of (asymptotic) second-order self-similarity for continuous-time processes can be given in a corresponding way. Note that for Gaussian processes, self-similarity and second-order self-similarity are equivalent, since their joint finite-dimensional distributions are fully determined by their first and second moments. Next we introduce the definition of long-range dependence.

Definition 1.3.3 *A second-order stationary stochastic process X is called* short-range dependent *if the autocorrelation function is summable, i.e., $\sum_{k=0}^{\infty} r(k) < \infty$. X is called* long-range dependent *if the autocorrelation function decays so slowly that $\sum_{k=0}^{\infty} r(k) = \infty$.*

The above definition can also be extended to continuous-time stochastic processes, cf. e.g., [28]. In general, self-similarity and long-range dependence

are not equivalent. For example, Brownian motion is self-similar in a distributional sense with Hurst parameter $H = 1/2$, but it is not long-range dependent. However, in the case of asymptotic second-order self-similarity with the restriction $1/2 < H < 1$, self-similarity is essentially equivalent to long-range dependence. The case $H \in (0, 1/2)$ is called *antipersistent* in the terminology of [77]. To be antipersistent is to tend to turn back constantly toward to the point one came from, hence to diffuse more slowly than the Brownian motion. This case is of minor interest for modeling purposes.

In order to get a better understanding of self-similarity and long-range dependence, we now show a few examples. A widely used family of self-similar processes is the class of fractional Brownian motions with Hurst parameter $H \in (0, 1)$, which was introduced in [78]. A normalized fractional Brownian motion $\{U_t : 0 \leq t < \infty\}$ with $H \in (0, 1)$ is characterized by the following properties, cf. e.g. [83].

- U_t has stationary increments;
- $U_0 = 0$, and $\mathrm{E}[U_t] = 0$ for all t;
- $\mathrm{E}[U_t^2] = t^{2H}$ for all t;
- U_t has continuous sample paths;
- U_t has a Gaussian distribution for all t.

Next we introduce a self-similar discrete-time stochastic process, for which we refer to [10]. Take $X = \{X_n : n \geq 1\}$ with $X_n = U_n - U_{n-1}$. Then the corresponding autocorrelation function $r(\cdot)$ is of the form

$$r(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}], \quad k \geq 1,$$

where $H \in (0, 1)$. In [87] the above formula is given as the definition of self-similar discrete-time processes. It can be checked that $r(k) = r^{(m)}(k)$ for all $m, k \geq 1$, where $r^{(m)}(\cdot)$ stands for the correlation function of $X^{(m)}$. Noticing that X is a Gaussian process, by Definition 1.2.1, X is self-similar. Note that for $H = 1/2$, X is uncorrelated, and for $H \in (1/2, 1)$, we have

$$r(k) \sim H(2H-1)k^{2H-2}, \quad k \to \infty,$$

which implies that $\sum_{k=1}^{\infty} r(k) = \infty$. Therefore, X exhibits long-range dependence. In this thesis we follow the notational convention that $f(t) \sim g(t)$ as $t \to c$ $(0 \leq c \leq \infty)$ stands for $\lim_{t \to c} f(t)/g(t) = 1$.

Stochastic modeling of long-range dependence

We now discuss two ways to model long-range dependence in an input process.

One possibility to introduce long-range dependence in an input process is to take fractional Brownian motion. This model was originally proposed by Norros [83], see also [84]. In [83], Norros studies a fluid queue fed by an input process $A(t) = mt + \sqrt{am}U_t$ where U_t is a normalized fractional Brownian motion with Hurst parameter H ($1/2 \leq H < 1$). One of his main contributions is to derive a lower bound for the complementary distribution function of the workload. Letting V denote the steady-state workload, Norros' result is as follows:

$$\mathbf{P}(V > x) \geq 1 - \Phi\left(\frac{1}{\sqrt{am}}\left(\frac{1-m}{H}\right)^H \left(\frac{x}{1-H}\right)^{1-H}\right),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. For the case $H = \frac{1}{2}$, $\Phi(\cdot)$ reduces to the exponential distribution and the lower bound coincides with the exact asymptotics; see Takács [105]. For the same model of [83], Massoulié and Simonian [80] prove an upper bound for the tail of the workload distribution by using extremal properties of Gaussian processes. Under a few assumptions, Narayan [82] proves that Norros' lower bound is an asymptotic expression for the workload distribution.

Another possibility to introduce long-range dependence is to take a single on/off source with on- and/or off-period distributions which do not have finite variance (see Roberts, Mocci and Virtamo [93]). They assume that the distribution of the on-period A has the following tail:

$$\mathbf{P}(A > t) \sim h_a t^{-a}, \quad \text{as } t \to \infty,$$

and/or the distribution of the off-period S has the following tail:

$$\mathbf{P}(S > t) \sim h_s t^{-s}, \quad \text{as } t \to \infty,$$

with $1 < a, s < 2$ and h_a, h_s positive constants. As pointed out in [93], the input process is indeed long-range dependent. Boxma and Dumas [28] establish a relation between the integrated covariance and the distribution functions of the on- and off-periods without any assumptions on the tails of the on- and off-period distributions, thus clarifying how long-range dependence occurs in the on/off source. The fluid queue with a single on/off source can be generalized to a system with N identical independent on/off sources, cf. Willinger et al. [114]. When taking $N \to \infty$, it is shown in [114] that the aggregate traffic, suitably normalized, is fractional Brownian motion.

1.4 Heavy tails and queues

The importance of queueing models with heavy-tailed distributions has triggered a large body of literature on this subject in recent years. This thesis focuses on the tail behavior of the waiting time or workload distributions. It should be pointed out that another asymptotic regime is to let the number of sources tend to infinity, see e.g. [46, 79]. In the remainder of this section, we review some results which are relevant to this thesis. We refer to Chapter 2 for the concepts of subexponential and regularly varying distributions.

In the following, we describe some results for the $G/G/1$ queue. We introduce some notation first. Denote by λ the arrival rate, by β and β_2 the first and second moment of the service time, and by $\rho := \lambda\beta$ the traffic load. Let B be the service time and B^{res} the residual service time which has density function $\mathbf{P}(B > t)/\beta$. Pakes [85] has proven that, in the $G/G/1$ queue with FCFS discipline, if the residual service time B^{res} has a subexponential distribution, then the tail of the steady-state waiting time W is related to the tail of the residual service time in the following way:

$$\mathbf{P}(W > t) \sim \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t), \quad t \to \infty. \tag{1.4.1}$$

Cohen [34] proves that the waiting time distribution has a regularly varying tail of index $1-\nu$ if and only if the service time distribution has a regularly varying tail of index $-\nu$ ($\nu > 1$). When the service time has a regularly varying tail, asymptotic results for the waiting time distribution in the $GI/GI/1$ queue with other service disciplines are available as well.

As mentioned earlier, different service disciplines result in different queueing performance. In the processor sharing $M/G/1$ queue, Zwart and Boxma [116] show that the sojourn time distribution has a regularly varying tail with the same index as the service time distribution tail. Zwart [115] generalizes the above result to the processor sharing queue with multiple customer classes. Boxma and Cohen [23] study the $M/G/1$ queue with the Last-Come-First-Served (LCFS) preemptive resume discipline as well as the LCFS nonpreemptive discipline. They obtain a similar result, i.e., the sojourn time tail in the LCFS nonpreemptive queue is regularly varying of index one higher than the service time (like FCFS) and the sojourn time tail in the LCFS preemptive resume queue is regularly varying of the same index as the service time. An $M/G/1$ queue with two priority classes and either the nonpreemptive or the preemptive resume discipline is considered by Abate and Whitt [3]. They study the effect of the service time distribution tail (which is light or heavy) on the tails of the waiting time distributions.

When both service time and interarrival time distributions have a finite second moment, a standard heavy-traffic limit theorem for the stationary waiting time W in the $G/G/1$ queue holds (cf. [68]):

$$\lim_{\rho\uparrow 1} \mathbf{P}(\Delta(\rho)W \leq t) = 1 - e^{-t},$$

where $\Delta(\rho) := 2\lambda(1-\rho)/[1+\lambda^2(\beta_2-\beta^2)]$. Boxma and Cohen [22] generalize the above results to the $G/G/1$ queue for which the interarrival and service times may have *infinite* second moments. One of their results is as follows: Assume that the tail of the service time distribution is regularly varying of index $-\nu$ $(1 < \nu < 2)$ and the tail of the interarrival time distribution is less heavy than that of the service time distribution. Then the distribution of the contracted waiting time $\Delta(\rho)W$ converges for $\rho \uparrow 1$ to the Mittag-Leffler distribution $R_{\nu-1}(t)$, which is specified by:

$$\int_0^\infty e^{-st} \mathrm{d}R_{\nu-1}(t) = \frac{1}{1+s^{\nu-1}}.$$

The coefficient of contraction $\Delta(\rho)$ is the unique solution to a contraction equation with the property that $\Delta(\rho) \downarrow 0$ for $\rho \uparrow 1$. In the case that the service time distribution is Pareto, the coefficient of contraction is given by $\Delta(\rho) = C(1-\rho)^{\frac{1}{\nu-1}}$ where C is a constant. Here the terms 'contracted', 'coefficient of contraction' and 'contraction equation' were pioneered by Boxma and Cohen [22]. The waiting time W tends to infinity as $\rho \uparrow 1$, while the product of W and $\Delta(\rho)$ tends to a finite random variable as $\rho \uparrow 1$. Therefore, $\Delta(\rho)W$ is called the 'contracted' waiting time. Boxma, Cohen and Deng [24] prove a similar result as in [22] for the low-priority waiting time distribution in the $M/G/1$ queue with priority classes.

A cyclic polling model with gated or exhaustive service is studied by Boxma, Deng and Resing [26]. They assume that at least one of the service time distributions is regularly varying of index $-\nu$ $(\nu > 1)$ and other service time distributions have a less heavy or similar tail behavior. It is shown that the waiting time distribution is regularly varying of index $1-\nu$.

Determining the tail behavior of the waiting time distribution in multi-server queues with heavy-tailed service time turns out to be a hard problem. Scheller-Wolf and Sigman [95, 96] attack this problem by studying the effect of the service time moments on the waiting time moments. For the stable FCFS $G/G/k$ $(k \geq 2)$ queue, they show that in order to have a finite kth moment of the waiting time, it is in general not necessary that the $(k+1)$th moment of the service time is finite. Their results weaken some classical conditions of Kiefer

and Wolfowitz [67] (which stated that a finite $(k+1)$th moment of the service time is sufficient for a finite kth moment of the waiting time). This suggests that the tail behavior of the waiting time may be less heavy than that of the residual service time. Based on previous work, Whitt [112] partly proves and partly conjectures bounds for the waiting time tail, which vary for different regimes of the traffic load. Since it is difficult to obtain exact asymptotics of the waiting time tail for the $G/G/k$ $(k \geq 2)$ queue, Boxma, Deng and Zwart [27] consider the $M/G/2$ queue with one exponential server and one general server. When the service time distribution at the general server has a regularly varying tail, they derive exact asymptotics of the waiting time, which indeed shows different behavior for different regimes of the traffic load. The $G/G/2$ queue with subexponential service time distributions at both queues is studied by Foss and Korshunov [53]. They present asymptotics for the waiting time tail in terms of a complicated expression involving the service time distributions. In the case of regularly varying service times, the expression can be simplified and thus exact asymptotics are obtained.

Queueing networks with heavy-tailed service time distributions have recently been studied in several papers. Anantharam [6] and Boxma and Dumas [29] obtain results regarding the propagation of long-range dependence in networks of (fluid) queues. Baccelli, Schlegel and Schmidt [9] consider tandem queues with a (Palm) stationary arrival process at the first node and *independent* service times at the various nodes, that have a subexponential distribution in at least one node. They derive lower and upper bounds for the tails of the sojourn time distributions; in some cases, these bounds coincide and hence the precise tail behavior is established. Partly building upon [9], Huang and Sigman [63] show that, in the two-node case, if the service time distribution at the second node is subexponential and the service time distribution at the first node has a lighter tail, then the tail behavior of the waiting time at the second node has the same asymptotics as if it were an ordinary $G/G/1$ queue in isolation. The same results can be extended to tandem queues with more nodes. Boxma and Deng [25] is also devoted to a tandem queue with heavy-tailed service time distributions.

1.5 Overview of the thesis

In the present chapter we have already described the motivation for studying queueing models with heavy-tailed service time distributions, and we have discussed some of the main developments in this area. In the remainder of this thesis, we analyze the following models with regularly varying (or heavy-tailed)

service times: the $M/G/1$ priority queue with nonpreemptive or preemptive resume discipline, the cyclic polling system with Poisson input, the $M/G/2$ queue with heterogeneous servers and the tandem queueing system with identical service times at both queues.

In Chapter 2 the basic properties of heavy-tailed distributions are discussed. We focus on two important subclasses: the class of subexponential distributions and the class of regularly varying distributions.

We study the $M/G/1$ queue with two priority classes in Chapter 3. The service times of the high- and/or low-priority customers are assumed to be regularly varying of index $-\nu$ $(1 < \nu < 2)$. Based on an expression for the Laplace-Stieltjes transforms (LST) of the low-priority waiting time distribution given by Abate and Whitt [3], we establish relations between the tail behavior of the waiting time distribution of the low-priority customers and that of the service time distributions. Furthermore, using similar techniques as in [22], we derive a heavy-traffic limit theorem for the waiting time distribution of the low-priority customers when the total traffic load $\rho \uparrow 1$. Chapter 3 builds on the analysis presented in Boxma, Cohen and Deng [24].

Chapters 4 and 5 are devoted to the cyclic polling system with Poisson arrival processes. In Chapter 4, we study a two-queue model with 1-limited service at one queue and exhaustive service at the other queue. Note that this model reduces to the $M/G/1$ queue with two priority classes of Chapter 3 if there is no switchover time. For the case in which there are switchover times and at least one of the service times and/or switchover times has an infinite variance, we derive a heavy-traffic limit theorem for the waiting time at the 1-limited-service queue. Finally we numerically test the approximation of the waiting time distribution at the 1-limited-service queue suggested by the heavy-traffic limit theorem.

In Chapter 5, we study the cyclic polling system with gated or exhaustive service at each queue. It is assumed that the service time distribution with the heaviest tail behavior has a regularly varying tail of index $-\nu$ $(\nu > 1)$. Based on an explicit expression for the LST of the waiting time distributions (cf. [13, 14]), we prove that the waiting time distribution at each queue is regularly varying of index $1 - \nu$. The analysis is based on Boxma, Deng and Resing [26].

Chapter 6 is devoted to the $M/G/2$ queue with one exponential server and one general server. Using the supplementary variable technique, we establish a set of differential equations satisfying some boundary condition. In the case that the LST of the service time distribution at the general server is rational, we can explicitly solve the differential equations and thus the LST of the steady-

state waiting time distribution follows. In the case that the service time at the general server has a regularly varying tail, we derive the tail behavior of the waiting time by using analytic methods. Furthermore, we provide intuitive arguments for the waiting time tail behavior. This chapter presents the results of Boxma, Deng and Zwart [27].

In Chapter 7, we turn to the tandem queueing system with identical service times at both queues. We focus on the steady-state sojourn time and workload at the second queue. Starting from explicit expressions for the distributions of the sojourn time and workload at the second queue (cf. Boxma [18]), we relate the tail behavior of the sojourn time distribution and the workload distribution at the second queue to that of the (residual) service time distribution. As a by-product, we prove that both the sojourn time distribution and the workload distribution at the second queue are regularly varying of index $1-\nu$, if the service time distribution is regularly varying of index $-\nu$ ($\nu > 1$), which coincides with the results we obtain by using intuitive arguments. Furthermore, in the latter case, we derive a heavy-traffic limit theorem for the sojourn time at the second queue when the traffic load $\rho \uparrow 1$. Chapter 7 presents the analysis in Boxma and Deng [25].

Chapter 2

Heavy-tailed distributions

2.1 Introduction

In this chapter we introduce some concepts and notation which we use throughout this thesis. In particular, we focus on some basic properties of heavy-tailed distributions.

Heavy-tailed distributions are widely used in the literature at present. We refer to [99] for a nice and brief introduction to heavy-tailed distributions. In this chapter we give the definition and discuss an important subclass: subexponential distributions. Furthermore, we shall focus on a subclass of subexponential distributions - the class of regularly varying distributions, and describe the main properties of this class which we shall use throughout this thesis. We assume all random variables in this thesis are nonnegative, unless indicated otherwise.

Definition 2.1.1 *A random variable X, with distribution function $F(\cdot)$, is called* heavy-tailed*, if for all real u:*

$$\lim_{t\to\infty} \frac{1-F(t+u)}{1-F(t)} = 1.$$

We denote the class of heavy-tailed distributions by $\mathcal{L}$ (and use the notation $F \in \mathcal{L}$ or $X \in \mathcal{L}$). In stark contrast to exponential distributions, heavy-tailed distributions satisfy the following property which was proved by Chistyakov [32] and Embrechts et al. [51].

Lemma 2.1.1 *If $X \in \mathcal{L}$, then for all $\epsilon > 0$, $e^{\epsilon t}\mathbf{P}(X > t) \to \infty$ as $t \to \infty$. In other words, for any $\epsilon > 0$, $\mathrm{E}[e^{\epsilon X}] = \infty$.*

Because of the above property of heavy-tailed distributions, we refer to any distribution function $F(\cdot)$ (or random variable X) as light-tailed if $\mathrm{E}[e^{\epsilon X}] < \infty$ for some $\epsilon > 0$. This includes the classical example, exponentially distributed random variables, as well as all bounded random variables.

2.2 Subexponential distributions

A very important subclass of $\mathcal{L}$ is the class of subexponential distributions, denoted by $\mathcal{S}$. We briefly discuss some properties of $\mathcal{S}$ next. For a detailed discussion and further references, we refer to Embrechts et al. [50] and Goldie and Klüppelberg [57].

Definition 2.2.1 *A random variable X, with distribution function $F(\cdot)$, is called* subexponential*, to be denoted by $X \in \mathcal{S}$ or $F \in \mathcal{S}$, if*

$$\lim_{t\to\infty} \frac{\mathbf{P}(X + X' > t)}{\mathbf{P}(X > t)} = 2,$$

where X and X' are i.i.d.

In the following examples, $F(\cdot)$ denotes the distribution function of random variable X.

Example

1. *(Pareto):* $F(t) = 1 - t^{-\alpha},\ t \geq 1$, with $\alpha > 0$. (There are many variations on this, such as $F(t) = 1 - (\frac{c}{c+t})^{\alpha},\ \ t \geq 0$, with $c > 0$ and $\alpha > 0$).

2. *(Lognormal):* Density

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(t) - \mu)^2}{2\sigma^2}\right), \quad t > 0,$$

   with $\sigma > 0$ and $\mu \in (-\infty, \infty)$. This is the distribution of the random variable $X = e^Y$ where Y has a normal distribution with mean μ and variance σ^2.

3. *(Heavy-tailed Weibull):* $F(t) = 1 - \exp(-\lambda t^{\alpha}),\ t \geq 0$, with $\lambda > 0$ and $0 < \alpha < 1$. Such a random variable X can be derived from an exponential random variable Y with $P(Y > t) = \exp(-\lambda t)$ via the transformation $X = Y^{1/\alpha}$, which immediately yields the interesting fact that the Weibull distribution possesses finite moments of all orders (i.e., it has infinite moment generating function).

4. *(Regularly varying tail):* With $\alpha > 0$, the tail of X is said to be regularly varying with index $-\alpha$ if $1 - F(t)$ is a regularly varying function, that is, if
$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}, \quad x > 0.$$
An alternative definition will be given in Section 2.3. Note that the Pareto distribution is a special case of the class of regularly varying distributions.

The class of subexponential distributions has the following nice properties. For a proof of Lemma 2.2.1, see e.g., Athreya and Ney [8]. For a proof of Lemma 2.2.2, see Pakes [85].

Lemma 2.2.1 *Let $\{X_n : n \geq 1\}$ be an i.i.d. sequence. If $X_1 \in \mathcal{S}$, then*
(i) for $n \geq 1$: $\mathbf{P}(X_1 + ... + X_n > t)/\mathbf{P}(X_1 > t) \to n$ as $t \to \infty$;
(ii) for any $\epsilon > 0$, there exists a constant $K > 0$, such that for any $t \geq 0$, $n \geq 1$, we have
$$\frac{\mathbf{P}(X_1 + ... + X_n > t)}{\mathbf{P}(X_1 > t)} \leq K(1 + \epsilon)^n.$$

Lemma 2.2.2 *If $X \in \mathcal{S}$ and $\mathbf{P}(Y > t) \sim K\mathbf{P}(X > t)$ (where K is a positive constant), then $Y \in \mathcal{S}$.*

Recall that $f(t) \sim g(t)$ means $\lim_{t\to\infty} f(t)/g(t) = 1$.

2.3 Regularly varying distributions

Unfortunately it is in general unknown how to conclude that a distribution is subexponential from its LST. In the following we discuss a subclass of subexponential distributions, called regularly varying distributions, for which there does exist a useful relation between the asymptotic behavior of its tail probabilities and the asymptotic behavior of its LST in the neighborhood of the origin.

Regular variation is an important concept in probability theory and various other fields. The main reference text is the book [12]. The definition of regular variation involves the notion of a slowly varying function. So we give the definition of a slowly varying function first. A measurable positive function $L(t)$ defined on some interval $[a, \infty)$ is called *slowly varying* if for all $x > 0$, $\lim_{t\to\infty} L(xt)/L(t) = 1$. For example, a constant and a logarithmic function are both slowly varying functions. In the remainder of this thesis, $L(\cdot)$ denotes a slowly varying function.

We explain some notation first. As usual,

$$f(s) = h(s) + \mathrm{o}(g(s)) \text{ as } s \downarrow 0$$

means that

$$\lim_{s\downarrow 0} \frac{f(s) - h(s)}{g(s)} = 0,$$

and

$$f(s) = h(s) + \mathrm{O}(g(s)) \text{ as } s \downarrow 0$$

means that

$$\limsup_{s\downarrow 0} \frac{f(s) - h(s)}{g(s)} < \infty.$$

Definition 2.3.1 *A random variable X, with distribution function $F(\cdot)$, is called* regularly varying of index $-\nu$ *($\nu > 0$), to be denoted by $X \in \mathcal{R}_{-\nu}$ or $F \in \mathcal{R}_{-\nu}$, if*

$$1 - F(t) \sim t^{-\nu} L(t), \quad t \to \infty, \tag{2.3.1}$$

where $L(\cdot)$ is a slowly varying function.

The Pareto distribution is a special case of a regularly varying distribution. For $X \in \mathcal{R}_{-\nu}$ where $m < \nu < m+1$ ($m \in \mathbb{N}$), the mth moment of X exists while the $(m+1)$th moment does not exist. Of particular interest to us is the case that $1 < \nu < 2$, i.e., X has a finite mean and infinite variance. Recall that the fluid queue with an activity period that is regularly varying of index in the interval $(-2, -1)$ exhibits long-range dependence, as mentioned in Section 1.3.

The following lemma (cf. Lemma 7.7 in [28]), which can be derived easily from Karamata's Theorem and the Monotone Density Theorem (cf. [12] Sections 1.5.6 and 1.7.3), shows the equivalence between the tail behavior of X and the tail behavior of X^{res}. In this thesis we follow the convention that X^{res} stands for the residual lifetime of X which has density function $(1-X(t))/\mathrm{E}X$.

Lemma 2.3.1 *For all $\nu > 0$, $X^{res} \in \mathcal{R}_{1-\nu}$ if and only if $X \in \mathcal{R}_{-\nu}$, and if either is the case then:*

$$\mathbf{P}(X^{res} > t) \sim \frac{t}{(\nu - 1)\mathrm{E}X} \mathbf{P}(X > t), \quad \text{as } t \to \infty.$$

The next lemma (see e.g. Kingman and Taylor [69], Theorem 12.6) establishes a relation between the finite moments of a random variable and the Taylor expansion of the corresponding LST near the origin.

Lemma 2.3.2 *Let X be a random variable with LST $f(s)$.*
(i) If X has finite moments $\phi_k := \mathrm{E}[X^k]$ of order k, $k = 0, 1, ..., n$, then

$$f_n(s) := (-1)^{n+1} \left(f(s) - \sum_{j=0}^{n} \phi_j \frac{(-s)^j}{j!} \right) = \mathrm{o}(s^n), \ s \downarrow 0. \tag{2.3.2}$$

(ii) If there exist finite constants a_j, $j = 0, ..., n$, such that

$$f(s) - \sum_{j=0}^{n} a_j s^j = \mathrm{o}(s^n), \ s \downarrow 0,$$

then $\phi_j = (-1)^j j! a_j < \infty$ for $j = 0, 1, ..., n$.

To simplify the notation, we introduce $\hat{f}_n(s) = s^{-(n+1)} f_n(s)$, which we will use in Chapter 3. Moreover, we have the following lemma, cf. Lemma 2 in [81].

Lemma 2.3.3 *If $\phi_n < \infty$ ($n \in \mathbb{N}$), then the following two statements hold:*
(i) $\hat{f}_n(s)$ is decreasing in s;
(ii) $s\hat{f}_n(s)$ is increasing in s.

The following lemma (cf. Lemma 2.2 in [28]), which is an extension of Theorem 8.1.6 (it is a special case of Karamata's Tauberian Theorem) in [12], links the regularly varying tail behavior of $\mathbf{P}(X > t)$ for $t \to \infty$ to the behavior of its LST $f(s)$. It plays a key role in the proof of our main results.

Lemma 2.3.4 *Let X be a random variable with LST $f(s)$, $L(\cdot)$ a slowly varying function, $\nu \in (n, n+1)$ ($n \in \mathbb{N}$) and $C \geq 0$. Then the following two statements are equivalent:*
(i) $\mathbf{P}(X > t) = [C + \mathrm{o}(1)]L(t)/t^\nu, \ t \to \infty$.
(ii) $\mathrm{E}[X^n] < \infty$ and $f_n(s) = (-1)^n \Gamma(1-\nu)[C + \mathrm{o}(1)]L(1/s)s^\nu, \ s \downarrow 0$.

Here $\Gamma(\cdot)$ denotes the Gamma function. The next lemma characterizes a property of slowly varying functions.

Lemma 2.3.5 *Let $L(x)$ be a slowly varying function, and $t(x)$ a positive function such that $\lim_{x\to\infty} t(x)/x = a$ where $0 < a < \infty$. Then for any constant ν ($\nu \in \mathbb{R}$), we have*

$$\lim_{x\to\infty} \frac{(t(x))^\nu L(t(x))}{x^\nu L(x)} = a^\nu.$$

Proof. We only need to prove

$$\lim_{x\to\infty} \frac{L(t(x))}{L(x)} = 1.$$

Define $\lambda(x) := \frac{t(x)}{x}$, so that $\lambda(x) \in [a/2, 2a]$ for sufficiently large x. Applying the Uniform Convergence Theorem, cf. Theorem 1.2.1 in [12], we obtain

$$\lim_{x\to\infty} \frac{L(\lambda(x)x)}{L(x)} = 1,$$

and the result follows. □

Some key formulas of this thesis involve iterated functions. The following result is useful in this respect; it is a consequence of Lemma 2.3.5.

Lemma 2.3.6 *Suppose $\phi(\cdot)$, $\psi(\cdot)$ can be written as*

$$\phi(x) = \sum_{i=1}^{n} \phi_i x^i + \phi_\nu x^\nu L(1/x) + \mathrm{o}(x^\nu L(1/x)), \quad \textit{for } x \downarrow 0, \tag{2.3.3}$$

$$\psi(x) = \sum_{i=1}^{n} \psi_i x^i + \psi_\nu x^\nu L(1/x) + \mathrm{o}(x^\nu L(1/x)), \quad \textit{for } x \downarrow 0, \tag{2.3.4}$$

where $\phi_1, \psi_1 > 0$, $n < \nu < n+1$ ($n \in \mathbb{N}$), $\phi_i, \psi_i < \infty$ for $i = 1, ..., n$ and $L(\cdot)$ is a slowly varying function. Then the asymptotic expansion of the function $\phi(\psi(x))$ at point 0 is given by

$$\phi(\psi((x)) = \sum_{i=1}^{n} \theta_i x^i + (\phi_1\psi_\nu + \phi_\nu\psi_1^\nu)x^\nu L(1/x) + \mathrm{o}(x^\nu L(1/x)), \quad \textit{for } x \downarrow 0,$$

where $\theta_i < \infty$ for $i = 1, ..., n$.

Proof. For $1 \le i \le n$, $(\psi(x))^i$ can be written as

$$(\psi(x))^i = p_i(x) + \sum_{j=1}^{i} (x^\nu L(1/x))^j q_{i,j}(x) + \mathrm{o}(x^\nu L(1/x)),$$

where

$$p_i(x) = \left(\sum_{j=1}^{n} \psi_j x^j\right)^i,$$

$$q_{i,j}(x) \quad = \quad \binom{i}{j} \phi_\nu^j (p_i(x))^{i-j}, \quad j = 1, ..., i.$$

Note that $q_{i,1}(x)$ are all equal to 0 if $x = 0$ for $2 \leq i \leq n$. Therefore, we have

$$\sum_{i=1}^{n} \phi_i(\psi(x))^i = \sum_{j=1}^{n} a_j x^j + \phi_1 \psi_\nu x^\nu L(1/x) + \mathrm{o}(x^\nu L(1/x)), \tag{2.3.5}$$

for some real numbers $a_j < \infty$ $(j = 1, ..., n)$. Since $\lim_{x \downarrow 0} \psi(x)/x = \psi_1$, it follows from Lemma 2.3.5 that

$$\lim_{x \downarrow 0} \frac{(\psi(x))^\nu L(1/\psi(x))}{x^\nu L(1/x)} = \psi_1^\nu,$$

which in combination with (2.3.3) and (2.3.5) leads to the desired statement. □

Remark 2.3.1 It should be noted that, despite the symmetry in (2.3.3) and (2.3.4), it is possible that $\phi(x)$ refers to a heavier-tailed function than $\psi(x)$ (or vice versa); for example, ψ_ν might be equal to zero.

Remark 2.3.2 Suppose $A_1, A_{1,i}$ $(i \geq 1)$ are i.i.d. random variables with distribution function $A_1(t)$ and $A_2, A_{2,i}$ $(i \geq 1)$ are i.i.d. random variables with distribution function $A_2(t)$. The tail of $A_1(t)$ is regularly varying of index $-\nu$ $(\nu \geq 1)$ and $A_2(t)$ has a lighter tail. Denote by $\alpha_j(s)$ the LSTs of $A_j(t)$ for $j = 1, 2$. Let $K(A_j)$ $(j = 1, 2)$ be the number of arrivals of an independent Poisson process with parameter λ_j during a time period of length A_j. Thus, the generating function of $K(A_j)$ is given by

$$\begin{aligned} \mathrm{E}[z^{K(A_j)}] &= \int_0^\infty \sum_{k=0}^{\infty} z^k \frac{(\lambda_j x)^k}{k!} e^{-\lambda_j x} \mathrm{d}A_j(x) \\ &= \alpha_j(\lambda_j - \lambda_j z), \quad |z| < 1, \end{aligned}$$

for $j = 1, 2$. Define the following random sums, which naturally arise in several queueing models (cf. Chapters 3, 4 and 5) ,

$$\begin{aligned} A^{(1)} &:= A_{1,1} + A_{1,2} + ... + A_{1,K(A_2)}, \\ A^{(2)} &:= A_{2,1} + A_{2,2} + ... + A_{2,K(A_1)}. \end{aligned}$$

Then we have

$$\mathrm{E}[e^{-sA^{(1)}}] = \alpha_2(\lambda_2 - \lambda_2 \alpha_1(s)),$$

$$\mathrm{E}[e^{-sA^{(2)}}] = \alpha_1(\lambda_1 - \lambda_1\alpha_2(s)).$$

Applying the above lemma and Theorem 8.1.6 of [12] yields that $A^{(1)}$ and $A^{(2)}$ both have a regularly varying tail at infinity of index $-\nu$. Actually the tail behavior of $K_1(A_1)$ is shown to be regularly varying at infinity of index $-\nu$, cf. Chapter 8 in [58]. More general results on the tail behavior of $K_1(A_1)$ can be found in [7] where it is assumed that A_1 has a heavy-tailed (subexponential) distribution. Furthermore, Sigman [99] provides some results on the tail behavior of a random sum of some random variables with subexponential tail.

Chapter 3

The $M/G/1$ queue with priority classes

3.1 Introduction

In communication networks often different traffic types can be distinguished, with different traffic characteristics and different performance requirements. One way to implement this is by imposing a priority structure. Abate and Whitt [3] consider an $M/G/1$ queue with two priority classes and either the nonpreemptive or the preemptive resume discipline. They study the effect of the service time distribution tails on the tails of the waiting time distributions under the assumption that the service times have finite variance. In this chapter we consider the same model. We are mainly interested in the heavy-tailed (infinite variance) case, and in particular in the heavy-traffic situation. This chapter is an extended version of [24].

Let us first describe the model. There is a single server in the system. Two classes of customers with different priorities arrive according to two independent Poisson processes (which may have different rates). Within each class, customers receive service according to the FCFS discipline. When the server starts to serve a new customer, it serves the high-priority customers first. For the preemptive resume discipline, the service of low-priority customers is interrupted by arrivals of high-priority customers. In that case, the low-priority service is resumed once the server finishes serving the high-priority customers in the system. For the nonpreemptive discipline, the service of low-priority customers is not interrupted by arrivals of high-priority customers. For a complete analysis of this model in terms of LSTs, we refer to Cohen [36], Chapter III.3.

Let us now introduce some notation. The high-priority class is indexed by 1 and the low-priority class by 2. Let $B_j(t)$ denote the service time distribution function of class-j with mean $\beta_j < \infty$ and second moment $\beta_{j2} \leq \infty$, λ_j the arrival rate of class-j and $\rho_j := \lambda_j \beta_j$ the traffic load of class-j for $j = 1, 2$. The arrival processes of the two classes are independent. We assume the stability condition holds: $\rho := \rho_1 + \rho_2 < 1$.

Let W_2 denote the steady-state waiting time of the low-priority customers until the start of the service (note that it has the same distribution for the nonpreemptive and the preemptive resume discipline). If $\beta_{j2} < \infty$ for $j = 1, 2$, i.e., both service times have finite variance, then the distribution of the contracted low-priority waiting time $\zeta(\rho_2)W_2$ converges for $\rho_2 \uparrow 1 - \rho_1$ to the unit exponential distribution where $\zeta(\rho_2) := \frac{2(1-\rho_1)(1-\rho)}{\rho_1\beta_{12}/\beta_1 + \rho_2\beta_{22}/\beta_2}$, cf. [3].

Boxma and Cohen [23] deal with the $G/G/1$ queue where the service time does *not* have finite variance. They derive a heavy-traffic limit theorem for the waiting time distribution (cf. Section 1.4). Their techniques are used in this chapter.

In the present chapter we consider the above-described $M/G/1$ queue with two priority classes, for the case that at least one of the service time distributions is regularly varying of index $-\nu$ with $1 < \nu < 2$, i.e., at least one of the service times does *not* have finite variance. It is shown for this heavy-tailed case that *the waiting time distribution of the low-priority customers is regularly varying of index one degree higher than that of the service time distribution with the heaviest tail.* We also prove a heavy-traffic limit theorem for the steady-state low-priority waiting time W_2. When the low-priority traffic load $\rho_2 \uparrow 1 - \rho_1$, the distribution of the contracted low-priority waiting time $\Delta(\rho_2)W_2/\beta_1$ converges to $R_{\nu-1}(t)$ where $\Delta(\rho_2)$ is a particular function of ρ_2 with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1 - \rho_1$. The heavy-traffic limit theorem gives rise to an approximation for the steady-state distribution of W_2, which is extensively tested numerically.

This chapter is organized as follows.

In Section 3.2 we characterize the service time distributions $B_j(\cdot)$ for $j = 1, 2$. We assume that at least one of the service time distributions has a regularly varying tail of index $-\nu$ where $1 < \nu < 2$. Moreover, we derive the asymptotic expansions of the LSTs of the service time distributions and the class-1 busy period distribution.

A representation for $\omega_2(s)$, the LST of the distribution of the class-2 waiting time W_2, given by Abate and Whitt [3], is used in Section 3.3 to derive the asymptotic expansion of $\omega_2(s)$ for $s \downarrow 0$. It is also shown that the class-2 waiting time distribution $W_2(t)$ has a regularly varying tail of index $1 - \nu$.

The aim of Section 3.4 is to show a reverse result; i.e., if $W_2(t)$ has a regularly varying tail with index $1-\nu$ where $\nu > 1$ and the class-1 service time distribution $B_1(t)$ has a tail which is less heavy than $t^{-\nu}$, then the class-2 service time distribution $B_2(t)$ is regularly varying with index $-\nu$.

The asymptotic expansions obtained in Section 3.3 are used in Section 3.5 to derive the main result of this chapter (the heavy-traffic limit theorem) which is described above.

In Section 3.6 we generalize the heavy-traffic limit theorem for the waiting time distribution of the lowest priority class to the $M/G/1$ queue with k ($k \geq 2$) priority classes. We obtain a similar result as the above-mentioned heavy-traffic limit theorem.

In Section 3.7 we make a comparison with a heavy-traffic limit theorem for the waiting time distribution in the $M/G/1$ queue *without* priority structure. This suggests approximating $\mathbf{P}(W_2 > t)$ by $\mathbf{P}((1-\rho_1)W > t)$, where W is the steady-state waiting time in the model without priority structure.

In Section 3.8 we propose an approximation for $\mathbf{P}(W_2 > t)$ based on the obtained heavy-traffic limit theorem, and we numerically investigate its accuracy as well as that of $\mathbf{P}((1-\rho_1)W > t)$. Both appear to perform very well over a wide range of ρ- and t-values.

3.2 On the service time distributions

In this section we describe the classes of distributions $B_1(\cdot)$ and $B_2(\cdot)$ for which we analyze the heavy-traffic behavior of the low-priority waiting time distribution. In this chapter, we assume the variable s is real, unless indicated otherwise. For $s \geq 0$ and $j = 1, 2$, define the LSTs of the service time distributions and of the residual service time distributions,

$$\beta_j(s) := \int_0^\infty e^{-st} \mathrm{d}B_j(t), \tag{3.2.1}$$

$$\beta_{je}(s) := \int_0^\infty e^{-st} \frac{1-B_j(t)}{\beta_j} \mathrm{d}t = \frac{1-\beta_j(s)}{\beta_j}. \tag{3.2.2}$$

Concerning the service time distributions $B_j(\cdot)$ for $j = 1, 2$, we only make assumptions about their tails, i.e. about $1 - B_j(t)$ for $t \to \infty$. It is assumed that one of the service time distributions has a regularly varying tail, and the other one has a less heavy tail, or both of the service time distributions have a regularly varying tail with the same index.

Assumption 3.2.1 *We assume that one of the following assumptions holds,*

$$(i) \qquad 1 - B_1(t) \sim -\frac{1}{\Gamma(1-\nu)}(t/\beta_1)^{-\nu} L(t/\beta_1) \text{ as } t \to \infty,$$

$$M_{2\mu} := \int_0^\infty t^\mu \mathrm{d}B_2(t) < \infty \quad \text{for a } \mu > \nu;$$

$$(ii) \qquad 1 - B_2(t) \sim -\frac{1}{\Gamma(1-\nu)}(t/\beta_2)^{-\nu} L(t/\beta_2) \text{ as } t \to \infty,$$

$$M_{1\mu} := \int_0^\infty t^\mu \mathrm{d}B_1(t) < \infty \quad \text{for a } \mu > \nu;$$

$$(iii) \quad 1 - B_j(t) \sim -\frac{1}{\Gamma(1-\nu)}(t/\beta_j)^{-\nu} L_j(t/\beta_j) \text{ as } t \to \infty, \text{ for } j = 1, 2,$$

$$L(t) := L_1(t) \quad \text{for } t \geq 0,$$

$$\alpha := \lim_{t\to\infty} \frac{L_2(t)}{L(t)} < \infty;$$

$$(iv) \quad 1 - B_j(t) \sim -\frac{1}{\Gamma(1-\nu)}(t/\beta_j)^{-\nu} L_j(t/\beta_j) \text{ as } t \to \infty, \text{ for } j = 1, 2,$$

$$L(t) := L_2(t) \text{ for } t \geq 0,$$

$$\lim_{t\to\infty} \frac{L(t)}{L_1(t)} = \infty,$$

where $1 < \nu < 2$ and $L(\cdot)$, $L_1(\cdot)$ and $L_2(\cdot)$ are slowly varying functions.

In order to simplify the notation, we define the function $L(\cdot)$ in (iii) and (iv) which determines the heaviest service time tail behavior. To obtain our heavy-traffic limit theorem, we assume that $L(t)$ is continuous for sufficiently large t. We use the same μ in (i) and (ii), since it does not make any difference if we use different notation. Without loss of generality, we may assume $\nu < \mu < 2$.

The next lemma relates the assumptions on the service time distributions to the corresponding LSTs.

Lemma 3.2.1 *(i) Assumption 3.2.1(i) implies that, as $s \downarrow 0$,*

$$\beta_1(s) = 1 - \beta_1 s + (\beta_1 s)^\nu L(1/\beta_1 s) + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big), \tag{3.2.3}$$

$$\beta_2(s) = 1 - \beta_2 s + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big); \tag{3.2.4}$$

(ii) Assumption 3.2.1(ii) implies that, as $s \downarrow 0$,

$$\beta_1(s) = 1 - \beta_1 s + \mathrm{o}\Big((\beta_1 s)^\mu\Big) \quad \textit{where } \nu < \mu < 2,$$

$$\beta_2(s) = 1 - \beta_2 s + (\beta_2 s)^\nu L(1/\beta_2 s) + \mathrm{o}\Big((\beta_2 s)^\nu L(1/\beta_2 s)\Big);$$

(iii) Assumption 3.2.1(iii) implies that, as $s \downarrow 0$,

$$\beta_1(s) = 1 - \beta_1 s + (\beta_1 s)^\nu L(1/\beta_1 s) + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big),$$

$$\beta_2(s) = 1 - \beta_2 s + \alpha(\beta_2 s)^\nu L(1/\beta_2 s) + \mathrm{o}\Big((\beta_2 s)^\nu L(1/\beta_2 s)\Big);$$

(iv) Assumption 3.2.1(iv) implies that, as $s \downarrow 0$,

$$\beta_1(s) = 1 - \beta_1 s + (\beta_1 s)^\nu L_1(1/\beta_1 s) + \mathrm{o}\Big((\beta_1 s)^\nu L_1(1/\beta_1 s)\Big),$$

$$\beta_2(s) = 1 - \beta_2 s + (\beta_2 s)^\nu L_2(1/\beta_2 s) + \mathrm{o}\Big((\beta_2 s)^\nu L_2(1/\beta_2 s)\Big),$$

where $lim_{t\to\infty} L_1(t)/L_2(t) = 0$.

Proof. We only prove (i), the proof for the rest is similar. Equality (3.2.3) immediately follows from Assumption 3.2.1(i), by using Lemma 2.3.4. Since

$$\int_0^\infty t^\mu \mathrm{d}B_2(t) < \infty,$$

it follows that

$$1 - B_2(t) = \mathrm{o}\left(\left(\frac{t}{\beta_2}\right)^{-\mu}\right).$$

Applying Lemma 2.3.2 to the above equality, we have for $s \downarrow 0$,

$$\beta_2(s) = 1 - \beta_2 s + \mathrm{o}\Big((\beta_2 s)^\mu\Big) \quad \text{where } \mu > \nu.$$

Moreover, it follows from Proposition 1.3.6 (v) in [12] that $L(1/s) = \mathrm{o}(s^{\nu-\mu})$. Hence, (3.2.4) follows. □

3.3 The class-2 waiting time distribution

Denote by $P_1(t)$ the busy period distribution in an $M/G/1$ queue with only class-1 customers and by $\eta_1(s)$ the LST of $P_1(t)$. Let W_2 be a random variable with distribution the steady-state waiting time distribution $W_2(t)$ of class-2 customers, and $\omega_2(s)$ the LST of $W_2(t)$ where $s \geq 0$. In this section we present the explicit expression for $\omega_2(s)$ and its asymptotic properties as $s \downarrow 0$, when one of the assumptions in Section 3.2 is satisfied. From (2.14) in [3] we have

$$\omega_2(s) = \frac{1-\rho}{1-\rho f(s)}, \tag{3.3.1}$$

where

$$f(s) := \frac{\rho_1}{\rho_1+\rho_2} h_0^{(1)}(s) + \frac{\rho_2}{\rho_1+\rho_2} \beta_{2e}(z(s)), \tag{3.3.2}$$

$$h_0^{(1)}(s) := \frac{1-\eta_1(s)}{\beta_1 s + \rho_1 - \rho_1 \eta_1(s)}, \tag{3.3.3}$$

$$z := z(s) = s + \lambda_1 - \lambda_1 \eta_1(s), \tag{3.3.4}$$

for $\beta_{2e}(s)$ in (3.2.2). Note that there are minor differences between the above formula and the formula which was obtained by Abate and Whitt in [3], caused by their choice of $\beta_1 = 1$. Denote by $F_2(t)$ the probability distribution function with LST $f_2(s) := \beta_{2e}(z)$.

As explained in [3], $h_0^{(1)}(s)$ is the LST of the high-priority server-occupancy distribution function $H_0^{(1)}(t)$, which is defined by

$$H_0^{(1)}(t) = (1 - P_{00}^{(1)}(t))/\rho,$$

where $P_{00}^{(1)}(t)$ is the high-priority emptiness probability, i.e., the probability that the system has no class-1 customers at time t given that it had none at time 0. Actually, an expression for $\omega_2(s)$ has been known for a long time, cf. Section III.3.6 of [36], but the representation in (3.3.1) found in [3], which is similar to the Pollaczek-Khintchine form for the ordinary $M/G/1$ waiting time transform, appears to be new and is a suitable starting point for our analysis.

For the sake of simplicity, let us use the convention that $\beta_{j,n}(s)$, $\beta_{je,n}(s)$, $\eta_{j,n}(s)$, $h_{0,n}^{(1)}(s)$ and $f_{2,n}(s)$ stand for the function defined in (2.3.2) with $f(s)$ replaced by $\beta_j(s)$, $\beta_{je}(s)$, $\eta_j(s)$, $h_0^{(1)}(s)$ and $f_2(s)$ respectively.

The following lemma establishes a relation between $\beta_1(s)$, $\beta_{1e}(s)$, $\eta_1(s)$ and $h_0^{(1)}(s)$.

Lemma 3.3.1 *For $n < \nu < n+1$ ($n \in \mathbb{N}$), $C \geq 0$, the following statements are equivalent,*

(i) $\beta_{1,n}(s) = [C + \mathrm{o}(1)](-1)^n \Gamma(1-\nu)(\beta_1 s)^\nu L(1/\beta_1 s)$ *for* $s \downarrow 0$*;*

(ii) $\beta_{1e,n-1}(s) = [C + \mathrm{o}(1)](-1)^{n-1} \Gamma(1-\nu)(\beta_1 s)^{\nu-1} L(1/\beta_1 s)$ *for* $s \downarrow 0$*;*

(iii) $\eta_{1,n}(s) = [C + \mathrm{o}(1)](-1)^n \frac{\Gamma(1-\nu)}{1-\rho_1} \left(\frac{\beta_1 s}{1-\rho_1}\right)^\nu L(1/\beta_1 s)$ *for* $s \downarrow 0$*;*

(iv) $h_{0,n-1}^{(1)}(s) = [C + \mathrm{o}(1)](-1)^{n-1} \Gamma(1-\nu) \left(\frac{\beta_1 s}{1-\rho_1}\right)^{\nu-1} L(1/\beta_1 s)$ *for* $s \downarrow 0$.

Proof. $(i) \Leftrightarrow (ii)$ follows immediately from the fact that $\beta_{1e}(s) = \frac{1-\beta_1(s)}{\beta_1 s}$.
$(i) \Leftrightarrow (iii)$ follows from the main result of De Meyer and Teugels [81] which links the busy period distribution tail and the service time distribution tail in the $M/G/1$ queue.
$(ii) \Leftrightarrow (iv)$. Since $\eta_1(s) = \beta_1(z)$, we have

$$h_0^{(1)}(s) = \frac{1-\eta_1(s)}{\beta_1 s + \rho_1 - \rho_1\eta_1(s)} = \frac{1-\beta_1(z)}{\beta_1 z} = \beta_{1e}(z).$$

Note that s can be represented as $s = z - \lambda_1(1-\beta_1(z))$. Applying Lemma 2.3.6 then yields the result. □

By the above lemma, we may deduce the asymptotic properties of $\eta_1(s)$ and $\beta_{2e}(s+\lambda_1-\lambda_1\eta_1(s))$ for $s \downarrow 0$, which appear in the expression (3.3.1) of $\omega_2(s)$ and completely determine the asymptotic behavior of $\omega_2(s)$ for $s \downarrow 0$.

Lemma 3.3.2 *(i) If Assumption 3.2.1(i) holds, then as $s \downarrow 0$,*

$$\eta_1(s) = 1 - \frac{\beta_1 s}{1-\rho_1} + \frac{(\beta_1 s)^\nu L(1/\beta_1 s)}{(1-\rho_1)^{\nu+1}} + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big), \quad (3.3.5)$$

$$\beta_{2e}(s+\lambda_1-\lambda_1\eta_1(s)) = 1 + \mathrm{o}\Big((\beta_1 s)^{\nu-1} L(1/\beta_1 s)\Big); \quad (3.3.6)$$

(ii) If Assumption 3.2.1(ii) or (iv) holds, then as $s \downarrow 0$,

$$\eta_1(s) = 1 - \tfrac{\beta_1 s}{1-\rho_1} + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big), \quad (3.3.7)$$

$$\beta_{2e}(s+\lambda_1-\lambda_1\eta_1(s)) = 1 - \left(\tfrac{\beta_2 s}{1-\rho_1}\right)^{\nu-1} L(1/\beta_1 s) + \mathrm{o}\Big((\beta_2 s)^{v-1} L(1/\beta_1 s)\Big); \quad (3.3.8)$$

(iii) If Assumption 3.2.1(iii) holds, then as $s \downarrow 0$,

$$\eta_1(s) = 1 - \frac{\beta_1 s}{1-\rho_1} + (\beta_1 s)^\nu L(1/\beta_1 s) + \mathrm{o}\Big((\beta_1 s)^\nu L(1/\beta_1 s)\Big), \quad (3.3.9)$$

$$\beta_{2e}(s+\lambda_1-\lambda_1\eta_1(s)) = 1 - \alpha\left(\frac{\beta_2 s}{1-\rho_1}\right)^{\nu-1} L(1/\beta_1 s) + \mathrm{o}\Big((\beta_2 s)^{\nu-1} L(1/\beta_1 s)\Big). \quad (3.3.10)$$

Proof. We only prove (i). Statements (ii) and (iii) follow by similar reasoning. Since (i) in (3.2.1) holds, it follows from the main theorem in [81] that

$$1 - P_1(t) \sim -\frac{1}{\Gamma(1-\nu)(1-\rho_1)^{1+\nu}}(t/\beta_1)^{-\nu} L(t/\beta_1), \quad t \to \infty, \quad (3.3.11)$$

where $P_1(t)$ is the class-1 busy-period distribution function. Using Lemma 2.3.4, (3.3.11) leads to (3.3.5) immediately. From (3.3.4) and (3.3.5), we have for $s \downarrow 0$,

$$\frac{z(s)}{s} = \frac{1}{1-\rho_1} + \mathrm{o}\Big((\beta_1 s)^{\nu-1} L(1/\beta_1 s)\Big).$$

By Lemma 3.2.1(i) and Lemma 3.3.1, we have

$$\beta_{2e}(s) = 1 - \mathrm{o}\Big((\beta_2 s)^{\nu-1} L(1/\beta_2 s)\Big), \quad s \downarrow 0. \tag{3.3.12}$$

We write

$$\frac{1-\beta_{2e}(z(s))}{(\beta_2 s)^{\nu-1} L(1/\beta_2 s)} = \frac{1-\beta_{2e}(z(s))}{(z(s))^{\nu-1} L(1/z(s))} \frac{(z(s))^{\nu-1} L(1/z(s))}{(\beta_2 s)^{\nu-1} L(1/\beta_2 s)}. \tag{3.3.13}$$

Taking the limit of the above equality for $s \downarrow 0$, and applying Lemma 2.3.5 and (3.3.12), we get

$$\lim_{s\downarrow 0} \frac{1-\beta_{2e}(z(s))}{(\beta_2 s)^{\nu-1} L(1/\beta_2 s)} = 0.$$

□

To obtain our heavy-traffic limit theorem, we rewrite $\omega_2(s)$ into the following form.

Lemma 3.3.3 *Let $\beta_{2e}(s)$, $h_0^{(1)}(s)$ and $z(s)$ be given by (3.2.2), (3.3.3) and (3.3.4) respectively.*
(i) Assumption 3.2.1(i) implies that $\omega_2(s)$ can be written as

$$\omega_2(s) = \left(1 + \frac{\rho_1 (\beta_1 s)^{\nu-1} L(1/\beta_1 s)}{(1-\rho)(1-\rho_1)^{\nu-1}} + \frac{H_1(s)}{1-\rho}\right)^{-1}, \tag{3.3.14}$$

with

$$H_1(s) = \rho_1[1 - h_0^{(1)}(s)] + \rho_2[1 - \beta_{2e}(z(s))] - \frac{\rho_1 (\beta_1 s)^{\nu-1} L(1/\beta_1 s)}{(1-\rho_1)^{\nu-1}}, \tag{3.3.15}$$

where

$$\lim_{s\downarrow 0} \frac{H_1(s)}{(\beta_1 s)^{\nu-1} L(1/\beta_1 s)} = 0. \tag{3.3.16}$$

(ii) Assumption 3.2.1(ii) or (iv) implies that $\omega_2(s)$ can be written as

$$\omega_2(s) = \left(1 + \frac{\rho_2 (\beta_2 s)^{\nu-1} L(1/\beta_2 s)}{(1-\rho)(1-\rho_1)^{\nu-1}} + \frac{H_2(s)}{1-\rho}\right)^{-1}, \tag{3.3.17}$$

where

$$H_2(s) = \rho_1[1 - h_0^{(1)}(s)] + \rho_2[1 - \beta_{2e}(z(s))] - \frac{\rho_2(\beta_2 s)^{\nu-1}L(1/\beta_2 s)}{(1-\rho_1)^{\nu-1}} \tag{3.3.18}$$

satisfies (3.3.16) and with $H_1(s)$ being replaced by $H_2(s)$.
(iii) Assumption 3.2.1(iii) implies that $\omega_2(s)$ can be written as

$$\omega_2(s) = \left(1 + \frac{\rho_1(\beta_1 s)^{\nu-1}L(1/\beta_1 s)}{(1-\rho)(1-\rho_1)^{\nu-1}} + \frac{\alpha\rho_2(\beta_2 s)^{\nu-1}L(1/\beta_1 s)}{(1-\rho)(1-\rho_1)^{\nu-1}} + \frac{H_3(s)}{1-\rho}\right)^{-1}, \tag{3.3.19}$$

where

$$\begin{aligned} H_3(s) &= \rho_1[1 - h_0^{(1)}(s)] + \rho_2[1 - \beta_{2e}(z(s))] \\ &\quad - \frac{\rho_1(\beta_1 s)^{\nu-1}L(1/\beta_1 s)}{(1-\rho_1)^{\nu-1}} - \frac{\alpha\rho_2(\beta_2 s)^{\nu-1}L(1/\beta_1 s)}{(1-\rho_1)^{\nu-1}} \end{aligned} \tag{3.3.20}$$

satisfies (3.3.16) and with $H_1(s)$ being replaced by $H_3(s)$.

Proof. We only prove (i). In a similar way, by using Lemma 3.3.2, we can show (ii) and (iii). By Lemma 3.3.2(i), Equalities (3.3.5) and (3.3.6) follow. Substituting (3.3.5) into (3.3.3), we get, as $s \downarrow 0$,

$$1 - h_0^{(1)}(s) = \frac{(\beta_1 s)^{\nu-1}L(1/\beta_1 s)}{(1-\rho_1)^{\nu-1}} + \mathrm{o}\Big((\beta_1 s)^{\nu-1}L(1/\beta_1 s)\Big). \tag{3.3.21}$$

Rewrite (3.3.1) as

$$\omega_2(s) = \left(1 + \frac{\rho_1}{1-\rho}[1 - h_0^{(1)}(s)] + \frac{\rho_2}{1-\rho}[1 - \beta_{2e}(z(s))]\right)^{-1}. \tag{3.3.22}$$

Replacing $H_1(s)$ in (3.3.14) with the right-hand side of (3.3.15) gives (3.3.22). Dividing $H_1(s)$ by $s^{\nu-1}L(1/s)$, substituting (3.3.21) and (3.3.6) into (3.3.15), and taking the limit for $s \downarrow 0$, we obtain (3.3.16). □

Note that in Equalities (3.3.14), (3.3.17) and (3.3.19), the factor $1/(1-\rho_1)$ does not occur in the function $H_j(s)$ for $j = 1, 2, 3$; this plays a key role in proving our heavy-traffic limit theorem. Actually, applying Lemmas 2.3.4, it is easy to derive the tail behavior of class-2 waiting time distribution from Lemma 3.3.3.

Theorem 3.3.1 *If Assumption 3.2.1 holds, then the stationary class-2 waiting time distribution $W_2(t)$ is regularly varying of index $1-\nu$, $1<\nu<2$, i.e.,*

$$1 - W_2(t) \sim M t^{1-\nu} L(t), \quad t \to \infty,$$

where M can be determined by (3.3.14), (3.3.17) or (3.3.22). E.g., if Assumption 3.2.1(i) holds, then

$$1 - W_2(t) \sim -\frac{\rho_1 (t/\beta_1)^{1-\nu} L(t/\beta_1)}{\Gamma(1-\nu)(1-\rho_1)^{\nu-1}(1-\rho)}, \quad t \to \infty.$$

The above theorem coincides with Corollary 4.3.1 in the next chapter. The intuitive arguments provided in Section 4.3 are also applicable to the priority queue, since the priority queue is a special case of the two-queue E/1−L model which we consider in Chapter 4.

Remark 3.3.1 One can prove similar statements as in Theorem 3.3.1 for the case $\nu \geq 2$. In fact, Theorem 9.3 in [3] provides similar results for the case of a regularly varying service time distribution of the class-2 customers. But the condition we require in our theorem is weaker than that in Theorem 9.3. There it is assumed that $B_2(\cdot)$ is regularly varying with index $\nu \geq 2$, and for the LST of $B_1(t)$, there exists an $s^* > 0$ such that $\beta_1(-s^*) = \infty$, and $\beta_1(s) < \infty$ for $s > -s^*$, i.e., the tail behavior of the high-priority class is less heavy than that of some negative exponential distribution.

3.4 Links between the service time and the waiting time

As we have proved, if one of the service time distributions has a regularly varying tail with index $-\nu$, $\nu > 1$, and the other one has a less heavy tail, then the stationary class-2 waiting time distribution $W_2(t)$ has a regularly varying tail with index $1-\nu$. Conversely, if $W_2(t)$ has a regularly varying tail with index $1-\nu$ where $\nu > 1$, and the class-1 service time distribution $B_1(t)$ has a "less heavy tail than $t^{-\nu}$", then the class-2 service time has a regularly varying tail with index $-\nu$. We shall prove this in Theorem 3.4.1. If the tail of the class-1 service time distribution is heavier, then the problem becomes harder. We will not deal with this problem in this thesis.

We introduce the inverse function of $z(s)$ defined by (3.3.4):

$$s(z) := z - \lambda_1 + \lambda_1 \beta_1(z). \tag{3.4.1}$$

This inverse function is unique for $s \in \mathbb{R}$ since $s + \lambda_1 - \lambda_1\eta_1(s)$ is increasing for $s \geq 0$.

In the following lemma, $g_1(\cdot)$ and $g_2(\cdot)$ are some arbitrary functions which later will be given a specific meaning.

Lemma 3.4.1 *Assume the nth moment of B_1 exists and $g_1(s) \equiv g_2(z)$ where z is defined by (3.3.4). Then*

(i) for $k = 1, \cdots, n-1$, the kth derivative of $g_1(\cdot)$ at point 0 exists if and only if the kth derivative of $g_2(\cdot)$ at point 0 exists.

(ii) if the kth derivative of $g_1(\cdot)$ at point 0 exists or the kth derivative of $g_2(\cdot)$ at point 0 exists, then there exist polynomials U_k and $P_{k,m}$ $(m = 1, \cdots, k)$ in s such that

$$s^{-(k+1)}(g_{1,k}(s) - g_{2,k}(z)) = U_k(s) + s^{-(k+1)} \sum_{m=1}^{k} (\eta_{1,k+1}(s))^m P_{k,m}(s). \tag{3.4.2}$$

Moreover, if $\eta_{1,k+1}(s) = \mathrm{o}(s^\mu)$ where $k+1 < \mu < k+2$, then

$$s^{-(k+1)}(g_{1,k}(s) - g_{2,k}(z)) = U_k(0) + \mathrm{o}(s^{\mu-k-1}) \quad \textit{for } s \downarrow 0. \tag{3.4.3}$$

Proof. First we prove (i). Assume that $g_1(\cdot)$ has a kth derivative at point 0. Hence, there exists a polynomial $\sum_{j=0}^{k} g_{1j}s^j$ such that

$$g_1(s) = \sum_{j=0}^{k} g_{1j}s^j + \mathrm{o}(s^k) \quad \text{for } s \downarrow 0. \tag{3.4.4}$$

We may write

$$s(z) = \sum_{j=1}^{k+1} \alpha_j z^j + (-1)^{(k+1)}\rho_1\beta_{1,k+1}(z),$$

which follows from (3.4.1) and the fact that B_1 has finite $(k+1)$th moment. In (3.4.4) replace $g_1(s)$ by $g_2(z)$ and s on the right-hand side by the right-hand side of the above equation, and rearrange it to obtain

$$g_2(z) = \sum_{j=0}^{k} g_{2j}z^j + \mathrm{o}(z^k),$$

which implies the result by using Lemma 2.3.2. The proof for the reverse direction is similar, by writing

$$z(s) = \sum_{j=0}^{k+1} c_j s^j + (-1)^{k+1} \rho_1 \eta_{1,k+1}(s), \tag{3.4.5}$$

where $\eta_{1,k+1}(s)$ is such that

$$\lim_{s \downarrow 0} s^{-(k+2)}((1-\rho_1)\eta_{1,k+1}(s) - \beta_{1,k+1}(z)) = 0,$$

which follows from Corollary 1 in [81]. We omit the proof.
Next we prove (ii). Since both $g_1(\cdot)$ and $g_2(\cdot)$ have a kth derivative at 0, we may write

$$g_{1,k}(s) - g_{2,k}(z) = (-1)^k \left(\sum_{j=0}^{k} g_{1j} s^j - \sum_{j=0}^{k} g_{2j} z^j \right).$$

Replace z in the above equation by (3.4.5) and rearrange slightly to obtain

$$g_{1,k}(s) - g_{2,k}(z) = Q_k(s) + \sum_{i=1}^{k} (\eta_{1,k+1}(s))^i P_{i,m}(s). \tag{3.4.6}$$

It follows from (2.3.2) that

$$\lim_{s \downarrow 0} s^{-k}(g_{1,k}(s) - g_{2,k}(z)) = 0. \tag{3.4.7}$$

Since $\lim_{s\downarrow 0} s^{-k}\eta_{1,k}(s) = 0$, it follows from (3.4.6) and (3.4.7) that $\lim_{s\downarrow 0} s^{-k} Q_k(s) = 0$, which implies that $U_k(s) = s^{-(k+1)} Q_k(s)$ is a polynomial. Multiplying (3.4.6) by $s^{-(k+1)}$ gives the result. From (3.4.2) we can derive (3.4.3) directly. □

The next theorem establishes a relation between the asymptotic behavior of the service time distributions and the class-2 waiting time distribution.

Theorem 3.4.1 *If $W_2(t)$ has a regularly varying tail with index $1-\nu$, specifically,*

$$1 - W_2(t) \sim \frac{\rho_2 (t/\beta_2)^{1-\nu} L(t/\beta_2)}{(\nu-1)(1-\rho_1)^{\nu-1}(1-\rho)} \quad \textit{for } t \to \infty, \tag{3.4.8}$$

where $\nu > 1$ and $L(t)$ is a slowly varying function, and $B_1(t)$ has a less heavy tail than $t^{-\nu}$, then

$$1 - B_2(t) \sim (t/\beta_2)^{-\nu} L(t/\beta_2) \quad \textit{for } t \to \infty. \tag{3.4.9}$$

Proof. Let $F(t)$ be the distribution function with LST $f(s)$ which is defined in (3.3.2). Obviously, (3.3.2) implies that

$$F(t) = \frac{\rho_1}{\rho_1+\rho_2} H_0^{(1)}(t) + \frac{\rho_2}{\rho_1+\rho_2} F_2(t), \tag{3.4.10}$$

where $F_2(t)$ is the distribution function with LST $f_2(s) = \beta_{2e}(z)$, as introduced in the first paragraph of Section 3.3. Applying Theorem 1 in [34], we obtain that (3.4.8) implies that

$$1 - F(t) \sim \frac{(\nu-1)\rho_2 (t/\beta_2)^{1-\nu} L(t/\beta_2)}{(\rho_1+\rho_2)(1-\rho_1)^{\nu-1}} \quad \text{for } t \to \infty. \tag{3.4.11}$$

Since $B_1(t)$ has a less heavy tail than $t^{-\nu}$, it follows that there exists a $\mu > \nu$ such that

$$\int_0^\infty t^\mu \mathrm{d}B_1(t) < \infty.$$

The above relation implies that

$$1 - B_1(t) = \mathrm{o}\Big((t/\beta_1)^{-\mu}\Big) \quad \text{for } t \to \infty. \tag{3.4.12}$$

Applying Lemmas 2.3.4 and 3.3.1 it follows from (3.4.12) that

$$1 - H_0^{(1)}(t) = \mathrm{o}\Big((t/\beta_1)^{-\mu}\Big),$$

which in combination with (3.4.10) and (3.4.11) yields that

$$1 - F_2(t) \sim \frac{(\nu-1)(t/\beta_2)^{1-\nu} L(t/\beta_2)}{(1-\rho_1)^{\nu-1}} \quad \text{for } t \to \infty. \tag{3.4.13}$$

We shall show that (3.4.9) holds, first for noninteger ν and subsequently for integer ν.

(i) ν is not an integer.
Hence, there exists an integer n such that $n < \nu < n+1$ where $n \geq 1$. Without loss of generality, we may assume that $\nu < \mu < n+1$. By Lemma 2.3.4 it follows from (3.4.12) that for $s \downarrow 0$,

$$\eta_{1,n}(s) = \mathrm{o}\Big((\beta_1 s)^\mu\Big). \tag{3.4.14}$$

We shall show that

$$\beta_{2e,n-1}(s) = (-1)^{n-1}\Gamma(1-\nu)\frac{(\beta_2 s)^{\nu-1} L(1/\beta_2 s)}{(1-\rho_1)^{\nu-1}} + \mathrm{o}\left((\beta_2 s)^{\nu-1} L(1/\beta_2 s)\right). \tag{3.4.15}$$

Since $f_2(s)$ is the LST of $F_2(t)$, by (3.4.13) and applying Lemma 2.3.4, we obtain

$$f_{2,n-1}(s) \sim (-1)^{n-1}\Gamma(1-\nu)\frac{(\beta_2 s)^{\nu-1}L(1/\beta_2 s)}{(1-\rho_1)^{\nu-1}}. \tag{3.4.16}$$

Because $f_2(s) = \beta_{2e}(z)$, applying Lemma 3.4.1 leads to (3.4.3) with $g_{1,n-1}(s)$ and $g_{2,n-1}(z)$ replaced by $f_{2,n-1}(s)$ and $\beta_{2e,n-1}(z)$, i.e.,

$$f_{2,n-1}(s) = \beta_{2e,n-1}(z) + \mathrm{O}\Big((\beta_2 s)^n\Big).$$

Dividing by $(\beta_2 z)^{\nu-1}L(1/\beta_2 z)$ on both sides of the above equation and noting that

$$\lim_{z\downarrow 0}\frac{s^{\nu-1}L(1/s)}{z^{\nu-1}L(1/z)} = (1-\rho_1)^{\nu-1},$$

it follows from (3.4.16) that for $z \downarrow 0$,

$$\beta_{2e,n-1}(z) \sim (-1)^{n-1}\Gamma(1-\nu)(\beta_2 z)^{\nu-1}L(1/\beta_2 z),$$

which implies that (3.4.9) holds, by the equivalence of (i) and (ii) in Lemma 3.3.1.

(ii) ν is an integer, $\nu = 2, 3, \cdots$.
First, we consider the case that $\nu \geq 3$. Recall that $\hat{g}_n(s)$ denotes $s^{-(n+1)}g_n(s)$. As proved, $1-F_2(t) \sim (\nu-1)(t/\beta_2)^{1-\nu}L(t/\beta_2)/(1-\rho_1)^{\nu-1}$ where $\nu \in \{3, 4, \cdots\}$, or equivalently, by De Haan's Theorem (cf. Theorem 3.7.3 in [12]) for $x > 1$,

$$\lim_{s\downarrow 0}[a(s)]^{-1}(\hat{f}_{2,\nu-2}(s) - \hat{f}_{2,\nu-2}(xs)) = \log x, \tag{3.4.17}$$

where we can take $a(s) = L(1/\beta_2 s)/((1-\rho_1)^{\nu-2}(\nu-2)!)$. To prove that (3.4.9) holds, it is sufficient to show that

$$\lim_{s\downarrow 0}[a(s)]^{-1}[\hat{f}_{2,\nu-2}(s) - (1-\rho_1)^{1-\nu}\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1}) - \theta_{\nu-2}] = 0, \tag{3.4.18}$$

for some constant $\theta_{\nu-2}$. If (3.4.18) holds, then we may write for $x > 1$,

$$\lim_{s\downarrow 0}[a(s)]^{-1}[\hat{f}_{2,\nu-2}(xs) - (1-\rho_1)^{1-\nu}\hat{\beta}_{2e,\nu-2}(\frac{xs}{1-\rho_1}) - \theta_{\nu-2}] = 0. \tag{3.4.19}$$

Subtracting (3.4.18) by (3.4.19) and using (3.4.17) yields

$$\lim_{s\downarrow 0}[a(s)(1-\rho_1)^{\nu-1}]^{-1}[\beta_{2e,\nu-2}(s) - \beta_{2e,\nu-2}(xs)] = \log x.$$

Applying the reverse statement of De Haan's Theorem (cf. Theorem 3.7.3 in [12]) to the above relation leads to (3.4.9).

To prove (3.4.18) we use the expression (3.4.2) for $k = \nu - 2$; the right-hand side of (3.4.2) will be abbreviated by $A_{\nu-2}(s)$ with $g_1(s)$ replaced by $f_2(s)$ and $g_2(z)$ replaced by $\beta_{2e}(z)$. Hence,

$$\hat{f}_{2,\nu-2}(s) = \hat{\beta}_{2e,\nu-2}(z)(z/s)^{\nu-1} + A_{\nu-2}(s). \tag{3.4.20}$$

This suggests that we might take $\theta_{\nu-2} = A_{\nu-2}(0)$. So define

$$J_{\nu-2}(s) = [a(s)]^{-1}[\hat{f}_{2,\nu-2}(s) - (1-\rho_1)^{1-\nu}\hat{\beta}_{2e,\nu-2}(s/(1-\rho_1)) - A^{\nu-2}(0)]. \tag{3.4.21}$$

We shall show that $\lim_{s\downarrow 0} J_{\nu-2}(s) = 0$ in the same way as De Meyer and Teugels [81], p. 810-811. Since $\eta_{1,0}(s)$ is decreasing by Lemma 2.3.3(i), we have

$$\hat{\mu}_{1,0}(s) \leq \lim_{s\downarrow 0} \frac{1-\eta_1(s)}{s} = \frac{\beta_1}{1-\rho_1},$$

so that

$$z = s[1 + \lambda\hat{\mu}_{1,0}(s)] \leq \frac{s}{1-\rho_1}.$$

Again, by Lemma 2.3.3, it follows that $\hat{\beta}_{2e,\nu-2}(s)$ is decreasing and $s\hat{\beta}_{2e,\nu-2}(s)$ is increasing, therefore

$$\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1}) \leq \hat{\beta}_{2e,\nu-2}(z) \leq \frac{s}{(1-\rho_1)z}\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1}). \tag{3.4.22}$$

By using (3.4.20) in (3.4.21) and subsequently applying the above relation, we have

$$\begin{aligned} J_{\nu-2}(s) &= [a(s)]^{-1}(\hat{\beta}_{2e,\nu-2}(z)(z/s)^{1-\nu} - (1-\rho_1)^{1-\nu}\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1}) \\ &\quad +[A_{\nu-2}(s) - A_{\nu-2}(0)]) \\ &\geq [a(s)]^{-1}(\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1})[(z/s)^{\nu-1} - (1-\rho_1)^{1-\nu}] \\ &\quad +[A_{\nu-2}(s) - A_{\nu-2}(0)]), \end{aligned} \tag{3.4.23}$$

and

$$\begin{aligned} J_{\nu-2}(s) &\leq [a(s)]^{-1}((1-\rho_1)^{-1}\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1})[(z/s)^{\nu-2} - (1-\rho_1)^{2-\nu}] \\ &\quad +[A_{\nu-2}(s) - A_{\nu-2}(0)]). \end{aligned} \tag{3.4.24}$$

By (3.3.4) it follows that

$$\frac{z(s)}{s} = \frac{1}{1-\rho_1} + (\beta_1 s)\mathrm{O}(1),$$

thus

$$(z/s)^{\nu-1} - (1-\rho_1)^{1-\nu} = (\nu-1)(\beta_1 s)\mathrm{O}(1). \tag{3.4.25}$$

Moreover, it follows from the definition of $A_{\nu-2}(s)$ and (3.4.3) that

$$\lim_{s\downarrow 0}[a(s)]^{-1}[A_{\nu-2}(s) - A_{\nu-2}(0)] = 0. \tag{3.4.26}$$

Multiplying (3.4.17) by s, it follows that

$$\lim_{s\downarrow 0}[a(s)]^{-1}[s\hat{f}_{2,\nu-2}(s) - s\hat{f}_{2,\nu-2}(xs)] = 0,$$

while on the other hand,

$$\lim_{s\downarrow 0}[a(s)]^{-1}[s\hat{f}_{2,\nu-2}(s) - s\hat{f}_{2,\nu-2}(xs)] = (1-1/x)\lim_{s\downarrow 0}[a(s)]^{-1}s\hat{f}_{2,\nu-2}(s).$$

The above two relations imply that $\lim_{s\downarrow 0}[a(s)]^{-1}s\hat{f}_{2,\nu-2}(s) = 0$. Consequently, it follows from (3.4.20) that $\lim_{s\downarrow 0}[a(s)]^{-1}s\beta_{2e,\nu-2}(z) = 0$, or equivalently,

$$\lim_{z\downarrow 0}[a(z)]^{-1}z\beta_{2e,\nu-2}(z) = 0. \tag{3.4.27}$$

Combining (3.4.25), (3.4.26) and (3.4.27), we have

$$\begin{aligned}
&\lim_{s\downarrow 0}[a(s)]^{-1}(\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1})[(z/s)^{\nu-1} - (1-\rho_1)^{1-\nu}] \\
&+[A_{\nu-2}(s) - A_{\nu-2}(0)]) = 0, \qquad (3.4.28)\\
&\lim_{s\downarrow 0}[a(s)]^{-1}((1-\rho_1)^{-1}\hat{\beta}_{2e,\nu-2}(\frac{s}{1-\rho_1})[(z/s)^{\nu-2} - (1-\rho_1)^{2-\nu}] \\
&+[A_{\nu-2}(s) - A_{\nu-2}(0)]) = 0. \qquad (3.4.29)
\end{aligned}$$

Therefore, combining (3.4.23), (3.4.24), (3.4.28) and (3.4.29) yields that

$$\lim_{s\downarrow 0} J_{\nu-2}(s) = 0.$$

Secondly, we consider the case that $\nu = 2$. Again we intend to show that (3.4.18) holds by taking $\theta_{\nu-2} = 0$. We define

$$J_0(s) := [a(s)]^{-1}[\hat{f}_{2,0}(s) - \frac{1}{1-\rho_1}\hat{\beta}_{2e,0}(\frac{s}{1-\rho_1})].$$

By the fact that $\hat{f}_{2,0}(s) = \hat{\beta}_{2e,0}(z)(z/s)$ and (3.4.22), it follows that

$$[a(s)]^{-1}(\hat{\beta}_{2e,0}(\frac{s}{1-\rho_1})(\frac{z}{s} - \frac{1}{1-\rho_1})) \leq J_0(s) \leq 0.$$

It follows from (3.4.25) and (3.4.27) that

$$\lim_{s\downarrow 0}[a(s)]^{-1}(\hat{\beta}_{2e,0}(\frac{s}{1-\rho_1})(\frac{z}{s} - \frac{1}{1-\rho_1})) = 0.$$

The above two relations lead to $\lim_{s\downarrow 0} J_0(s) = 0$, which implies that (3.4.18) is satisfied for $\nu = 2$ and thus (3.4.9) follows. □

3.5 The $M/G/1$ queue with two priority classes

In [22] Boxma and Cohen obtained heavy-traffic limit theorems for the $G/G/1$ queue (cf. Section 1.4). In this section we apply a similar method as in [22] to derive a heavy-traffic limit theorem for the low-priority waiting time in the queueing model with two types of customers. We assume that $\rho_2 \uparrow 1-\rho_1,\ 0 < \rho_1 < 1$ and that Assumption 3.2.1 is satisfied.

Consider the contraction equation

$$\frac{Kx^{\nu-1}L(1/x)}{1-\rho} = 1, \quad x > 0, \tag{3.5.1}$$

where K is a function of both ρ_1 and ρ_2 such that $K > c$ for some positive constant c, $L(x)$ is a slowly varying function, and denote by $\Delta(\rho_2)$ the unique root of (3.5.1) such that

$$\Delta(\rho_2) \downarrow 0 \quad \text{for } \rho_2 \uparrow 1-\rho_1, \tag{3.5.2}$$

cf. [22].

We say that the solution $\Delta(\rho)$ to the contraction equation is the unique solution with the property that $\Delta(\rho) \downarrow 0$ for $\rho \uparrow 1$, if for two solutions to the contraction equation $\Delta_j(\rho)$ $(j = 1, 2)$ such that $\Delta_j(\rho) \downarrow 0$ for $\rho \uparrow 1$, the limit of the ratio of the two solutions for $\rho \uparrow 1$ is equal to 1, i.e.,

$$\lim_{\rho\uparrow 1} \frac{\Delta_1(\rho)}{\Delta_2(\rho)} = 1.$$

In the following we provide a lemma which characterizes a property of the solution to the contraction equation (3.5.1).

Lemma 3.5.1 *If $L(t)$ is continuous, then there exists a unique solution $\Delta(\rho_2)$ to the contraction equation (3.5.1) with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_1$.*

Proof. Since

$$\lim_{s\downarrow 0} s^{\nu-1}L(1/s) = 0,$$

by the continuity of $L(1/s)$, it follows that, for $\rho_2 < 1-\rho_1$, there exists at least one solution $\xi(\rho_2)$ to the equation

$$Kx^{\nu-1}L(1/x) = 1-\rho. \tag{3.5.3}$$

Put

$$\Delta(\rho_2) = \inf\{\xi(\rho_2) : K\xi(\rho_2)^{\nu-1}L(1/\xi(\rho_2)) = 1-\rho\}.$$

By continuity of $L(1/x)$, $\Delta(\rho_2)$ is also a solution to Equation (3.5.3). Next we show that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_1$. Assume, to the contrary, that there exists a sequence (ρ_{2n}) $(n = 1, 2, ...)$ which tends to $1-\rho_1$ such that, for all n, $\Delta(\rho_{2n}) > \epsilon$ for some positive constant ϵ. If n is large enough, then $1-\rho_1-\rho_{2n}$ is arbitrarily small. Thus for sufficiently large n, there exists at least one solution $\xi(\rho_{2n})$ to Equation (3.5.3) such that $\xi(\rho_{2n}) < \epsilon$. On the other hand, by the definition of $\Delta(\rho_2)$,

$$\xi(\rho_{2n}) \geq \Delta(\rho_{2n}) > \epsilon,$$

which contradicts the fact that $\xi(\rho_{2n}) < \epsilon$. Hence,

$$\lim_{\rho_2\uparrow 1-\rho_1} \Delta(\rho_2) = 0.$$

Now we shall prove the uniqueness of the solution $\Delta(\rho_2)$ to Equation (3.5.3) with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_1$. Let $\Delta_j(\rho_2)$ be two solutions to Equation (3.5.3) with the property that $\Delta_j(\rho_2) \to 0$ for $\rho_2 \to 1-\rho_1$, $j = 1, 2$. It is sufficient to show that if

$$\lim_{n\to\infty} \rho_{2n} = 1-\rho_1,$$

$$\lim_{n\to\infty} \Delta_j(\rho_{2n}) = 0, \quad \text{for } j = 1, 2,$$

$$a = \lim_{n\to\infty} \frac{\Delta_1(\rho_{2n})}{\Delta_2(\rho_{2n})} \quad \text{where } 0 \leq a \leq \infty,$$

then $a = 1$. Since $\Delta_j(\rho_2)$ $(j = 1, 2)$ are solutions to Equation (3.5.3), we can write

$$K\Delta_j(\rho_2)^{\nu-1}L(1/\Delta_j(\rho_2)) = 1-\rho.$$

It follows that

$$\frac{\Delta_1(\rho_2)^{\nu-1}L(1/\Delta_1(\rho_2))}{\Delta_2(\rho_2)^{\nu-1}L(1/\Delta_2(\rho_2))} = 1. \tag{3.5.4}$$

If $0 < a < \infty$, then it follows from Lemma 2.3.5 that

$$\lim_{\rho_2 \to 1-\rho_1} \frac{\Delta_1(\rho_2)^{\nu-1}L(1/\Delta_1(\rho_2))}{\Delta_2(\rho_2)^{\nu-1}L(1/\Delta_2(\rho_2))} = a^{\nu-1}.$$

Combining the above equality and (3.5.4) gives $a = 1$. Next we prove that $a \neq 0$. Set

$$b_n := \frac{\Delta_1(\rho_{2n})}{\Delta_2(\rho_{2n})}.$$

Assume $\lim_{n\to\infty} b_n = 0$. Choose ϵ positive and small enough. Applying Theorem 1.5.2 in [12], we have

$$\frac{\Delta_1(\rho_{2n})^{\nu-1}L(1/b_n\Delta_2(\rho_{2n}))}{\Delta_2(\rho_{2n})^{\nu-1}L(1/\Delta_2(\rho_{2n}))} < b_n^{\nu-1} + \epsilon$$

for sufficiently large n, which contradicts (3.5.4). Hence, $a \neq 0$. Similarly,

$$\lim_{n\to\infty} \frac{\Delta_2(\rho_{2n})}{\Delta_1(\rho_{2n})} \neq 0.$$

Consequently, $a \neq \infty$. □

If explicit representations (as in [22], cf. Remark 3.5.2 below) for the service time distributions are given, i.e., the LSTs $\beta_j(s)$ $(j = 1, 2)$ can be represented by (3.5.15), then one can prove that there is a unique root with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1 - \rho_1$.

Lemma 3.5.2 *If Assumption 3.2.1 is satisfied, then there exists a contraction coefficient $\Delta(\rho_2)$ satisfying (3.5.1) such that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1 - \rho_1$, and*

$$\lim_{\rho_2 \uparrow 1-\rho_1} \omega_2(\Delta(\rho_2)s/\beta_1) = \frac{1}{1 + s^{\nu-1}}. \tag{3.5.5}$$

Proof. We prove the case in which Assumption 3.2.1(i) holds. Let $\Delta(\rho_2)$ be the root of the contraction equation

$$\frac{\rho_1 x^{\nu-1}L(1/x)}{(1-\rho)(1-\rho_1)^{\nu-1}} = 1 \tag{3.5.6}$$

for which (3.5.2) holds. It follows from (3.5.2) and (3.5.6) that, for $s \geq 0$,

$$\lim_{\rho_2 \uparrow 1-\rho_1} \frac{\rho_1 (s\Delta(\rho_2))^{\nu-1} L(1/\Delta(\rho_2)s}{(1-\rho)(1-\rho_1)^{\nu-1}} = s^{\nu-1}. \tag{3.5.7}$$

By Lemma 3.3.2(i) we have

$$\omega_2(s) = \left(1 + \frac{\rho_1 (\beta_1 s)^{\nu-1} L(1/\beta_1 s)}{(1-\rho)(1-\rho_1)^{\nu-1}} + \frac{H_1(s)}{1-\rho}\right)^{-1}, \tag{3.5.8}$$

where $H_1(s)$ is such that

$$\lim_{s \downarrow 0} \frac{H_1(s)}{(\beta_1 s)^{\nu-1} L(1/\beta_1 s)} = 0.$$

By (3.5.7) and the above relation,

$$\lim_{\rho_2 \uparrow 1-\rho_1} \frac{H_1(\Delta(\rho_2)s/\beta_1)}{1-\rho} = 0, \tag{3.5.9}$$

where $H_1(s)$ is given by (3.3.15). Substituting $\Delta(\rho_2)s/\beta_1$ in $\omega_2(s)$ and taking the limit for $\rho_2 \uparrow 1-\rho_1$ yields (3.5.5). □

The analysis given above leads to the following theorem.

Theorem 3.5.1 *For the stable $M/G/1$ queue with two priority classes, the service time distributions $B_1(t)$ and $B_2(t)$ satisfying Assumption 3.2.1, the contracted waiting time $\Delta(\rho_2)W_2/\beta_1$ converges in distribution for $\rho_2 \uparrow 1-\rho_1$, and the limit distribution is given by: for $t \geq 0$,*

$$R_{\nu-1}(t) = 1 - \sum_{n=0}^{\infty} (-1)^n \frac{t^{n(\nu-1)}}{\Gamma(n(\nu-1)+1)}. \tag{3.5.10}$$

The coefficient of contraction $\Delta(\rho_2)$ is that root of Equation (3.5.1) with the property that $\Delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_1$, and with $K = K_1, ..., K_4$ corresponding to Assumption 3.2.1(i),...,(iv) respectively, where $K_1 = \frac{\rho_1}{(1-\rho_1)^{\nu-1}}$, $K_2 = \frac{\rho_2(\beta_2/\beta_1)^{\nu-1}}{(1-\rho_1)^{\nu-1}}$, $K_3 = \frac{\rho_1+\rho_2\alpha(\beta_2/\beta_1)^{\nu-1}}{(1-\rho_1)^{\nu-1}}$ and $K_4 = K_2$. Moreover, the LST of $R_{\nu-1}(t)$ is

$$\int_{0-}^{\infty} e^{-st} \mathrm{d}R_{\nu-1}(t) = \frac{1}{1+s^{\nu-1}}, \quad s \geq 0. \tag{3.5.11}$$

Proof. By Lemma 3.5.2, the LST of the distribution of the contracted waiting time $\Delta(\rho_2)W_2/\beta_1$ converges, for $\rho_2 \uparrow 1-\rho_1$, and the limit function is given by

$$\lim_{\rho_2\uparrow 1-\rho_1} \mathrm{E}[e^{-\Delta(\rho_2)W_2 s/\beta_1}] = \lim_{\rho_2\uparrow 1-\rho_1} \omega_2(\Delta(\rho_2)s/\beta_1) = \frac{1}{1+s^{\nu-1}}. \tag{3.5.12}$$

Since $1/(1+s^{\nu-1}) \uparrow 1$ for $s \downarrow 0$, using the convergence theorem of Feller for LSTs, cf. [52], it follows that there exists a proper distribution function $R_{\nu-1}(t)$ which has LST $1/(1+s^{\nu-1})$. Relation (3.5.12) implies that the distribution of the contracted waiting time of $\Delta(\rho_2)W_2/\beta_1$ converges to $R_{\nu-1}(t)$ for $\rho_2 \uparrow 1-\rho_1$. For this distribution $R_{\nu-1}(t)$ we have

$$\int_0^\infty e^{-st}(1-R_{\nu-1}(t))\mathrm{d}t = \frac{s^{\nu-2}}{1+s^{\nu-1}}, \ s \geq 0. \tag{3.5.13}$$

By applying Theorem 2 of [43] , Vol. II, p. 175, it is readily seen that: for $t \geq 0$,

$$1-R_{\nu-1}(t) = \sum_{n=0}^{\infty}(-1)^n \frac{t^{n(\nu-1)}}{\Gamma(n(\nu-1)+1)}, \tag{3.5.14}$$

from which we know that $R_{\nu-1}(t)$ is continuous on $[0,\infty)$. □

Remark 3.5.1 Applying Lemma 2.3.4, we obtain the tail behavior of the distribution $R_{\nu-1}(t)$:

$$1-R_{\nu-1}(t) \sim \frac{t^{1-\nu}}{\Gamma(2-\nu)} \quad \text{for } t \to \infty.$$

It can also be derived from the asymptotic expansion of $\omega_2(\Delta(\rho_2)s/\beta_1)$ for $s \downarrow 0$, and Lemma 2.3.4, that

$$\mathbf{P}(\Delta(\rho_2)W_2/\beta_1 \geq t) \sim \frac{t^{1-\nu}}{\Gamma(2-\nu)} \quad \text{for } t \to \infty.$$

Remark 3.5.2 The heavy-traffic limit theorem for the steady-state waiting time in the $G/G/1$ queue with heavy-tailed service and/or interarrival time distribution was obtained by Boxma and Cohen; see [22]. In [22] it is assumed that the LST of the service time distribution $\beta(s)$ can be represented as: for Re $s \geq 0$,

$$1-\frac{1-\beta(s)}{\beta s} = g(\beta s) + c_0(\beta s)^{\nu-1}L(1/\beta s), \tag{3.5.15}$$

where

(i) $c_0 > 0$ is a constant;

(ii) $1 < \nu \leq 2$;

(iii) $g(\beta s)$ is a regular function of s for Re $s > -\epsilon$ ($\epsilon > 0$), $g(0) = 0$;

(iv) $L(1/\beta s)$ is regular for Re $s > 0$, and continuous for Re $s \geq 0$, except possibly at $s = 0$;
$L(1/\beta s) \to b > 0$ for $|s| \to 0$, Re $s \geq 0$, with $b = \infty$ if $\nu = 2$,
$\lim_{x \downarrow 0} \frac{L(1/\beta s x)}{L(1/\beta x)} = 1$ for Re $s \geq 0, s \neq 0$;

(v) For a $\mu \in (1,\ \nu)$:

$$\int_0^\infty t^\mu \mathrm{d}B(t) < \infty.$$

More generally, the LST of the service time distribution can be represented as

$$1 - \frac{1 - \beta(s)}{\beta s} = \sum_{i=1}^{\infty} c_i (\beta s)^{\nu_i - 1} L_i(1/\beta s) + g(\beta s),$$

where $1 < \nu_1 < \cdots < \nu_n < \cdots$, $L_i(1/s)$ satisfies (iv) in (3.5.15), c_i is a constant and $g(s)$ satisfies (iii) in (3.5.15).

Remark 3.5.3 For the $M/G/1$ queue, which is a special case of the $G/G/1$ queue, the heavy-traffic limit theorem in [22] can simply be obtained from a theorem formulated in [56], p. 38, concerning geometric sums of i.i.d. random variables. Using the notation in Remark 3.5.2 and letting ρ denote the traffic load, notice that the waiting time W has the same distribution as the geometric sum

$$(1 - \rho) \sum_{n=0}^{\infty} \rho^n (B_1^{res} + ... + B_n^{res})$$

with $B_1^{res}, B_2^{res}, ...$ i.i.d. residual service times (hence with LST $(1-\beta(s))/(\beta s)$).

Remark 3.5.4 Resnick and Samorodnitsky [92] prove the same result as in [22] by directly analyzing the weak convergence of a sequence of negative drift random walks with heavy right tail and the associated all time maxima of these random walks. Their result implies Theorem 3.5.1 because it follows from (3.3.1) that W_2 can be viewed as the waiting time in an $M/G/1$ queue.

3.6 The $M/G/1$ queue with k priority classes

In this section we consider the $M/G/1$ queue with k priority classes where $k \geq 2$. Let the jth priority class be indexed by j for $1 \leq j \leq k$. Denote by ρ_j the traffic load generated by class-j, λ_j the arrival rate of class-j, $B_j(t)$ the service time distribution of class-j, W_j the steady-state class-j waiting time for $1 \leq j \leq k$. We assume $\sum_{j=1}^{k} \rho_j < 1$ to ensure that the steady-state class-k waiting time distribution exists.

Suppose one of the service time distributions has the following heavy tail behavior:

$$1 - B_i(t) \sim L(t)t^{-\nu}, \quad \text{as } t \to \infty, \tag{3.6.1}$$

with $L(t)$ a slowly varying function and $1 < \nu < 2$, the other service time distributions being such that, for $j \neq i$, $1 \leq j \leq k$,

$$\int_0^\infty t^{\mu_j} \mathrm{d}B_j(t) < \infty \quad \text{where } \mu_j > \nu,$$

or

$$1 - B_j(t) \sim L_j(t)t^{-\nu}$$

with $\lim_{t\to\infty} L_j(t)/L(t) < \infty$. Obviously, only class-$k$ customers experience heavy traffic for $\rho_k \uparrow 1 - \sum_{j=1}^{k-1} \rho_j$. We can solve this problem by viewing this queueing model with k priority classes as a queueing model with two priority classes. Subsequently, we use the result in Section 3.5 to get the heavy-traffic limit theorem for the generalized model. Let the first $k-1$ classes be the high-priority class, class-k the low-priority class in a queueing model with two priority classes. The service time distributions of the two classes in the new model are given by

$$\begin{aligned} \tilde{B}_1(t) &= \frac{\sum_{j=1}^{k-1} \lambda_j B_j(t)}{\sum_{j=1}^{k-1} \lambda_j}, \\ \tilde{B}_2(t) &= B_k(t). \end{aligned}$$

The above assumptions imply that Assumption 3.2.1 holds for $\tilde{B}_1(t)$, $\tilde{B}_2(t)$. Hence, the following heavy-traffic limit theorem holds.

Theorem 3.6.1 *For the stable $M/G/1$ queue with k ($k \geq 2$) priority classes, the above assumptions for the service time distributions $B_j(t)$, $1 \leq j \leq k$, holding, the contracted waiting time $\Delta(\rho_k)W_k/\beta_1$ converges in distribution for $\rho_k \uparrow 1 - \sum_{j=1}^{k-1} \rho_j$; the limit distribution $R_{\nu-1}(t)$ is given by (3.5.10), and the coefficient of contraction $\Delta(\rho_k)$ is that root of Equation (3.5.1) with the property that $\Delta(\rho_k) \downarrow 0$ for $\rho_k \uparrow 1 - \sum_{j=1}^{k-1} \rho_j$.*

3.7 The $M/G/1$ queue without priorities

The heavy-traffic theorem for the waiting time in the ordinary $M/G/1$ queue is well-known, cf. [22]. In this section we use Theorem 3.5.1 to derive a heavy-traffic theorem for the $M/G/1$ queue without priorities. We compare the low-priority waiting time W_2 in the $M/G/1$ queueing model with two priority classes and the waiting time W in the same model without priorities.

For the $M/G/1$ queueing model without priorities, the traffic load ρ, the service time distribution $B(t)$, the LST $\beta(s)$ of $B(t)$ and the mean β of the service time are given by

$$\rho = \rho_1 + \rho_2, \tag{3.7.1}$$

$$B(t) = \frac{\lambda_1}{\lambda_1+\lambda_2}B_1(t) + \frac{\lambda_2}{\lambda_1+\lambda_2}B_2(t), \tag{3.7.2}$$

$$\beta(s) = \frac{\lambda_1}{\lambda_1+\lambda_2}\beta_1(s) + \frac{\lambda_2}{\lambda_1+\lambda_2}\beta_2(s), \tag{3.7.3}$$

$$\beta = \frac{\rho_1+\rho_2}{\lambda_1+\lambda_2}, \tag{3.7.4}$$

from which it follows that

$$\beta_e(s) = \frac{\rho_1\beta_{1e}(s)}{\rho_1+\rho_2} + \frac{\rho_2\beta_{2e}(s)}{\rho_1+\rho_2}. \tag{3.7.5}$$

To get a heavy-traffic limit theorem for this model, we take $\tilde{\rho}_1 = 0$, $\tilde{\rho}_2 = \rho$, and assume that $\tilde{B}_1(t)$ is exponentially distributed and $\tilde{B}_2(t) = B(t)$. Applying Theorem 3.5.1 then yields the following heavy-traffic limit theorem for the classical $M/G/1$ queue without priority discipline.

Theorem 3.7.1 *For the stable $M/G/1$ queue with FCFS discipline, the service time distribution $B(t)$ being given by (3.7.2) where $B_1(t)$ and $B_2(t)$ satisfy Assumption 3.2.1, the contracted waiting time $\delta(\rho_2)W/\beta_1$ converges in distribution for $\rho_2 \uparrow 1-\rho_1$; the limit distribution $R_{\nu-1}(t)$ is given by (3.5.10), and the coefficient of contraction $\delta(\rho_2)$ is that root of Equation (3.5.1) with the property that $\delta(\rho_2) \downarrow 0$ for $\rho_2 \uparrow 1-\rho_1$, and with $K = K_1, ..., K_4$ corresponding to Assumption 3.5.1(i),...,(iv) respectively, where $K_1 = \frac{\rho_1}{\rho_1+\rho_2}$, $K_2 = \frac{\rho_2(\beta_2/\beta_1)^{\nu-1}}{\rho_1+\rho_2}$, $K_3 = \frac{\rho_1+\rho_2 a(\beta_2/\beta_1)^{\nu-1}}{\rho_1+\rho_2}$ and $K_4 = K_2$.*

The above result has been proven in [22].

Theorems 3.5.1 and 3.7.1 show that the distributions of both $\Delta(\rho_2)W_2/\beta_1$ and $\delta(\rho_2)W/\beta_1$ converge to $R_{\nu-1}(t)$ for $\rho_2 \uparrow 1-\rho_1$. The following lemma shows

the relation between $\Delta(\rho_2)$ and $\delta(\rho_2)$. One can prove the lemma by applying a similar method as in the proof of Lemma 3.5.1. We omit the details.

Lemma 3.7.1 *If Assumption 3.2.1 is satisfied, then*

$$\lim_{\rho_2 \uparrow 1-\rho_1} \frac{\Delta(\rho_2)}{\delta(\rho_2)} = 1 - \rho_1.$$

Apparently, the effect of introducing priorities is (cf. the difference in the constants K_i in Theorems 3.5.1 and 3.7.1): class-1 customers do not experience heavy traffic and class-2 customers have a similar heavy-traffic waiting time tail behavior as in the case without priorities, apart from a scaling factor $1-\rho_1$. ρ_1 is the fraction of time that the server is occupied by class-1 customers and $1-\rho_1$ is the fraction of time that the server is available for class-2 customers. Actually, the following approximation seems useful:

$$1 - W_2(t) \approx 1 - W((1-\rho_1)t), \quad t \geq 0.$$

First of all, this approximation satisfies the heavy-traffic behavior indicated above. Secondly, it yields the correct mean waiting time $\mathrm{E}[W_2] = \mathrm{E}[W]/(1-\rho_1)$; cf. Cohen [36], Formula (II.3.64). Thirdly, it gives the correct behavior at $t = 0$ (unlike the heavy-traffic approximation). And finally, it follows from Theorem 3.3.1 (see also Remark 3.3.1) that, if Assumption 3.2.1 holds, then

$$1 - W_2(t) \sim M t^{1-\nu} L(t), \quad t \to \infty, \tag{3.7.6}$$

where M is a constant. It is well-known (cf. Cohen [34] and Pakes [85]) that, in the $M/G/1$ queue with regularly varying (even subexponential) service time distribution $B(\cdot)$,

$$1 - W(t) \sim \frac{\rho}{1-\rho} \int_t^\infty \frac{1 - B(x)}{\beta} \mathrm{d}x, \quad t \to \infty.$$

One can easily verify that this yields exactly the same tail behavior for $1 - W((1-\rho_1)t)$ as in (3.7.6).

In fact, even in the $M/G/1$ queueing model with two priority classes and the nonpreemptive discipline, only class-2 customers experience heavy traffic. This is easily seen from the following expression for $\omega_1(s)$, for $\rho_1 + \rho_2 < 1$,

$$\omega_1(s) = \frac{1 - \rho_1 - \rho_2 + \rho_2 \beta_{2e}(s)}{1 - \rho_1 \beta_{1e}(s)},$$

and for $\rho_1 + \rho_2 \geq 1$ but $\rho_1 < 1$,

$$\omega_1(s) = \frac{(1-\rho_1)\beta_{2e}(s)}{1-\rho_1\beta_{1e}(s)},$$

cf. Section III.3.8 in Cohen [36]. Generally, in the $M/G/1$ queueing model with k ($k \geq 2$) priority classes, nonpreemptive or preemptive resume discipline, only the lowest priority class suffers from heavy traffic unless $\rho_k \downarrow 0$.

3.8 Applications of the heavy-traffic limit theorem

Theorem 3.5.1, the heavy-traffic limit theorem, suggests the following heavy-traffic approximation for the stationary class-2 waiting time distribution: for $0 < 1-\rho_1-\rho_2 << 1-\rho_1$, and with $\Delta(\rho_2)$ specified by the contraction equation (3.5.1),

$$\mathbf{P}(\frac{\Delta(\rho_2)W_2}{\beta_1} > t) \approx 1 - R_{\nu-1}(t), \quad t > 0, \tag{3.8.1}$$

or equivalently,

$$1 - W_2(t) = \mathbf{P}(W_2 > t) \approx 1 - R_{\nu-1}(\Delta(\rho_2)t/\beta_1), \quad t > 0. \tag{3.8.2}$$

According to the heavy-traffic theorem, this approximation should perform very well when ρ is sufficiently close to 1. In this section we investigate whether this approximation is still useful when ρ is not very close to 1. We follow a similar procedure as [21], where such a heavy-traffic approximation is numerically investigated for the waiting time distribution of the $M/G/1$ case without priorities. We suppose that the service time distributions are of the following form,

$$B_j(t) = 1 - \frac{1}{\Gamma(2-\nu_j)} \int_0^\infty e^{-\theta} \frac{\theta}{(\theta+t)^{\nu_j}} \mathrm{d}\theta, \quad j = 1, 2, \tag{3.8.3}$$

with $1 < \nu_j < 2$. Note that

$$B_j(0+) = 0,$$
$$\beta_j = \int_0^\infty t \mathrm{d}B_j(t) = \frac{2-\nu_j}{\nu_j - 1};$$

the second moment of B_j is infinite. As shown in [21], the explicit expression and the LST of $B_j(t)$ as given in (3.8.3) are characterized by: for Re $s \geq 0$,

$$1 - B_j(t) \quad = \quad \frac{a_j}{1-a_j} \left(\frac{1}{a_j}(1-a_j+t)e^t - \frac{t^{a_j}}{\Gamma(1+a_j)} \right.$$

$$-\frac{1-a_j+t}{\Gamma(1+a_j)}e^t\sum_{n=0}^{\infty}\frac{(-1)^n t^{a_j+n}}{n!(a_j+n)}\Bigg), \quad \text{with } a_j := 2-\nu_j,$$

$$\frac{1-\beta_j(s)}{\beta_j s} = \frac{\omega}{\omega-1}\left[1-\frac{1}{2-\nu_j}\frac{\omega^{2-\nu_j}-1}{\omega-1}\right], \quad \text{with } \omega := \frac{1}{s},$$

for $j = 1, 2$. Thus we have

$$1-\frac{1-\beta_j(s)}{\beta_j s} \sim (\beta_j s)^{\nu-1}L(1/\beta_j s).$$

In determining $\Delta(\cdot)$, we have taken $L(\cdot) \equiv 1$. Put, cf. (3.8.2), for $\rho_2 \in (0, 1-\rho_1)$ (with HT denoting Heavy Traffic),

$$1 - W_{HT}(t) := 1 - R_{\nu-1}(\Delta(\rho_2)t/\beta_1).$$

As proved in Theorem 3.3.1, we have

$$1 - W_2(t) \sim \frac{(\nu-1)K}{\Gamma(1-\nu)(1-\rho)}(t/\beta_1)^{1-\nu}, \quad t \to \infty,$$

where K is given in Theorem 3.5.1. Define

$$1 - W_{RV}(t) := \frac{(\nu-1)K}{\Gamma(1-\nu)(1-\rho)}(t/\beta_1)^{1-\nu}L(t/\beta_1).$$

From (3.5.1) we obtain that $1-W_{HT}(t)$ exhibits the same asymptotic behavior:

$$1 - W_{HT}(t) \sim \frac{(\nu-1)K}{\Gamma(1-\nu)(1-\rho)}(t/\beta_1)^{1-\nu}L(t/\beta_1).$$

As observed in the previous section, we can also use $1 - W((1-\rho_1)t)$ to approximate $1-W_2(t)$ where $W(t)$ is the waiting time distribution in the same $M/G/1$ model without priority structure, i.e., the $M/G/1$ queue with FCFS discipline. As remarked there, $1 - W((1-\rho_1)t)$ has the same tail behavior as $1 - W_2(t)$ in the regularly varying case. Therefore, for $0 < \rho < 1$:

$$1 - W((1-\rho_1)t) \sim 1 - W_2(t) \sim 1 - W_{HT}(t) \sim 1 - W_{RV}(t), \quad t \to \infty.$$

We have tested the approximations $1 - W_{HT}(t)$, $1 - W((1-\rho_1)t)$ and $1 - W_{RV}(t)$ for three cases: (i) the class-1 service time distribution $B_1(t)$ is specified by (3.8.3) and the class-2 service time distribution $B_2(t)$ is exponentially distributed with mean 1; (ii) the class-1 service time distribution $B_1(t)$ is exponentially distributed with mean 1 and the class-2 service time distribution

$B_2(t)$ is specified by (3.8.3); (iii) both of the class-1 and class-2 service time distributions are specified by (3.8.3) with $\nu_1 = \nu_2$. The exact class-2 waiting time distribution is calculated by inverting the LST (see [2] for a description of the algorithm used).

In view of the very large number of parameter combinations, we only indicate maximal relative errors over certain t-regions. We distinguish between three t-regions: "small" t indicates t-values such that $0.1\rho < 1 - W_2(t) \leq 0.5\rho$; "medium" t indicates t-values such that $0.01\rho < 1 - W_2(t) \leq 0.1\rho$; "large" t indicates t-values such that $1 - W_2(t) \leq 0.01\rho$. Note that $W_{HT}(0) = 0$ while $W_2(0) = 1 - \rho$, so that t-values close to zero always yield large errors. We compare the errors of $1 - W_{HT}(t)$, $1 - W_{RV}(t)$ and $1 - W((1-\rho_1)t)$; the latter is referred to as the FCFS case. In the error columns, we consider the absolute value of the largest relative error in a region. Let "$- - -$" denote that this largest error exceeds 20%; "$--$" that it is between 10% and 20%; "$-$" that it is between 5% and 10%; "$+$" that it is between 1% and 5%; "$++$" that it is between 0.1% and 1%; "$+ + +$" that it is less than 0.1%. Denote Case (i) by "exp/RV", Case (ii) by "RV/exp" and Case (iii) by "RV/RV".

The numerical results are gathered in Tables 3.1-3.6. Table 3.1 displays cases with $\rho = 0.9$ and one of the service time distributions being given by (3.8.3) and another service time being exponentially distributed with mean 1 or both service time distributions being given by (3.8.3) with $\nu = 1.25$. Table 3.2 does the same except that $\nu = 1.75$. Table 3.3 shows cases with $\rho = 0.5$ and $\nu = 1.25$; Table 3.4 does the same except that $\nu = 1.75$; Table 3.5 presents cases with $\rho = 0.1$ and $\nu = 1.25$; Table 3.6 displays cases with $\rho = 0.1$ and $\nu = 1.75$.

The main conclusions from the numerical work are as follows.

1. All the approximations $1 - W_{HT}(t)$, $1 - W_{RV}(t)$ and $1 - W((1-\rho_1)t)$ provide extremely accurate approximations for t large.

2. $1 - W_{HT}(t)$ performs much better for $\nu = 5/4$ (the case with a heavier tail) than for $\nu = 7/4$.

3. $1 - W((1-\rho_1)t)$ provides a very good approximation even for small t, better than $1 - W_{HT}(t)$ and $1 - W_{RV}(t)$; when ρ_1 is small, it performs the best.

4. In heavy traffic (ρ is sufficiently large), $1 - W_{HT}(t)$ yields much better results than $1 - W_{RV}(t)$ does; $1 - W_{RV}(t)$ is almost useless here (it is not a heavy-traffic approximation).

5. In light traffic, $1 - W_{HT}(t)$ still provides surprisingly accurate results, when t is not too small and ν is small.

6. In the case of light traffic and $\nu = 1.75$, the accuracy of $1 - W_{HT}(t)$ is almost the same as that of $1 - W_{RV}(t)$ or even worse.

Remark 3.8.1 If ρ_1 in Case (i) equals ρ_2 in Case (ii) and ρ in both Cases (i) and (ii) are the same, i.e., both cases have the same traffic load for the class with heavy-tailed service time distribution and the same total traffic load, then they have the same contraction coefficients and thus the approximation $1 - W_{HT}(t)$ is exactly the same.

Remark 3.8.2 $1 - W((1 - \rho_1)t)$ performs particularly well for ρ_1 small, because $\lim_{\rho_1 \to 0}(1 - W_2(t))/(1 - W((1 - \rho_1)t)) = 1$. When ρ_1 is small, the busy period of class-1 customers will not have much effect on the class-2 waiting time distribution.

Table 3.1: Approximations for class-2 waiting time tails.

$\rho = 0.9,\ \nu = 1.25$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.8, 0.1)	$--$	$---$	$+$	$+++$	$--$	$+++$	$+++$	$+$	$+++$
exp/RV:(0.45, 0.45)	$++$	$---$	$+++$	$+++$	$--$	$+++$	$+++$	$+$	$+++$
exp/RV:(0.1, 0.8)	$+++$	$---$	$+++$	$+++$	$--$	$+++$	$+++$	$+$	$+++$
RV/exp:(0.8, 0.1)	$+$	$---$	$+$	$++$	$-$	$++$	$+++$	$+$	$+++$
RV/exp:(0.45, 0.45)	$+$	$---$	$+$	$++$	$-$	$++$	$+++$	$+$	$+++$
RV/exp:(0.1, 0.8)	$--$	$---$	$+$	$++$	$-$	$++$	$+++$	$++$	$+++$
RV/RV:(0.8, 0.1)	$+$	$---$	$+$	$++$	$--$	$++$	$+++$	$+$	$+++$
RV/RV:(0.45, 0.45)	$+$	$---$	$+$	$+$	$-$	$++$	$+++$	$+$	$+++$
RV/RV:(0.1, 0.8)	$++$	$---$	$++$	$+++$	$-$	$+++$	$+++$	$+$	$+++$

Table 3.2: Approximations for class-2 waiting time tails.

$\rho = 0.9,\ \nu = 1.75$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.8, 0.1)	$---$	$---$	$--$	$++$	$-$	$++$	$+++$	$++$	$+++$
exp/RV:(0.45, 0.45)	$---$	$--$	$++$	$--$	$--$	$++$	$++$	$+$	$+++$
exp/RV:(0.1, 0.8)	$---$	$--$	$+++$	$-$	$--$	$+++$	$++$	$+$	$+++$
RV/exp:(0.8, 0.1)	$---$	$--$	$-$	$+$	$--$	$+$	$++$	$+$	$++$
RV/exp:(0.45, 0.45)	$---$	$--$	$+$	$-$	$--$	$++$	$++$	$+$	$++$
RV/exp:(0.1, 0.8)	$-$	$--$	$++$	$++$	$+$	$++$	$+++$	$++$	$+++$
RV/RV:(0.8, 0.1)	$---$	$--$	$-$	$-$	$--$	$+$	$++$	$+$	$++$
RV/RV:(0.45, 0.45)	$---$	$--$	$++$	$-$	$--$	$++$	$++$	$+$	$++$
RV/RV:(0.1, 0.8)	$---$	$--$	$++$	$-$	$--$	$+++$	$++$	$+$	$+++$

Table 3.3: Approximations for class-2 waiting time tails.

$\rho = 0.5,\ \nu = 1.25$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.4, 0.1)	$---$	$---$	$--$	$+++$	$+$	$+++$	$+++$	$++$	$+++$
exp/RV:(0.25, 0.25)	$+$	$---$	$+++$	$+++$	$+$	$+++$	$+++$	$++$	$+++$
exp/RV:(0.1, 0.4)	$++$	$---$	$+++$	$+++$	$+$	$+++$	$+++$	$++$	$+++$
RV/exp:(0.4, 0.1)	$-$	$---$	$-$	$+$	$--$	$-$	$++$	$++$	$++$
RV/exp:(0.25, 0.25)	$---$	$-$	$+$	$--$	$+$	$++$	$++$	$++$	$+++$
RV/exp:(0.1, 0.4)	$---$	$---$	$++$	$-$	$+$	$+++$	$++$	$++$	$+++$
RV/RV:(0.4, 0.1)	$+$	$---$	$+$	$++$	$+$	$++$	$+++$	$++$	$+++$
RV/RV:(0.25, 0.25)	$+$	$---$	$+$	$++$	$+$	$++$	$+++$	$++$	$+++$
RV/RV:(0.1, 0.4)	$+$	$---$	$++$	$++$	$+$	$++$	$+++$	$++$	$+++$

Table 3.4: Approximations for class-2 waiting time tails.

$\rho = 0.5,\ \nu = 1.75$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.4, 0.1)	$-$	$---$	$-$	$+$	$--$	$-$	$++$	$++$	$++$
exp/RV:(0.25, 0.25)	$---$	$---$	$+$	$+$	$+$	$+++$	$++$	$++$	$+++$
exp/RV:(0.1, 0.4)	$---$	$---$	$++$	$-$	$+$	$++$	$++$	$+$	$++$
RV/exp:(0.4, 0.1)	$---$	$---$	$-$	$--$	$+$	$++$	$++$	$+$	$++$
RV/exp:(0.25, 0.25)	$---$	$---$	$-$	$--$	$+$	$+$	$++$	$+$	$++$
RV/exp:(0.1, 0.4)	$-$	$---$	$+$	$-$	$-$	$+$	$++$	$+$	$++$
RV/RV:(0.4, 0.1)	$---$	$---$	$-$	$-$	$+$	$+$	$+++$	$+$	$++$
RV/RV:(0.25, 0.25)	$---$	$--$	$+$	$--$	$+$	$++$	$+++$	$+$	$++$
RV/RV:(0.1, 0.4)	$---$	$---$	$+$	$--$	$+$	$++$	$++$	$+$	$++$

Table 3.5: Approximations for class-2 waiting time tails.

$\rho = 0.1,\ \nu = 1.25$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.09, 0.01)	$---$	$---$	$+$	$--$	$--$	$-$	$+++$	$+++$	$+++$
exp/RV:(0.05, 0.05)	$-$	$-$	$++$	$++$	$++$	$++$	$+++$	$+++$	$+++$
exp/RV:(0.01, 0.09)	$+$	$-$	$+++$	$++$	$++$	$+++$	$+++$	$++$	$+++$
RV/exp:(0.09, 0.01)	$+$	$-$	$+$	$++$	$++$	$++$	$+++$	$+++$	$+++$
RV/exp:(0.05, 0.05)	$---$	$---$	$++$	$++$	$++$	$++$	$+++$	$+++$	$+++$
RV/exp:(0.01, 0.09)	$---$	$---$	$++$	$++$	$+$	$+++$	$+++$	$+++$	$+++$
RV/RV:(0.09, 0.01)	$+$	$+$	$++$	$++$	$++$	$++$	$+++$	$+++$	$+++$
RV/RV:(0.05, 0.05)	$-$	$-$	$++$	$+++$	$++$	$+++$	$+++$	$++$	$+++$
RV/RV:(0.01, 0.09)	$+$	$-$	$+++$	$+++$	$++$	$+++$	$+++$	$++$	$+++$

Table 3.6: Approximations for class-2 waiting time tails.

$\rho = 0.1,\ \nu = 1.25$									
	"small" t			"medium" t			"large" t		
	HT	RV	FCFS	HT	RV	FCFS	HT	RV	FCFS
exp/RV:(0.09, 0.01)	$---$	$---$	$--$	$---$	$---$	$--$	$+$	$+$	$++$
exp/RV:(0.05, 0.05)	$---$	$---$	$+$	$-$	$-$	$++$	$++$	$++$	$+++$
exp/RV:(0.01, 0.09)	$---$	$---$	$+++$	$+$	$+$	$+++$	$++$	$+++$	$+++$
RV/exp:(0.09, 0.01)	$---$	$---$	$+$	$-$	$-$	$++$	$++$	$++$	$++$
RV/exp:(0.05, 0.05)	$---$	$---$	$+$	$-$	$-$	$++$	$++$	$+++$	$+++$
RV/exp:(0.01, 0.09)	$---$	$---$	$++$	$+$	$+$	$++$	$++$	$++$	$+++$
RV/RV:(0.09, 0.01)	$---$	$---$	$+$	$+$	$+$	$++$	$+++$	$+++$	$+++$
RV/RV:(0.05, 0.05)	$---$	$---$	$++$	$+$	$+$	$++$	$++$	$+++$	$+++$
RV/RV:(0.01, 0.09)	$---$	$---$	$++$	$+$	$+$	$++$	$+++$	$+++$	$+++$

Chapter 4

The two-queue $E/1$-L polling model

4.1 Introduction

Cyclic polling systems are queueing systems in which a single server visits several queues in cyclic order. They have a wide range of applications in, e.g., computer communications, manufacturing and road traffic. The abundant literature on polling systems (see [102, 103]) contains the exact analysis of polling systems for a large number of service disciplines, like the 1-limited, exhaustive and gated disciplines.

Both this chapter and the next one are devoted to (the asymptotic analysis of) polling systems. The basic polling system consists of K queues, $Q_1, ..., Q_K$, attended by a single server S. Customers arrive at Q_k, $k = 1, ..., K$, according to a Poisson process with rate λ_k and require a generally distributed service time B_k having distribution function $B_k(\cdot)$, finite first moment β_k and LST $\beta_k(\cdot)$. In the sequel, customers arriving at Q_k are also referred to as type-k customers. The server visits the queues in a strictly cyclic order, i.e., $Q_1, ..., Q_K, Q_1, ..., Q_K, Q_1, ...$ The service policy at each queue is either 1-limited, gated or exhaustive; we do allow mixtures, like gated service at Q_1 and Q_3 and exhaustive service at the other queues. In 1-limited service, the server serves at most one customer at a queue before switching to the next queue. In gated service, the server serves at a queue exactly those customers that are present at the start of his visit to the queue. In exhaustive service, the server continues to work at a queue until it becomes empty.

We consider both the model with and without switchover times. In the model with switchover times, when moving from Q_k to $Q_{(k \bmod K)+1}$, the server

incurs a generally distributed switchover time S_k, having distribution function $S_k(\cdot)$ with finite first moment σ_k and LST $\sigma_k(\cdot)$. The server continues switching even when the whole system is empty. In the model without switchover times, when the system becomes empty the server makes a full cycle (i.e. passes all the queues at least once) and subsequently stops right before Q_1. When the first new customer arrives, S cycles along the queues to that new customer. The various interarrival, service and switchover times are independent.

Let us denote by $\lambda := \sum_{k=1}^{K} \lambda_k$ the total arrival intensity of customers, by $\rho_k := \lambda_k \beta_k$ the traffic load at Q_k, by $\rho := \sum_{k=1}^{K} \rho_k$ the total traffic load, and by $\sigma := \sum_{k=1}^{K} \sigma_k$ the mean of the total switchover time in one cycle.

Tail probabilities, which are especially helpful in understanding the performance of different polling disciplines, have received some attention in recent years, and they are also the main topic of this chapter and the next one. We now describe some works which are relevant to the present and next chapters. For models with Poisson arrivals, general service and switchover time distributions, and various service disciplines, Choudhury and Whitt [33] have developed efficient iterative algorithms to compute the exact tail behavior of, for example, the steady-state waiting time W, of the form,

$$\mathbf{P}(W > t) \sim a t^b e^{-ct}, \quad t \to \infty, \tag{4.1.1}$$

with $c > 0$ and $a > 0$. Such tail behavior occurs when the service and switchover time distributions have finite moment generating functions, i.e., there is some positive real number s^* such that $\beta_k(-s^*) < \infty$ and $\sigma_k(-s^*) < \infty$; here $\beta_k(\cdot)$ and $\sigma_k(\cdot)$ are the LSTs of the service and switchover time distributions at the kth queue, respectively. Motivated by [33], and using analytic methods, Duffield [45] explores the relationship between the exponents b and c in (4.1.1) and their dependence (and sometimes independence) on the service and switchover time distributions.

In view of the central role of polling in computer-communication networks and the often observed occurrence of heavy-tailed traffic in computer-communications (see Section 1.2), it is of importance to study the effect of heavy-tailed service and/or switchover time distributions on the waiting time tail behavior in polling systems. In this chapter and the next one, we study the waiting time tail behavior in cyclic polling systems with various service policies.

In this chapter we consider a polling system as described above with $K = 2$ queues. The service policy is exhaustive at Q_1 and 1-limited at Q_2. The present chapter is based on [42]. Note that the model with zero switchover times coincides with the $M/G/1$ queue with two priority classes and the nonpreemptive

discipline which we considered in Chapter 3. That provides some of our motivation for the study in the present chapter. In the next chapter, we consider general polling systems with gated or exhaustive service. In that sense, the present chapter can be viewed as a bridge which links Chapters 3 and 5.

It is well-known [59] that the stability condition is: $\rho+\lambda_2\sigma < 1$. We assume this condition to hold in the remainder of this chapter.

The model with nonzero switchover times has been studied by Groenendijk [59] and Ibe [64]. They derived the explicit LSTs of the waiting time distributions. Based on their results, we investigate (i) the tail asymptotics of the waiting times at both queues when at least one of the service and/or switchover times has a regularly varying tail and (ii) the waiting time at Q_2 in the heavy-traffic situation when at least one of the service and/or switchover times does not have a finite second moment. Finally we show some numerical results to test the accuracy of the approximation for the waiting time distribution at Q_2 suggested by the heavy-traffic limit theorem.

4.2 Preliminaries

In this section we first introduce the expressions for the LSTs of the waiting time distributions, cf. [59]. Next we make some assumptions on the service and switchover times.

Let us start with some notation. Denote by $\eta_1(\cdot)$ the LST of the length of the busy period at Q_1 starting with one customer. Let W_k be the steady-state waiting time at Q_k with distribution function $W_k(\cdot)$ and LST $\omega_k(\cdot)$ for $k = 1, 2$. From (6.70) in [59], $\omega_k(s)$ $(k = 1, 2)$ are, for Re $s \geq 0$, given by

$$\begin{aligned}
\omega_1(s) &= \frac{\sigma_1(s)\sigma_2(s)\beta_2(s) - 1}{\lambda_1 - s - \lambda_1\beta_1(s)}\frac{1-\rho}{\sigma} \\
&\quad + \frac{1-\rho-\lambda_2\sigma}{\sigma}\frac{\sigma_1(\lambda_2+s)\sigma_2(s)}{\sigma_1(\lambda_2)}\frac{1-\beta_2(s)}{\lambda_1 - s - \lambda_1\beta_1(s)}, \qquad (4.2.1) \\
\omega_2(s) &= \frac{1-\rho-\lambda_2\sigma}{\sigma}\sigma_1(s)\frac{\sigma_1(\lambda_2+\lambda_1(1-\eta_1(s)))}{\sigma_1(\lambda_2)}\sigma_2(f(s)) \\
&\quad \frac{\lambda_2 - s - \lambda_2\beta_2(f(s))}{\lambda_2 - s - \lambda_2\sigma_1(f(s))\sigma_2(f(s))\beta_2(f(s))}\frac{1}{\lambda_2 - s} \\
&\quad - \frac{1}{\lambda_2 - s}\frac{1-\rho-\lambda_2\sigma}{\sigma}, \qquad (4.2.2)
\end{aligned}$$

where

$$f(s) := s + \lambda_1(1 - \eta_1(s)).$$

Concerning the service and switchover times, we are only interested in the properties of their tail behavior, i.e., the behavior of $1 - B_k(t)$ and $1 - S_k(t)$ for $t \to \infty$. For future reference, we make assumptions on the service and switchover time distributions for the general cyclic polling systems with K ($K \in \mathbb{N}$) queues. We assume the following holds:

Assumption 4.2.1 *For the service and switchover time distributions, we have:*

$$1 - B_k(t) = [b_k + \mathrm{o}(1)]t^{-\nu}L(t), \quad t \to \infty, \tag{4.2.3}$$

$$1 - S_k(t) = [s_k + \mathrm{o}(1)]t^{-\nu}L(t), \quad t \to \infty, \tag{4.2.4}$$

where $b_k, s_k \geq 0$, $L(\cdot)$ is a slowly varying function and $k = 1, ..., K$. Here we assume $\sum_{k=1}^{K}(s_k + b_k) > 0$, i.e., at least one of the service and/or switchover times has a regularly varying tail of index $-\nu$.

For ease of presentation, we take the same function $L(\cdot)$ for all distributions, but one can easily change this into different slowly varying functions for different distributions. Note that the possibility that $b_k = 0$ or $s_k = 0$ implies that we do allow the possibility that some of the service and switchover time distributions have an exponential tail, or regularly varying of index strictly smaller than $-\nu$. According to Lemma 2.3.4, the tail behavior of the service and switchover time distributions as given in (4.2.3) and (4.2.4) is equivalent with the following behavior of their LSTs $\beta_k(s)$ (of the service time distributions) and $\sigma_k(s)$ (of the switchover time distributions):

$$1 - \beta_k(s) = \sum_{j=1}^{m}(-1)^{j+1}\beta_{k,j}s^j + (-1)^m\beta_{k,\nu}s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \tag{4.2.5}$$

$$1 - \sigma_k(s) = \sum_{j=1}^{m}(-1)^{j+1}\sigma_{k,j}s^j + (-1)^m\sigma_{k,\nu}s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \tag{4.2.6}$$

where $m < \nu < m + 1$ ($m \in \mathbb{N}$), $\beta_{k,j} > 0$ and $\sigma_{k,j} > 0$ for $j = 1, ..., m$, $k = 1, ..., K$. Note that $\beta_{k,1} = \beta_k$, $\beta_{k,\nu} = (-1)^m\Gamma(1-\nu)b_k$, $\sigma_{k,1} = \sigma_k$, and $\sigma_{k,\nu} = (-1)^m\Gamma(1-\nu)s_k$ for $k = 1, ..., K$.

It follows from the main result of [81] that the asymptotic behavior of the LST $\eta_k(s)$ of the length of the busy period in the 'corresponding' isolated $M/G/1$ queue of Q_k is given by

$$1 - \eta_k(s) = \sum_{j=1}^{m}(-1)^{j+1}\eta_{k,j}s^j + (-1)^m\eta_{k,\nu}s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \tag{4.2.7}$$

where $\eta_{k,1} = \beta_k/(1-\rho_k)$ and $\eta_{k,\nu} = \beta_{k,\nu}/(1-\rho_k)^{\nu+1}$ and $\eta_{k,j} > 0$ for $j = 1, ..., m$, $k = 1, ..., K$. Here the 'corresponding' isolated $M/G/1$ queue of Q_k stands for the single-server queue with the same arrival rate and service time distributions as Q_k.

4.3 The tail behavior of the waiting time distributions

In this section we derive the asymptotic behavior of the waiting times when at least one of the service and/or switchover times has a regularly varying tail.

Let us first consider the asymptotic expansions of the functions $\sigma_1(f(s))$, $\sigma_2(f(s))$ and $\beta_2(f(s))$ in the neighborhood of the origin, because these functions appear in (4.2.2). By using Lemma 2.3.6, we immediately get, for $k = 1, 2$,

$$\begin{aligned}\sigma_k(f(s)) &= 1 + \sum_{j=1}^{m} g_{k,j} s^j + (-1)^{m+1} \left(\frac{\lambda_1 \beta_{1,\nu} \sigma_k}{(1-\rho_1)^{\nu+1}} + \frac{\sigma_{k,\nu}}{(1-\rho_1)^{\nu}} \right) s^\nu L(1/s) \\ &\quad + o(s^\nu L(1/s)), \quad s \downarrow 0, \qquad (4.3.1)\\ \beta_2(f(s)) &= 1 + \sum_{j=1}^{m} g_{3,j} s^j + (-1)^{m+1} \left(\frac{\lambda_1 \beta_{1,\nu} \beta_2}{(1-\rho_1)^{\nu+1}} + \frac{\beta_{2,\nu}}{(1-\rho_1)^{\nu}} \right) s^\nu L(1/s) \\ &\quad + o(s^\nu L(1/s)), \quad s \downarrow 0, \qquad (4.3.2)\end{aligned}$$

where $g_{k,j}$ $(k = 1, 2, 3,\ j = 1, ..., m)$ are some constants. Moreover, for $k = 1, 2$, $g_{k,1} = -\dfrac{\sigma_k}{1-\rho_1}$ and $g_{3,1} = -\dfrac{\beta_2}{1-\rho_1}$.

Again, applying Theorem 2.3.6 to (4.2.1) and (4.2.2), straightforward calculations lead to

$$\begin{aligned}\omega_1(s) &= 1 + \sum_{j=1}^{m-1} (-1)^j \omega_{1,j} s^j \\ &+ (-1)^m \left(\frac{\lambda_1 \beta_{1,\nu} + \lambda_2 \beta_{2,\nu}}{1-\rho_1} + \frac{(1-\rho)(\sigma_{1,\nu} + \sigma_{2,\nu})}{\sigma(1-\rho_1)} \right) s^{\nu-1} L(1/s) \\ &+ o(s^{\nu-1} L(1/s)), \quad s \downarrow 0, \qquad (4.3.3)\\ \omega_2(s) &= 1 + \sum_{j=1}^{m-1} (-1)^j \omega_{2,j} s^j\end{aligned}$$

$$
\begin{aligned}
&+(-1)^m \left[\frac{1-\rho}{1-\rho-\lambda_2\sigma} \left(\frac{\lambda_1\beta_{1,\nu}(\rho_2+\lambda_2\sigma)}{(1-\rho_1)^{\nu+1}} + \frac{\lambda_2(\sigma_{1,\nu}+\sigma_{2,\nu}+\beta_{2,\nu})}{(1-\rho_1)^{\nu}} \right) \right. \\
&\left. + \frac{\lambda_1\beta_{1,\nu}\rho_2}{(1-\rho_1)^{\nu+1}} + \frac{\lambda_2\beta_{2,\nu}}{(1-\rho_1)^{\nu}} \right] s^{\nu-1}L(1/s) \\
&+\mathrm{o}(s^{\nu-1}L(1/s)), \quad s \downarrow 0, \qquad (4.3.4)
\end{aligned}
$$

where $j!\omega_{k,j}$ $(k=1,2,\ j=1,...,m-1)$ equals the jth moment of the waiting time W_k. Applying Lemma 2.3.4 to (4.3.3) and (4.3.4), we then get the following theorem.

Theorem 4.3.1 *If Assumption 4.2.1 holds, then the waiting times at both queues have a regularly varying tail of index which is one higher than the heaviest of the service and switchover times. In particular, we have*

$$
\begin{aligned}
1-W_1(t) &\sim \frac{1}{\nu-1}\left(\frac{\lambda_1 b_1+\lambda_2 b_2}{1-\rho_1} + \frac{(1-\rho)(s_1+s_2)}{(1-\rho_1)\sigma} \right) t^{1-\nu}L(t), \quad t\to\infty, \\
1-W_2(t) &\sim \frac{1}{\nu-1}\left(\frac{\lambda_1 b_1+\lambda_2(s_1+s_2+b_2)}{(1-\rho_1)^{\nu-1}(1-\rho-\lambda_2\sigma)} + \frac{s_1+s_2}{(1-\rho_1)^{\nu-1}\sigma} \right) t^{1-\nu}L(t), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t\to\infty.
\end{aligned}
$$

We now relate the waiting time distribution to the residual service and switchover time distributions. Applying Lemma 2.3.1 to (4.2.3) and (4.2.4), we obtain the asymptotic behavior of the residual service times B_k^{res} and the residual switchover times S_k^{res}. For $k=1,2$, if $b_k,\ s_k>0$, then we have

$$
\begin{aligned}
\mathbf{P}(B_k^{res}>t) &\sim \frac{b_k}{(\nu-1)\beta_k} t^{1-\nu}L(t), \quad t\to\infty, \\
\mathbf{P}(S_k^{res}>t) &\sim \frac{s_k}{(\nu-1)\sigma_k} t^{1-\nu}L(t), \quad t\to\infty,
\end{aligned}
$$

which in combination with Theorem 4.3.1 implies the following corollary.

Corollary 4.3.1 *If Assumption 4.2.1 holds, then for $t\to\infty$,*

$$
\begin{aligned}
1-W_1(t) \sim\ & \frac{\rho_1 I_{\{b_1>0\}}}{1-\rho_1}\mathbf{P}(B_1^{res}>t) + \frac{\rho_2 I_{\{b_2>0\}}}{1-\rho_1}\mathbf{P}(B_2^{res}>t) \\
&+\frac{(1-\rho)\sigma_1 I_{\{s_1>0\}}}{(1-\rho_1)\sigma}\mathbf{P}(S_1^{res}>t) \\
&+\frac{(1-\rho)\sigma_2 I_{\{s_2>0\}}}{(1-\rho_1)\sigma}\mathbf{P}(S_2^{res}>t),
\end{aligned}
$$

$$
\begin{aligned}
1 - W_2(t) \quad \sim \quad & \frac{\rho_1 I_{\{b_1>0\}}}{1-\rho-\lambda_2\sigma}\mathbf{P}(B_1^{res} > (1-\rho_1)t) \\
& + \frac{\rho_2 I_{\{b_2>0\}}}{1-\rho-\lambda_2\sigma}\mathbf{P}(B_2^{res} > (1-\rho_1)t) \\
& + \frac{(1-\rho)\sigma_1 I_{\{s_1>0\}}}{(1-\rho-\lambda_2\sigma)\sigma}\mathbf{P}(S_1^{res} > (1-\rho_1)t) \\
& + \frac{(1-\rho)\sigma_2 I_{\{s_2>0\}}}{(1-\rho-\lambda_2\sigma)\sigma}\mathbf{P}(S_2^{res} > (1-\rho_1)t),
\end{aligned}
$$

where $I_{\{A\}}$ is the indicator function of event $\{A\}$.

In the following we give a heuristic explanation of the above corollary. These heuristic arguments are similar to those in [91] for a fluid queue with $M/G/\infty$ input. We should point out that the heuristic arguments below (and those in Sections 6.5 and 7.6) are not rigorous in a mathematical sense, and do not really give a strict proof, but only identify a possible way for the desired event to occur, and thus provide a lower bound for the corresponding probability. However, the fact that the lower bound coincides with the formula we found analytically, implies that the probability of any other scenario is negligible. Hence, the scenario that we identify must actually represent the only plausible way in which the event occurs. For more complicated models, like the $M/G/k$ queue, this technique may be a starting point to find the exact waiting time tail behavior.

The heuristic arguments below (and those in Sections 6.5 and 7.6) are based on the following two preliminary observations:

1. At the scale of large t, one may think of the evolution of the workload as approximately linear.

2. Due to the PASTA property, the waiting time has the same distribution as the virtual waiting time (which is equal to the workload in some cases) at any time.

We consider a special case, $b_1 > 0$, $b_2 = s_1 = s_2 = 0$, i.e., the service time B_1 at Q_1 has the heaviest tail. The general case allows a similar intuitive explanation. We use heuristic arguments to verify

$$\mathbf{P}(W_1 > t) \quad \sim \quad \frac{\rho_1}{1-\rho_1}\mathbf{P}(B_1^{res} > t), \quad t \to \infty, \tag{4.3.5}$$

$$\mathbf{P}(W_2 > t) \quad \sim \quad \frac{\rho_1}{1-\rho-\lambda_2\sigma}\mathbf{P}(B_1^{res} > (1-\rho_1)t), \quad t \to \infty. \tag{4.3.6}$$

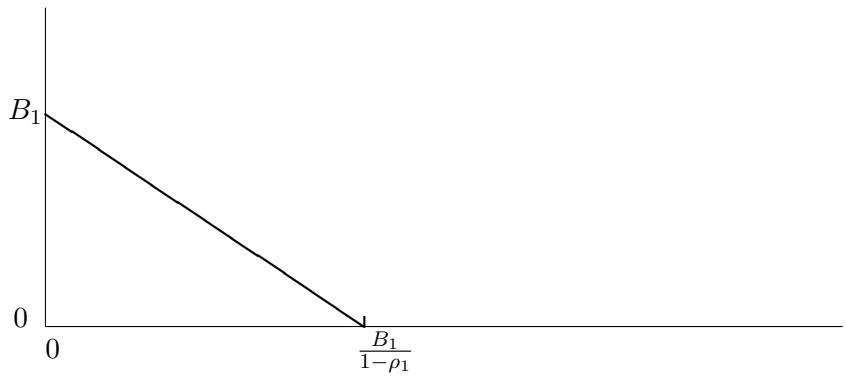

Figure 4.1: Evolution of the workload at Q_1.

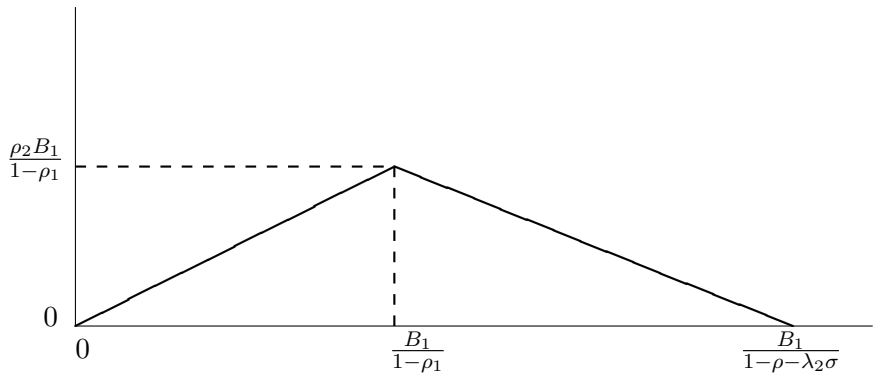

Figure 4.2: Evolution of the workload at Q_2.

Suppose a customer with a large service time B_1 enters Q_1 in steady state at time 0. Assume that the total workloads at both queues are very small compared to B_1. So at time 0 the workload at Q_1 is roughly B_1 and the workload at Q_2 is roughly 0. The workload at Q_1 decreases at rate $1-\rho_1 > 0$ until it becomes 0 at time $\frac{B_1}{1-\rho_1}$, see Figure 4.1.

Now we consider the workload at Q_2. During the time interval $(0, \frac{B_1}{1-\rho_1})$, the server stays at Q_1. Therefore, the workload at Q_2 increases at rate ρ_2. Notice that: When Q_2 is not empty, the long term fraction that the server stays at Q_2 is $\frac{\rho_2}{\rho_2+\lambda_2\sigma}(1-\rho_1)$ because the server incurs a vacation time (which

is the sum of the switchover time and the period that the server stays at Q_1) for every service provided at Q_2. Hence, after time $\frac{B_1}{1-\rho_1}$, the service speed at Q_2 is $\frac{\rho_2}{\lambda_2\sigma+\rho_2}(1-\rho_1)$. The workload decreases at rate $\frac{\rho_2}{\lambda_2\sigma+\rho_2}(1-\rho-\lambda_2\sigma) > 0$ until time $\frac{B_1}{1-\rho-\lambda_2\sigma}$. After time $\frac{B_1}{1-\rho-\lambda_2\sigma}$, the effect of the customer with the large service time B_1 has disappeared, see Figure 4.2.

Suppose we observe the system at time y ($y \geq 0$). The virtual waiting time at Q_1 is large, i.e., $W_1 > t$ (t large), because at time 0 a customer with a large service time B_1 entered Q_1. The arrival rate of customers at Q_1 is λ_1. Consider Figure 4.1. In order to make $W_1 > t$, it is necessary to require $0 < (1-\rho_1)y < B_1 - t$. So,

$$\begin{aligned}
\mathbf{P}(W_1 > t) &\approx \int_{y=0}^{\infty} \mathbf{P}(B_1 > (1-\rho_1)y + t)\lambda_1 \mathrm{d}y \\
&= \frac{\lambda_1}{1-\rho_1} \int_{y=t}^{\infty} \mathbf{P}(B_1 > y)\mathrm{d}y \\
&= \frac{\rho_1}{1-\rho_1} \mathbf{P}(B_1^{res} > t),
\end{aligned}$$

which coincides with (4.3.5).

Now consider Figure 4.2. For the waiting time at Q_2 to become large, there are two possibilities:

1. $0 < y < \frac{B_1}{1-\rho_1}$. Note that the service speed is $\frac{\rho_2}{\rho_2+\lambda_2\sigma}(1-\rho_1)$ instead of 1. In this case, the waiting time can be represented in terms of y as
$$W_2 = \frac{B_1}{1-\rho_1} - y + \frac{(\rho_2+\lambda_2\sigma)y}{1-\rho_1} = \frac{B_1}{1-\rho_1} - \frac{1-\rho-\lambda_2\sigma}{1-\rho_1}y.$$
2. $\frac{B_1}{1-\rho_1} < y < \frac{B_1}{1-\rho-\lambda_2\sigma}$. In this case, the waiting time W_2 is related to y as
$$W_2 = \frac{B_1}{1-\rho_1} - \frac{1-\rho-\lambda_2\sigma}{1-\rho_1}y.$$

In both scenarios, note that when the customer arrives $\mathrm{d}y$ (here $\mathrm{d}y$ stands for a small positive number) time units later, the waiting time is reduced by

$$\mathrm{d}y - \frac{\rho_2 \mathrm{d}y}{\frac{\rho_2}{\rho_2+\lambda_2\sigma}(1-\rho_1)} = \frac{1-\rho-\lambda_2\sigma}{1-\rho_1}\mathrm{d}y,$$

which explains why the waiting time behaves the same way in both cases. So in order to get $W_2 > t$, we need $B_1 > (1-\rho_1)t + (1-\rho-\lambda_2\sigma)y$. The tail behavior of the waiting time distribution is given thus by

$$\mathbf{P}(W_2 > t)$$

$$\begin{aligned}
&\approx \int_{y=0}^{\infty} \mathbf{P}(B_1 > (1-\rho-\lambda_2\sigma)y + (1-\rho_1)t)\lambda_1 \mathrm{d}y \\
&= \frac{\lambda_1}{1-\rho-\lambda_2\sigma} \int_{y=(1-\rho_1)t}^{\infty} \mathbf{P}(B_1 > y)\mathrm{d}y \\
&= \frac{\rho_1}{1-\rho-\lambda_2\sigma} \mathbf{P}(B_1^{res} > (1-\rho_1)t),
\end{aligned}$$

which coincides with (4.3.6).

Remark 4.3.1 Ibe [64] studies a more general model, a K-station mixed polling system in which station 1 is served exhaustively and stations $2, ..., K$ are served according to the 1-limited policy. He indicates a way to calculate the mean waiting times. Suppose at least one of the service and/or switchover times is regularly varying. By using the above heuristic arguments, it is not difficult to obtain the waiting time asymptotics for this K-station polling model.

4.4 A heavy-traffic limit theorem

This section is devoted to a heavy-traffic limit theorem for the waiting time distribution at Q_2 when at least one of the service and/or switchover times does not have a finite second moment.

In the sequel, we assume $1 < \nu < 2$, i.e., indeed at least one of the service and/or switchover times does not have a finite second moment. Then (4.3.1) and (4.3.2) reduce to, for $k = 1, 2$,

$$\begin{aligned}
\sigma_k(f(s)) &= 1 - \frac{\sigma_k}{1-\rho_1}s + \left(\frac{\lambda_1\beta_{1,\nu}\sigma_k}{(1-\rho_1)^{\nu+1}} + \frac{\sigma_{k,\nu}}{(1-\rho_1)^{\nu}}\right) s^{\nu}L(1/s) \\
&\quad + \mathrm{o}(s^{\nu}L(1/s)), \qquad (4.4.1) \\
\beta_2(f(s)) &= 1 - \frac{\beta_2}{1-\rho_1}s + \left(\frac{\lambda_1\beta_{1,\nu}\beta_2}{(1-\rho_1)^{\nu+1}} + \frac{\beta_{2,\nu}}{(1-\rho_1)^{\nu}}\right) s^{\nu}L(1/s) \\
&\quad + \mathrm{o}(s^{\nu}L(1/s)). \qquad (4.4.2)
\end{aligned}$$

Just like the priority queue which we discussed in Chapter 3, it is easy to see that the waiting time W_1 at Q_1 is not subject to heavy traffic when $\rho + \lambda_2\sigma \uparrow 1$ unless $\rho_1 \uparrow 1$. In the following we derive a heavy-traffic limit theorem for the waiting time W_2 at Q_2.

Theorem 4.4.1 *If Assumption 4.2.1 holds with $1 < \nu < 2$, then the contracted waiting time $\Delta(\lambda_1, \lambda_2)W_2$ at Q_2 converges in distribution for $\rho + \lambda_2\sigma \uparrow$*

1. *The limit distribution function $R_{\nu-1}(t)$ has the following LST:*

$$\int_0^\infty e^{-st}\mathrm{d}R_{\nu-1}(t) = \frac{1}{1+s^{\nu-1}}, \quad s>0, \tag{4.4.3}$$

and the coefficient of contraction $\Delta(\lambda_1,\lambda_2)$ is the unique root of the following equation

$$x^{\nu-1}L(1/x) = -\frac{(1-\rho_1)^{\nu-1}}{\Gamma(1-\nu)(\lambda_1 b_1+\lambda_2 s_1+\lambda_2 s_2+\lambda_2 b_2)}(1-\rho-\lambda_2\sigma), \tag{4.4.4}$$

with the property that $\Delta(\lambda_1,\lambda_2)\downarrow 0$ for $\rho+\lambda_2\sigma\uparrow 1$.

Proof. We have

$$\begin{aligned}&\frac{1-\rho-\lambda_2\sigma}{(\lambda_2-s)\sigma}\left(\frac{\lambda_2-s-\lambda_2\beta_2(f(s))}{\lambda_2-s-\lambda_2\sigma_1(f(s))\sigma_2(f(s))\beta_2(f(s))}-1\right)\\ =\ &\frac{\lambda_2\beta_2(f(s))}{(\lambda_2-s)\sigma}\frac{1-\sigma_1(f(s))\sigma_2(f(s))}{s}\frac{1-\rho-\lambda_2\sigma}{1-\lambda_2\frac{1-\sigma_1(f(s))\sigma_2(f(s))\beta_2(f(s))}{s}}.\end{aligned} \tag{4.4.5}$$

For ease of notation, we define

$$H_1(s) := \sigma_1(s)\frac{\sigma_1(\lambda_2+\lambda_1(1-\eta_1(s)))}{\sigma_1(\lambda_2)}\sigma_2(s+\lambda_1(1-\eta_1(s))), \tag{4.4.6}$$

$$H_2(s) := \frac{\lambda_2\beta_2(f(s))}{(\lambda_2-s)\sigma}\frac{1-\sigma_1(f(s))\sigma_2(f(s))}{s}, \tag{4.4.7}$$

$$H_3(s) := \frac{1-\rho-\lambda_2\sigma}{1-\lambda_2\frac{1-\sigma_1(f(s))\sigma_2(f(s))\beta_2(f(s))}{s}}. \tag{4.4.8}$$

Inserting (4.4.5), ..., (4.4.8) into (4.2.2), we may rewrite $\omega_2(s)$ as

$$\omega_2(s) = H_1(s)H_2(s)H_3(s) + \frac{1-\rho-\lambda_2\sigma}{(\lambda_2-s)\sigma}(H_1(s)-1). \tag{4.4.9}$$

Let $\Delta(\lambda_1,\lambda_2)$ be the solution of Equation (4.4.4) with the property that $\Delta(\lambda_1,\lambda_2)\downarrow 0$ for $\rho+\lambda_2\sigma\uparrow 1$. As has been proved in Lemma 3.5.1, the solution $\Delta(\lambda_1,\lambda_2)$ with that property exists and is unique. To simplify the notation, we make the convention that Δ stands for $\Delta(\lambda_1,\lambda_2)$. It is observed that, for a small number $\delta>0$, there exists a large number $N>0$, such that for any $0<s<\delta$,

$$|H_1(s)-1| < Ns \quad \text{and} \quad \left|H_2(s)-\frac{1}{1-\rho_1}\right| < Ns$$

hold uniformly for $0 < \rho + \lambda_2\sigma < 1$. Therefore, we have

$$\lim_{\rho+\lambda_2\sigma\uparrow 1} H_1(\Delta s) = 1, \tag{4.4.10}$$

$$\lim_{\rho+\lambda_2\sigma\uparrow 1} H_2(\Delta s) = \frac{1}{1-\rho_1}. \tag{4.4.11}$$

By (4.4.9), (4.4.10) and (4.4.11), in order to show that

$$\lim_{\rho+\lambda_2\sigma\uparrow 1} \omega_2(\Delta s) = \frac{1}{1+s^{\nu-1}}, \tag{4.4.12}$$

it remains to prove that $\lim_{\rho+\lambda_2\sigma\uparrow 1} H_3(\Delta s) = \frac{1-\rho_1}{1+s^{\nu-1}}$. From (4.4.1) and (4.4.2), we may write

$$\begin{aligned}\frac{1-\sigma_1(f(s))\sigma_2(f(s))\beta_2(f(s))}{s} &= \frac{\sigma+\beta_2}{1-\rho_1} \\ &- \left(\frac{\lambda_1\beta_{1,\nu}(\sigma+\beta_2)}{(1-\rho_1)^{\nu+1}} + \frac{\sigma_{1,\nu}+\sigma_{2,\nu}+\beta_{2,\nu}}{(1-\rho_1)^{\nu}}\right) s^{\nu-1}L(1/s) + G(s)s,\end{aligned} \tag{4.4.13}$$

where $G(s)$ is a function of s. For simplicity, we omit the expression for $G(s)$ here. One can easily prove that there exist a large number $N > 0$ and a small number $\epsilon > 0$ such that for any $0 < s < \epsilon$,

$$|G(s)| < N.$$

Note that N is independent of $\rho + \lambda_2\sigma$. Thus, inserting (4.4.13) into (4.4.8) gives

$$\begin{aligned}H_3(s) &= (1-\rho_1)\left[1 + \frac{1-\rho_1}{1-\rho-\lambda_2\sigma}\left(\frac{\lambda_1\beta_{1,\nu}(\lambda_2\sigma+\rho_2)}{(1-\rho_1)^{\nu+1}}\right.\right. \\ &\left.\left. + \frac{\lambda_2\sigma_{1,\nu}+\lambda_2\sigma_{2,\nu}+\lambda_2\beta_{2,\nu}}{(1-\rho_1)^{\nu}}\right) s^{\nu-1}L(1/s) + \frac{\lambda_2(1-\rho_1)}{1-\rho-\lambda_2\sigma}G(s)s\right]^{-1}.\end{aligned} \tag{4.4.14}$$

Since in the neighborhood of the origin $G(s)$ is uniformly bounded for $0 < \rho + \lambda_2\sigma < 1$, it is easy to see that

$$\lim_{\rho+\lambda_2\sigma\uparrow 1} \frac{1-\rho_1}{1-\rho-\lambda_2\sigma} G(\Delta s)\Delta s = 0.$$

Therefore, replacing s in (4.4.14) by Δs, we get

$$\lim_{\rho+\lambda_2\sigma\uparrow 1} H_3(\Delta s) = \frac{1-\rho_1}{1+s^{\nu-1}}. \tag{4.4.15}$$

Let $R_{\nu-1}(t)$ denote the distribution which has LST $\dfrac{1}{1+s^{\nu-1}}$. Using the convergence theorem of Feller for Laplace-Stieltjes transforms, cf. [52], it follows from (4.4.12) that ΔW_2 converges in distribution and the limit distribution $R_{\nu-1}(t)$ satisfies (4.4.3). □

4.5 Application of the heavy-traffic limit theorem

In this section, we numerically test the accuracy of the approximation suggested by the heavy-traffic limit theorem. We conclude that this approximation is useful in some cases.

Theorem 4.4.1 suggests the following heavy-traffic approximation for the waiting time distribution $W_2(t)$ at Q_2: For $\rho + \lambda_2\sigma < 1$,

$$1 - W_2(t) = \mathbf{P}(W_2 > t) \approx 1 - R_{\nu-1}(\Delta(\lambda_1, \lambda_2)t), \quad t > 0,$$

where $\Delta(\lambda_1, \lambda_2)$ is specified by Equation (4.4.4). According to the heavy-traffic limit theorem, this approximation should perform very well when $\rho + \lambda_2\sigma$ is sufficiently close to 1. In the following, we follow the same procedure as in Section 3.8 to numerically investigate the accuracy of this heavy-traffic approximation for different values of ν and $\rho + \lambda_2\sigma$. Theorem 4.3.1 suggests the following asymptotic approximation for $1 - W_2(t)$:

$$1 - W_{RV}(t) := \min\{1, C_{RV} t^{1-\nu} L(t)\}$$

with

$$C_{RV} := \frac{1}{\nu-1}\left(\frac{\lambda_1 b_1 + \lambda_2(s_1+s_2+b_2)}{(1-\rho_1)^{\nu-1}(1-\rho-\lambda_2\sigma)} + \frac{s_1+s_2}{(1-\rho_1)^{\nu-1}\sigma}\right),$$

and Theorem 4.4.1 suggests the following heavy-traffic approximation:

$$1 - W_{HT}(t) := 1 - R_{\nu-1}(\Delta(\lambda_1, \lambda_2)t), \quad t > 0.$$

Suppose the service time distributions are of the form (3.8.3) and the switchover times are exponentially distributed. Hence, we have

$$1 - W_2(t) \sim 1 - W_{HT}(t) \sim 1 - W_{RV}(t), \quad t \to \infty.$$

We have tested the approximations $1 - W_{HT}(t)$ and $1 - W_{RV}(t)$ for a large number of parameter combinations. Tables 4.1-4.6 below display numerical results for the following 6 cases: (i) $\nu = 1.25$, $\rho + \lambda_2\sigma = 0.9$, (ii) $\nu = 1.25$, $\rho + \lambda_2\sigma = 0.5$, (iii) $\nu = 1.5$, $\rho + \lambda_2\sigma = 0.9$, (iv) $\nu = 1.5$, $\rho + \lambda_2\sigma = 0.5$, (v) $\nu = 1.75$, $\rho + \lambda_2\sigma = 0.9$, and (vi) $\nu = 1.75$, $\rho + \lambda_2\sigma = 0.5$. Again, we use the Fourier-series method for inverting transforms of probability distributions (cf. [2]) to compute $1 - W_2(t)$. Similar conclusions as in Section 3.8 can be made:

(i) When t is large, e.g., $t \geq 50000$, the heavy-traffic approximation $1 - W_{HT}(t)$ is very accurate in all cases; while the asymptotic approximation $1 - W_{RV}(t)$ performs much worse when ν is small.

(ii) The larger the value of $\rho+\lambda_2\sigma$, the better the heavy-traffic approximation $1 - W_{HT}(t)$ performs.

(iii) When ν is small and $\rho + \lambda_2\sigma$ is large, e.g., $\nu \leq 1.5$ and $\rho + \lambda_2\sigma \geq 0.9$, the heavy-traffic approximation $1 - W_{HT}(t)$ is very good even for small t.

(iv) When ν is large, e.g., $\nu \geq 1.75$, the heavy-traffic approximation $1 - W_{HT}(t)$ performs poorly for small t; it is not better than the asymptotic approximation $1 - W_{RV}(t)$.

Table 4.1: Approximations for the waiting time tails at Q_2.

$\nu = 1.25;\ \rho + \lambda_2\sigma = 0.9$ $\nu_1 = \nu_2 = 1.25;\ \sigma_1 = \sigma_2 = 0.05;\ \lambda_1 = \lambda_2 = 0.1475$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.886	0.925	4.40	1	12.88
2	0.878	0.912	3.89	1	13.94
5	0.863	0.891	3.30	1	15.88
10	0.849	0.873	2.93	1	17.85
20	0.831	0.853	2.60	1	20.33
50	0.803	0.821	2.26	1	24.52
100	0.778	0.794	2.04	1	28.52
200	0.750	0.764	1.86	1	33.38
500	0.707	0.719	1.67	1	41.37
1000	0.672	0.682	1.54	1	48.85
2000	0.634	0.643	1.42	1	57.82
5000	0.580	0.587	1.27	1	72.46
10000	0.537	0.544	1.17	1	57.82
20000	0.494	0.499	1.07	0.937	89.72
50000	0.437	0.441	0.94	0.745	70.68
100000	0.394	0.398	0.84	0.627	59.00
200000	0.353	0.356	0.75	0.527	49.26
500000	0.302	0.304	0.64	0.419	38.81

Table 4.2: Approximations for the waiting time tails at Q_2.

$\nu = 1.25;\ \rho + \lambda_2\sigma = 0.5$ $\nu_1 = \nu_2 = 1.25;\ \sigma_1 = \sigma_2 = 0.05;\ \lambda_1 = \lambda_2 = 0.082$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.462	0.551	19.34	1	116.50
2	0.441	0.507	14.95	0.966	118.97
5	0.406	0.448	10.34	0.768	88.99
10	0.376	0.405	7.71	0.646	71.80
20	0.343	0.363	5.74	0.543	58.25
50	0.299	0.310	3.92	0.432	44.57
100	0.266	0.274	2.99	0.363	36.64
200	0.234	0.240	2.33	0.305	30.26
500	0.196	0.200	1.73	0.243	23.63
1000	0.171	0.173	1.41	0.204	19.67
2000	0.148	0.149	1.17	0.172	16.41
5000	0.121	0.122	0.92	0.137	12.94
10000	0.104	0.104	0.78	0.115	10.83
20000	0.089	0.089	0.65	0.097	9.07
50000	0.072	0.072	0.52	0.077	7.18
100000	0.061	0.061	0.44	0.065	6.02
200000	0.052	0.052	0.37	0.054	5.05
500000	0.041	0.042	0.37	0.043	4.08

Table 4.3: Approximations for the waiting time tails at Q_2.

$\nu = 1.5;\ \rho + \lambda_2\sigma = 0.9$ $\nu_1 = \nu_2 = 1.5;\ \sigma_1 = \sigma_2 = 0.15;\ \lambda_1 = \lambda_2 = 0.3913$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.869	0.946	8.90	1	15.10
2	0.846	0.925	9.40	1	18.25
5	0.808	0.886	9.61	1	23.76
10	0.771	0.844	9.53	1	29.73
20	0.724	0.791	9.22	1	38.09
50	0.646	0.700	8.37	1	54.83
100	0.574	0.616	7.38	1	74.19
200	0.494	0.524	6.15	0.800	62.13
500	0.382	0.398	4.32	0.506	32.62
1000	0.300	0.310	3.01	0.358	19.11
2000	0.229	0.233	1.92	0.253	10.74
5000	0.153	0.154	0.95	0.160	4.73
10000	0.110	0.111	0.52	0.113	2.47
20000	0.079	0.079	0.29	0.080	1.28
50000	0.050	0.050	0.12	0.051	0.51
100000	0.036	0.036	0.08	0.036	0.28
200000	0.025	0.025	0.20	0.025	0.30
500000	0.016	0.016	-0.17	0.016	-0.13

Table 4.4: Approximations for the waiting time tails at Q_2.

$\nu = 1.5;\ \rho + \lambda_2\sigma = 0.5$ $\nu_1 = \nu_2 = 1.5;\ \sigma_1 = \sigma_2 = 0.15;\ \lambda_1 = \lambda_2 = 0.2174$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.420	0.611	45.37	1	138.10
2	0.368	0.518	40.73	0.784	112.90
5	0.301	0.393	30.43	0.496	64.72
10	0.249	0.305	22.35	0.351	40.85
20	0.199	0.229	15.18	0.248	24.79
50	0.140	0.151	8.09	0.157	11.99
100	0.104	0.109	4.64	0.111	6.59
200	0.076	0.078	2.53	0.078	3.50
500	0.049	0.049	1.08	0.050	1.46
1000	0.035	0.035	0.55	0.035	0.74
2000	0.025	0.025	0.28	0.025	0.38
5000	0.016	0.016	0.12	0.016	0.15
10000	0.011	0.011	0.06	0.011	0.08
20000	0.008	0.008	0.03	0.008	0.04
50000	0.005	0.005	0.03	0.005	0.03
100000	0.003	0.004	0.35	0.004	0.36
200000	0.002	0.002	-0.15	0.002	-0.15
500000	0.002	0.002	0.14	0.002	0.14

Table 4.5: Approximations for the waiting time tails at Q_2.

$\nu = 1.75;\ \rho + \lambda_2\sigma = 0.9$ $\nu_1 = \nu_2 = 1.75;\ \sigma_1 = \sigma_2 = 0.45;\ \lambda_1 = \lambda_2 = 0.5745$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.893	0.942	5.49	1	11.98
2	0.818	0.904	10.50	1	22.25
5	0.691	0.821	18.90	1	44.72
10	0.577	0.723	25.33	0.881	52.85
20	0.452	0.589	30.40	0.524	15.97
50	0.288	0.379	31.56	0.264	-8.53
100	0.184	0.232	25.92	0.157	-15.05
200	0.110	0.128	16.48	0.093	-15.15
500	0.053	0.056	6.47	0.047	-10.93
1000	0.030	0.031	2.89	0.028	-7.49
2000	0.017	0.018	1.29	0.017	-4.86
5000	0.009	0.009	0.47	0.008	-2.62
10000	0.005	0.005	0.55	0.005	-1.29
20000	0.003	0.003	0.01	0.003	-1.09
50000	0.001	0.001	0.00	0.001	-0.34
100000	0.001	0.001	0.00	0.001	-0.03
200000	0.001	0.001	0.00	0.001	0.00
500000	0.000	0.000	0.00	0.000	0.00

Table 4.6: Approximations for the waiting time tails at Q_2.

$\nu = 1.75$; $\rho + \lambda_2\sigma = 0.5$ $\nu_1 = \nu_2 = 1.75$; $\sigma_1 = \sigma_2 = 0.45$; $\lambda_1 = \lambda_2 = 0.3191$					
t	$1 - W_2(t)$	$1 - W_{HT}(t)$	$\%\text{error}_{HT}$	$1 - W_{RV}(t)$	$\%\text{error}_{RV}$
1	0.538	0.582	8.16	0.511	-5.05
2	0.327	0.423	29.41	0.304	-7.10
5	0.154	0.226	47.01	0.153	-0.64
10	0.092	0.124	34.36	0.091	-1.61
20	0.056	0.066	18.56	0.054	-3.33
50	0.028	0.030	6.86	0.027	-3.67
100	0.017	0.017	3.08	0.016	-3.03
200	0.010	0.010	1.36	0.010	-2.23
500	0.005	0.005	0.44	0.005	-1.35
1000	0.003	0.003	0.18	0.003	-0.88
2000	0.002	0.002	0.07	0.002	-0.56
5000	0.001	0.001	0.02	0.001	-0.29
10000	0.001	0.001	0.00	0.001	-0.15
20000	0.000	0.000	0.00	0.000	0.03
50000	0.000	0.000	0.00	0.000	0.00
100000	0.000	0.000	0.00	0.000	0.00
200000	0.000	0.000	0.00	0.000	0.00
500000	0.000	0.000	0.00	0.000	0.00

Chapter 5

Polling systems with gated or exhaustive service

5.1 Introduction

In this chapter we consider a cyclic polling system consisting of K ($K \geq 2$) queues. At each queue, the service policy is either gated or exhaustive (for a detailed model description, and the definitions of 'gated' and 'exhaustive' service, we refer to Section 4.1). We investigate the tail behavior of the waiting time distributions at the various queues in the case that at least one of the service and/or switchover time distributions has a regularly varying tail. This chapter is based on Boxma, Deng and Resing [26].

As has been stated for example by Eisenberg [49], Fricker and Jaibi [54] and Resing [90], the condition $\rho < 1$ is a necessary and sufficient condition for ergodicity of a cyclic polling system with gated or exhaustive service. From now on, we assume that this ergodicity condition is satisfied.

We consider the waiting time (and, briefly, workload) tail behavior for cyclic polling systems with Poisson arrivals, general independent service times, general independent switchover times, and the gated or exhaustive service discipline. At each queue, customers are served in the order of arrival. The main result in this chapter is the following. *If at least one of the service and/or switchover time distributions has a regularly varying tail of index $-\nu$ ($\nu > 1$) and the others have a lighter tail, then the waiting time distribution at each queue is regularly varying of index $1-\nu$, i.e.,*

$$\mathbf{P}(W_k > t) \sim \alpha_k t^{1-\nu} L(t), \quad t \to \infty,$$

for $\alpha_k > 0$, where W_k is the steady-state waiting time at the kth queue.

The rest of this chapter is organized as follows. Section 5.2 presents an explicit formula for the LST of the waiting time distribution given in [13, 14], which will be the starting point of the tail investigation. This distribution is expressed in terms of the generating functions of particular queue length distributions; those are discussed in Section 5.3. The main theorems are provided in Section 5.4, in which the tail behavior of the waiting time distribution is given under the assumption that at least one of the service and/or switchover time distributions is regularly varying. Section 5.5 is devoted to the proof of Theorem 5.4.1 which describes the tail behavior of the intervisit time distribution in the case of exhaustive service and the tail behavior of the cycle time distribution in the case of gated service. Section 5.6 summarizes the results in this chapter and gives some suggestions for further research. Finally, the Appendix gives some results on the first-moment matrix which play a key role in the proof of our main result.

5.2 The waiting time distribution

Let W_k denote the stationary waiting time of type-k customers, with distribution function $W_k(\cdot)$ and LST $\omega_k(\cdot)$. In this section we give an explicit formula for $\omega_k(s)$, as provided by Borst and Boxma [14] (see also [100]). Let $W_{k|M/G/1}$, with LST $\omega_{k|M/G/1}(\cdot)$, denote the waiting time of an arbitrary customer in the 'corresponding' isolated $M/G/1$ queue of Q_k (see the end of Section 4.2 for a definition of the 'corresponding' queue) and let $N_{k|I}$, with pgf (probability generating function) $n_{k|I}(\cdot)$, denote the queue length at Q_k at an arbitrary epoch in an intervisit period for Q_k. The formula for $\omega_k(s)$ is based on the following decomposition, cf. Keilson and Servi [65],

$$\omega_k(s) = \mathrm{E}[e^{-sW_k}] = \omega_{k|M/G/1}(s) n_{k|I}(1 - s/\lambda_k), \quad \mathrm{Re}\ s \geq 0. \tag{5.2.1}$$

By the Pollaczek-Khintchine formula, $\omega_{k|M/G/1}(s)$ is given by

$$\omega_{k|M/G/1}(s) = \mathrm{E}[e^{-sW_{k|M/G/1}}] = \frac{1-\rho_k}{1-\rho_k \frac{1-\beta_k(s)}{\beta_k s}}, \quad \mathrm{Re}\ s \geq 0.$$

Introduce X_k, the queue length at Q_k at the beginning of a visit to Q_k, and Y_k, the queue length at Q_k at the end of a visit to Q_k. Borst and Boxma [14] relate $n_{k|I}(1 - s/\lambda_k)$ to X_k and Y_k as follows:

$$n_{k|I}(1 - s/\lambda_k) = \mathrm{E}[(1 - s/\lambda_k)^{N_{k|I}}] = \frac{y_k(s) - x_k(s)}{s(\mathrm{E}X_k - \mathrm{E}Y_k)/\lambda_k}, \tag{5.2.2}$$

with

$$x_k(s) := \mathrm{E}[(1 - s/\lambda_k)^{X_k}], \qquad y_k(s) := \mathrm{E}[(1 - s/\lambda_k)^{Y_k}]. \tag{5.2.3}$$

To get a better understanding of the function $n_{k|I}(1 - s/\lambda_k)$, we introduce the following random variables. For $k = 1, ..., K$, let C_k denote the cycle time of Q_k, i.e. the time between two successive arrivals of the server to Q_k, D_k the station time of Q_k, i.e. the time between an arrival of the server to Q_k and the next departure of the server from Q_k, and I_k the intervisit time of Q_k, i.e. the time between the departure of the server from Q_k and his next arrival to Q_k. The distribution functions of C_k, D_k and I_k are denoted by $C_k(\cdot), D_k(\cdot), I_k(\cdot)$, respectively. Furthermore, let I_k^* be the residual intervisit time, with probability density function $\frac{1-I_k(t)}{\mathrm{E}I_k}$. As has been pointed out in [5], the ergodicity condition implies the stationarity of the cycle times, the station times and the intervisit times.

In the case of exhaustive service, by definition $Y_k = 0$ and thus $\mathrm{E}[y^{Y_k}]$ =1. On the other hand, the customers at the beginning of a visit must have arrived during the previous intervisit time. Since the arrival process at Q_k is a Poisson process with rate λ_k, the generating function of the distribution of the number of arrivals during the previous intervisit time is related to the LST of the intervisit time distribution function as follows:

$$\mathrm{E}[y^{X_k}] = \mathrm{E}[e^{-\lambda_k(1-y)I_k}].$$

Thus it follows from the above relation and (5.2.3) that $x_k(s) = \mathrm{E}[e^{-sI_k}]$, which in combination with (5.2.2) implies that $n_{k|I}(1 - s/\lambda_k) = \mathrm{E}[e^{-sI_k^*}]$, i.e., $n_{k|I}(1 - s/\lambda_k)$ is in fact the LST of the residual intervisit time. Or equivalently, $n_{k|I}(z) = \mathrm{E}[e^{-\lambda_k(1-z)I_k^*}]$, which may also easily be seen to hold from the definition of $N_{k|I}$. Furthermore, the following well-known result is implied by (5.2.1) (with $\overset{\mathrm{d}}{=}$ denoting equality in distribution):

$$W_k \overset{\mathrm{d}}{=} W_{k|M/G/1} + I_k^*, \tag{5.2.4}$$

where $W_{k|M/G/1}$ and I_k^* are independent.

In the case of gated service, using similar arguments as in [37], X_k and Y_k can be considered as the number of type-k arrivals during a time interval of length C_k and D_k, respectively, and thus $x_k(s) = \mathrm{E}[e^{-sC_k}]$ and $y_k(s) = \mathrm{E}[e^{-sD_k}]$. It is readily seen that

$$n_{k|I}(1 - s/\lambda_k) = \mathrm{E}[e^{-sU_k}], \tag{5.2.5}$$

for some nonnegative random variable U_k with density function $\frac{D_k(t)-C_k(t)}{\mathrm{E}C_k-\mathrm{E}D_k}$. Thus it follows from (5.2.1) that

$$W_k \stackrel{\mathrm{d}}{=} W_{k|M/G/1} + U_k,$$

where $W_{k|M/G/1}$ and U_k are independent. The probabilistic meaning of U_k is: $U_k \stackrel{\mathrm{d}}{=} D_k + I_k^*$, cf. [11] for the case of nonzero switchover times; it is the sum of the station time D_k and the subsequent excess intervisit time I_k^* (which depends on D_k).

5.3 Joint queue length distribution

In order to obtain an explicit expression for the LST $\omega_k(s)$ of the waiting time distribution, we need an expression for the generating functions of the queue lengths X_k and Y_k at the beginning and end, respectively, of a server visit to Q_k (see (5.2.1) and (5.2.2)). We first concentrate on an expression for $F_k(\mathbf{z})$ and $G_k(\mathbf{z})$, $\mathbf{z} = (z_1, ..., z_K)^T$, $|z_j| \leq 1$, $j = 1, ..., K$, the pgf of the *joint* queue length vector at visit beginning and visit completion epochs, respectively. Here, we follow the approach of Resing [90]. In fact, in [90] a more general class of service disciplines called Bernoulli-type service is considered, which contains the gated and exhaustive service disciplines. This class of service disciplines satisfies the following property.

Property 5.3.1 *If the server arrives at Q_k to find n_k customers there, then during the course of the server's visit, each of these n_k customers is effectively replaced in an i.i.d. manner by a random population having pgf $h_k(\mathbf{z})$.*

The gated and exhaustive service discipline both satisfy this property. In these cases the functions $h_k(\mathbf{z})$ are, respectively, given by

$$h_k(\mathbf{z}) = \beta_k(\sum_{j=1}^{K} \lambda_j(1 - z_j)),$$

which is the pgf of the joint distribution of the numbers of arrivals at all queues during one service time at Q_k, and

$$h_k(\mathbf{z}) = \eta_k(\sum_{j \neq k} \lambda_j(1 - z_j)), \tag{5.3.1}$$

where $\eta_k(\cdot)$ denotes the LST of the length of a busy period in an isolated $M/G/1$ queue with arrival rate λ_k and service time distribution $B_k(\cdot)$. In the

case of exhaustive service, the function $h_k(\mathbf{z})$ represents the pgf of the joint distribution of the numbers of arrivals at all other queues during a busy period of Q_k when this queue was in isolation.

In the remainder of this section, we consider the queue length pgf for the class of service disciplines that satisfy Property 5.3.1; in the next section we use these results for the waiting time asymptotics, then restricting ourself to gated and exhaustive service. It may be worthwhile to investigate whether/how the asymptotic results of the present chapter can be generalized to the case of that general class. For service disciplines satisfying Property 5.3.1, the pgf's $G_k(\mathbf{z})$ (queue length at departure epochs of the server from Q_k) can be nicely related to the pgf's $F_k(\mathbf{z})$ (queue length at arrival epochs of the server at Q_k), for $k = 1, ..., K$, by

$$G_k(\mathbf{z}) = F_k(z_1, ..., z_{k-1}, h_k(\mathbf{z}), z_{k+1}, ..., z_K). \tag{5.3.2}$$

Next, define for $|z_j| \leq 1, j = 1, ..., K$, the functions

$$\mathbf{f}(\mathbf{z}) := (f_1(\mathbf{z}), ..., f_K(\mathbf{z}))^T, \tag{5.3.3}$$

with

$$f_k(\mathbf{z}) := h_k(z_1, ..., z_k, f_{k+1}(\mathbf{z}), ..., f_K(\mathbf{z})), \tag{5.3.4}$$

and the iterates

$$\begin{aligned} \mathbf{f}^{(0)}(\mathbf{z}) &:= \mathbf{z}, \\ \mathbf{f}^{(i)}(\mathbf{z}) &:= \mathbf{f}(\mathbf{f}^{(i-1)}(\mathbf{z})), \quad i \geq 1. \end{aligned}$$

In the following we distinguish between the case with and the case without switchover times.

Zero switchover times

In the sequel we add a superscript 0 for the case of zero switchover times, in order to distinguish its quantities from those for the case with switchover times. The pgf's $F_k^0(\cdot)$ and $G_{k-1}^0(\cdot)$ are related by

$$F_k^0(\mathbf{z}) = G_{k-1}^0(\mathbf{z}), \quad \text{for } k = 2, ..., K, \tag{5.3.5}$$

$$F_1^0(\mathbf{z}) = G_K^0(\mathbf{z}) - F_1^0(\mathbf{0})(\sum_{j=1}^{K} \frac{\lambda_j}{\lambda}(1 - z_j)), \tag{5.3.6}$$

where $\mathbf{0}$ stands for the K-dimensional vector with all components equal to zero. Equation (5.3.6) is obtained by using the special convention that when

the system is empty at the start of a visit to Q_1, the next visit does not take place until a customer has arrived. In fact, $F_1^0(\mathbf{z})$ satisfies the functional equation

$$F_1^0(\mathbf{z}) = F_1^0(\mathbf{f}(\mathbf{z})) - F_1^0(\mathbf{0}) \sum_{k=1}^{K} \frac{\lambda_k}{\lambda}(1 - z_k),$$

the solution of which, after iteration, is given by

$$F_1^0(\mathbf{z}) = 1 - \frac{F_1^0(\mathbf{0})}{\lambda} \sum_{i=1}^{\infty} \sum_{k=1}^{K} \lambda_k (1 - f_k^{(i)}(\mathbf{z})), \tag{5.3.7}$$

with

$$F_1^0(\mathbf{0}) = \left[1 + \frac{1}{\lambda} \sum_{i=1}^{\infty} \sum_{k=1}^{K} \lambda_k (1 - f_k^{(i)}(\mathbf{0})) \right]^{-1}.$$

The infinite sum $\sum_{i=1}^{\infty} \sum_{k=1}^{K} \lambda_k (1 - f_k^{(i)}(\mathbf{0}))$ is convergent when the ergodicity condition is fulfilled. Once we know $F_1^0(\mathbf{z})$, we immediately get, by using (5.3.2) and (5.3.5), the pgf $F_k^0(\mathbf{z})$ of the joint queue length distribution at a visit beginning epoch,

$$F_k^0(\mathbf{z}) = F_{k-1}^0(z_1, ..., z_{k-2}, h_{k-1}(\mathbf{z}), z_k, ..., z_K), \quad \text{for } k = 2, ..., K.$$

Furthermore, by the Relations (5.3.5) and (5.3.6) we get an expression for $G_k^0(\mathbf{z})$,

$$\begin{aligned} G_k^0(\mathbf{z}) &= F_{k+1}^0(\mathbf{z}), \quad \text{for } k = 1, ..., K-1, \\ G_K^0(\mathbf{z}) &= F_1^0(\mathbf{z}) + \frac{F_1^0(0)}{\lambda} (\sum_{j=1}^{K} \lambda_j (1 - z_j)). \end{aligned}$$

Nonzero switchover times

In the case of nonzero switchover times, the following equations relate $F_k(\mathbf{z})$ to $G_{k-1}(\mathbf{z})$:

$$F_{(k \bmod K)+1}(\mathbf{z}) = G_k(\mathbf{z}) \sigma_k (\sum_{j=1}^{K} \lambda_j (1 - z_j)), \quad \text{for } k = 1, ..., K. \tag{5.3.8}$$

Together with Equation (5.3.2) this leads to the functional equation

$$F_1(\mathbf{z}) = F_1(\mathbf{f}(\mathbf{z})) g(\mathbf{z}), \tag{5.3.9}$$

with

$$g(\mathbf{z}) = \prod_{k=1}^{K} \sigma_k(\sum_{j=1}^{k} \lambda_j(1-z_j) + \sum_{j=k+1}^{K} \lambda_j(1-f_j(\mathbf{z}))).$$

The solution of (5.3.9) is given by

$$\begin{aligned} F_1(\mathbf{z}) &= \prod_{i=1}^{\infty} g(\mathbf{f}^{(i)}(\mathbf{z})) \\ &= \prod_{i=1}^{\infty}\prod_{k=1}^{K} \sigma_k(\sum_{j=1}^{k} \lambda_j(1-f_j^{(i)}(\mathbf{z})) + \sum_{j=k+1}^{K} \lambda_j(1-f_j^{(i+1)}(\mathbf{z}))). \end{aligned} \tag{5.3.10}$$

Again, the infinite product is convergent when the ergodicity condition is fulfilled.

To make the obtained queue length pgf expressions suitable for the analysis of the waiting time tail behavior, we have to slightly rewrite them (we want to move from pgf asymptotics near 1 to LST asymptotics near 0).
Put $\mathbf{r} := (r_1, ..., r_K)^T$, where $0 \le r_k \le \lambda_k$, and relate $\mathbf{z}$ to $\mathbf{r}$ by $\mathbf{z}(\mathbf{r}) = (1 - r_1/\lambda_1, ..., 1 - r_K/\lambda_K)^T$. If we define $\tilde{F}_k(\mathbf{r}) := F_k(\mathbf{z}(\mathbf{r}))$ and $\tilde{G}_k(\mathbf{r}) := G_k(\mathbf{z}(\mathbf{r}))$, then it follows from (5.3.2) that

$$\tilde{G}_k(\mathbf{r}) = \tilde{F}_k(r_1, ..., r_{k-1}, \tilde{h}_k(\mathbf{r}), r_{k+1}, ..., r_K), \tag{5.3.11}$$

with

$$\begin{aligned} \tilde{h}_k(\mathbf{r}) &:= \lambda_k(1 - h_k(\mathbf{z}(\mathbf{r}))) \\ &= \begin{cases} \lambda_k(1 - \beta_k(\sum_{j=1}^{K} r_j)), & \text{for gated service,} \\ \lambda_k(1 - \eta_k(\sum_{j\neq k} r_j)), & \text{for exhaustive service.} \end{cases} \end{aligned} \tag{5.3.12}$$

Furthermore, similarly as in (5.3.3) and (5.3.4) we define the functions

$$\tilde{\mathbf{f}}(\mathbf{r}) := (\tilde{f}_1(\mathbf{r}), ..., \tilde{f}_K(\mathbf{r}))^T,$$

with

$$\tilde{f}_k(\mathbf{r}) := \tilde{h}_k(r_1, ..., r_k, \tilde{f}_{k+1}(\mathbf{r}), ..., \tilde{f}_K(\mathbf{r})), \tag{5.3.13}$$

and the iterates

$$\begin{aligned} \tilde{\mathbf{f}}^{(0)}(\mathbf{r}) &:= \mathbf{r}, \\ \tilde{\mathbf{f}}^{(i)}(\mathbf{r}) &:= \tilde{\mathbf{f}}(\tilde{\mathbf{f}}^{(i-1)}(\mathbf{r})),\ i \ge 1. \end{aligned}$$

In the following we distinguish again between the cases of zero switchover times and nonzero switchover times.

Zero switchover times
The following equations relate $\tilde{F}_k^0(\mathbf{r})$ to $\tilde{G}_{k-1}^0(\mathbf{r})$. It follows from (5.3.5) and (5.3.6) that

$$\tilde{F}_k^0(\mathbf{r}) = \tilde{G}_{k-1}^0(\mathbf{r}), \quad \text{for } k = 2, ..., K, \tag{5.3.14}$$

$$\tilde{F}_1^0(\mathbf{r}) = \tilde{G}_K^0(\mathbf{r}) - \frac{\tilde{F}_1^0(\mathbf{\Lambda})}{\lambda} \sum_{j=1}^{K} r_j, \tag{5.3.15}$$

where $\mathbf{\Lambda} = (\lambda_1, ..., \lambda_K)$. Introduce, for $0 < r_k < \lambda_k$, $k = 1, ..., K$,

$$H(\mathbf{r}) = \sum_{i=1}^{\infty} \sum_{k=1}^{K} \tilde{f}_k^{(i)}(\mathbf{r}),$$

which is well-defined if the ergodicity condition is fulfilled. Then, by (5.3.7), we can write

$$\tilde{F}_1^0(\mathbf{r}) := F_1^0(\mathbf{z}) = 1 - \tilde{F}_1^0(\mathbf{\Lambda}) H(\mathbf{r})/\lambda. \tag{5.3.16}$$

By using (5.3.11), (5.3.14) and (5.3.15), one can derive expressions for $\tilde{F}_k^0(\mathbf{r})$ and $\tilde{G}_k^0(\mathbf{r})$, $k = 1, ..., K$.

Nonzero switchover times
It follows from (5.3.8) that

$$\tilde{F}_{(k \bmod K)+1}(\mathbf{r}) = \tilde{G}_k(\mathbf{r}) \sigma_k(\sum_{j=1}^{K} r_j), \quad \text{for } k = 1, ..., K. \tag{5.3.17}$$

Put

$$\tilde{g}(\mathbf{r}) := \prod_{k=1}^{K} \sigma_k(\sum_{j=1}^{k} r_j + \sum_{j=k+1}^{K} \tilde{f}_j(\mathbf{r})).$$

Replacing $\mathbf{z}(\mathbf{r}) = (1 - r_1/\lambda_1, ..., 1 - r_K/\lambda_K)$ into (5.3.10), we obtain

$$\begin{aligned} \tilde{F}_1(\mathbf{r}) &= \prod_{i=1}^{\infty} \tilde{g}(\tilde{\mathbf{f}}^{(i)}(\mathbf{r})) \\ &= \prod_{i=1}^{\infty} \prod_{k=1}^{K} \sigma_k(\sum_{j=1}^{k} \tilde{f}_j^{(i)}(\mathbf{r}) + \sum_{j=k+1}^{K} \tilde{f}_j^{(i+1)}(\mathbf{r})), \end{aligned} \tag{5.3.18}$$

the infinite product being convergent for $0 \leq r_k \leq \lambda_k$, $k = 1, ..., K$, when the ergodicity condition is fulfilled. By using (5.3.11), (5.3.17), one can immediately derive expressions for $\tilde{G}_1(\mathbf{r})$, $\tilde{G}_k(\mathbf{r}), \tilde{F}_k(\mathbf{r})$ $(k = 2, ..., K)$.

Marginal queue length pgf
As a final step towards deriving the waiting time LST (see (5.2.1) and (5.2.2)), we now obtain the pgf of the marginal distributions of the queue lengths X_k and Y_k at the beginning and end of a server visit to Q_k, respectively. For ease of notation, we define $\mathbf{e} := (1, ..., 1)^T$ and for $k = 1, ..., K$, $\mathbf{e}_k := (0, ..., 0, 1, 0, ..., 0)^T$ with the kth component being 1. Taking $\mathbf{r} = \mathbf{e}_k s$ in (5.3.16) for the case of zero switchover times (add a superscript "0") or in (5.3.18) for the case of nonzero switchover times, we get

$$x_k(s) := \mathrm{E}[(1 - s/\lambda_k)^{X_k}] = \tilde{F}_k(\mathbf{e}_k s), \tag{5.3.19}$$

$$y_k(s) := \mathrm{E}[(1 - s/\lambda_k)^{Y_k}] = \tilde{G}_k(\mathbf{e}_k s). \tag{5.3.20}$$

5.4 The main result

In this section we present our main result: *If at least one of the service and/or switchover times is regularly varying of index $-\nu$ ($\nu > 1$) and the other service and/or switchover times have a lighter tail, then the waiting time distribution at each queue is regularly varying of index $1 - \nu$.* As a by-product, we also show that the intervisit time distribution at Q_k $(k = 1, ..., K)$ in the case of exhaustive service and the cycle time and station time distributions at Q_k $(k = 1, ..., K)$ in the case of gated service are all regularly varying of index $-\nu$.

As pointed out in Section 5.2, W_k can be represented as the sum of two independent random variables $W_{k|M/G/1}$ and V_k where $V_k = I_k^*$ is the residual intervisit time in the case of exhaustive service and $V_k = U_k$ in the case of gated service where the LST of U_k is given by (5.2.5). The relation between $1 - W_{k|M/G/1}(t)$ and $1 - B_k(t)$ for $t \to \infty$ is already well-known if the residual service time has a subexponential tail (this contains the case of a regularly varying tail),

$$1 - W_{k|M/G/1}(t) \sim \frac{\lambda}{1 - \rho} \int_{x=t}^{\infty} (1 - B_k(x)) \mathrm{d}x, \quad t \to \infty, \tag{5.4.1}$$

cf. [85]. In the following we first investigate the tail behavior of the distribution of V_k by analyzing the asymptotic behavior of its LST $n_{k|I}(1 - s/\lambda_k)$ for $s \downarrow 0$, and we subsequently derive the tail behavior of $1 - W_k(t)$ for $t \to \infty$.

Without loss of generality, we only analyze the explicit expression (5.2.2) for $n_{k|I}(1-s/\lambda_k)$ for $k=1$. Combining (5.3.11), (5.3.19) and (5.3.20) yields

$$y_1(s) = x_1(\tilde{h}_1(\mathbf{e}_1 s)). \tag{5.4.2}$$

If the asymptotic behavior of $x_1(s)$ for $s \downarrow 0$ is known, then we can obtain the asymptotic behavior of $y_1(s)$ for $s \downarrow 0$ immediately by using Lemma 2.3.6.

Concerning the tail behavior of the service and switchover time distributions, we assume that Assumption 4.2.1 holds. In order to simplify the proof of Theorem 5.4.1 below in Section 5.5, we assume, without loss of generality, that $s^\nu L(1/s)$ (for $L(\cdot)$, see Assumption 4.2.1) is a non-decreasing function for $s > 0$.

Theorem 5.4.1 *If Relations (4.2.5) and (4.2.6) hold, then*

$$x_1(s) = \sum_{j=0}^{m} (-1)^j x_{1,j} s^j + (-1)^{m+1} x_{1,\nu} s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \tag{5.4.3}$$

where $x_{1,j} \geq 0$ for $j = 1, ..., m$ and $x_{1,\nu} \geq 0$. Moreover, $x_{1,\nu} = 0$ if and only if $\sum_{k=1}^{K} (\beta_{k,\nu} + \sigma_{k,\nu}) = 0$.

Proof. See Section 5.5. □

The next corollary, which follows immediately from Theorem 5.4.1 and Relation (5.4.2) by using Lemma 2.3.6, characterizes the asymptotic behavior of $y_1(s)$ for $s \downarrow 0$ in the gated case. Remember that $y_1(s) \equiv 1$ if the service discipline at Q_1 is exhaustive.

Corollary 5.4.1 *In the case of gated service at Q_1, if (4.2.5) and (4.2.6) hold, then*

$$y_1(s) = \sum_{j=0}^{m} (-1)^j y_{1,j} s^j + (-1)^{m+1} y_{1,\nu} s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \tag{5.4.4}$$

where $y_{1,j} \geq 0$ for $j = 1, ..., m$ and $y_{1,\nu} \geq 0$. Moreover, $y_{1,\nu} = x_{1,1}\beta_{1,\nu} + x_{1,\nu}\rho_1^\nu$.

It is now easy to give the asymptotic expansion of $n_{1|I}(1-s/\lambda_1)$ for $s \downarrow 0$.

Corollary 5.4.2 *If (4.2.5) and (4.2.6) hold, then*

$$n_{1|I}(1-s/\lambda_1) \quad = \quad \sum_{j=1}^{m-1} (-1)^j n_{1|I,j} s^j + (-1)^m n_{1|I,\nu} s^\nu L(1/s)$$

$$+\mathrm{o}(s^{\nu}L(1/s)), \tag{5.4.5}$$

where $n_{1|I,j} > 0$ *for* $j = 1, ..., m-1$ *and* $n_{1|I,\nu} \geq 0$. *Moreover, if* $\sum_{k=1}^{K}(\beta_{k,\nu} + \sigma_{k,\nu}) > 0$, *then* $n_{1|I,\nu} > 0$.

Proof. By (5.2.2), (5.4.3) and (5.4.4), (5.4.5) follows. As shown in Section 5.2, $n_{1|I}(1 - s/\lambda_1)$ is the LST of some nonnegative random variable. Thus $n_{1|I,j} > 0$, $n_{1|I,\nu} \geq 0$ in (5.4.5). Again by (5.2.2), $n_{1|I,\nu} = \lambda_1(x_{1,\nu} - y_{1,\nu})/(\mathrm{E}X_1 - \mathrm{E}Y_1)$ where $y_{1,\nu} = x_{1,1}\beta_{1,\nu} + x_{1,\nu}\rho_1^{\nu}$. By using Formula (5.5.5) from the next section for the case of zero switchover time or Formula (5.5.34) for the case of nonzero switchover time, we can prove that $n_{1|I,\nu} > 0$ if $\sum_{k=1}^{K}(\beta_{k,\nu} + \sigma_{k,\nu}) > 0$. □

Applying Lemma 2.3.4, the above results yield the following theorem on the relation between the tail behavior of the service and switchover time distribution and that of the intervisit time, cycle time, station time and waiting time distributions.

Theorem 5.4.2 *If Assumption 4.2.1 holds, then in the case of exhaustive service at* Q_1, *the tail behavior of the intervisit time and waiting time at* Q_1 *satisfies the following relations:*

$$1 - I_1(t) = [c_1 + \mathrm{o}(1)]t^{-\nu}L(t), \quad t \to \infty, \tag{5.4.6}$$

$$1 - W_1(t) = [c_2 + \mathrm{o}(1)]t^{1-\nu}L(t), \quad t \to \infty; \tag{5.4.7}$$

in the case of gated service at Q_1, *the tail behavior of (i) the cycle time, (ii) the station time, (iii)* U_1 *with LST given by (5.2.5) and (iv) the waiting time at* Q_1 *is respectively given by*

$$1 - C_1(t) = [c_3 + \mathrm{o}(1)]t^{-\nu}L(t), \quad t \to \infty, \tag{5.4.8}$$

$$1 - D_1(t) = [c_4 + \mathrm{o}(1)]t^{-\nu}L(t), \quad t \to \infty, \tag{5.4.9}$$

$$1 - U_1(t) = [c_5 + \mathrm{o}(1)]t^{1-\nu}L(t), \quad t \to \infty, \tag{5.4.10}$$

$$1 - W_1(t) = [c_6 + \mathrm{o}(1)]t^{1-\nu}L(t), \quad t \to \infty, \tag{5.4.11}$$

where the c_r *are nonnegative constants for* $r = 1, ..., 6$. *Moreover, if* $\sum_{k=1}^{K}(b_k + s_k) = 0$, *then* $c_r = 0$ *for* $r = 1, ..., 6$; *if* $\sum_{k=1}^{K}(b_k + s_k) > 0$, *then* $c_r > 0$ *for* $r = 1, ..., 6$.

Proof. In the case of exhaustive service at Q_1, applying Theorem 5.4.1 and Lemma 2.3.4, (5.4.6) follows immediately. Combining (5.2.4), (5.4.1), (5.4.6) and using Lemma 7.7 in [29] yields (5.4.7) where

$$c_2 = \frac{\lambda_1 b_1}{(1-\rho_1)(\nu-1)} + \frac{c_1}{\mathrm{E}I_1(\nu-1)}.$$

In the case of gated service at Q_1, Relations (5.4.8) and (5.4.9) follow immediately from Theorem 5.4.1, Corollary 5.4.1 and Lemma 2.3.4. Relation (5.4.10) follows from Corollary 5.4.2. Combining (5.2.4), (5.4.1), (5.4.6) and using Lemma 7.7 in [29] yields (5.4.11) where

$$c_6 = \frac{\lambda_1 b_1}{(1-\rho_1)(\nu-1)} + \frac{c_5}{\nu-1}.$$

It is easy to see that if $\sum_{k=1}^{K}(b_k + s_k) = 0$, then $c_r = 0$ for $r = 1, ..., 6$; if $\sum_{k=1}^{K}(b_k + s_k) > 0$, then $c_r > 0$ for $r = 1, ..., 6$. □

By symmetry, Theorem 5.4.1 and Corollaries 5.4.1, 5.4.2 hold for $k = 1, ..., K$. Thus Theorem 5.4.2 also holds for $k = 1, ..., K$.

Remark 5.4.1 In order to get explicit expressions for c_r in the above theorem in terms of b_k and s_k for $k = 1, ..., K$, one can refer to Relation (5.5.5) below in the case of zero switchover times or Relation (5.5.34) in the case of nonzero switchover times.

Remark 5.4.2 Consider the $M/G/1$ queue with repeated vacations. The server continues serving until the system has become empty, and then takes a vacation V. If the system is still empty after this vacation, then he takes another vacation, etc.; successive vacations are independent and identically distributed. Fuhrmann and Cooper [55] have proven the following decomposition result:

$$W_{with} \stackrel{\mathrm{d}}{=} W_{M/G/1} + V^*,$$

where W_{with} ($W_{M/G/1}$) denotes waiting time in the model with (without) vacations and V^* has the equilibrium (residual lifetime) distribution of V; $W_{M/G/1}$ and V^* are independent.

This vacation queue is a special case of the polling model with switchover times and exhaustive service; take $K = 1$. Asmussen et al. [7] have proven the following result for the $M/G/1$ vacation queue with residual vacation or residual service time distributions that belong to the class $\mathcal{S}$ of subexponential distributions (which contains the class of regularly varying distributions):
(i) If the equilibrium service time $S^* \in \mathcal{S}$ and if $\mathbf{P}(V^* > t) = \mathrm{o}(\mathbf{P}(S^* > t))$ as $t \to \infty$, then

$$\mathbf{P}(W_{with} > t) \sim \frac{\rho}{1-\rho}\mathbf{P}(S^* > t), \quad t \to \infty;$$

(ii) If the equilibrium vacation time $V^* \in \mathcal{S}$ and if $\mathbf{P}(S^* > t) = \mathrm{o}(\mathbf{P}(V^* > t))$ as $t \to \infty$, then

$$\mathbf{P}(W_{with} > t) \sim \mathbf{P}(V^* > t), \quad t \to \infty;$$

(iii) If the equilibrium service time $S^* \in S$ and if $\mathbf{P}(V^* > t) \sim c\mathbf{P}(S^* > t)$ as $t \to \infty$ for some $c \geq 0$, then

$$\mathbf{P}(W_{with} > t) \sim (c + \frac{\rho}{1-\rho})\mathbf{P}(S^* > t), \quad t \to \infty.$$

In the polling model one might also try to prove that the waiting time distribution is *subexponential* in the case of subexponential service and/or switchover time distributions. However, at least in the case of exhaustive service at some queue Q_k, this requires the solution of the following open problem (cf. [7]): Is the busy period distribution of $M/G/1$ queue Q_k in isolation subexponential, when its service time distribution is subexponential? (Notice that the busy period distribution of Q_k appears prominently in $h_k(\mathbf{z})$ and $\tilde{f}_k(\mathbf{r})$, cf. (5.3.1) and (5.3.13)). Regarding the busy period distribution in the $G/G/1$ queue with regularly varying service times, we refer to Zwart [117].

Remark 5.4.3 In the present chapter we have concentrated on the tails of the waiting time distributions. It is slightly easier to study the tail behavior of the total workload distribution in a polling system. Boxma and Groenendijk [30] (cf. also Boxma [20] for generalizations) have proven the following workload decomposition for a broad category of multiclass queueing systems with Poisson arrivals and server vacations – a category which includes cyclic polling systems with switchover times:

$$U \stackrel{\mathrm{d}}{=} U_{M/G/1} + Z,$$

$U_{M/G/1}$ and Z being independent. Here U is the steady-state workload in the system, $U_{M/G/1}$ is the steady-state workload in the corresponding $M/G/1$ queue to which the multiclass system reduces when there are no switchover times, and Z is the steady-state workload at an arbitrary time during a vacation. Takagi et al. [104] provide an expression for the LST of the distribution of Z in the case of either exhaustive or gated service at all queues. Using that expression and the above decomposition result, one can apply the technique in Section 5.5 of this chapter to obtain similar tail behavior results for the workload as for the individual waiting times.

Remark 5.4.4 In the $M/G/1$ FCFS queue, if the service time distribution is regularly varying of index $-\nu$ ($\nu > 1$), then the waiting time distribution is regularly varying of index $1 - \nu$. However, the $M/G/1$ queue with the LCFS preemptive resume discipline has the attractive feature that the waiting time distribution is regularly varying of index just $-\nu$ [23]. In the polling system

with a LCFS preemptive resume discipline *within a queue visit of the server*, customers may have to wait a residual cycle time (in the case of gated service) or a residual intervisit time (in the case of exhaustive service), and these are regularly varying of index $1-\nu$. Thus one cannot expect to get a 'better' index than $1-\nu$ by providing LCFS preemptive resume service within a queue visit of the server.

Note that the above discussions are only valid for polling systems with more than one queue. For the $M/G/1$ queue with repeated vacations (cf. Remark 5.4.2), the intervisit time is the vacation time, which is independent of the service time. From the observation that the customer may have to wait for a residual vacation time, we conclude that if the vacation time has a lighter tail than $t^{-\nu}$, then the LCFS preemptive resume discipline implies that the waiting time distribution has a lighter tail than $t^{1-\nu}$. Otherwise, the LCFS preemptive resume discipline does not lead to a 'better' tail behavior of the waiting time distribution.

5.5 Proof of Theorem 5.4.1

We treat the cases of zero and nonzero switchover times separately. We restrict ourself mainly to the case in which all queues are served according to the *same* discipline (either gated or exhaustive); the proofs require only minor adaptations in the case of mixtures of these disciplines.

1. Zero switchover times
Using (5.3.16) and (5.3.19) we have

$$x_1(s) = \tilde{F}_1^0(\mathbf{e}_1 s) = 1 - \tilde{F}_1^0(\mathbf{\Lambda})H(\mathbf{e}_1 s)/\lambda. \tag{5.5.1}$$

In the following we concentrate on determining the asymptotic behavior of $H(\mathbf{e}_1 s)$ for $s \downarrow 0$. We prove that

$$H(\mathbf{e}_1 s) = \sum_{j=1}^{m} H_{1,j} s^j + (-1)^m H_{1,\nu} s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \quad s \downarrow 0, \tag{5.5.2}$$

for some constants $H_{1,j}$ where $j = 1, ..., m$ and $H_{1,\nu} \geq 0$. The proof of Relation (5.5.2) is divided into three steps. In the first step, we construct a new function $P(\cdot)$ which has a similar structure as $H(\cdot)$. In the second step, we show that the asymptotic expansion of this function is given by

$$P(\mathbf{e}_1 s) = \sum_{j=1}^{m} P_{1,j} s^j + \mathrm{O}(s^{m+1}), \quad \text{for } s \downarrow 0. \tag{5.5.3}$$

Finally, in the third step we will show that

$$\lim_{s\downarrow 0} \frac{H(\mathbf{e}_1 s) - P(\mathbf{e}_1 s)}{s^\nu L(1/s)} = (-1)^m H_{1,\nu}, \tag{5.5.4}$$

for some nonnegative constant $H_{1,\nu}$. Clearly, Relation (5.5.2) follows by combining (5.5.3) and (5.5.4). Once we have proven (5.5.2), the proof of Theorem 5.4.1 is almost completed. Substituting (5.5.2) into (5.5.1) and noting that $x_1(s)$ is the LST of some nonnegative random variable (the cycle time if the service discipline at Q_1 is gated or the intervisit time if the service discipline at Q_1 is exhaustive) yields Formula (5.4.3) of Theorem 5.4.1, where

$$x_{1,1} = \tilde{F}_1^0(\Lambda) P_{1,1}/\lambda, \qquad x_{1,\nu} = \tilde{F}_1^0(\Lambda) H_{1,\nu}/\lambda, \tag{5.5.5}$$

$P_{1,1}$ and $H_{1,\nu}$ being given by (5.5.11) and (5.5.27), respectively.

Step 1: Similarly as we constructed the function $H(\cdot)$ in Section 5.3, we now construct the function $P(\cdot)$. So, define:

$$\xi_k(\mathbf{r}) := \begin{cases} \lambda_k \sum_{j=1}^m (-1)^{j+1} \beta_{k,j} (\sum_{i=1}^k r_i)^j, & \text{for gated service,} \\ \\ \lambda_k \sum_{j=1}^m (-1)^{j+1} \eta_{k,j} (\sum_{i\neq k} r_i)^j, & \text{for exhaustive service,} \end{cases} \tag{5.5.6}$$

(cf. (5.3.12), and notice that we take the first m terms in the righthand sides of (4.2.5) and (4.2.7) multiplied by λ_k). Furthermore, we define:

$$\mathbf{p}(\mathbf{r}) := (p_1(\mathbf{r}), ..., p_K(\mathbf{r}))^T, \tag{5.5.7}$$

with

$$p_k(\mathbf{r}) := \xi_k(r_1, \ldots, r_k, p_{k+1}(\mathbf{r}), \ldots, p_K(\mathbf{r})),$$

and the iterates

$$\begin{aligned} \mathbf{p}^{(0)}(\mathbf{r}) &:= \mathbf{r}, \\ \mathbf{p}^{(i)}(\mathbf{r}) &:= \mathbf{p}(\mathbf{p}^{(i-1)}(\mathbf{r})), \quad i \geq 1. \end{aligned}$$

The function $P(\cdot)$ is defined by

$$P(\mathbf{r}) := \sum_{i=1}^{\infty} \sum_{k=1}^{K} p_k^{(i)}(\mathbf{r}). \tag{5.5.8}$$

In Lemma 5.5.2 we prove that the infinite sum in (5.5.8) is well-defined. Before we can do that, we first need to prove Lemma 5.5.1. In the following we make the convention that $|\mathbf{v}| = (|v_1|, ..., |v_n|)^T$ where $\mathbf{v}$ is an n-dimensional vector and $\mathbf{v} \leq \mathbf{u}$ if and only if $v_k \leq u_k$ for all $k = 1, ..., n$. For the definition of $\tilde{\mathbf{M}}$, we refer to the Appendix.

Lemma 5.5.1 *There exists a* $\delta_1 > 0$ *such that* $\mathbf{p}(\mathbf{r}) \leq \tilde{\mathbf{M}}\mathbf{r}$ *for* $\mathbf{0} \leq \mathbf{r} \leq \delta_1 \mathbf{e}$.

Proof. For $k = 1, ..., K$, it is easy to check that

$$\frac{\mathrm{d}}{\mathrm{d}s}\xi_k(\mathbf{e}_1 s) \leq \begin{cases} \left[\frac{\mathrm{d}}{\mathrm{d}s}\lambda_k(1-\beta_k(s))\right]_{s=0} = \rho_k, & \text{for gated service,} \\ \left[\frac{\mathrm{d}}{\mathrm{d}s}\lambda_k(1-\eta_k(s))\right]_{s=0} = \dfrac{\rho_k}{1-\rho_k}, & \text{for exhaustive service,} \end{cases}$$

for $0 < s < \delta_1$ where δ_1 is some positive constant. Therefore, we have

$$\begin{aligned} p_k(\mathbf{r}) &= \xi_k(r_1, ..., r_k, p_{k+1}(\mathbf{r}), , ..., p_K(\mathbf{r})) \\ &\leq \begin{cases} \rho_k[r_1 + ... + r_k + p_{k+1}(\mathbf{r}) + ... + p_K(\mathbf{r})], \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for gated service,} \\ \dfrac{\rho_k}{1-\rho_k}[r_1 + ... + r_{k-1} + p_{k+1}(\mathbf{r}) + ... + p_K(\mathbf{r})], \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for exhaustive service.} \end{cases} \end{aligned}$$

Rewriting the above inequalities in terms of matrices, we obtain

$$\mathbf{p}(\mathbf{r}) \leq \mathbf{B}\mathbf{r} + \mathbf{A}\mathbf{p}(\mathbf{r}),$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are given by (5.7.4) and (5.7.5) in the Appendix, respectively. Then it follows from the fact that $(\mathbf{I} - \mathbf{A})^{-1}$ is a nonnegative matrix that

$$\mathbf{p}(\mathbf{r}) \leq (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r} = \tilde{\mathbf{M}}\mathbf{r}.$$

□

Now we are able to prove that the infinite sum in (5.5.8) is well-defined.

Lemma 5.5.2 *There exists a* $\delta_1 > 0$ *such that* $P(\mathbf{e}_1 s) < \infty$ *for* $0 < s < \delta_1$.

Proof. It follows from Lemma 5.5.1 that there exists a $\delta_1 > 0$ such that for $0 < s < \delta_1$,

$$\mathbf{p}(\mathbf{e}_1 s) \leq \tilde{\mathbf{M}}\mathbf{e}_1 s. \tag{5.5.9}$$

Iterating (5.5.9) leads to

$$\mathbf{p}^{(i)}(\mathbf{e}_1 s) \leq \tilde{\mathbf{M}}^i \mathbf{e}_1 s, \quad i = 1, 2, ...$$

Summing the above relations, we get

$$\sum_{i=1}^{\infty} \mathbf{p}^{(i)}(\mathbf{e}_1 s) \leq (\mathbf{I} - \tilde{\mathbf{M}})^{-1}\tilde{\mathbf{M}}\mathbf{e}_1 s,$$

which implies that

$$P(\mathbf{e}_1 s) = \sum_{k=1}^{K} \sum_{i=1}^{\infty} p_k^{(i)}(\mathbf{e}_1 s) \leq \mathbf{e}^T (\mathbf{I} - \tilde{\mathbf{M}})^{-1} \tilde{\mathbf{M}} \mathbf{e}_1 s < \infty. \tag{5.5.10}$$

□

Actually, dividing by s in (5.5.9) and taking the limit for $s \downarrow 0$, we obtain

$$\left[\frac{\mathrm{d}}{\mathrm{d}s} \mathbf{p}(\mathbf{e}_1 s) \right]_{s=0} = \tilde{\mathbf{M}} \mathbf{e}_1.$$

Equality is seen to hold because the first inequality in the proof of Lemma 5.5.1 also reduces to an equality for $s \downarrow 0$. By using similar arguments as in the proof of Lemma 5.5.2, it is easy to derive from (5.5.10) that

$$P_{1,1} = \mathbf{e}^T (\mathbf{I} - \tilde{\mathbf{M}})^{-1} \tilde{\mathbf{M}} \mathbf{e}_1. \tag{5.5.11}$$

This relation is used in (5.5.5).

Step 2: The asymptotic expansion (5.5.3) is proved in the following lemma.

Lemma 5.5.3 *The function $P(\mathbf{e}_1 s)$ defined by (5.5.8) has the following expansion in the neighborhood of the origin,*

$$P(\mathbf{e}_1 s) = \sum_{j=1}^{m} P_{1,j} s^j + \mathrm{O}(s^{m+1}), \quad \textit{for } s \downarrow 0. \tag{5.5.12}$$

Proof. First, we observe that for all $k = 1, ..., K$ and all $i = 1, 2, ...,$ the functions $p_k^{(i)}(\mathbf{e}_1 s)$ are polynomials in s, i.e.,

$$p_k^{(i)}(\mathbf{e}_1 s) = \sum_{j=1}^{n_k^{(i)}} p_{k,j}^{(i)} s^j,$$

where $n_k^{(i)} = m^{Ki-k+1}$. It remains to prove that

$$\sum_{i=1}^{\infty} \sum_{k=1}^{K} \sum_{j=1}^{n_k^{(i)}} |p_{k,j}^{(i)}| s^j < \infty, \tag{5.5.13}$$

for $0 \leq s \leq \delta_2$. Observe that, if Equation (5.5.13) holds, we can interchange the order of summation as follows,

$$P(\mathbf{e}_1 s) = \sum_{i=1}^{\infty} \sum_{k=1}^{K} \sum_{j=1}^{n_k^{(i)}} p_{k,j}^{(i)} s^j = \sum_{j=1}^{\infty} \left(\sum_{k=1}^{K} \sum_{\{i: n_k^{(i)} \geq j\}} p_{k,j}^{(i)} \right) s^j = \sum_{j=1}^{\infty} P_{1,j} s^j, \tag{5.5.14}$$

for $0 \leq s \leq \delta_2$. Therefore, the expansion (5.5.12) immediately follows from (5.5.14).

In order to prove (5.5.13), we define a function $\mathbf{q} : \mathbb{R}_K \mapsto \mathbb{R}_K$,

$$\begin{aligned} \mathbf{q}(\mathbf{r}) &:= (q_1(\mathbf{r}), ..., q_K(\mathbf{r}))^T, \\ q_k(\mathbf{r}) &:= -\xi_k(-r_1, ..., -r_k, -q_{k+1}(\mathbf{r}), ..., -q_K(\mathbf{r})), \end{aligned}$$

and its iterates

$$\begin{aligned} \mathbf{q}^{(0)}(\mathbf{r}) &:= \mathbf{r}, \\ \mathbf{q}^{(i)}(\mathbf{r}) &:= \mathbf{q}(\mathbf{q}^{(i-1)}(\mathbf{r})), \quad i \geq 1. \end{aligned}$$

Next we show that the infinite sum $\sum_{i=1}^{\infty} \sum_{k=1}^{K} q_k^{(i)}(\mathbf{r})$ converges in a neighborhood of the origin. From the definition of $\mathbf{q}(\mathbf{r})$, using similar arguments as in the proof of Lemma 5.5.1, it follows that for any $\epsilon > 0$, there exists a $\delta_1 > 0$ such that, for $\mathbf{0} \leq \mathbf{r} \leq \delta_1 \mathbf{e}$,

$$\mathbf{0} \leq \mathbf{q}(\mathbf{r}) \leq (1+\epsilon)\tilde{\mathbf{M}}\mathbf{r},$$

where the entries of $\tilde{\mathbf{M}}$ are given by (5.7.1). Let $a_{\max} < 1$ be the maximal eigenvalue of $\tilde{\mathbf{M}}$. If we take $\epsilon = (1/a_{\max} - 1)/2$, then the maximal eigenvalue of $(1+\epsilon)\tilde{\mathbf{M}}$ is also less than 1. Thus, applying similar arguments as in the proof of Lemma 5.5.2, it follows that

$$\sum_{i=1}^{\infty} \sum_{k=1}^{K} q_k^{(i)}(\mathbf{r}) < \infty, \tag{5.5.15}$$

for $\mathbf{0} \leq \mathbf{r} \leq \delta_2 \mathbf{e}$ for some $\delta_2 > 0$. Similarly as was observed for $p_k^{(i)}(\mathbf{e}_1 s)$, we see that also the functions $q_k^{(i)}(\mathbf{e}_1 s)$, for all k and i, are polynomials in s, i.e.,

$$q_k^{(i)}(\mathbf{e}_1 s) = \sum_{j=1}^{n_k^{(i)}} q_{k,j}^{(i)} s^j.$$

Furthermore, from the definition of $q_k^{(i)}(\mathbf{e}_1 s)$, it is easy to see that

$$|p_{k,j}^{(i)}| \leq q_{k,j}^{(i)}, \tag{5.5.16}$$

for $k = 1, ..., K$, $i = 1, 2, ...$ and $j = 1, ..., n_k^{(i)}$. So, finally, (5.5.13) follows from (5.5.15) and (5.5.16). □

Step 3: Having proven (5.5.3), we must now prove (5.5.4). For this we need the following lemma.

Lemma 5.5.4 *For any $\epsilon > 0$, there exists a $\delta_1 > 0$ such that for $\mathbf{0} \leq \mathbf{r}, \mathbf{u} \leq \delta_1 \mathbf{e}$,*

$$|\tilde{\mathbf{f}}(\mathbf{u}) - \mathbf{p}(\mathbf{r})| \leq (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{D} + \epsilon \mathbf{I})\mathbf{d} + \tilde{\mathbf{M}}|\mathbf{u} - \mathbf{r}|, \tag{5.5.17}$$

where $\mathbf{A}$ is given by (5.7.4) in the Appendix, and

$$\mathbf{D} = \begin{cases} \text{diag}(\dfrac{\lambda_1 \beta_{1,\nu}}{\rho_1^\nu}, ..., \dfrac{\lambda_K \beta_{K,\nu}}{\rho_K^\nu}), & \text{for gated service,} \\ \\ \text{diag}(\dfrac{\lambda_1 \beta_{1,\nu}}{(1-\rho_1)\rho_1^\nu}, ..., \dfrac{\lambda_K \beta_{K,\nu}}{(1-\rho_K)\rho_K^\nu}), & \text{for exhaustive service,} \end{cases} \tag{5.5.18}$$

$$\mathbf{d} = ((\phi_1)^\nu L(1/\phi_1), ..., (\phi_K)^\nu L(1/\phi_K))^T, \tag{5.5.19}$$

with ϕ_k being the kth component of $\tilde{\mathbf{M}}\mathbf{u}$ ($k = 1, ..., K$).

Proof. We only prove the case of gated service. By similar arguments, one can obtain the result for exhaustive service. For $k = 1, ..., K$, recall that $\tilde{h}_k(\cdot)$ and $\xi_k(\cdot)$ are defined by (5.3.12) and (5.5.6), respectively. Then we have, for $\mathbf{0} < \mathbf{u}, \mathbf{r} < \delta_1 \mathbf{e}$, where δ_1 is some positive constant,

$$\begin{aligned} & |\tilde{f}_k(\mathbf{u}) - p_k(\mathbf{r})| \\ \leq\ & |\tilde{h}_k(u_1, ..., u_k, \tilde{f}_{k+1}(\mathbf{u}), ..., \tilde{f}_K(\mathbf{u})) - \xi_k(u_1, ..., u_k, \tilde{f}_{k+1}(\mathbf{u}), ..., \tilde{f}_K(\mathbf{u}))| \\ & + |\xi_k(u_1, ..., u_k, \tilde{f}_{k+1}(\mathbf{u}), ..., \tilde{f}_K(\mathbf{u})) - \xi_k(r_1, ..., r_k, p_{k+1}(\mathbf{r}), ..., p_K(\mathbf{r}))| \\ \leq\ & (\frac{\lambda_k \beta_{k,\nu}}{\rho_k^\nu} + \epsilon)(\rho_k u_1 + ... + \rho_k u_k + \rho_k \tilde{f}_{k+1}(\mathbf{u}) + ... + \rho_k \tilde{f}_K(\mathbf{u}))^\nu \\ & L(1/(\rho_k u_1 + ... + \rho_k u_k + \rho_k \tilde{f}_{k+1}(\mathbf{u}) + ... + \rho_k \tilde{f}_K(\mathbf{u}))) + \rho_k |u_1 - r_1| \\ & + ... + \rho_k |u_k - r_k| + \rho_k |\tilde{f}_{k+1}(\mathbf{u}) - p_{k+1}(\mathbf{r})| + ... + \rho_k |\tilde{f}_K(\mathbf{u}) - p_K(\mathbf{r})|, \end{aligned} \tag{5.5.20}$$

where the last inequality in (5.5.20) follows from the fact that, cf. (4.2.5),

$$|1-\beta_k(s)-\sum_{j=1}^{m}(-1)^{j+1}\beta_{k,j}s^j| \le \beta_{k,\nu}s^\nu L(1/s), \quad 0<s<\delta,$$

δ being a positive constant. By similar arguments as in the proof of Lemma 5.5.1, one can easily prove that, for $\mathbf{0}<\mathbf{u}<\delta\mathbf{e}$,

$$\tilde{\mathbf{f}}(\mathbf{u}) \le \tilde{\mathbf{M}}\mathbf{u}.$$

Thus, it follows that

$$\begin{aligned}\mathbf{B}\mathbf{u}+\mathbf{A}\tilde{\mathbf{f}}(\mathbf{u}) \le \mathbf{B}\mathbf{u}+\mathbf{A}\tilde{\mathbf{M}}\mathbf{u} &= (\mathbf{B}+\mathbf{A}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B})\mathbf{u}\\ &= (\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{u}=\tilde{\mathbf{M}}\mathbf{u},\end{aligned}$$

which implies that

$$\rho_k u_1+...+\rho_k u_k+\rho_k\tilde{f}_{k+1}(\mathbf{u})+...+\rho_k\tilde{f}_K(\mathbf{u}) \le \phi_k. \tag{5.5.21}$$

Rewriting the inequality (5.5.20) in terms of matrices and combining with (5.5.21), we obtain

$$|\tilde{\mathbf{f}}(\mathbf{u})-\mathbf{p}(\mathbf{r})| \le (\mathbf{D}+\epsilon\mathbf{I})\mathbf{d}+\mathbf{B}|\mathbf{u}-\mathbf{r}|+\mathbf{A}|\tilde{\mathbf{f}}(\mathbf{u})-\mathbf{p}(\mathbf{r})|, \tag{5.5.22}$$

$\mathbf{D}$ and $\mathbf{d}$ being given by (5.5.18) and (5.5.19) respectively. Since $(\mathbf{I}-\mathbf{A})^{-1}$ is a nonnegative matrix, (5.5.17) immediately follows from (5.5.22). □

Lemma 5.5.5 *There exists a nonnegative constant $H_{1,\nu}$ such that*

$$\lim_{s\downarrow 0}\frac{H(\mathbf{e}_1 s)-P(\mathbf{e}_1 s)}{s^\nu L(1/s)}=(-1)^m H_{1,\nu}. \tag{5.5.23}$$

The constant $H_{1,\nu}=0$ if and only if $\sum_{k=1}^K \beta_{k,\nu}=0$.

Proof. To simplify the notation, denote by r_{ik} the kth component of the vector $\tilde{\mathbf{M}}^i\mathbf{e}_1$, put

$$\mathbf{v}_i(s):=((r_{i1}s)^\nu L(1/r_{i1}s),...,(r_{iK}s)^\nu L(1/r_{iK}s))^T, \tag{5.5.24}$$

and let $v_{ik}(s)$ denote the kth component of $\mathbf{v}_i(s)$ where $k=1,...,K$, $i=1,2,...$. By Lemma 5.5.4 it follows that

$$|\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s)-\mathbf{p}^{(\mathbf{i})}(\mathbf{e}_1 s)| \le (\mathbf{I}-\mathbf{A})^{-1}(\mathbf{D}+\epsilon\mathbf{I})\mathbf{v_i}(s)$$

$$+\tilde{\mathbf{M}}|\tilde{\mathbf{f}}^{(i-1)}(\mathbf{e}_1 s) - \mathbf{p}^{(i-1)}(\mathbf{e}_1 s)|,$$

for $i = 1, 2, ...$ Iterating the above relations, we get for $i = 1, 2, ...,$

$$|\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s) - \mathbf{p}^{(i)}(\mathbf{e}_1 s)| \leq \sum_{j=1}^{i} \tilde{\mathbf{M}}^{i-j}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{D} + \epsilon\mathbf{I})\mathbf{v}_j(s).$$

Summing the above inequalities yields

$$\begin{aligned}
& \frac{\sum_{i=1}^{\infty} |\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s) - \mathbf{p}^{(i)}(\mathbf{e}_1 s)|}{s^\nu L(1/s)} \\
\leq \; & \sum_{i=1}^{\infty}\sum_{j=1}^{i} \tilde{\mathbf{M}}^{i-j}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{D} + \epsilon\mathbf{I})\frac{\mathbf{v}_j(s)}{s^\nu L(1/s)} \\
= \; & \sum_{j=1}^{\infty}\sum_{i=j}^{\infty} \tilde{\mathbf{M}}^{i-j}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{D} + \epsilon\mathbf{I})\frac{\mathbf{v}_j(s)}{s^\nu L(1/s)} \\
= \; & (\mathbf{I} - \tilde{\mathbf{M}})^{-1}(\mathbf{I} - \mathbf{A})^{-1}(\mathbf{D} + \epsilon\mathbf{I})\sum_{j=1}^{\infty} \frac{\mathbf{v}_j(s)}{s^\nu L(1/s)}, \qquad (5.5.25)
\end{aligned}$$

where the last identity follows from (5.7.6) in the Appendix.

Next we prove that the infinite sum $\sum_{i=1}^{\infty} \mathbf{v}_i(s)/(s^\nu L(1/s))$ converges. By using Potter's theorem (cf. Theorem 1.5.6 in [12]), it follows from the fact that $\lim_{i\to\infty} r_{ik} = 0$ for $k = 1, ..., K$ that $\dfrac{r_{ik}^{\nu-1} L(1/r_{ik}s)}{L(1/s)}$ converges to 0 uniformly in s for $s > 0$ as $i \to \infty$. Thus there exists an N_0 such that for $i \geq N_0$, $k = 1, ..., K$,

$$\frac{v_{ik}(s)}{s^\nu L(1/s)} \leq r_{ik}.$$

From the definition of $v_{ik}(s)$ and r_{ik}, we have for $k = 1, ..., K$, $0 < s < \delta$ where δ is some positive constant,

$$\begin{aligned}
& \sum_{i=N_0}^{\infty}\sum_{k=1}^{K} \frac{v_{ik}(s)}{s^\nu L(1/s)} \leq \sum_{i=N_0}^{\infty}\sum_{k=1}^{K} r_{ik} \leq \sum_{i=1}^{\infty}\sum_{k=1}^{K} r_{ik} \\
= \; & \sum_{i=1}^{\infty} \mathbf{e}^T \tilde{\mathbf{M}}^k \mathbf{e}_1 = \mathbf{e}^T(\mathbf{I} - \tilde{\mathbf{M}})^{-1}\tilde{\mathbf{M}}\mathbf{e}_1 < \infty.
\end{aligned}$$

Hence, applying the Dominated Convergence Theorem, it follows that

$$\lim_{s\downarrow 0} \frac{H(\mathbf{e}_1 s) - P(\mathbf{e}_1 s)}{s^\nu L(1/s)}$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} \lim_{s\downarrow 0} \frac{\mathbf{e}^T(\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s) - \mathbf{p}^{(i)}(\mathbf{e}_1 s))}{s^\nu L(1/s)} \\
&= (-1)^m \mathbf{e}^T(\mathbf{I}-\tilde{\mathbf{M}})^{-1}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{D} \sum_{i=1}^{\infty} \lim_{s\downarrow 0} \frac{\mathbf{v}_i(s)}{s^\nu L(1/s)} \\
&= (-1)^m \mathbf{e}^T(\mathbf{I}-\tilde{\mathbf{M}})^{-1}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{D} \sum_{i=1}^{\infty} \mathbf{u}_i < \infty,
\end{aligned}
$$

where

$$\mathbf{u}_i := (r_{i1}^\nu, ..., r_{iK}^\nu), \tag{5.5.26}$$

and the last identity follows from Lemma 2.3.5. Put

$$H_{1,\nu} = \mathbf{e}^T(\mathbf{I}-\tilde{\mathbf{M}})^{-1}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{D} \sum_{i=1}^{\infty} \mathbf{u}_i, \tag{5.5.27}$$

and subsequently (5.5.23) follows. Noticing that $\mathbf{D} = \mathbf{0}$ if and only if $\sum_{k=1}^K \beta_{k,\nu} = 0$, we conclude that $H_{1,\nu} = 0$ if and only if $\sum_{k=1}^K \beta_{k,\nu} = 0$. □

2. Nonzero switchover times

Again we wish to prove (5.4.3) for $x_1(s)$. As shown in Section 5.3, $x_1(s) = \tilde{F}_1(\mathbf{e}_1 s)$ where $\tilde{F}_1(\cdot)$ is given by (5.3.18). Put

$$
\mathbf{C} := \begin{pmatrix} 1 & 0 & ... & 0 & 0 \\ 1 & 1 & ... & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & ... & 1 & 0 \\ 1 & 1 & ... & 1 & 1 \end{pmatrix}, \quad \mathbf{G} := \begin{pmatrix} 0 & 1 & ... & 1 & 1 \\ 0 & 0 & ... & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & 0 & 1 \\ 0 & 0 & ... & 0 & 0 \end{pmatrix},
$$

and subsequently define

$$
\begin{aligned}
\hat{\mathbf{f}}^{(i)}(\mathbf{r}) &:= \mathbf{C}\tilde{\mathbf{f}}^{(i)}(\mathbf{r}) + \mathbf{G}\tilde{\mathbf{f}}^{(i+1)}(\mathbf{r}), \quad i = 1, 2, ..., \\
\hat{F}(\mathbf{r}) &:= \sum_{i=1}^{\infty}\sum_{k=1}^{K} \ln(\sigma_k(\hat{f}_k^{(i)}(\mathbf{r}))),
\end{aligned} \tag{5.5.28}
$$

with $\hat{f}_k^{(i)}(\mathbf{r})$ being the kth component of $\hat{\mathbf{f}}^{(i)}(\mathbf{r})$ for $k = 1, ..., K$. So we may rewrite (5.3.18) as

$$\tilde{F}(\mathbf{r}) = \exp(\hat{F}(\mathbf{r})), \tag{5.5.29}$$

where $\hat{F}(\mathbf{r})$ is given by (5.5.28). To prove (5.4.3), it is sufficient to show that

$$\hat{F}(\mathbf{e}_1 s) = \sum_{j=1}^{m} \hat{F}_{1,j} s^j + (-1)^{m+1} \hat{F}_{1,\nu} s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \quad s \downarrow 0. \quad (5.5.30)$$

We shall use similar arguments as in the proof for zero switchover times to obtain (5.5.30). We have again divided the proof into three steps. In the first step, we construct a new function $\hat{P}(\mathbf{r})$ which has a similar structure as $\hat{F}(\mathbf{r})$. In the second step, we shall use similar arguments as in the proof of the results for zero switchover times to show that

$$\hat{P}(\mathbf{e}_1 s) = \sum_{j=1}^{m} \hat{P}_{1,j} s^j + (-1)^{m+1} \hat{P}_{1,\nu} s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \quad (5.5.31)$$

where $\hat{P}_{1,j}$ are some constants for $j = 1, ..., m$ and $\hat{P}_{1,\nu} \geq 0$. Moreover, $\hat{P}_{1,\nu} = 0$ if and only if $\sum_{k=1}^{K} \sigma_{k,\nu} = 0$. In the third step, we verify that

$$\lim_{s \downarrow 0} \frac{\hat{F}(\mathbf{e}_1 s) - \hat{P}(\mathbf{e}_1 s)}{s^\nu L(1/s)} = (-1)^{m+1} g_{1,\nu}, \quad (5.5.32)$$

where $g_{1,\nu} \geq 0$. Moreover, $g_{1,\nu} = 0$ if and only if $\sum_{k=1}^{K} (\beta_{k,\nu} + \sigma_{k,\nu}) = 0$. Obviously, combining (5.5.31) and (5.5.32) yields (5.5.30) where

$$\hat{F}_{1,\nu} = \hat{P}_{1,\nu} + g_{1,\nu}, \quad (5.5.33)$$

with $\hat{P}_{1,\nu}$ and $g_{1,\nu}$ being given by (5.5.38) and (5.5.39), respectively. Then applying Lemma 2.3.6, and noting that $x_1(s) = \tilde{F}_1(\mathbf{e}_1 s)$ is the LST of some nonnegative random variable, Formula (5.4.3) of Theorem 5.4.1 follows from (5.5.29) and (5.5.30) with

$$x_{1,1} = \hat{P}_{1,1}, \quad x_{1,\nu} = \hat{F}_{1,\nu}, \quad (5.5.34)$$

$\hat{P}_{1,1}$ and $\hat{F}_{1,\nu}$ being given by (5.5.36) and (5.5.33), respectively.

Step 1: Define:

$$\begin{aligned} \hat{\mathbf{p}}^{(i)}(\mathbf{r}) &:= \mathbf{C}\mathbf{p}^{(i)}(\mathbf{r}) + \mathbf{G}\mathbf{p}^{(i+1)}(\mathbf{r}), \quad i = 1, 2, ..., \quad (5.5.35) \\ \hat{P}(\mathbf{r}) &:= \sum_{i=1}^{\infty} \sum_{k=1}^{K} \ln(\sigma_k(\hat{p}_k^{(i)}(\mathbf{r}))), \end{aligned}$$

$\mathbf{p}^{(i)}(\cdot)$ being given by (5.5.7) and $\hat{p}_k^{(i)}(\cdot)$ denoting the kth component of $\hat{\mathbf{p}}^{(i)}(\cdot)$. By Lemma 5.5.1, we have

$$\begin{aligned}\hat{\mathbf{p}}^{(i)}(\mathbf{e}_1 s) &= \mathbf{C}\mathbf{p}^{(i)}(\mathbf{e}_1 s) + \mathbf{G}\mathbf{p}^{(i+1)}(\mathbf{e}_1 s)\\ &\leq \mathbf{C}\tilde{\mathbf{M}}^i\mathbf{e}_1 s + \mathbf{G}\tilde{\mathbf{M}}^{i+1}\mathbf{e}_1 s\\ &= (\mathbf{C}+\mathbf{G}\tilde{\mathbf{M}})\tilde{\mathbf{M}}^i\mathbf{e}_1 s,\end{aligned}$$

where the above inequality follows from (5.5.9). Using the fact that $\ln(\sigma_k(x)) < \sigma_k x$ for small x, one can easily prove that $\hat{P}(\mathbf{r})$ is well-defined in some neighborhood of the origin. It is not difficult to see that

$$\lim_{s\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}s}\hat{\mathbf{p}}^{(i)}(\mathbf{e}_1 s) = (\mathbf{C}+\mathbf{G}\tilde{\mathbf{M}})\tilde{\mathbf{M}}^i\mathbf{e}_1.$$

It follows that

$$\hat{P}_{1,1} = \mathbf{e}^T\mathbf{H}(\mathbf{C}+\mathbf{G}\tilde{\mathbf{M}})(\mathbf{I}-\tilde{\mathbf{M}})^{-1}\tilde{\mathbf{M}}\mathbf{e}_1. \tag{5.5.36}$$

Step 2: We prove (5.5.31) by using similar arguments as in the proof for the case with zero switchover times. We omit some of the details here. First, by using Lemma 2.3.6, we may write

$$\ln(\sigma_k(x)) = \sum_{j=1}^{m} a_{k,j}x^j + (-1)^{m+1}a_{k,\nu}x^\nu L(1/x) + \mathrm{o}(x^\nu L(1/x)), \quad x\downarrow 0,$$

with $a_{k,\nu} = \sigma_{k,\nu}$. Applying similar arguments as in the proof of Lemma 5.5.3, one can easily verify that

$$A(s) := \sum_{i=1}^{\infty}\sum_{k=1}^{K}\sum_{j=1}^{m} a_{k,j}(\hat{p}_k^{(i)}(\mathbf{e}_1 s))^j = \sum_{j=1}^{\infty} A_j s^j. \tag{5.5.37}$$

For any $\epsilon > 0$, there exists a $\delta > 0$ such that for $0 < s < \delta$,

$$\begin{aligned}\frac{|\hat{P}(\mathbf{e}_1 s) - A(s)|}{s^\nu L(1/s)} &\leq \sum_{i=1}^{\infty}\sum_{k=1}^{K}\frac{(\sigma_{k,\nu}+\epsilon)(\hat{p}_k^{(i)}(\mathbf{e}_1 s))^\nu L(1/(\hat{p}_k^{(i)}(\mathbf{e}_1 s)))}{s^\nu L(1/s)}\\ &\leq \sum_{i=1}^{\infty}\sum_{k=1}^{K}\frac{(\sigma_{k,\nu}+\epsilon)(\alpha_{ik}s)^\nu L(1/\alpha_{ik}s)}{s^\nu L(1/s)} < \infty,\end{aligned}$$

α_{ik} denoting the kth component of the vector $(\mathbf{C}+\mathbf{G}\tilde{\mathbf{M}})\tilde{\mathbf{M}}^i\mathbf{e}_1$. By the Dominated Convergence Theorem, it can be shown that

$$\lim_{s\downarrow 0}\frac{\hat{P}(\mathbf{e}_1 s) - A(s)}{s^\nu L(1/s)} = (-1)^{m+1}\hat{P}_{1,\nu}, \tag{5.5.38}$$

with

$$\hat{P}_{1,\nu} = \sum_{i=1}^{\infty}\sum_{k=1}^{K} \sigma_{k,\nu}\alpha_{ik}^{\nu}.$$

Notice that $\hat{P}_{1,\nu} = 0$ if and only if $\sum_{k=1}^{K} \sigma_{k,\nu} = 0$. Combining (5.5.38) and (5.5.37) leads to (5.5.31).

Step 3: The proof of (5.5.32) is similar to that of Lemma 5.5.5. Here we omit some of the details. For ease of notation, define

$$\mathbf{H} := \mathrm{diag}(\sigma_1, ..., \sigma_K).$$

By the definitions of $\hat{F}(\mathbf{e}_1 s)$ and $\hat{P}(\mathbf{e}_1 s)$, we have

$$\begin{aligned}
& |\hat{F}(\mathbf{e}_1 s) - \hat{P}(\mathbf{e}_1 s)| \\
= & \sum_{i=1}^{\infty}\sum_{k=1}^{K} |\ln(\sigma_k(\hat{p}_k^{(i)}(\mathbf{e}_1 s))) - \ln(\sigma_k(\hat{f}_k^{(i)}(\mathbf{e}_1 s)))| \\
\leq & \sum_{i=1}^{\infty}\sum_{k=1}^{K} \sigma_k |\hat{p}_k^{(i)}(\mathbf{e}_1 s) - \hat{f}_k^{(i)}(\mathbf{e}_1 s)| \\
\leq & \sum_{i=1}^{\infty}\sum_{k=1}^{K} (\sigma_k \sum_{j=1}^{k} |\tilde{f}_k^{(i)}(\mathbf{e}_1 s) - p_k^{(i)}(\mathbf{e}_1 s)| \\
& + \sigma_k \sum_{j=k+1}^{K} |\tilde{f}_k^{(i+1)}(\mathbf{e}_1 s) - p_k^{(i+1)}(\mathbf{e}_1 s)|) \\
= & \sum_{i=1}^{\infty} \mathbf{e}^T \mathbf{H}(\mathbf{C}|\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s) - \mathbf{p}^{(i)}(\mathbf{e}_1 s)| + \mathbf{G}|\tilde{\mathbf{f}}^{(i+1)}(\mathbf{e}_1 s) - \mathbf{p}^{(i+1)}(\mathbf{e}_1 s)|) \\
\leq & \mathbf{e}^T \mathbf{H}(\mathbf{C} + \mathbf{G}) \sum_{i=1}^{\infty} |\tilde{\mathbf{f}}^{(i)}(\mathbf{e}_1 s) - \mathbf{p}^{(i)}(\mathbf{e}_1 s)|,
\end{aligned}$$

which in combination with (5.5.25) yields

$$\begin{aligned}
& \frac{|\hat{F}(\mathbf{e}_1 s) - \hat{P}(\mathbf{e}_1 s)|}{s^{\nu} L(1/s)} \\
\leq & \mathbf{e}^T \mathbf{H}(\mathbf{C} + \mathbf{G})(\mathbf{I} - \tilde{\mathbf{M}})^{-\mathbf{1}}(\mathbf{I} - \mathbf{A})^{-\mathbf{1}}(\mathbf{D} + \epsilon \mathbf{I}) \sum_{i=1}^{\infty} \frac{\mathbf{v}_j(s)}{s^{\nu} L(1/s)} < \infty,
\end{aligned}$$

where $\mathbf{v}_i(s)$ is defined by (5.5.24). Again, using the Dominated Convergence Theorem, we obtain

$$\lim_{s\downarrow 0} \frac{\hat{F}(\mathbf{e}_1 s) - \hat{P}(\mathbf{e}_1 s)}{s^\nu L(1/s)} = (-1)^{m+1} g_{1,\nu},$$

with

$$g_{1,\nu} = \mathbf{e}^T \mathbf{H}(\mathbf{C} + \mathbf{G}\tilde{\mathbf{M}})(\mathbf{I} - \tilde{\mathbf{M}})^{-1}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{D} \sum_{i=1}^{\infty} \mathbf{u}_i, \tag{5.5.39}$$

where $\mathbf{u}_i$ is given in (5.5.26). Notice that $\mathbf{H} = \mathbf{D} = \mathbf{0}$ if and only if $\sum_{k=1}^{K}(\beta_{k,\nu} + \sigma_{k,\nu}) = 0$, thus $g_{1,\nu} = 0$ if and only if $\sum_{k=1}^{K}(\beta_{k,\nu} + \sigma_{k,\nu}) = 0$.

5.6 Summary

In this chapter we have investigated the tail behavior of the waiting time distributions in cyclic polling systems with gated or exhaustive service. Under the assumption that at least one of the service and/or switchover time distributions has a regularly varying tail, the waiting time distributions at all queues are shown to be regularly varying of index one higher than the heaviest tail of the service and switchover time distributions. This result gives important insight into the effect of heavy-tailed service or switchover time distributions on the performance of a large class of polling systems. We expect the same result to be true for non-cyclic polling systems, and for a larger class of arrival processes and service disciplines. For the class of service disciplines satisfying Property 5.3.1, it may be possible to prove this along similar lines as in the present chapter. For almost all polling systems in which the service discipline in at least one queue does not satisfy Property 5.3.1, no explicit expression for the waiting time LSTs is known, so that the approach via Lemma 2.3.4 does not work. An exception is provided by the 2-queue polling system with exhaustive service at Q_1 and 1-limited service at Q_2, as studied in the previous chapter.

5.7 Appendix: On the first-moment matrix

Consider the mean matrix $\mathbf{M} = (m_{kj} : k, j = 1, ..., K)$, where

$$m_{kj} := \frac{\partial f_k}{\partial z_j}(1, ..., 1),$$

is the mean number of type-j customers that are direct descendants of a single type-k customer. As shown in [90], the matrix $\mathbf{M}$ plays an essential role in proving that $\rho < 1$ is sufficient for ergodicity in the case of gated or exhaustive service. In this appendix we shall derive some properties of the matrix $\tilde{\mathbf{M}} = (\tilde{m}_{kj} : k, j = 1, ..., K)$, where

$$\tilde{m}_{kj} := \frac{\partial \tilde{f}_k}{\partial r_j}(0, ..., 0). \tag{5.7.1}$$

The following lemma relates the eigenvalues and eigenvectors of $\mathbf{M}$ and $\tilde{\mathbf{M}}$.

Lemma 5.7.1 *The eigenvalues of* $\mathbf{M}$ *and* $\tilde{\mathbf{M}}$ *are identical. Moreover, if* $\mathbf{v} = (v_1, ..., v_K)^T$ *is a right eigenvector of* $\mathbf{M}$ *w.r.t. eigenvalue* a*, then* $\mathbf{u} = (\lambda_1 v_1, ..., \lambda_k v_K)^T$ *is a right eigenvector of* $\tilde{\mathbf{M}}$ *w.r.t.* a*.*

Proof. Using the fact that

$$\tilde{m}_{kj} = \frac{\lambda_k}{\lambda_j} m_{kj}, \tag{5.7.2}$$

which follows from the relation (see (5.3.4), (5.3.12) and (5.3.13)), we obtain

$$\tilde{f}_k(r) = \lambda_k(1 - f_k(z)).$$

□

Furthermore, we can derive an explicit formula for $\tilde{\mathbf{M}}$. It follows from (5.3.13) that

$$\tilde{\mathbf{M}} = \mathbf{B} + \mathbf{A}\tilde{\mathbf{M}}, \tag{5.7.3}$$

where

$$\mathbf{A} = \begin{pmatrix} 0 & \frac{\partial \tilde{h}_1}{\partial r_2}(\mathbf{0}) & ... & \frac{\partial \tilde{h}_1}{\partial r_{K-1}}(\mathbf{0}) & \frac{\partial \tilde{h}_1}{\partial r_K}(\mathbf{0}) \\ 0 & 0 & ... & \frac{\partial \tilde{h}_2}{\partial r_{K-1}}(\mathbf{0}) & \frac{\partial \tilde{h}_2}{\partial r_K}(\mathbf{0}) \\ \vdots & \vdots & ... & \vdots & \vdots \\ 0 & 0 & ... & 0 & \frac{\partial \tilde{h}_{K-1}}{\partial r_K}(\mathbf{0}) \\ 0 & 0 & ... & 0 & 0 \end{pmatrix}, \tag{5.7.4}$$

$$\mathbf{B} = \begin{pmatrix} \frac{\partial \tilde{h}_1}{\partial r_1}(\mathbf{0}) & 0 & ... & 0 & 0 \\ \frac{\partial \tilde{h}_2}{\partial r_1}(\mathbf{0}) & \frac{\partial \tilde{h}_2}{\partial r_2}(\mathbf{0}) & ... & 0 & 0 \\ \vdots & \vdots & ... & \vdots & \vdots \\ \frac{\partial \tilde{h}_{K-1}}{\partial r_1}(\mathbf{0}) & \frac{\partial \tilde{h}_{K-1}}{\partial r_2}(\mathbf{0}) & ... & \frac{\partial \tilde{h}_{K-1}}{\partial r_{K-1}}(\mathbf{0}) & 0 \\ \frac{\partial \tilde{h}_K}{\partial r_1}(\mathbf{0}) & \frac{\partial \tilde{h}_K}{\partial r_2}(\mathbf{0}) & ... & \frac{\partial \tilde{h}_K}{\partial r_{K-1}}(\mathbf{0}) & \frac{\partial \tilde{h}_K}{\partial r_K}(\mathbf{0}) \end{pmatrix}. \tag{5.7.5}$$

For the case of gated service at all queues, $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A}_{\text{gat}} = \begin{pmatrix} 0 & \rho_1 & \dots & \rho_1 & \rho_1 \\ 0 & 0 & \dots & \rho_2 & \rho_2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \rho_{K-1} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

$$\mathbf{B}_{\text{gat}} = \begin{pmatrix} \rho_1 & 0 & \dots & 0 & 0 \\ \rho_2 & \rho_2 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \rho_{K-1} & \rho_{K-1} & \cdots & \rho_{K-1} & 0 \\ \rho_K & \rho_K & \cdots & \rho_K & \rho_K \end{pmatrix},$$

and for the case of exhaustive service at all queues, $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A}_{\text{exh}} = \begin{pmatrix} 0 & \frac{\rho_1}{1-\rho_1} & \dots & \frac{\rho_1}{1-\rho_1} & \frac{\rho_1}{1-\rho_1} \\ 0 & 0 & \dots & \frac{\rho_2}{1-\rho_2} & \frac{\rho_2}{1-\rho_2} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \frac{\rho_{K-1}}{1-\rho_{K-1}} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

$$\mathbf{B}_{\text{exh}} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \frac{\rho_2}{1-\rho_2} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \frac{\rho_{K-1}}{1-\rho_{K-1}} & \frac{\rho_{K-1}}{1-\rho_{K-1}} & \dots & 0 & 0 \\ \frac{\rho_K}{1-\rho_K} & \frac{\rho_K}{1-\rho_K} & \dots & \frac{\rho_K}{1-\rho_K} & 0 \end{pmatrix}.$$

From Equation (5.7.3) we get that

$$\tilde{\mathbf{M}} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}.$$

If $\rho < 1$ then the largest eigenvalue $a_{\max} < 1$ (see [90]) and it can be readily shown that $\lim_{n\to\infty} \tilde{\mathbf{M}}^n = \mathbf{0}$. Thus, applying Lemma B.1 in [97], we have

$$(\mathbf{I} - \tilde{\mathbf{M}})^{-1} = \sum_{i=0}^{\infty} \tilde{\mathbf{M}}^i, \tag{5.7.6}$$

which is a nonnegative matrix.

Chapter 6

The $M/G/2$ queue with heterogeneous servers

6.1 Introduction

In this chapter we consider a heterogeneous $M/G/2$ queue. Customers arrive according to a Poisson process with rate λ. The queueing discipline is FCFS, where we make the additional convention that when a customer arrives and there is no other customer in the system, he receives service from server 1 immediately. The service time distribution of a customer depends on the server who serves him. The service times at server 1 are exponentially distributed with rate μ, and at server 2 they have a general distribution $B(\cdot)$ with mean β. It is easily verified that $\lambda < \mu + 1/\beta$ is necessary and sufficient for stability. In the sequel, we assume this stability condition to hold. In the present chapter, we are interested in the steady-state waiting time distribution of the above-described $M/G/2$ queue, in particular in its asymptotic behavior. This chapter is based on Boxma, Deng and Zwart [27].

For the classical $G/G/1$ queue, it is well-known [34] that the waiting time tail is regularly varying of index $1-\nu$ if (and only if) the service time tail is regularly varying of index $-\nu$. In fact, Pakes [85] establishes Relation (1.4.1) for the larger class of subexponential residual service times.

The tail behavior of the waiting time in *multi*-server queues with heavy-tailed service times is an almost completely open problem. Recent results suggest that the waiting time tail may not always be as heavy as the tail of the residual service time. For example, Scheller-Wolf and Sigman [95, 96] indicate that the tail of waiting time distribution may be less heavy than that of the residual service time if the offered traffic to the k-server queue is less

than $k-1$. Bounds for the waiting time tail, which were partially proven and partially conjectured in [112], also point in the direction of different waiting time tail behavior for different regimes of the traffic load ρ. Foss and Korshunov [53] derive asymptotic lower and upper bounds for the waiting time distribution in the $G/G/2$ queue with heavy-tailed service times at both servers. Again, the crucial role of the traffic load is striking.

In Chapter 4 of his PhD thesis, Daniëls [41] obtains different tail behavior (of the buffer content distribution) for different traffic loads in a particular multi-server queue. He considers a discrete-time $DBMAP/D/k$ queue with a mix of short-range and long-range dependent traffic. If the mean arrival rate of the short-range dependent background traffic is less than $k-1$, then the tail probabilities decay exponentially; if that mean arrival rate is larger than $k-1$, then they decay according to a power law. Dumas and Simonian [48] describe a similar phenomenon for fluid queues. The present study confirms this kind of behavior for a two-server queue.

For the two-server queue with one exponential server and one server with regularly varying service time distribution, we are able to *prove* the following: The waiting time tail is *semi-exponential* [3] if the arrival rate λ is less than the service rate μ of the exponential server (so the exponential server would be able to handle all offered traffic on his own); and the waiting time tail is regularly varying of index $1-\nu$ if $\lambda > \mu$. More precisely, our main asymptotic results are the following. If the service time at the general server is regularly varying of index $-\nu$, i.e., (2.3.1) holds, then for $t \to \infty$:
(i) if $\lambda > \mu$:

$$\mathbf{P}(W > t) \sim C_1 t^{1-\nu} L(t), \tag{6.1.1}$$

C_1 being specified in Theorem 6.5.1;
(ii) if $\lambda < \mu$:

$$\mathbf{P}(W > t) \sim C_2 t^{1-\nu} e^{(\lambda-\mu)t}, \tag{6.1.2}$$

C_2 being specified in Theorem 6.5.2.

Besides proving (6.1.1) and (6.1.2), we also provide heuristics in both cases that explain and interpret the occurrence of each term. These heuristics might be of independent interest as they suggest ways to generalize the above results and to generate bounds in more complicated systems.

For the moment it suffices to provide a global interpretation of the two different asymptotic regimes. In Case (i) the exponential server is not able to handle all the traffic on its own. The most likely way for a long waiting time to occur is due to a long service time at server 2 which drives the system into

temporary instability. Below (6.5.11) we argue that

$$\mathbf{P}(W > t) \sim C_1^* \mathbf{P}\left(B^{res} > \frac{\lambda t}{\lambda - \mu}\right).$$

In Case (ii) the exponential server is able to handle all the traffic on its own. The most likely way for a long waiting time to occur is to arrive during a long service time at server 2 which causes the system to behave as an $M/M/1$ queue; moreover, this service time must be long enough for the deviant behavior of an $M/M/1$ queue to occur at server 1. Below (6.5.25) we argue that, with obvious notation,

$$\mathbf{P}(W > t) \sim C_2^* \mathbf{P}\left(B^{past} > \frac{\lambda t}{\mu - \lambda}, B^{res} > t\right) \mathbf{P}(W_{M/M/1} > t),$$

where B^{past} and B^{res} refer to the same B, thus they are dependent.

The remainder of this chapter is organized as follows. In Section 6.2 we derive an expression for the steady-state distribution of the number of customers in the system. This expression still involves an unknown function $Q_1(x)$, which is related to the probability of having one customer in the system, at server 2. We express the waiting time distribution in terms of the former distribution in Section 6.3. In Section 6.4 we show how $Q_1(x)$ can be determined in case the service time distribution at server 2 has a rational LST. Unfortunately, we are not able to determine $Q_1(x)$ in the general case, and regularly varying distributions do not have a rational LST. However, in Section 6.5 we show that, in the latter case, the expression for the waiting time LST, which was obtained in Section 6.2, is still sufficient for determining the tail behavior of the waiting time distribution. We provide explicit asymptotics for this tail behavior, distinguishing between $\lambda < \mu$ and $\lambda > \mu$.

6.2 The number of customers in the system

The goal of this section is to compute the generating function of the steady-state number of customers in the system. To accomplish this goal, we use the supplementary variable technique. We refer to Section II.6.2 in [36] for an application of this technique to the $M/G/1$ queue. We consider the process $(X_t, \zeta_t)_{t \geq 0}$, with X_t the number of customers at time t and ζ_t the past service time of the customer in service at the second server. The second server is idle at time t if and only if $\zeta_t = 0$. It is easy to see that $(X_t, \zeta_t)_{t \geq 0}$ is a Markov process. To be able to apply the supplementary variable technique, it will be

assumed that the service time distribution $B(t)$ is absolutely continuous. Also, we assume that $X_0 = 0$. Define for $t \geq 0$:

$$\begin{aligned} R_{0,t} &= \mathbf{P}(X_t = 0), \\ R_{1,t} &= \mathbf{P}(X_t = 1, \text{server 2 idle}), \\ R_{j,t}(u)\mathrm{d}u &= \mathbf{P}(X_t = j, u \leq \zeta_t < u + \mathrm{d}u), \quad u > 0, j = 1, 2, \ldots . \end{aligned}$$

As stated in Section 6.1, it is assumed that the stability condition $\lambda < \mu + 1/\beta$ is satisfied.

Before we proceed, we introduce some additional notation. Let $\beta(s)$ be the LST of the general service time distribution. We also need the LST of the residual service time B^{res}, which is given by

$$\beta_e(s) := \frac{1 - \beta(s)}{\beta s}, \quad \text{Re } s \geq 0. \tag{6.2.1}$$

Denote by X the steady-state number of customers in the system and by ζ the steady-state past service time of the customer in service at server 2. If the system is stable, then $R_{0,t}, R_{1,t}$ and $R_{j,t}(u)$ converge for $t \to \infty$ to R_0, R_1 and $R_j(u)$, which correspond to the distribution of (X, ζ). It can easily be verified that R_0, R_1 and $R_j(u)$ satisfy the following differential equations: For $u > 0$,

$$\begin{aligned} \lambda R_0 &= \mu R_1 + \int_0^\infty \frac{R_1(x)}{1 - B(x)} \mathrm{d}B(x), \\ (\mu + \lambda) R_1 &= \lambda R_0 + \int_0^\infty \frac{R_2(x)}{1 - B(x)} \mathrm{d}B(x), \\ R_1'(u) &= -\left(\lambda + \frac{B'(u)}{1 - B(u)}\right) R_1(u) + \mu R_2(u), \\ R_j'(u) &= -\left(\lambda + \frac{B'(u)}{1 - B(u)}\right) R_j(u) + \lambda R_{j-1}(u) + \mu R_{j+1}(u), \\ & \qquad\qquad\qquad\qquad\qquad\qquad j = 2, 3, \ldots, \\ R_1(0+) &= 0, \\ R_2(0+) &= \int_0^\infty \frac{R_3(x)}{1 - B(x)} \mathrm{d}B(x) + \lambda R_1, \\ R_j(0+) &= \int_0^\infty \frac{R_{j+1}(x)}{1 - B(x)} \mathrm{d}B(x), \quad j \geq 3. \end{aligned}$$

These equations can be derived in the same way as in the ordinary $M/G/1$ queue, see Section II.6.2 in [36]. We also transform these differential equations

in a similar way as in [36]: Define $Q_0 = R_0, Q_1 = R_1$, and for $j \geq 1$, $u > 0$,

$$Q_j(u) = \frac{R_j(u)}{1 - B(u)}.$$

Q_0, Q_1 and $Q_j(u)$ satisfy

$$\begin{aligned}
\lambda Q_0 &= \mu Q_1 + \int_0^\infty Q_1(x)\mathrm{d}B(x), \\
(\mu + \lambda)Q_1 &= \lambda Q_0 + \int_0^\infty Q_2(x)\mathrm{d}B(x), \\
Q_1'(u) &= -\lambda Q_1(u) + \mu Q_2(u), \quad u > 0, \\
Q_j'(u) &= -(\lambda + \mu)Q_j(u) + \lambda Q_{j-1}(u) + \mu Q_{j+1}(u), \quad j \geq 2,\ u > 0, \\
Q_1(0+) &= 0, \\
Q_2(0+) &= \int_0^\infty Q_3(x)\mathrm{d}B(x) + \lambda Q_1, \\
Q_j(0+) &= \int_0^\infty Q_{j+1}(x)\mathrm{d}B(x), \quad j \geq 3.
\end{aligned}$$

Define for $0 \leq p \leq \max(1, \mu/\lambda)$, $u \geq 0$,

$$\begin{aligned}
G(p, u) &:= \sum_{j=1}^\infty Q_j(u)p^j, \\
f(p) &:= \lambda(1 - p) + \mu\left(1 - \frac{1}{p}\right).
\end{aligned}$$

If $\mu > \lambda$, then it is not difficult to see that $G(p, u)$ is well-defined for $1 \leq p < \mu/\lambda$ by using similar arguments as in the proof of Theorem 6.5.2 below. From the last set of differential equations we get

$$\frac{\partial}{\partial u}G(p, u) = \mu(p - 1)Q_1(u) - f(p)G(p, u), \tag{6.2.2}$$

which satisfies the following boundary condition,

$$\begin{aligned}
G(p, 0+) &= \frac{1}{p}\int_0^\infty G(p, u)\mathrm{d}B(u) + \lambda Q_1 p^2 - [(\mu + \lambda)Q_1 - \lambda Q_0]p \\
&\quad -(\lambda Q_0 - \mu Q_1).
\end{aligned} \tag{6.2.3}$$

The general solution to (6.2.2) is given by

$$G(p, u) = e^{-f(p)u}\left[c_1(p) - \mu(1 - p)\int_0^u e^{f(p)x}Q_1(x)\mathrm{d}x\right], \tag{6.2.4}$$

where $c_1(p)$ is independent of u. It is easy to see that

$$c_1(p) = G(p, 0+) = \sum_{j=2}^{\infty} Q_j(0+)p^j. \tag{6.2.5}$$

We now derive two different expressions for $c_1(p)$ which will be used for the two respective cases in which $f(p) < 0$ and $f(p) \geq 0$. Subsequently we can get expressions for $G(p, u)$ which do not contain $c_1(p)$.

If $0 < p < \min(1, \mu/\lambda)$, then $f(p) < 0$ and $G(p, u) \leq G(1, u)$. From (6.2.2) we know that $G(1, u)$ is a constant which is not related to u (in fact, its interpretation is: $G(1, u) = \frac{1}{\beta}\mathbf{P}(X \geq 1, \text{server 2 busy})$; notice that the density of ζ is $\frac{1-B(u)}{\beta}$). Multiplying by $e^{f(p)u}$ on both sides of (6.2.4) and taking the limit for $u \to \infty$, we obtain

$$\lim_{u\to\infty}\left[c_1(p) - \mu(1-p)\int_0^u e^{f(p)x}Q_1(x)\mathrm{d}x\right] = \lim_{u\to\infty}[G(p, u)e^{f(p)u}] = 0,$$

which implies that, for $0 < p < \min(1, \mu/\lambda)$,

$$c_1(p) = \mu(1-p)\int_0^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x. \tag{6.2.6}$$

Substitute (6.2.6) into (6.2.4) to get

$$G(p, u) = \mu(1-p)e^{-f(p)u}\int_{x=u}^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x. \tag{6.2.7}$$

If $\min(1, \mu/\lambda) \leq p \leq \max(1, \mu/\lambda)$, then $f(p) \geq 0$. Substituting (6.2.4) into (6.2.3), we obtain

$$\begin{aligned} c_1(p) &= \lambda Q_1 p^2 - [(\mu+\lambda)Q_1 - \lambda Q_0]p - (\lambda Q_0 - \mu Q_1) \\ &\quad + \frac{c_1(p)}{p}\int_0^{\infty} e^{-f(p)u}\mathrm{d}B(u) \\ &\quad - \frac{\mu(1-p)}{p}\int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\int_{u=x}^{\infty} e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x, \end{aligned}$$

which implies that

$$\begin{aligned} c_1(p) &= \frac{1}{p-\beta(f(p))}\Bigg[p(1-p)[(\mu-\lambda p)Q_1 - \lambda Q_0] \\ &\quad -\mu(1-p)\int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\int_{u=x}^{\infty} e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x\Bigg]. \end{aligned} \tag{6.2.8}$$

Using (6.2.1) we may rewrite (6.2.8) as: For $\min(1,\mu/\lambda) \le p \le \max(1,\mu/\lambda)$,

$$\begin{aligned} c_1(p) &= \frac{1}{1-\frac{(\lambda p-\mu)\beta}{p}\beta_e(f(p))}\Big[p[(\lambda p-\mu)Q_1+\lambda Q_0] \\ &\quad +\mu\int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\int_{u=x}^{\infty} e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x\Big]. \end{aligned} \tag{6.2.9}$$

We are now ready to calculate the generating function $X(p) := \mathrm{E}[p^X]$ of the steady-state number of customers in the system. We have

$$\begin{aligned} X(p) &= \sum_{j=0}^{\infty} p^j \mathbf{P}(X=j) = R_0 + R_1 p + \sum_{j=1}^{\infty} p^j \int_0^{\infty} R_j(u)\mathrm{d}u \\ &= Q_0 + Q_1 p + \int_0^{\infty} G(p,u)(1-B(u))\mathrm{d}u. \end{aligned} \tag{6.2.10}$$

Recalling that $G(1,u)$ is a constant, from (6.2.10) we can derive that $G(1,u) = (1-Q_0-Q_1)/\beta$. For the ease of presentation, put

$$\tilde{Q}_1 := \mathbf{P}(X=1, \text{server 1 idle}) = \int_0^{\infty} Q_1(u)(1-B(u))\mathrm{d}u. \tag{6.2.11}$$

Taking $p=1$ in (6.2.9) and noting that $c_1(1) = G(1,u) = (1-Q_0-Q_1)/\beta$, we get the following equation:

$$\frac{1}{\beta} + \mu - \lambda = \frac{1}{\beta}Q_0 + \mu Q_0 + \frac{1}{\beta}Q_1 + \mu\tilde{Q}_1. \tag{6.2.12}$$

If $0 < p < \min(1,\mu/\lambda)$, substitute (6.2.7) into (6.2.10) to get

$$\begin{aligned} X(p) &= Q_0 + Q_1 p + \mu(1-p)\int_{u=0}^{\infty}(1-B(u))e^{-f(p)u}\int_{x=u}^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x\mathrm{d}u \\ &= Q_0 + Q_1 p + \mu(1-p)\int_{u=0}^{\infty}\int_{t=u}^{\infty}\int_{x=u}^{\infty} e^{f(p)(x-u)}Q_1(x)\mathrm{d}x\mathrm{d}B(t)\mathrm{d}u \\ &= Q_0 + Q_1 p + \mu(1-p)\Big[\int_{t=0}^{\infty}\int_{x=0}^{t}\int_{u=0}^{x} e^{f(p)(x-u)}Q_1(x)\mathrm{d}u\mathrm{d}x\mathrm{d}B(t) \\ &\quad + \int_{t=0}^{\infty}\int_{x=t}^{\infty}\int_{u=0}^{t} e^{f(p)(x-u)}Q_1(x)\mathrm{d}u\mathrm{d}x\mathrm{d}B(t)\Big] \\ &= Q_0 + Q_1 p - \frac{\mu(1-p)}{f(p)}\Big[\int_0^{\infty} Q_1(x)(1-B(x))\mathrm{d}x \end{aligned}$$

$$-\int_0^\infty e^{f(p)x}Q_1(x)\mathrm{d}x + \int_{x=0}^\infty e^{f(p)x}Q_1(x)\int_{u=0}^x e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x\Bigg],$$

which in combination with (6.2.11) and (6.2.15) below leads to

$$\begin{aligned} X(p) &= Q_0 + Q_1 p - \frac{p}{\lambda p - \mu}\Bigg[\mu\tilde{Q}_1 + \lambda Q_0 p - (\mu - \lambda p)pQ_1 \\ &\quad -\mu(1-p)\int_{x=0}^\infty e^{f(p)x}Q_1(x)\mathrm{d}x\Bigg]. \end{aligned} \tag{6.2.13}$$

If $\min(1, \mu/\lambda) \le p \le \max(1, \mu/\lambda)$, substitute (6.2.4) and (6.2.9) into (6.2.10) to get

$$\begin{aligned} X(p) &= Q_0 + Q_1 p + c_1(p)\frac{1-\beta(f(p))}{f(p)} \\ &\quad -\mu(1-p)\int_{u=0}^\infty e^{-f(p)u}(1-B(u))\int_{x=0}^u e^{f(p)x}Q_1(x)\mathrm{d}x\mathrm{d}u \\ &= Q_0 + Q_1 p + c_1(p)\frac{1-\beta(f(p))}{f(p)} \\ &\quad -\frac{\mu(1-p)}{f(p)}\Bigg[\int_0^\infty Q_1(x)(1-B(x))\mathrm{d}x \\ &\quad -\int_0^\infty e^{f(p)x}Q_1(x)\int_{u=x}^\infty e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x\Bigg] \\ &= Q_0 + Q_1 p + \frac{p^2\beta[(\lambda p - \mu)Q_1 + \lambda Q_0]\beta_e(f(p))}{p - (\lambda p - \mu)\beta\beta_e(f(p))} \\ &\quad +\frac{\mu p}{\lambda p - \mu}\Bigg[\frac{p}{p - (\lambda p - \mu)\beta\beta_e(f(p))} \\ &\quad \int_{x=0}^\infty e^{f(p)x}Q_1(x)\int_{u=x}^\infty e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x - \tilde{Q}_1\Bigg]. \end{aligned} \tag{6.2.14}$$

As we can see, the expressions (6.2.13) and (6.2.14) for $X(p)$ contain an unknown function $Q_1(x)$. Replacing $G(p,u)$ in (6.2.3) by (6.2.7), we derive an equation for $Q_1(x)$ which is given by: For $0 < p < \min(1, \mu/\lambda)$,

$$\begin{aligned} \mu\int_0^\infty e^{f(p)x}Q_1(x)\mathrm{d}x &= \frac{\mu}{p}\int_{x=0}^\infty e^{f(p)x}Q_1(x)\int_{u=0}^x e^{-f(p)u}\mathrm{d}B(u)\mathrm{d}x \\ &\quad +(\mu - \lambda p)Q_1 - \lambda Q_0. \end{aligned} \tag{6.2.15}$$

Unfortunately, we are not able to obtain $Q_1(x)$, and hence $X(p)$, for the case of a completely general service time distribution. In case $B(\cdot)$ has a rational LST,

we can determine $X(p)$ completely; this is done in Section 6.4. In Section 6.5, in the case of regularly varying $B(\cdot)$, we perform an asymptotic analysis of the waiting time distribution.

6.3 The waiting time distribution

In this section we establish a link between the waiting time distribution and the queue length generating function $X(p)$. In order to get an explicit formula for the LST $\omega(s)$ of the waiting time distribution, we introduce the queue length N which is the number of customers who are waiting in the system, and its probability generating function $N(p) := \mathrm{E}[p^N]$. By the distributional form of Little's law (cf. [60]), $\omega(s)$ is related to $N(p)$ as follows:

$$\omega(s) = N(1 - s/\lambda), \quad 0 \le s \le \lambda. \tag{6.3.1}$$

Since $N = \max(X - 2, 0)$, it follows that

$$N(p) = Q_0 + Q_1 + \tilde{Q}_1 + \frac{1}{p^2}\left[X(p) - Q_0 - Q_1 p - \tilde{Q}_1 p\right],$$

which in combination with (6.3.1) implies

$$\begin{aligned} \omega(s) &= Q_0 + Q_1 + \tilde{Q}_1 + \frac{1}{(1 - s/\lambda)^2}[X(1 - s/\lambda) - Q_0 \\ &\quad -(1 - s/\lambda)Q_1 - (1 - s/\lambda)\tilde{Q}_1]. \end{aligned} \tag{6.3.2}$$

For the ease of notation, put

$$\hat{f}(s) := f(1 - s/\lambda) = \frac{s(\lambda - \mu - s)}{\lambda - s}. \tag{6.3.3}$$

If $\max(0, \lambda - \mu) < s < \lambda$, then $0 < 1 - s/\lambda < \min(1, \mu/\lambda)$. So we can substitute (6.2.13) into (6.3.2) to get

$$\omega(s) = \frac{(\mu + s)Q_0}{\lambda - \mu - s} + \frac{(\mu + s)\tilde{Q}_1}{\lambda - \mu - s} - \frac{\mu s}{(\lambda - s)(\lambda - \mu - s)} \int_0^\infty e^{\hat{f}(s)x} Q_1(x)\mathrm{d}x. \tag{6.3.4}$$

If $\min(0, \lambda - \mu) \le s \le \max(0, \lambda - \mu)$, then $\min(1, \mu/\lambda) \le 1 - s/\lambda \le \max(1, \mu/\lambda)$. Substitute (6.2.14) into (6.3.2) to get

$$\omega(s) = Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s}\tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))}$$

$$+\frac{\mu}{(\lambda-\mu-s)[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]}$$
$$\int_{x=0}^{\infty} e^{\hat{f}(s)x}Q_1(x)\int_{u=x}^{\infty} e^{-\hat{f}(s)u}\mathrm{d}B(u)\mathrm{d}x. \tag{6.3.5}$$

The above formulas are quite useful in deriving the asymptotic behavior of the waiting time distribution in the case that $B(t)$ has a regularly varying tail, even though they contain an unknown function $Q_1(x)$.

6.4 Rational service time distribution

In this section we assume that the service time distribution at server 2 has a rational LST. Notice that distributions with rational LSTs are light-tailed instead of heavy-tailed. Suppose $\beta(s)$ can be written as $\beta(s) = \frac{\beta_1(s)}{\beta_2(s)}$ where $\beta_1(s)$ and $\beta_2(s)$ are relatively prime polynomials, the degree of $\beta_2(s)$ being higher than that of $\beta_1(s)$. Without loss of generality we can write

$$\beta_2(s) \;=\; \prod_{j=1}^{n}(s-s_j)^{m_j},$$

where $s_1,\dots,s_n$ are distinct, $m_j \in \{1,2,\dots\}$ and Re $s_j < 0$, $j = 1,\dots,n$ (because $\beta(s)$ is analytic for Re $s \geq 0$). We outline how, in this case of a rational service time LST, one can obtain the Laplace transform, $q_1(s) := \int_0^\infty e^{-sx}Q_1(x)\mathrm{d}x$, of $Q_1(x)$. This enables the calculation of the LST $\omega(s)$ of the waiting time distribution.

Step 1: Obtaining an expression for $q_1(-f(p))$.
Replace $Q_1(x)$ in (6.2.15) by the following representation of the inverse of $q_1(s)$:

$$Q_1(x) \;=\; \frac{1}{2\pi\mathrm{i}}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} e^{sx}q_1(s)\mathrm{d}s.$$

Formula (6.2.15) then becomes, after some interchange of integrals and division by μ: For $0 < p < \min(1,\mu/\lambda)$,

$$q_1(-f(p)) \;=\; -\frac{1}{p}\frac{1}{2\pi\mathrm{i}}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\frac{q_1(s)}{f(p)+s}\beta(-s)\mathrm{d}s \;+\; (1-\frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0. \tag{6.4.1}$$

The integral in the righthand side can be handled by observing that all the poles of its integrand are located in the righthalf plane: $-s_1,\dots,-s_n$, and also $-f(p) > 0$ since $0 < p < \min(1,\mu/\lambda)$. Consider the semi-circle with center

in the origin and radius R in the righthalf plane. Choose R so large that all $n+1$ above-mentioned poles are inside the semi-circle. Then the integral along the line segment from $-\mathrm{i}R$ to $+\mathrm{i}R$ and then along the semi-circle back to $-\mathrm{i}R$ equals minus the sum of the residues of those poles. Since the integral along the semi-circle disappears when $R \to \infty$ (remember that the degree of $\beta_2(s)$ is larger than that of $\beta_1(s)$), we have:

$$\begin{aligned} \tfrac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \tfrac{q_1(s)}{f(p)+s} \beta(-s)\mathrm{d}s \;&=\; -q_1(-f(p))\beta(f(p)) \;+ \\ \textstyle\sum_{j=1}^{n} \tfrac{(-1)^{m_j}}{(m_j-1)!} \tfrac{\mathrm{d}^{m_j-1}}{\mathrm{d}a^{m_j-1}} \{ \tfrac{q_1(a)}{f(p)+a} \tfrac{\beta_1(-a)}{\prod_{i\neq j}(-a-s_i)^{m_i}} \}|_{a=-s_j}. \end{aligned} \tag{6.4.2}$$

Formula (6.4.1) thus reduces to: For $0 < p < \min(1, \mu/\lambda)$,

$$\begin{aligned} q_1(-f(p)) \;&=\; \left[\frac{1}{p} \sum_{j=1}^{n} \frac{(-1)^{m_j-1}}{(m_j-1)!} \frac{\mathrm{d}^{m_j-1}}{\mathrm{d}a^{m_j-1}} \{ \frac{q_1(a)}{f(p)+a} \frac{\beta_1(-a)}{\prod_{i\neq j}(-a-s_i)^{m_i}} \}|_{a=-s_j} \right. \\ &\quad \left. +(1-\frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0 \right] \Big/ \left[1 - \frac{\beta(f(p))}{p} \right]. \end{aligned} \tag{6.4.3}$$

The numerator of (6.4.3) contains $\sum_{j=1}^{n} m_j + 2$ unknown constants: Q_0, Q_1, and the $\sum_{j=1}^{n} m_j$ terms relating to the ith derivative of $q_1(s)$ at $s = -s_j$, $i = 0, \ldots, m_j - 1$, $j = 1, \ldots, n$.

Step 2: Determining the unknown constants.
Noting that $q_1(-f(p)) = \frac{c_1(p)}{\mu(1-p)} = \frac{G(p,0+)}{\mu(1-p)}$ (cf. (6.2.5) and (6.2.6)), it follows by analytic continuation that the righthand side of (6.4.3) is analytic inside the unit circle. Let us consider the poles of the denominator of (6.4.3). Multiply numerator and denominator of the righthand side of (6.4.3) by $p^{\sum_{i=1}^{n} m_i}\beta_2(f(p))$. We will prove that

$$p^{\sum_{i=1}^{n} m_i}\beta_2(f(p))(1 - \frac{\beta(f(p))}{p}) \tag{6.4.4}$$

has $\sum_{i=1}^{n} m_i - 1$ different roots inside the unit circle. Remember that $f(p) = \lambda(1-p)+\mu(1-1/p)$. For any $\epsilon > 0$, one can easily check that $\mathrm{Re}\, f(p) \geq f(|p|)$ for $1 \leq |p| \leq 1+\epsilon$, and hence $|\beta(f(p))| \leq |\beta(\mathrm{Re}\, f(p))| \leq |\beta(f(|p|))|$. For p such that $|p| = 1+\epsilon$ we then have:

$$|p| > |\beta(f(|p|))| \geq |\beta(f(p))|,$$

the inequality sign holding because $\lambda < \mu + 1/\beta$ (the ergodicity condition). The resulting inequality $|p| > |\beta(f(p))|$ is equivalent with: For $|p| = 1+\epsilon$,

$$|p^{\sum_{i=1}^{n} m_i}\beta_2(f(p))| > |p^{\sum_{i=1}^{n} m_i - 1}\beta_1(f(p))|.$$

Notice that the multiplication by suitable powers of p has led to functions that are analytic inside $|p| = 1+\epsilon$. Application of Rouché's theorem (cf. Titchmarsh [107]) now implies that $p^{\sum_{i=1}^n m_i}\beta_2(f(p))\ (1 - \frac{\beta(f(p))}{p})$ has the same number of zeros as $p^{\sum_{i=1}^n m_i}\beta_2(f(p))$ inside the circle $|p| = 1 + \epsilon$.
We will prove that the latter number of zeros is $\sum_{i=1}^n m_i$; observe that there is one zero $p = 1$, but we need the $\sum_{i=1}^n m_i - 1$ zeros *inside* the unit circle. Consider

$$p^{\sum_{i=1}^n m_i}\beta_2(f(p)) \ = \ \prod_{j=1}^n (\lambda p(1-p) + \mu(p-1) - ps_j)^{m_j}.$$

Each factor $\lambda p(1-p)+\mu(p-1)-ps_j$ has exactly one zero inside the unit circle, and one zero outside of it. This can be seen, e.g., by another application of Rouché's theorem. In fact, comparison with the expression for the LST of the busy period P of an $M/M/1$ queue with arrival rate λ and service rate μ ($\lambda < \mu$) reveals that the zero inside the unit circle is $\mathrm{E}[e^{s_j P}]$.

The analyticity of $q_1(-f(p))$ inside the unit circle implies that the $\sum_{j=1}^n m_j - 1$ zeros of the denominator of (6.4.3) inside the unit circle should also be zeros of the numerator. This yields $\sum m_i - 1$ linear equations. As remarked at the end of Step 1, we have $\sum m_i + 2$ unknowns. Two additional equations result from the fact that $p = 0$ is a double root of the righthand side of (6.4.3) which follows from the observation that (see (6.2.5)) $c_1(0) = 0$ and $c_1'(0) = 0$. Noticing that $\tilde{Q}_1$ can be represented by a linear combination of the $\sum m_i$ terms which relate to the ith derivative of $q_1(s)$ at $s = -s_j$, $i = 0, \ldots, m_j - 1$, $j = 1, \ldots, n$, the final equation is provided by (6.2.12). Solution of the resulting $\sum m_i + 2$ linear equations yields the $\sum m_i + 2$ unknowns, and finally $q_1(-f(p))$. Once $q_1(-f(p))$ and hence $c_1(p) = \mu(1-p)q_1(-f(p))$ have been obtained, the generating function $X(p)$ of the number of customers follows from (6.2.13) and (6.2.14) and the waiting time LST follows from (6.3.4). Note that (6.3.4) and (6.3.5) are equivalent when the service time at server 2 has a rational LST.

The Erlang-n and hyperexponential distributions are examples of distributions with rational LST. In the special case that $B(t) = 1 - e^{-t/\beta}$, (6.4.4) reduces to $(p-1)(\mu + 1/\beta - \lambda p)$, which does not have a zero in $|p| < 1$. The numerator of the righthand side of (6.4.3) reduces to

$$-\frac{\tilde{Q}_1/\beta}{pf(p) + p/\beta} + (1 - \frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0.$$

Noting that $p = 0$ is a double root of this function, it follows that

$$\frac{\tilde{Q}_1}{\mu\beta} + Q_1 - \frac{\lambda}{\mu}Q_0 = 0,$$

$$\frac{(\lambda+\mu+1/\beta)\tilde{Q}_1}{\mu\beta} - \lambda Q_1 = 0.$$

Combining the above two equations and (6.2.12) leads to

$$\begin{aligned}
Q_1 &= \frac{\lambda Q_0(\lambda+\mu+1/\beta)}{\mu(2\lambda+\mu+1/\beta)}, \\
\tilde{Q}_1 &= \frac{\lambda^2\beta Q_0}{2\lambda+\mu+1/\beta}, \\
Q_0 &= \frac{\mu(2\lambda+\mu+1/\beta)(1-\frac{\lambda}{\mu+1/\beta})}{\lambda\beta[\lambda(\mu+1/\beta)+1/\beta^2+\mu/\beta]+\mu(1-\frac{\lambda}{\mu+1/\beta})(2\lambda+\mu+1/\beta)}.
\end{aligned}$$

Remark 6.4.1 In [73] Knessl et al. also study the $M/G/2$ queue with heterogeneous servers. Using a supplementary variable approach, they derive integral equations for the joint steady-state distribution of numbers of customers and elapsed service times. Their integral equations hold for general service time distributions. But they can only construct the solution of these equations for several mixtures of exponential, Erlang and hyperexponential service time distributions.

6.5 Main asymptotic results

In this section we present our main results: the asymptotic behavior of the waiting time distribution under the assumption that $B(t)$ has a regularly varying tail. First of all, it is useful to recall that Abate, Choudhury and Whitt [1] and Abate and Whitt [3] divide probability distributions on the positive halfline into three classes according to the rightmost singularity of the LST and the value of the LST at this singularity. Let $G(t)$ be a probability distribution function with LST $\gamma(s)$ and let $-s^*$ be the rightmost singularity of $\gamma(s)$, with $-s^* = -\infty$ if $\gamma(s)$ is analytic everywhere. In this setting $G(t)$ and its LST $\gamma(s)$ are classified as follows:

$$\begin{aligned}
&\text{class I}: && s^* > 0 \text{ and } \gamma(-s^*) = \infty, \\
&\text{class II}: && s^* > 0 \text{ and } 1 < \gamma(-s^*) < \infty, \\
&\text{class III}: && s^* = 0 \text{ and } \gamma(-s^*) = 1.
\end{aligned}$$

As indicated in [1, 3], class-I distributions are the 'well behaved' distributions and class-III distributions are the long-tailed distributions. Class-II distributions are called *semi-exponential distributions*, because they are dominated by an exponential, i.e., $\lim_{t\to\infty} e^{\alpha t}(1-G(t)) = 0$ for all real $\alpha < s^*$. Since

$\gamma(-s^*) < \infty$, the rightmost singularity $-s^*$ is necessarily a branch point singularity, not a pole (note that $-s^*$ can still be a branch point when G is in class I). A typical example is the busy period distribution $P(\cdot)$ in the stationary $M/M/1$ queue, which has the following asymptotic behavior:

$$1 - P(t) \sim a_1 t^{-3/2} e^{-b_1 t}, \quad t \to \infty,$$

where $a_1, b_1 > 0$, cf. (II.2.33) in [36]. For more discussions of class I, II and III, see [1, 3].

The results in this section show that, when $1 - B(t)$ is regularly varying, the waiting time distribution $W(t)$ can be in all classes considered above. In particular, we show that $W(t)$ belongs to

- class I if $\lambda < \mu$ and $\beta_2 = \infty$,
- class II if $\lambda < \mu$ and $\beta_2 < \infty$,
- class III if $\lambda > \mu$.

Next, we consider the cases $\lambda > \mu$ and $\lambda < \mu$. In both cases, we analyze the tail $1 - W(t)$ of the waiting time distribution, using the expression for $\omega(s)$ developed in Section 6.2 and an appropriate Tauberian theorem. In the case $\lambda > \mu$ we have $s^* = 0$ and apply Lemma 2.3.4. The case $\lambda < \mu$ is more intricate and here we apply a theorem of Sutton (cf. [101]).

For convenience, we provide the theorem of Sutton [101] in the following lemma. We need to consider the Laplace transform $\phi(s)$ of some positive function $g(t)$ $(t \geq 0)$, e.g.,

$$\phi(s) = \int_0^\infty e^{-st} g(t) \mathrm{d}t, \quad \mathrm{Re}\ s \geq a,$$

for $s = x + \mathrm{i}y$ in the complex plane. Let $\phi(s)/s = \psi(s)$. Then the result in [101] is as follows.

Lemma 6.5.1 *If*
(i) $\psi(s)$ is analytic for $x \geq a - \delta$ $(\delta > 0)$, except at k points $s_1, ..., s_j, ..., s_k$ on $x = a$;
(ii) near each such point s_j, we have

$$(s - s_j)\psi(s) = \sum_{n=0}^{\infty} a_{nj}(s - s_j)^n + (s - s_j)^{\alpha_j} \sum_{n=0}^{\infty} b_{nj}(s - s_j)^n,$$

where $0 < \alpha_j < 1$, and the series converges for

$$|s - s_j| < r \quad (r > 0);$$

(iii) $\psi(s) \to 0$ *as* $y \to \pm\infty$*, uniformly in* x *for* $a - \delta \leq x \leq c$ $(c > a)$*, and in such a manner that* $\int |\psi(s)| \mathrm{d}y$ *converges at* $y = \pm\infty$*,*
then, for $t > 0$*,*

$$g(t) \sim \sum_{j=1}^{k} e^{s_j t} \left(a_{0j} + \sum_{n=0}^{\infty} (-1)^n \frac{b_{nj}}{\Gamma(1 - n - \alpha_j)} t^{-\alpha_j - n} \right).$$

The regime $\lambda > \mu$

When $\lambda > \mu$, the exponential server alone cannot cope with all the traffic: The second server is necessary for stability of the system. This makes it plausible that the heavy-tailed service times at the second server give rise to a heavy-tailed waiting time. In fact, we have:

Theorem 6.5.1 *Suppose that* $\lambda > \mu$ *and*

$$1 - B(t) \sim t^{-\nu} L(t), \quad t \to \infty.$$

$\nu \in (m, m+1)$ $(m \in \mathbb{N})$ *and* $L(t)$ *is a slowly varying function. Then*

$$1 - W(t) \sim \frac{1 - Q_0 - Q_1}{(\nu - 1)(1 - \lambda\beta + \mu\beta)\beta} \left(\frac{\lambda - \mu}{\lambda} \right)^{\nu - 1} t^{1-\nu} L(t), \quad t \to \infty.$$

Proof. By using Lemma 2.3.4, we have

$$\beta(s) = 1 + \sum_{i=1}^{m} (-1)^i \frac{b_i}{i!} s^i + (-1)^{m+1} \Gamma(1 - \nu) s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \quad (6.5.1)$$

for $s \downarrow 0$, where b_i $(i = 1, ..., m)$ stands for the ith moment of the service time and $\Gamma(\cdot)$ is the Gamma function. Remember that $b_1 = \beta$. Again, by using Lemma 2.3.4, it is sufficient to prove that $\omega(s)$ can be written as

$$\begin{aligned} \omega(s) &= 1 + \sum_{i=1}^{m-1} (-1)^i d_i s^i \\ &\quad + (-1)^m \frac{\Gamma(1 - \nu)(1 - Q_0 - Q_1)}{(1 - \lambda\beta + \mu\beta)\beta} \left(\frac{\lambda - \mu}{\lambda} \right)^{\nu - 1} s^{\nu - 1} L(1/s) \\ &\quad + \mathrm{o}(s^{\nu - 1} L(1/s)), \quad s \downarrow 0, \end{aligned} \quad (6.5.2)$$

where $d_i > 0$ for $i = 1, ..., m-1$. For $0 \le s \le \lambda - \mu$, $\omega(s)$ is given by (6.3.5). The expression (6.3.5) contains the function

$$\hat{q}_1(s) := \int_{x=0}^{\infty} e^{\hat{f}(s)x} Q_1(x) \int_{u=x}^{\infty} e^{-\hat{f}(s)u} \mathrm{d}B(u) \mathrm{d}x$$

of which the asymptotic expansion in the neighborhood of the origin is not known. From (6.2.9), we observe that $\hat{q}_1(s)$ for $0 < s < \lambda - \mu$ can be expressed in terms of $c_1(\frac{\mu - \hat{f}(s) + s}{\lambda})$ and $\beta(\hat{f}(s))$. We will analyze the behavior of the latter functions in the origin. Taking $p = 1 - s/\lambda$ in (6.2.9), we obtain

$$c_1(1 - s/\lambda) = \frac{(1 - s/\lambda)[(\lambda - \mu - s)Q_1 + \lambda Q_0] + \mu \hat{q}_1(s)}{1 - \frac{(\lambda - \mu - s)\beta}{1 - s/\lambda} \beta_e(\hat{f}(s))}. \tag{6.5.3}$$

For $s \downarrow 0$, there exists an $s_1 = s_1(s) \uparrow \lambda - \mu$ such that $\hat{f}(s) = \hat{f}(s_1)$ where $\hat{f}(s)$ is given by (6.3.3). It is not difficult to see that $s_1 = \lambda - \mu + \hat{f}(s) - s$. Using (6.5.3), we may write, for $0 \le s \le \lambda - \mu$,

$$\begin{aligned}
\hat{q}_1(s) &= \frac{1}{\mu} c_1(1 - s_1/\lambda) \left[1 - \frac{(\lambda - \mu - s_1)\beta}{1 - s_1/\lambda} \beta_e(\hat{f}(s)) \right] \\
&\quad - \frac{1}{\mu}(1 - s_1/\lambda)[(\lambda - \mu - s_1)Q_1 + \lambda Q_0] \\
&= \frac{1}{\mu} c_1(\frac{\mu - \hat{f}(s) + s}{\lambda}) \left[1 - \frac{\beta\lambda(s - \hat{f}(s))\beta_e(\hat{f}(s))}{\mu + s - \hat{f}(s)} \right] \\
&\quad - \frac{\mu - \hat{f}(s) + s}{\lambda\mu}[(s - \hat{f}(s))Q_1 + \lambda Q_0].
\end{aligned} \tag{6.5.4}$$

Then, replacing $\hat{q}_1(s)$ in the last term of (6.3.5) by the expression in (6.5.4) gives, for $0 \le s \le \lambda - \mu$,

$$\begin{aligned}
\omega(s) &= Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s} \tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))} \\
&\quad + \frac{1}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]} \\
&\quad \left[c_1(\frac{\mu - \hat{f}(s) + s}{\lambda}) \left(1 - \frac{\beta\lambda(s - \hat{f}(s))\beta_e(\hat{f}(s))}{\mu + s - \hat{f}(s)} \right) \right. \\
&\quad \left. - \frac{1}{\lambda}(\mu - \hat{f}(s) + s)[(s - \hat{f}(s))Q_1 + \lambda Q_0] \right]
\end{aligned}$$

$$
\begin{aligned}
= \quad & Q_0 + Q_1 - \frac{\mu+s}{\lambda-\mu-s}\tilde{Q}_1 \\
& + \frac{1}{(\lambda-\mu-s)[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]} \\
& \Bigg[c_1(\frac{\mu-\hat{f}(s)+s}{\lambda}) - \frac{1}{\lambda}(\mu-\hat{f}(s)+s)[(s-\hat{f}(s))Q_1+\lambda Q_0] \\
& + \Bigg((\lambda-\mu-s)[(\lambda-\mu-s)Q_1+\lambda Q_0]\beta \\
& - \frac{\beta\lambda(s-\hat{f}(s))}{\mu-\hat{f}(s)+s} c_1(\frac{\mu-\hat{f}(s)+s}{\lambda}) \Bigg) \beta_e(\hat{f}(s)) \Bigg]. \qquad (6.5.5)
\end{aligned}
$$

Letting $s = 0$ in (6.5.4), the lefthand side is equal to $\tilde{Q}_1$. Therefore, we get

$$c_1(\frac{\mu}{\lambda}) = \mu(\tilde{Q}_1 + Q_0). \qquad (6.5.6)$$

Since $c_1(p)$ is well-defined for $|p| \leq 1$, the Taylor expansion of $c_1(p)$ in the neighborhood of μ/λ exists which is given as follows:

$$c_1(\frac{\mu+s}{\lambda}) = \mu(\tilde{Q}_1 + Q_0) + \sum_{i=1}^{\infty} c_{\mu/\lambda,i} s^i, \quad |s| < \lambda - \mu,$$

where $c_{\mu/\lambda,i}$ are constants for $i = 1, 2, ...$ Because $\hat{f}(s)$ also has a Taylor expansion in the neighborhood of the origin and $\hat{f}(0) = 0$, we may write

$$c_1(\frac{\mu+s-\hat{f}(s)}{\lambda}) = \mu(\tilde{Q}_1 + Q_0) + \sum_{i=1}^{\infty} \tilde{c}_{\mu/\lambda,i} s^i, \quad \text{for } |s| < \delta, \qquad (6.5.7)$$

where δ is some positive constant and $\tilde{c}_{\mu/\lambda,i}$ are all constants for $i = 1, 2, ...$ By (6.2.1) and (6.5.1), we have for $s \downarrow 0$:

$$
\begin{aligned}
\beta_e(s) \quad = \quad & 1 + \sum_{i=1}^{m-1} (-1)^i \frac{b_{i+1}}{(i+1)!\beta} s^i + (-1)^m \frac{\Gamma(1-\nu)}{\beta} s^{\nu-1} L(1/s) \\
& + o(s^{\nu-1} L(1/s)). \qquad (6.5.8)
\end{aligned}
$$

From (6.5.8) we get

$$\frac{1}{(\lambda-\mu-s)[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]}$$

$$
\begin{aligned}
&= \frac{1}{(\lambda-\mu)[1-(\lambda-\mu)\beta]} + g_1(s)s \\
&\quad +(-1)^m \frac{\Gamma(1-\nu)}{(1+\mu\beta-\lambda\beta)^2}\left(\frac{\lambda-\mu}{\lambda}\right)^{\nu-1} s^{\nu-1}L(1/s) \\
&\quad +\mathrm{o}(s^{\nu-1}L(1/s)), \quad s\downarrow 0, \qquad (6.5.9)
\end{aligned}
$$

and

$$
\begin{aligned}
&\Bigg[c_1\Big(\frac{\mu-\hat{f}(s)+s}{\lambda}\Big) - \frac{1}{\lambda}(\mu-\hat{f}(s)+s)[(s-\hat{f}(s))Q_1+\lambda Q_0] \\
&+\Bigg((\lambda-\mu-s)[(\lambda-\mu-s)Q_1+\lambda Q_0]\beta \\
&-\frac{\beta\lambda(s-\hat{f}(s))}{\mu-\hat{f}(s)+s}c_1\Big(\frac{\mu-\hat{f}(s)+s}{\lambda}\Big)\Bigg)\beta_e(\hat{f}(s))\Bigg] \\
&= \mu\tilde{Q}_1 + (\lambda-\mu)(\lambda Q_1-\mu Q_1+\lambda Q_0)\beta + (-1)^m\Gamma(1-\nu)(\lambda-\mu) \\
&\quad (\lambda Q_1-\mu Q_1+\lambda Q_0)\left(\frac{\lambda-\mu}{\lambda}\right)^{\nu-1} s^{\nu-1}L(1/s) \\
&\quad +g_2(s)s + \mathrm{o}(s^{\nu-1}L(1/s)), \quad s\downarrow 0, \qquad (6.5.10)
\end{aligned}
$$

where $g_i(s)$ $(i=1,2)$ are polynomials of degree $m-1$. In combination with (6.5.5), (6.5.9), (6.5.10) and (6.5.7), this leads to (6.5.2). □

An alternative characterization of the tail of $W(t)$ is

$$
\mathbf{P}(W>t) \sim \frac{1-Q_0-Q_1}{1-\lambda\beta+\mu\beta}\mathbf{P}\left(B^{res} > \frac{\lambda t}{\lambda-\mu}\right), \qquad (6.5.11)
$$

when $t\to\infty$.

It is possible to give a heuristic explanation of (6.5.11) by identifying a possible way (which we claim is the most probable way) for W to become large. The heuristics given in the following are similar to those in Section 4.3. They are also based on the preliminary observations in Section 4.3.

First, we make the following two observations:

1. The long-term fraction of customers served by server 2 equals $\frac{1-Q_0-Q_1}{\lambda\beta}$ (note that the mean number of customers handled by server 2 per time unit equals $\frac{1-Q_0-Q_1}{\beta}$).

2. If both servers are busy (i.e. if the waiting time is larger than zero), then the fraction of customers that go to server 1 equals $\frac{\mu}{\mu+\beta^{-1}} = \frac{\beta\mu}{1+\beta\mu}$. Hence, the workload then decreases at rate

$$\frac{\lambda}{\mu}\frac{\beta\mu}{1+\beta\mu} + \lambda\beta\frac{1}{1+\beta\mu} - 2 < 0.$$

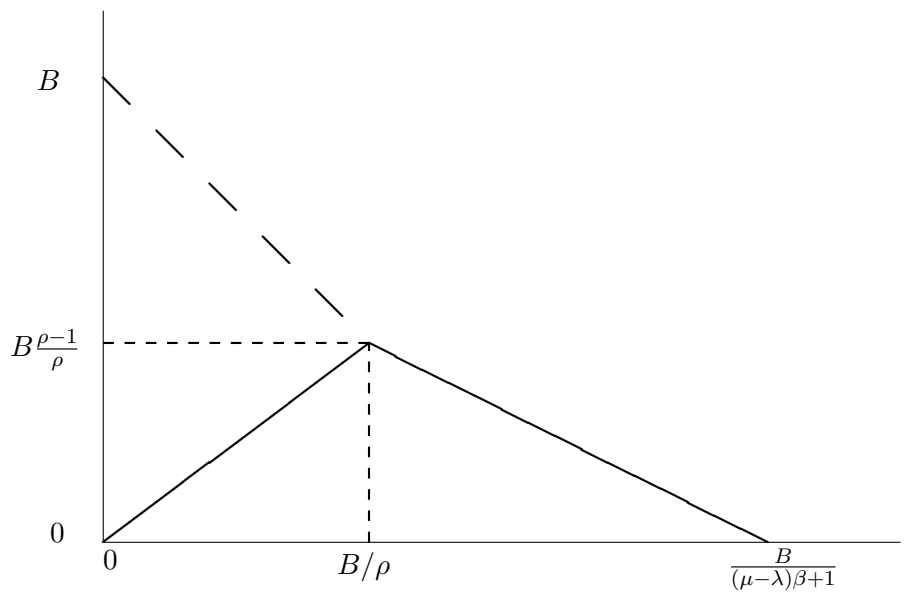

Figure 6.1: Evolution of the waiting time.

We now start the heuristic explanation of (6.5.11). For ease of notation, we introduce $\rho := \lambda/\mu$. Suppose a customer enters the system in steady state at time 0 and is served by server 2. This happens with probability $\frac{1-Q_0-Q_1}{\lambda\beta}$ (due to PASTA and observation 1). Let the service time of this customer be equal to B. Assume that the total workload in the system is very small compared to B. Then, the workload at the second server is roughly equal to B and the workload at server 1 is approximately 0. This means that all incoming traffic will be allocated to server 1, implying that the workload at server 1 will increase linearly with rate $\rho - 1 = \lambda/\mu - 1 > 0$. As no work is allocated to the second server, the workload of server 2 decreases with rate 1. This continues until both workloads are the same, which happens at time B/ρ, see Figure 6.1.

After time B/ρ, the waiting time decreases at rate $1-\frac{\lambda}{\mu+\beta^{-1}}$, by observation 2. Hence, at time $\frac{B}{(\mu-\lambda)\beta+1}$ the effect of the large customer entering the system at time 0 has vanished, see again Figure 6.1.

Suppose that we observe the system at time y and that $W > t$, t large. Our claim is that the waiting time is large because at time 0, a customer entered the system and got served by server 2. This customer had a large service time B. Keeping Figure 6.1 in mind, it is necessary to require $t/(\rho-1) < y < \frac{B-(1+\mu\beta)t}{1+(\mu-\lambda)\beta}$ to get $W > t$. This condition can be rewritten into

$$B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y, \quad y > \frac{t}{\rho-1}.$$

To summarize, the event $W > t$ occurs at time $y > t/(\rho-1)$ if at time 0 a customer enters the system which is served by server 2 and has a service time $B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y$. By observation 1, the probability that the customer is served by server 2 equals $\frac{1-Q_0-Q_1}{\lambda\beta}$. We conclude, after a straightforward computation, that

$$\begin{aligned}\mathbf{P}(W > t) &\approx \int_{\frac{t}{\rho-1}}^{\infty} \frac{1-Q_0-Q_1}{\lambda\beta}\mathbf{P}(B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y)\lambda \mathrm{d}y \\ &= \frac{1-Q_0-Q_1}{1+(\mu-\lambda)\beta}\frac{1}{\beta}\int_{\frac{\rho t}{\rho-1}}^{\infty} \mathbf{P}(B > z)\mathrm{d}z,\end{aligned}$$

which is equal to (6.5.11).

The heuristic arguments do not depend on the service time distribution of server 1 (as long as its tail is lighter than that of the service time distribution of server 2) and can thus be extended to more general multi-server queues.

In addition to the heuristics given above one may try to find a different way of explaining (6.5.11), namely by relating the $M/G/2$ queue to an $M/G/1$ queue with arrival rate $\mu-\lambda$ and service time B. This might lead to some type of *reduced-load* equivalence, as is often the case in fluid queues under the FIFO [4] or GPS [15] discipline: Those studies show that under certain conditions, in the asymptotic analysis one can ignore exponentially tailed sources, apart from reducing the outflow rate by the load offered by those exponential sources.

The regime $\lambda < \mu$

Now we turn to the case $\lambda < \mu$. This case is more intricate. If one wants to apply a similar technique as in the proof of Theorem 6.5.1, one needs to consider the function $e^{s^*t}(1-W(t))$ (which has LST $\omega(s-s^*)$). However, this

function need not be monotone, so a standard Tauberian theorem does not work.

Instead, we shall use Lemma 6.5.1 (cf. Sutton [101]). This lemma does not need a monotonicity assumption, but requires other regularity conditions. In order to meet these conditions, we make the following assumptions on the general service time distribution $B(t)$; these assumptions are similar to the ones in Section 2 of [22], see also [21]. It is assumed that $\beta(s)$ can be represented as: for Re $s \geq 0$,

$$1 - \frac{1-\beta(s)}{\beta s} = h(s) + s^{\nu-1} l(s), \tag{6.5.12}$$

where

(i) $m < \nu < m+1$ $(m \in \mathbb{N})$;

(ii) $h(s)$ is analytic in s for Re $s > -\epsilon_0$ $(\epsilon_0 > 0)$, $h(0) = 0$;

(iii) $l(s)$ is analytic in $s \in \{s:$ Re s $> 0,$ or $|s| < \epsilon_0\}$ $(\epsilon_0 > 0)$ and continuous for Re $s \geq 0$, $l(0) \neq 0$.

The conditions above are satisfied by various distributions, like the distributions considered in Examples (i) and (ii) in Section 3 of [22]. From Lemma 2.3.4, it is easily shown that the assumptions above imply that $1 - B(t)$ is regularly varying with index $-\nu$. Note that our assumptions are slightly stronger than those in [22]. For example, a logarithmic function $l(\cdot)$ in [22] does not satisfy the above condition.

Here is our main result for the case $\lambda < \mu$:

Theorem 6.5.2 *Suppose $\lambda < \mu$ and (6.5.12) holds. Then*

$$1 - W(t) \sim \frac{\lambda l(0)(1 - Q_0 - Q_1)}{\mu \Gamma(2-\nu)} \left(\frac{\mu - \lambda}{\mu}\right)^{\nu-1} t^{1-\nu} e^{(\lambda-\mu)t}, \quad t \to \infty. \tag{6.5.13}$$

Proof. Since $\lambda < \mu$, the ordinary $M/M/1$ queue with input rate λ and service rate μ is stable. Denote by $W_{M/M/1}$ the waiting time in this ordinary $M/M/1$ queue. It is easy to see that W is stochastically smaller than $W_{M/M/1}$, i.e.,

$$1 - W(t) \leq 1 - W_{M/M/1}(t) = \frac{\lambda}{\mu} e^{(\lambda-\mu)t}, \quad t > 0, \tag{6.5.14}$$

which implies that the rightmost singularity is $-s^* \leq \lambda - \mu$.

Next, we shall show that $\omega(s)$ is analytic in the region $\{s :$ Re $s \geq \lambda - \mu - \delta\} \backslash \{\lambda - \mu\}$ for some $\delta > 0$. By (6.5.14), we know that $\omega(s)$ is an analytic

function in the region $\{s : \text{Re } s \geq \lambda - \mu + \epsilon\}$ for any $\epsilon > 0$. So it is sufficient to show that $\omega(s)$ is an analytic function for $s \in \{s : \lambda - \mu - \delta \leq \text{Re } s < 0\} \backslash \{\lambda - \mu\}$. Noting that $\beta_e(s)$ is analytic in the region $\{s : \text{Re } s > 0 \text{ or } |s| < \epsilon_0\} \backslash \{0\}$, we may continue $\omega(s)$ as given by (6.3.5) analytically into $\{s : \text{Re } \hat{f}(s) \geq 0 \text{ or } |\hat{f}(s)| < \epsilon_0\} \cap \{s : \text{Re } s > \lambda - \mu - \delta\}$:

$$\begin{aligned}\omega(s) &= Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s}\tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))} \\ &\quad + \frac{\mu \int_{x=0}^{\infty} e^{\hat{f}(s)x} Q_1(x) \int_{u=x}^{\infty} e^{-\hat{f}(s)u} \mathrm{d}B(u)\mathrm{d}x}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]}. \qquad (6.5.15)\end{aligned}$$

For $s \in \{s : \text{Re } \hat{f}(s) > 0 \text{ and Re } s < 0\}$, we have

$$1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s)) = \frac{\lambda - s}{s}(\beta(\hat{f}(s)) - 1 + s/\lambda) \neq 0.$$

Since

$$\Big[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))\Big]_{s=\lambda-\mu} = \mu/\lambda > 0,$$

it follows that there exists an $\epsilon_1 > 0$ such that, for $|s - \lambda + \mu| < \epsilon_1$, we have

$$\text{Re } [1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))] > 0.$$

Hence, from the analytic continuation of $\omega(s)$ as given in (6.5.15), we conclude that $\omega(s)$ is analytic in the region $\{s : \text{Re } \hat{f}(s) > 0 \text{ and Re } s < 0\} \cup \{s : |s - \lambda + \mu| < \epsilon_1\} \backslash \{\lambda - \mu\}$. Taking $s = x + y\mathrm{i}$, we have

$$\hat{f}(s) = \frac{x(x - \lambda + \mu)(x - \lambda) + (x + \mu)y^2}{(\lambda - x)^2 + y^2} + \mathrm{i}\frac{y^3 + y((\lambda - x)^2 - \lambda\mu)}{(\lambda - x)^2 + y^2}.$$

It is easy to check that there exists a $\delta > 0$ such that $\{s : \lambda - \mu - \delta < \text{Re } s < 0\} \subseteq \{s : \text{Re } \hat{f}(s) \geq 0\} \cup \{s : |s - \lambda + \mu| < \epsilon_1 \text{ and } |\hat{f}(s)| < \epsilon_0\}$.

In order to apply Lemma 6.5.1, we define

$$\tilde{\omega}(s) := \frac{1 - \omega(s)}{s} - \frac{1 - Q_0 - Q_1 - \tilde{Q}_1}{s - \lambda + \mu}. \qquad (6.5.16)$$

We may write, for $t > 0$ and some real a,

$$\begin{aligned}1 - W(t) &= \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{ts}\frac{1 - \omega(s)}{s}\mathrm{d}s \qquad (6.5.17) \\ &= \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{ts}\tilde{\omega}(s)\mathrm{d}s + (1 - Q_0 - Q_1 - \tilde{Q}_1)e^{(\lambda-\mu)t}.\end{aligned}$$

By (6.3.5) and (6.5.16), we have

$$\begin{aligned}\tilde{\omega}(s) &= \frac{(\lambda-\mu)(1-Q_0-Q_1)+\mu\tilde{Q}_1}{s(\lambda-\mu-s)} - \frac{\beta[(\lambda-\mu-s)Q_1+\lambda Q_0]\beta_e(\hat{f}(s))}{s[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]} \\ &\quad -\frac{\mu\int_{x=0}^{\infty} e^{\hat{f}(s)x}Q_1(x)\int_{u=x}^{\infty} e^{-\hat{f}(s)u}\mathrm{d}B(u)\mathrm{d}x}{s(\lambda-\mu-s)[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]}. \qquad (6.5.18)\end{aligned}$$

It is not difficult to check from (6.5.18) that $\tilde{\omega}(s)\to 0$ as $y\to\pm\infty$, uniformly in x for $\lambda-\mu-\delta\le x\le\frac{\lambda-\mu}{2}$, and in such a manner that $\int_0^\infty|\tilde{\omega}(s)|\mathrm{d}y<\infty$.

In the following we shall concentrate on the asymptotic behavior of $\omega(s)$ in the neighborhood of $\lambda-\mu$. We apply similar arguments as in the proof of Theorem 6.5.1. In order to simplify the notation, we introduce $z:=s-\lambda+\mu$. There exists

$$z_1(z)=\mu-\lambda-z+\hat{f}(z+\lambda-\mu)$$

such that $\hat{f}(z+\lambda-\mu)=\hat{f}(z_1+\lambda-\mu)$. Taking $p=1-s/\lambda=(\mu-z)/\lambda$ in (6.2.9), we obtain, for $|z-\mu|<\mu$,

$$\begin{aligned}c_1(\mu/\lambda-z/\lambda) &= \frac{1}{1+\frac{\lambda\beta z}{\mu-z}\beta_e(\hat{f}(z+\lambda-\mu))}\Bigg[(\mu-z)(\lambda Q_0-Q_1z)/\lambda \\ &\quad +\mu\int_{x=0}^{\infty} e^{\hat{f}(z+\lambda-\mu)x}Q_1(x)\int_{u=x}^{\infty} e^{-\hat{f}(z+\lambda-\mu)u}\mathrm{d}B(u)\mathrm{d}x\Bigg].\end{aligned}$$

Using the above relation we may write, for $z\in\{z:|z|\le\mu-\lambda,\ |\lambda+z-\hat{f}(z+\lambda-\mu)|<\mu\}$,

$$\begin{aligned}&\int_{x=0}^{\infty} e^{\hat{f}(z+\lambda-\mu)x}Q_1(x)\int_{u=x}^{\infty} e^{-\hat{f}(z+\lambda-\mu)u}\mathrm{d}B(u)\mathrm{d}x \qquad (6.5.19)\\ &= \frac{1}{\mu}c_1(\mu/\lambda-z_1/\lambda)\left[1+\frac{\lambda\beta z_1}{\mu-z_1}\beta_e(\hat{f}(z_1+\lambda-\mu))\right] \\ &\quad -\frac{1}{\lambda\mu}(\mu-z_1)(\lambda Q_0-Q_1z_1) \\ &= \frac{1}{\mu}c_1(1+z/\lambda-\hat{f}(z+\lambda-\mu)/\lambda) \\ &\quad \left[1+\frac{(\mu-\lambda-z+\hat{f}(z+\lambda-\mu))\lambda\beta}{\lambda+z-\hat{f}(z+\lambda-\mu)}\beta_e(\hat{f}(z+\lambda-\mu))\right] \\ &\quad -\frac{1}{\lambda\mu}(\lambda+z-\hat{f}(z+\lambda-\mu))[\lambda Q_0-(\mu-\lambda-z+\hat{f}(z+\lambda-\mu))Q_1].\end{aligned}$$

Noting that $c_1(p)$ is analytic in $|p| < \mu/\lambda$ and $c_1(1) = (1 - Q_0 - Q_1)/\beta_1$, we may write: for $|p-1| < (\mu-\lambda)/\lambda$,

$$c_1(p) = (1 - Q_0 - Q_1)/\beta + \sum_{i=1}^{\infty} c_{1,i}(p-1)^i,$$

where $c_{1,i}$ $(i = 1, 2, ...)$ are real constants. Again, since $\hat{f}(z+\lambda-\mu)$ is analytic in the region $|z| < \mu-\lambda$ and $\hat{f}(\lambda-\mu) = 0$, it follows that $c_1(\frac{\lambda+z-\hat{f}(z+\lambda-\mu)}{\lambda})$ is also analytic in the region $\{z : |z+(\mu-\lambda-z)z/(\mu-z)| < \min(\lambda, \mu-\lambda), |z| < \mu-\lambda\}$. Thus, $c_1(\frac{\lambda+z-\hat{f}(z+\lambda-\mu)}{\lambda})$ can be represented as:

$$c_1(\frac{\lambda+z-\hat{f}(z+\lambda-\mu)}{\lambda}) = (1 - Q_0 - Q_1)/\beta + z\tilde{c}_1(z), \tag{6.5.20}$$

where $\tilde{c}_1(z)$ is analytic in the region $|z| < \delta$ for some $\delta > 0$. Substituting (6.5.19) into (6.5.15) yields, for $|z| < \delta$,

$$\begin{aligned}
&\omega(s) = \omega(z+\lambda-\mu) \qquad (6.5.21)\\
= \;& Q_0 + Q_1 + \tilde{Q}_1 + \frac{(\lambda Q_0 - Q_1 z)\beta\beta_e(\hat{f}(z+\lambda-\mu))}{\mu/\lambda - z/\lambda + z\beta\beta_e(\hat{f}(z+\lambda-\mu))} + \frac{\lambda}{z}\tilde{Q}_1\\
&- \frac{c_1(1+z/\lambda-\hat{f}(z+\lambda-\mu)/\lambda)}{z[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z+\lambda-\mu))]}\\
&\left[1 + \frac{(\mu-\lambda-z+\hat{f}(z+\lambda-\mu))\lambda\beta}{\lambda+z-\hat{f}(z+\lambda-\mu)}\beta_e(\hat{f}(z+\lambda-\mu))\right]\\
&+ \frac{(1+z/\lambda-\hat{f}(z+\lambda-\mu)/\lambda)[\lambda Q_0 - (\mu-\lambda-z+\hat{f}(z+\lambda-\mu))Q_1]}{z[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z+\lambda-\mu))]}.
\end{aligned}$$

By using (6.2.12) and (6.5.20), one can easily check that

$$\begin{aligned}
A(z) \;:=\; & \frac{1}{z}\Bigg[\mu\tilde{Q}_1 - \tilde{Q}_1 z + \lambda\beta\tilde{Q}_1 z\beta_e(\hat{f}(z+\lambda-\mu))\\
&+(1+z/\lambda-\hat{f}(z+\lambda-\mu)/\lambda)\\
&[\lambda Q_0 - (\mu-\lambda-z+\hat{f}(z+\lambda-\mu))Q_1]\\
&-c_1(1+z/\lambda-\hat{f}(z+\lambda-\mu)/\lambda)\\
&\left(1 + \frac{(\mu-\lambda-z+\hat{f}(z+\lambda-\mu))\lambda\beta}{\lambda+z-\hat{f}(z+\lambda-\mu)}\beta_e(\hat{f}(z+\lambda-\mu))\right)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= h_1(z) + h_2(z)\beta_e(\hat{f}(z+\lambda-\mu)) \\
&\quad + \frac{\lambda(\mu-\lambda)(1-Q_0-Q_1)}{\lambda+z-\hat{f}(z+\lambda-\mu)} \frac{1-\beta_e(\hat{f}(z+\lambda-\mu))}{z},
\end{aligned} \tag{6.5.22}
$$

where $h_j(z)$ $(j = 1, 2)$ are both analytic functions for $|z| < \delta_1$ with δ_1 some positive constant. Combining (6.5.21) and (6.5.22), we obtain, for $|z| < \delta_1$,

$$
\begin{aligned}
&\omega(z+\lambda-\mu) \\
&= Q_0 + Q_1 + \tilde{Q}_1 + \frac{A(z) + (\lambda\beta Q_0 - \beta Q_1 z)\beta_e(\hat{f}(z+\lambda-\mu))}{\mu/\lambda - z/\lambda + z\beta\beta_e(\hat{f}(z+\lambda-\mu))} \\
&= Q_0 + Q_1 + \tilde{Q}_1 + \frac{h_1(z) + (h_2(z) + \lambda\beta Q_0 - \beta Q_1 z)\beta_e(\hat{f}(z+\lambda-\mu))}{\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z+\lambda-\mu))} \\
&\quad + \frac{\lambda(\mu-\lambda)(1-Q_0-Q_1)}{(\lambda+z-\hat{f}(z+\lambda-\mu))[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z+\lambda-\mu))]} \\
&\quad \frac{1-\beta_e(\hat{f}(z+\lambda-\mu))}{z}.
\end{aligned} \tag{6.5.23}
$$

From (6.5.12), (6.5.16) and (6.5.23), we conclude that

$$
\tilde{\omega}(s) = \sum_{j=-1}^{\infty} d_j(s-\lambda+\mu)^{r_j}, \quad (-1 = r_{-1} < r_0 < r_1 < ...), \quad |s-\lambda+\mu| < \delta_2, \tag{6.5.24}
$$

where δ_2 is some positive constant. If $1 < \nu < 2$, we have

$$
\begin{aligned}
&r_{-1} = -1, \quad d_{-1} = -(1-Q_0-Q_1-\tilde{Q}_1), \\
&r_0 = \nu - 2, \quad d_0 = l(0)(1-Q_0-Q_1)\frac{\lambda}{\mu}\left(\frac{\mu-\lambda}{\mu}\right)^{\nu-1},
\end{aligned}
$$

where $l(\cdot)$ is given in (6.5.12); if $m < \nu < m+1$ $(m \geq 2)$, we have

$$
\begin{aligned}
&r_{-1} = -1, \quad d_{-1} = -(1-Q_0-Q_1-\tilde{Q}_1), \\
&r_j = j, \quad j = 0, ..., m-2, \\
&r_{m-1} = \nu - 2, \quad d_{m-1} = l(0)(1-Q_0-Q_1)\frac{\lambda}{\mu}\left(\frac{\mu-\lambda}{\mu}\right)^{\nu-1}.
\end{aligned}
$$

Therefore, applying Lemma 6.5.1, it follows from (6.5.17) that

$$
1 - W(t) - (1-Q_0-Q_1-\tilde{Q}_1)e^{(\lambda-\mu)t}
$$

$$\approx \quad -(1-Q_0-Q_1-\tilde{Q}_1)e^{(\lambda-\mu)t}+\sum_{j=0}^{\infty}\frac{d_j}{\Gamma(-r_j)}t^{-r_j-1}e^{(\lambda-\mu)t},$$

$$\left(\frac{1}{\Gamma(-r_j)}=0 \text{ for } r_j=0,1,2,...\right),$$

which implies that (6.5.13) holds. □

Like in the case $\lambda > \mu$, we may rewrite (6.5.13) in a different form. A straightforward computation (using Lemma 2.3.4 to obtain the tail behavior of the distribution of B^{res} from (6.5.12)) shows that, for $t \to \infty$,

$$1-W(t) \sim (1-Q_0-Q_1)\mathbf{P}\left(B^{res} > \frac{\mu t}{\mu-\lambda}\right)\mathbf{P}(W_{M/M/1} > t). \qquad (6.5.25)$$

This result has the following intuitive interpretation: A large waiting time W occurs as a consequence of a large service time at server 2, which causes the system to behave as an $M/M/1$ queue. It is well-known from standard large deviations theory that the most probable way for the workload in an $M/M/1$ queue ($W_{M/M/1}$) to get large is in a linear fashion, with a positive drift of $\mu/\lambda - 1$ (see e.g. p. 276 of [98]). Hence, the time it takes until $W_{M/M/1} > t$ (given that this event occurs) is equal to $\lambda t/(\mu-\lambda)$.

In order for the deviant behavior of the $M/M/1$ queue to occur, server 2 needs to be occupied (which has probability $1-Q_0-Q_1$) and the past service time B^{past} of the customer must be larger than $\lambda t/(\mu-\lambda)$. Finally, the residual service time B^{res} of the customer at server 2 must be larger than t. Standard renewal theory (see e.g. [36], p. 113) gives

$$\begin{aligned}\mathbf{P}\left(B^{past} > \frac{\lambda t}{\mu-\lambda}, B^{res} > t\right) &= \frac{1}{\beta}\int_{\frac{\lambda t}{\mu-\lambda}+t}^{\infty}\mathbf{P}(B>u)\mathrm{d}u \\ &= \mathbf{P}\left(B^{res} > \frac{\mu t}{\mu-\lambda}\right).\end{aligned}$$

Combining all these observations yields (6.5.25). The above interpretation shows an interesting feature of this model: Most likely, the waiting time becomes very large by the simultaneous occurrence of two events: A very long waiting time at an exponential server ($M/M/1$ large deviations) and a large service time of a heavy-tailed server.

6.6 Conclusions

The main results of our study of the heterogeneous $M/G/2$ queue with one exponential and one general server are: (i) an exact analysis of the queue length and waiting time distribution if the general service time distribution has a rational LST, and (ii) an asymptotic analysis of the waiting time tail if the general service time distribution is regularly varying. The analysis of (i) may be extended to the case of an $M/G/k$ queue with $k-1 \geq 2$ exponential servers and one general server with rational service time LST. The exact and heuristic analysis in (ii) should form just the beginning of an investigation of waiting time asymptotics in multi-server queues with heavy-tailed service times. In the two-server case, we have not yet been able to handle the intricate case $\lambda = \mu$; another line of research would be to generalize the class of service time distributions for server 1.

It would be interesting to investigate whether the asymptotic behavior that was observed in the present chapter holds more generally for a related class of light-tailed single server queues with a heavy-tailed 'random environment', that affects the behavior of the queue and where the queue may in its turn affect the random environment. Recent examples of such systems for which similar results have been obtained as in Section 6.5, can be found in [31] and [16]. Boxma and Kurkova [31] consider the $M/G/1$ queue with the special feature that the speed of the server alternates between two constant values, where the high-speed periods are exponentially distributed and the low-speed periods have a general distribution. They present an exact analysis for the case that the distribution of the low-speed periods has a rational LST and an asymptotic analysis for the case that the distributions of the low-speed periods and/or the service times are regularly varying. Borst, Boxma and van Uitert [16] deal with a system with two heterogeneous traffic classes, one having light-tailed characteristics, the other one exhibiting heavy-tailed properties. When both classes are backlogged, the two corresponding queues are each served according to a certain nominal rate. However, when one queue empties, the service rate for the other class increases. For this model, they obtain the asymptotic workload behavior of both traffic classes.

Chapter 7

The tandem queueing system

7.1 Introduction

In this chapter we consider a queueing system consisting of two single-server queues Q_1 and Q_2 in series with infinite waiting space at each queue. Customers arrive at Q_1 according to a Poisson process; Q_1 is an ordinary $M/G/1$ queue. The special feature of the model is that the service time experienced by any customer at Q_2 is *exactly equal to* the one he experienced at Q_1. We are in particular interested in the asymptotic behavior of the steady-state sojourn time and workload distributions at Q_2, paying special attention to the case of a heavy-tailed service time distribution. This chapter is based on Boxma and Deng [25].

The tandem system with identical service times at both nodes is interesting for some practical communication nets, as it reflects the situation in which a message retains the same length while being transmitted through various communication channels. The two-node case has been studied in detail in [18]. A nice feature of this model is, that it allows explicit expressions for the sojourn time and workload distributions at the second node (without taking recourse to LSTs). These explicit expressions play an important role in proving the main results of this chapter.

This chapter studies the tail behavior of key performance measures in tandem queues for *general* service time distributions. More precisely, we establish direct relations between the tail behavior of the (residual) service time distribution and the sojourn time and workload distributions at the second queue, see Theorems 7.4.1 and 7.5.1. By using those relations we obtain asymptotic results for the sojourn time and workload distributions in the case of a service time distribution with regularly varying tail, see Theorems 7.4.2 and 7.5.2. In

particular, both the sojourn time distribution and the workload distribution at Q_2 are shown to be regularly varying of index $1-\nu$, if the service time distribution is regularly varying of index $-\nu$. Finally a heavy-traffic limit theorem, see Theorem 7.7.1, is provided. It states that if the service time distribution is regularly varying of index $-\nu$ $(1 < \nu < 2)$, and the traffic load $\rho \uparrow 1$, then the contracted sojourn time $\Delta(\rho)S^{(2)}$ converges in distribution for an appropriately chosen coefficient of contraction $\Delta(\rho)$, and the limit distribution function is given by $H(t) = \dfrac{\exp(-t^{1-\nu})}{1+\nu t^{1-\nu}}$.

We now describe some related work. Vinogradov [109] considers a tandem system consisting of an arbitrary number of queues, with identical service times at all queues. He studies the joint steady-state distribution of the sojourn time at the first queue and the total sojourn time at the remaining queues, in the case of heavy traffic. He assumes that the service time distribution has a finite third moment. In [76], a tandem queueing system with identical service times at both nodes is considered for various service disciplines (e.g., FCFS at the first queue and LCFS preemptive resume at the second queue) in the case of heavy traffic. It is assumed that the service time distribution has a finite second moment.

While the tail behavior of the waiting time, sojourn time and workload distributions with heavy-tailed service time distributions is presently receiving considerable attention in performance analysis, hardly any *network* results have been obtained. Anantharam [6] and Boxma and Dumas [29] obtain results regarding the propagation of long-range dependence in networks of (fluid) queues. Baccelli et al. [9] and Huang and Sigman [63] consider tandem queues with renewal input process at the first node and *independent* service times at the various nodes, which have a subexponential distribution in at least one node. They obtain several tail results (for details, see Section 1.4).

Heavy-traffic limit theorems for the $G/G/1$ queue with regularly varying interarrival and/or service time, which has an infinite variance, have been obtained in [22], and for the $M/G/1$ queue with priority classes in Chapter 3. In the present chapter, such a heavy-traffic limit theorem is obtained for the sojourn time distribution at Q_2.

The remainder of this chapter is organized as follows. Section 7.2 summarizes the notation and the main results from [17, 18, 19] that will be used in the sequel. In Section 7.3 we obtain tail asymptotics for some performance measures for Q_1. These results are used in Sections 7.4 and 7.5 to obtain the tail behavior of the sojourn time and workload distributions at Q_2, respectively. In Section 7.6 we derive the asymptotic results in Sections 7.4 and 7.5 by using heuristic arguments. The tail behavior of the total sojourn time distribution

is also obtained. In Section 7.7 we derive a heavy-traffic limit theorem for the sojourn time distribution at Q_2, in the case of a regularly varying service time distribution with *infinite* or *finite* variance.

7.2 The basic equations

First we introduce some notation. λ denotes the arrival intensity, B the service time, $B(\cdot)$ the service time distribution and $\beta(\cdot)$ the LST of $B(\cdot)$. Note that when an arbitrary customer arrives at Q_1, his service time is a random variable with distribution $B(\cdot)$; when he enters Q_2, his service time is *identical* to his previous service time at the first queue. We assume that $B(\cdot)$ has a finite first moment β and that the traffic load $\rho := \lambda\beta < 1$. This ensures [18] that steady-state distributions of the sojourn time and workload at both queues exist.

Let $S^{(j)}$ be a random variable with distribution the steady-state distribution of the sojourn time at Q_j, $j = 1, 2$; the sojourn time distributions are denoted by $S^{(j)}(\cdot)$ and their LST by $s^{(j)}(\cdot)$, for $j = 1, 2$. To introduce an explicit expression for $S^{(2)}(\cdot)$, we need the following distributions. $M(\cdot)$ denotes the steady-state distribution function of the supremum, M, of the service times of customers during a busy period of Q_1; $G(\cdot)$ denotes the steady-state distribution function of the supremum, G, of the service times of an arbitrary customer C and of those customers who have arrived before C and belong to the same busy period at Q_1 as C. As shown in [17, 19], $M(t)$ is the unique zero inside the unit circle of the following equation,

$$M(t) = \int_0^t \exp(-\lambda(1 - M(t))x)\mathrm{d}B(x), \quad t > 0, \tag{7.2.1}$$

and $G(t)$ is given by

$$G(t) = (1 - \rho)\frac{1 - M(t)}{1 - B(t)}B(t), \quad t > 0. \tag{7.2.2}$$

Let X be the supremum of the service times of those customers who arrived before an arbitrary customer C and belong to the same busy period at Q_1 as C; $X = 0$ if C is the first customer during a busy period. Let $X(\cdot)$ be the distribution function of X. We have

$$G(t) = X(t)B(t), \quad t > 0, \tag{7.2.3}$$

which in combination with (7.2.2) implies that

$$X(t) = (1 - \rho)\frac{1 - M(t)}{1 - B(t)}, \quad t > 0. \tag{7.2.4}$$

$Y(\cdot)$ denotes the steady-state distribution function of the amount of work, Y, at Q_2 at the epoch that a busy cycle at Q_1 starts. From Theorem 6.1 in [18] we know that

$$Y(t) = \exp\left(-\lambda \int_t^\infty (1 - M(x))\mathrm{d}x\right), \quad t > 0. \tag{7.2.5}$$

Now we turn to the sojourn time distributions. The LST of the sojourn time distribution in the $M/G/1$ queue Q_1 follows immediately from the Pollaczek-Khintchine formula:

$$s^{(1)}(s) = \frac{(1-\rho)\beta(s)}{1 - \rho\frac{1-\beta(s)}{\beta s}}.$$

A probabilistic reasoning, cf. Theorem 6.4 in [18], shows that the steady-state sojourn time $S^{(2)}$ at Q_2 is the maximum of two independent random variables with distribution $G(\cdot)$ and $Y(\cdot)$, i.e.: for $t > 0$,

$$S^{(2)}(t) = G(t)Y(t) = (1-\rho)\frac{1 - M(t)}{1 - B(t)}B(t)\exp\left(-\lambda \int_t^\infty (1 - M(x))\mathrm{d}x\right). \tag{7.2.6}$$

7.3 Preliminaries

In this section we investigate the asymptotic behavior of $1 - X(t)$, $1 - G(t)$ and $1 - Y(t)$ for $t \to \infty$. These asymptotics will turn out to play a key role in the asymptotic behavior of the sojourn time and workload distributions at Q_2, cf. Sections 7.4 and 7.5. In this chapter we assume that service time is unbounded, i.e., $1 - B(t) > 0$ for $t > 0$.

Lemma 7.3.1

$$1 - X(t) = \mathbf{P}(X > t) \sim \frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x) \quad \textit{for } t \to \infty. \tag{7.3.1}$$

Proof. Rewrite (7.2.1) as

$$1 - M(t) = 1 - B(t) + \int_0^t (1 - \exp(-\lambda(1 - M(t))x))\mathrm{d}B(x). \tag{7.3.2}$$

It follows from the fact that $X(t)$ is a proper probability distribution that

$$\lim_{t\to\infty} \frac{1 - M(t)}{1 - B(t)} = \frac{1}{1-\rho}, \tag{7.3.3}$$

cf. (7.2.4). This implies that

$$\lim_{t\to\infty} t(1-M(t)) = 0,$$

since $\lim_{t\to\infty} t(1-B(t)) = 0$, which follows from the fact that $B(\cdot)$ has a finite first moment. Hence, for any $0 < \epsilon < 1$, if t is sufficiently large, then for $0 < x < t$,

$$\begin{aligned} & \lambda(1-M(t))x - (1+\epsilon)\frac{\lambda^2}{2}(1-M(t))^2x^2 \\ < \; & 1-\exp(-\lambda(1-M(t))x) \\ < \; & \lambda(1-M(t))x - (1-\epsilon)\frac{\lambda^2}{2}(1-M(t))^2x^2. \end{aligned} \tag{7.3.4}$$

For the ease of presentation, we define

$$F(t) = \frac{\lambda^2(1-M(t))^2}{2(1-B(t))}\int_0^t x^2 \mathrm{d}B(x). \tag{7.3.5}$$

Dividing both sides of (7.3.2) by $1-B(t)$ and applying (7.3.4) gives

$$\begin{aligned} & 1+\lambda\frac{1-M(t)}{1-B(t)}\int_0^t x\mathrm{d}B(x) - (1+\epsilon)F(t) \\ < \; & \frac{1-M(t)}{1-B(t)} \\ < \; & 1+\lambda\frac{1-M(t)}{1-B(t)}\int_0^t x\mathrm{d}B(x) - (1-\epsilon)F(t). \end{aligned}$$

Subtract $\lambda\frac{1-M(t)}{1-B(t)}\int_0^t x\mathrm{d}B(x)$ and multiply by $(1-\rho)/(1-\lambda\int_0^t x\mathrm{d}B(x))$ on both sides of the above inequality to obtain

$$\begin{aligned} \frac{1-(1+\epsilon)F(t)}{1+\frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x)} & < (1-\rho)\frac{1-M(t)}{1-B(t)} \\ & < \frac{1-(1-\epsilon)F(t)}{1+\frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x)}. \end{aligned} \tag{7.3.6}$$

Since $(1-\rho)(1-M(t))/(1-B(t)) = X(t)$, cf. (7.2.4), it follows from the above equality that

$$\frac{\frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x) + (1-\epsilon)F(t)}{1+\frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x)}$$

$$
\begin{aligned}
&< 1 - X(t) \\
&< \frac{\frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x) + (1+\epsilon)F(t)}{1 + \frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x)}. \qquad (7.3.7)
\end{aligned}
$$

We now investigate the asymptotic behavior of $F(t)$ for $t \to \infty$. We show that the first term in the numerator of the left- and righthand sides of (7.3.7) dominates the second term. Since $X(t)$ is a proper probability distribution, we have from (7.2.4) for every $t > 0$,

$$
\begin{aligned}
\frac{F(t)}{\int_t^\infty x\mathrm{d}B(x)} &= \frac{\lambda^2(1-M(t))^2\int_0^t x^2\mathrm{d}B(x)}{2(1-B(t))\int_t^\infty x\mathrm{d}B(x)} \\
&\le \frac{\lambda^2(1-M(t))^2\int_0^t x^2\mathrm{d}B(x)}{2(1-B(t))^2 t} \\
&\le \frac{\lambda^2\int_0^t x^2\mathrm{d}B(x)}{2(1-\rho)^2 t} \\
&= \lambda^2\frac{-t^2(1-B(t)) + 2\int_0^t(1-B(x))x\mathrm{d}x}{2(1-\rho)^2 t}. \qquad (7.3.8)
\end{aligned}
$$

Since

$$\int_0^\infty (1-B(x))\mathrm{d}x < \infty,$$

by the Dominated Convergence Theorem, it follows that

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t (1-B(x))x\mathrm{d}x = 0.$$

Combining the above equation and (7.3.8) yields

$$\lim_{t\to\infty}\frac{F(t)}{\int_t^\infty x\mathrm{d}B(x)} = 0. \qquad (7.3.9)$$

The result follows from the above relation and (7.3.7). □

From (7.2.3) we can immediately derive the asymptotic behavior of $1-G(t)$ for $t \to \infty$.

Lemma 7.3.2

$$1 - G(t) = \mathbf{P}(G > t) \sim \frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x) \quad \textit{for } t \to \infty. \qquad (7.3.10)$$

Proof. It follows from (7.2.3) that

$$1 - G(t) = 1 - X(t) + 1 - B(t) - (1 - X(t))(1 - B(t)).$$

By Lemma 7.3.1, we have

$$\lim_{t\to\infty} \frac{1 - B(t)}{1 - X(t)} = 0.$$

Combining the above two relations yields that

$$1 - G(t) \sim 1 - X(t), \quad t \to \infty,$$

which in combination with Lemma 7.3.1 implies (7.3.10). □

It should be noted that the latter relation can also be written as (with $I_{\{B>t\}}$ the indicator function of the event $\{B > t\}$):

$$\mathbf{P}(G > t) \sim \frac{\lambda}{1-\rho}\mathrm{E}[BI_{\{B>t\}}] \text{ for } t \to \infty.$$

Lemma 7.3.3

$$1 - Y(t) = \mathbf{P}(Y > t) \sim \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t) \text{ for } t \to \infty,$$

where $Y(t)$ is given by (7.2.5) and B^{res} is the residual service time which has density function $(1 - B(t))/\beta$.

Proof. As seen in (7.2.4), $(1-\rho)(1-M(t))/(1-B(t))$ is the probability distribution of a proper random variable X, the supremum of the service times of the customers who arrived before an arbitrary customer C in the same busy period at Q_1 as C. Hence, cf. also (7.3.3), with an arbitrary $\epsilon > 0$ and for t large enough,

$$\left(\frac{1}{1-\rho} - \epsilon\right)(1 - B(t)) \le 1 - M(t) \le \frac{1}{1-\rho}(1 - B(t)).$$

Thus

$$\begin{aligned}\left(\frac{1}{1-\rho} - \epsilon\right)\int_t^\infty (1 - B(x))\mathrm{d}x &\le \int_t^\infty (1 - M(x))\mathrm{d}x \\ &\le \frac{1}{1-\rho}\int_t^\infty (1 - B(x))\mathrm{d}x,\end{aligned}$$

which implies that

$$\lim_{t\to\infty} \frac{\int_t^\infty (1-M(x))\mathrm{d}x}{\int_t^\infty (1-B(x))\mathrm{d}x} = \frac{1}{1-\rho}.$$

Hence it follows that

$$\begin{aligned}\lim_{t\to\infty} \frac{1-Y(t)}{\int_t^\infty (1-B(x))\mathrm{d}x} &= \lim_{t\to\infty} \left(\frac{1-Y(t)}{\int_t^\infty (1-M(x))\mathrm{d}x} \frac{\int_t^\infty (1-M(x))\mathrm{d}x}{\int_t^\infty (1-B(x))\mathrm{d}x} \right) \\ &= \frac{\lambda}{1-\rho}.\end{aligned}$$

□

7.4 Asymptotic behavior of the sojourn time distribution

In this section we apply the lemmas that were obtained in the previous section to derive the asymptotic behavior of $1-S^{(2)}(t)$ for $t\to\infty$. Moreover, we show how $1-S^{(2)}(t)$ behaves for $t\to\infty$ if the service time distribution is regularly varying. In fact, if the service time distribution is regularly varying of index $-\nu$ ($\nu>1$), then the sojourn time in the second queue is shown to be regularly varying of index $1-\nu$, which is one degree higher than that of the service time distribution.

Theorem 7.4.1 *For $t\to\infty$,*

$$1-S^{(2)}(t) = \mathbf{P}(S^{(2)}>t) \sim \frac{\lambda}{1-\rho} t\mathbf{P}(B>t) + \frac{2\rho}{1-\rho}\mathbf{P}(B^{res}>t). \tag{7.4.1}$$

Proof. Since $\lim_{t\to\infty} t(1-B(t))=0$, it follows that

$$\int_t^\infty x\mathrm{d}B(x) = t(1-B(t)) + \int_t^\infty (1-B(x))\mathrm{d}x. \tag{7.4.2}$$

By applying Lemmas 7.3.2 and 7.3.3, it follows from (7.2.6) that

$$\begin{aligned}1-S^{(2)}(t) &\sim 1-G(t)+1-Y(t) \\ &\sim \frac{\lambda}{1-\rho}\int_t^\infty x\mathrm{d}B(x) + \frac{\rho}{1-\rho}\int_t^\infty \frac{1-B(x)}{\beta}\mathrm{d}x \\ &= \frac{\lambda t}{1-\rho}\mathbf{P}(B>t) + \frac{2\rho}{1-\rho}\mathbf{P}(B^{res}>t), \quad t\to\infty,\end{aligned}$$

where the last equation follows from (7.4.2). □

Theorem 7.4.1 gives a precise expression for the tail behavior of the sojourn time distribution at Q_2 in terms of the tail of the (residual) service time distribution, for arbitrary service time distributions. Below we specify this sojourn time tail behavior (and that of M, G, X and Y) for the case of a regularly varying service time distribution.

Theorem 7.4.2 *Let $\nu > 1$. If $\mathbf{P}(B > t)$ is regularly varying of index $-\nu$, then $\mathbf{P}(M > t)$ is regularly varying of index $-\nu$, while $\mathbf{P}(G > t)$, $\mathbf{P}(X > t)$, $\mathbf{P}(Y > t)$ and $\mathbf{P}(S^{(2)} > t)$ are regularly varying of index $1-\nu$. More precisely, if*

$$\mathbf{P}(B > t) \sim t^{-\nu} L(t), \quad t \to \infty, \tag{7.4.3}$$

then for $t \to \infty$,

$$\mathbf{P}(M > t) \sim \frac{1}{1-\rho} t^{-\nu} L(t), \tag{7.4.4}$$

$$\mathbf{P}(X > t) \sim \mathbf{P}(G > t) \sim \frac{\lambda}{1-\rho} \frac{\nu}{\nu-1} t^{1-\nu} L(t), \tag{7.4.5}$$

$$\mathbf{P}(Y > t) \sim \frac{\lambda}{1-\rho} \frac{1}{\nu-1} t^{1-\nu} L(t), \tag{7.4.6}$$

$$\mathbf{P}(S^{(2)} > t) \sim \frac{\lambda}{1-\rho} \frac{\nu+1}{\nu-1} t^{1-\nu} L(t). \tag{7.4.7}$$

Proof. It follows from Lemma 2.3.1 that $\mathbf{P}(B^{res} > t) \sim \dfrac{t^{1-\nu}}{(\nu-1)\beta} L(t)$ as $t \to \infty$. The results now follow immediately from (7.3.3), Lemmas 7.3.1, 7.3.2, 7.3.3 and Theorem 7.4.1. □

Remark 7.4.1 If the residual service time distribution is a Weibull distribution, i.e., $\mathbf{P}(B^{res} > t) = \exp(-t^\delta)$, $0 < \delta < 1$, then Theorem 7.4.1 implies that

$$\mathbf{P}(S^{(2)} > t) \sim \frac{\rho}{1-\rho} \delta t^\delta \exp(-t^\delta), \quad t \to \infty. \tag{7.4.8}$$

In this case, the second term in the righthand side of (7.4.1) becomes negligible compared to the first one.

Remark 7.4.2 The Weibull distribution of the previous remark is a subexponential distribution, cf. [12]. Pakes [85] has proven for the $G/G/1$ queue

that, if the residual service time distribution is subexponential, then the tail of the sojourn time distribution is asymptotically equivalent to the tail of the residual service time distribution, up to a multiplicative factor $\rho/(1-\rho)$. So in this subexponential case we have, for the $M/G/1$ queue Q_1:

$$\mathbf{P}(S^{(1)} > t) \sim \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t), \quad t \to \infty.$$

Hence, in the case of the above-mentioned Weibull distribution, the following holds for the $M/G/1$ queue Q_1:

$$\mathbf{P}(S^{(1)} > t) \sim \frac{\rho}{1-\rho}\exp(-t^{\delta}), \quad t \to \infty,$$

which is less heavy than the tail $\mathbf{P}(S^{(2)} > t)$ as given in (7.4.8).
In the case of a regularly varying service time distribution, (7.4.7) implies that

$$\mathbf{P}(S^{(2)} > t) \sim (\nu+1)\mathbf{P}(S^{(1)} > t), \quad t \to \infty. \tag{7.4.9}$$

7.5 Asymptotic behavior of the workload distribution

Let $V^{(2)}$ denote the steady-state workload at Q_2. It is shown in [18] that

$$\mathbf{P}(V^{(2)} < t) = (1-\rho)Y(t) + \lambda\int_{u=0}^{\infty}(1-B(u))\frac{1-M(t+u)}{1-B(t+u)}(1-\rho)Y(t+u)\mathrm{d}u,$$

for $t > 0$, which can be rewritten as (in the sequel, B^{res} and B_1^{res} will denote independent residual service times):

$$\mathbf{P}(V^{(2)} < t) = (1-\rho)\mathbf{P}(Y < t) + \rho\mathbf{P}(X < B_1^{res} + t, Y < B_1^{res} + t), \quad t > 0.$$

Hence,

$$\begin{aligned}\mathbf{P}(V^{(2)} > t) &= (1-\rho)\mathbf{P}(Y > t) + \rho\mathbf{P}(X > B_1^{res} + t) + \rho\mathbf{P}(Y > B_1^{res} + t)\\ &\quad -\rho\mathbf{P}(X > B_1^{res} + t, Y > B_1^{res} + t), \quad t > 0.\end{aligned}$$

Noting that

$$\begin{aligned}\mathbf{P}(X > B_1^{res} + t, Y > B_1^{res} + t) &\le \mathbf{P}(X > t, Y > B_1^{res} + t)\\ &= \mathbf{P}(X > t)\mathbf{P}(Y > B_1^{res} + t)\\ &= \mathrm{o}(\mathbf{P}(Y > B_1^{res} + t)), \quad t \to \infty,\end{aligned}$$

and using Lemma 7.3.3, we obtain that, as $t \to \infty$,

$$\mathbf{P}(V^{(2)} > t) \sim \rho[\mathbf{P}(B^{res} > t) + \mathbf{P}(X > B_1^{res} + t) + \mathbf{P}(Y > B_1^{res} + t)].$$

Using Lemma 7.3.3 again, for any $\epsilon > 0$, there exist a $T > 0$ such that for any $t > T$, we have

$$\left|\mathbf{P}(Y > t) - \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t)\right| \leq \epsilon \mathbf{P}(B^{res} > t).$$

It follows that, for $t > T$,

$$\begin{aligned} &\left|\int_{z=0}^{\infty} (\mathbf{P}(Y > z+t) - \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > z+t))\mathrm{d}B_1^{res}(z)\right| \\ \leq \quad & \epsilon \int_{z=0}^{\infty} \mathbf{P}(B^{res} > z+t)\mathrm{d}B_1^{res}(z), \end{aligned}$$

i.e., as $t \to \infty$,

$$\mathbf{P}(Y > B_1^{res} + t) \sim \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > B_1^{res} + t).$$

Thus we obtain for $t \to \infty$,

$$\begin{aligned} \mathbf{P}(V^{(2)} > t) \quad \sim \quad & \rho\Big[\mathbf{P}(B^{res} > t) + \mathbf{P}(X > B_1^{res} + t) \\ & + \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > B_1^{res} + t)\Big]. \end{aligned} \tag{7.5.1}$$

Applying similar arguments as above and using (7.3.1) and (7.4.2), we can prove, for $t \to \infty$,

$$\begin{aligned} \mathbf{P}(X > B_1^{res} + t) \quad \sim \quad & \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > B_1^{res} + t) \\ & + \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta}(t+z)(1-B(t+z))\mathrm{d}z. \end{aligned} \tag{7.5.2}$$

Now combining (7.5.1) and (7.5.2), we obtain for $t \to \infty$,

$$\begin{aligned} \mathbf{P}(V^{(2)} > t) \quad \sim \quad & \rho\Bigg[\mathbf{P}(B^{res} > t) + \frac{2\rho}{1-\rho}\mathbf{P}(B^{res} > B_1^{res} + t) \\ & + \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta}(t+z)(1-B(t+z))\mathrm{d}z\Bigg]. \end{aligned}$$

Slightly rewriting this result, we have proven the following:

Theorem 7.5.1 *For $t \to \infty$,*

$$\begin{aligned}\mathbf{P}(V^{(2)} > t) \quad \sim \quad & \rho[\mathbf{P}(B^{res} > t) + \frac{2\rho}{1-\rho}\mathbf{P}(B^{res} > B_1^{res} + t) \\ & + \frac{\lambda}{1-\rho} t\mathbf{P}(B > B_1^{res} + t) \\ & + \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta} z(1-B(t+z))\mathrm{d}z]. \qquad (7.5.3)\end{aligned}$$

For general service time distributions, we have now expressed the tail behavior of the distribution of the workload at Q_2 in terms of the (residual) service time distribution. It would be easy to specify the workload tail behavior for specific service time distributions with an exponential tail. Instead, we now restrict our attention to the case that the service time distribution is *heavy-tailed*, cf. Definition 2.1.1. The class of heavy-tailed distributions contains the class of subexponential distributions, which in turn contains the class of regularly varying distributions. It is easy to prove that, if $\mathbf{P}(B > t)$ is heavy-tailed and D is any nonnegative random variable that is independent of B, then $\dfrac{\mathbf{P}(B - D > t)}{\mathbf{P}(B > t)} \to 1$ as $t \to \infty$. We can apply this rule to replace $\mathbf{P}(B > B_1^{res} + t)$ by $\mathbf{P}(B > t)$ in (7.5.3). Actually, the fact that $\mathbf{P}(B > t)$ is heavy-tailed implies that $\mathbf{P}(B^{res} > t)$ is heavy-tailed by using l'Hospital's rule (but the reverse is not true in general, cf. [72]). Therefore, the second term in the righthand side of (7.5.3) can be replaced by $\dfrac{2\rho}{1-\rho}\mathbf{P}(B^{res} > t)$. If furthermore $\mathrm{E}B^2 < \infty$, then the last term in the righthand side of (7.5.3) can be replaced by $\dfrac{\lambda}{1-\rho}\dfrac{\mathrm{E}B^2}{2\beta}\mathbf{P}(B > t)$. One can prove this by applying the Dominated Convergence Theorem,

$$\begin{aligned} & \lim_{t\to\infty} \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta}\frac{1-B(t+z)}{1-B(t)} z\mathrm{d}z \\ = \quad & \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta} \lim_{t\to\infty}\left(\frac{1-B(t+z)}{1-B(t)}\right) z\mathrm{d}z \\ = \quad & \frac{\lambda}{1-\rho}\frac{\mathrm{E}B^2}{2\beta}. \end{aligned}$$

Below we restrict ourself to the subclass of regularly varying service time distributions. It then follows from Theorem 7.4.2 that $\mathbf{P}(X > t)$ is regularly

varying, hence heavy-tailed; therefore,

$$\mathbf{P}(X > B_1^{res} + t) \sim \mathbf{P}(X > t), \quad t \to \infty.$$

We can now conclude from (7.5.1) that the following result holds.

Theorem 7.5.2 *Let $\nu > 1$. If $\mathbf{P}(B > t)$ is regularly varying of index $-\nu$, then $\mathbf{P}(V^{(2)} > t)$ is regularly varying of index $1 - \nu$. More precisely, if*

$$\mathbf{P}(B > t) \sim t^{-\nu} L(t), \quad t \to \infty, \tag{7.5.4}$$

then

$$\mathbf{P}(V^{(2)} > t) \sim \frac{1}{\nu - 1} \frac{\lambda}{1 - \rho} (1 + \rho\nu) t^{1-\nu} L(t), \quad t \to \infty. \tag{7.5.5}$$

Remark 7.5.1 Under the conditions of Theorem 7.5.2, the tail of the waiting time distribution at Q_1 is regularly varying of index $1 - \nu$; this tail behavior in fact coincides with that of the distribution of $S^{(1)}$. Moreover, in the $M/G/1$ queue Q_1, the steady-state workload $V^{(1)}$ has the same distribution as the steady-state waiting time. Hence:

$$\mathbf{P}(S^{(1)} > t) \sim \mathbf{P}(V^{(1)} > t) \sim \frac{1}{\nu - 1} \frac{\lambda}{1 - \rho} t^{1-\nu} L(t), \quad t \to \infty,$$

which should be compared with (7.4.7) and (7.5.5).

7.6 Heuristics

In this section, we assume the service time B is regularly varying of index $-\nu$ ($\nu > 1$). We first give heuristic explanations of (7.4.9) and (7.5.5) by identifying two possible ways in which $S^{(2)}$ and $V^{(2)}$ may become large. These heuristic arguments are similar to those in Section 4.3 and those in Section 6.5 for the regime $\lambda > \mu$. Using these heuristic arguments, we also derive the asymptotic behavior of the total sojourn time distribution in the system. Note that the heuristic arguments are based on the preliminary observations made in Section 4.3.

Suppose a customer with a large service time B enters Q_1 in steady state at time 0. Assume that the total workloads at both queues are very small compared to B. So at time 0, the workload at Q_1 is roughly B and the workload at Q_2 is roughly 0. The workload at Q_1 decreases at rate $1 - \rho$ until

it becomes 0 at approximately time $\frac{B}{1-\rho}$ when the busy period at Q_1 which started at time 0 ends. At time B, this customer with service time B enters Q_2. During the time interval $(B, \frac{B}{1-\rho})$, Q_1 is still in its busy period. It means that the departure rate from Q_1 is $1/\beta$, which is the same as the service rate at Q_2. Thus, the workload at Q_2 stays at approximately the same level B. After time $\frac{B}{1-\rho}$, the effect of the large customer which entered Q_1 at time 0 has vanished at Q_1. Therefore, the workload at Q_2 decreases at rate $1-\rho > 0$ until it becomes 0 at time $\frac{2B}{1-\rho}$, see Figure 7.1.

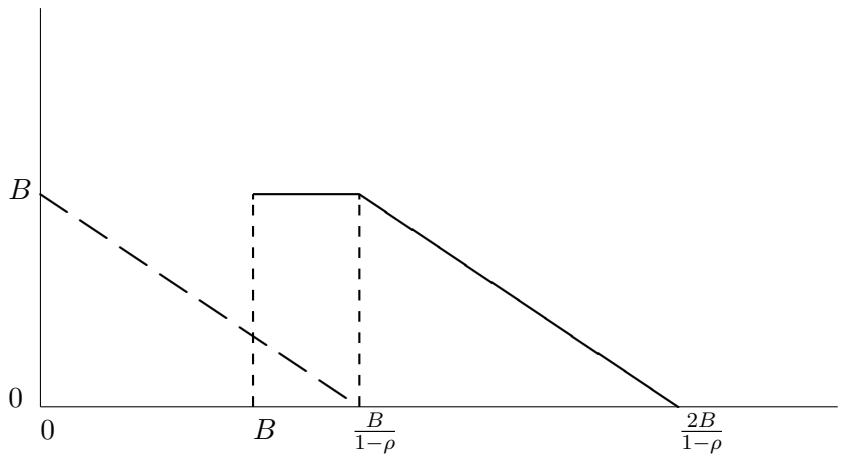

Figure 7.1: Evolution of the workload at Q_2.

Let us first consider the tail behavior of $S^{(2)}$. Suppose we observe the system at time y ($y \geq 0$) and a customer arrives at Q_1. We claim that the sojourn time at Q_2 of this particular customer is large, i.e., $S^{(2)} > t$ (t large), because at time 0, a customer with a large service time B entered Q_1. The arrival rate of the customers is λ. Keeping Figure 7.1 in mind, there are two possibilities:

1. $0 < y < \frac{B}{1-\rho}$, and $B > t$. The sojourn time at Q_2 of the customers who arrive at Q_1 in the time interval $(0, \frac{B}{1-\rho})$ is roughly B. So, we have

$$\begin{aligned} & \mathbf{P}(S^{(2)} > t, \text{ possibility 1 happens}) \\ \approx & \int_{y=0}^{\infty} \mathbf{P}(B > (1-\rho)y, B > t)\lambda \mathrm{d}y \end{aligned}$$

$$
\begin{aligned}
&= \lambda \int_{y=0}^{\frac{t}{1-\rho}} \mathbf{P}(B > t)\mathrm{d}y + \lambda \int_{y=\frac{t}{1-\rho}}^{\infty} \mathbf{P}(B > (1-\rho)y)\mathrm{d}y \\
&= \frac{\lambda t \mathbf{P}(B > t)}{1-\rho} + \frac{\lambda}{1-\rho} \int_{z=t}^{\infty} \mathbf{P}(B > z)\mathrm{d}z \\
&\sim \frac{\nu\rho}{1-\rho} \mathbf{P}(B^{res} > t) \sim \nu \mathbf{P}(S^{(1)} > t), \quad t \to \infty. \qquad (7.6.1)
\end{aligned}
$$

2. $\frac{B}{1-\rho} < y < \frac{2B}{1-\rho} - \frac{t}{1-\rho}$, and $B > t$. The customers who arrive at Q_1 during the time interval $(\frac{B}{1-\rho}, \frac{2B}{1-\rho} - \frac{t}{1-\rho})$ almost immediately enter Q_2. Thus, the sojourn time of those customers is roughly the same as the workload at Q_2. As can be seen from Figure 7.1, the workload at Q_2 during the time interval $(\frac{B}{1-\rho}, \frac{2B}{1-\rho})$ is approximately the same as the workload at Q_1 during the time interval $(0, \frac{B}{1-\rho})$. Thus, we may write

$$
\begin{aligned}
& \mathbf{P}(S^{(2)} > t, \text{ possibility 2 happens}) \\
\approx\ & \mathbf{P}(V^{(1)} > t) \sim \mathbf{P}(S^{(1)} > t), \quad t \to \infty. \qquad (7.6.2)
\end{aligned}
$$

Summing (7.6.1) and (7.6.2), we get the desired result:

$$
\mathbf{P}(S^{(2)} > t) \sim (\nu + 1)\mathbf{P}(S^{(1)} > t), \quad t \to \infty,
$$

which coincides with (7.4.9).

Now we turn to the tail behavior of the workload distribution at Q_2. Suppose we observe the system at time y ($y \geq 0$) and $V^{(2)} > t$, t large. We claim that at time y, the workload at Q_2 is large because at time 0, a customer with a large service time B entered Q_1. The arrival rate is λ. Keeping Figure 7.1 in mind, there are two possibilities:

1. $t < B < y < \frac{B}{1-\rho}$. Thus, we may write

$$
\begin{aligned}
& \mathbf{P}(V^{(2)} > t, \text{ possibility 1 happens}) \\
\approx\ & \int_{y=0}^{\infty} \mathbf{P}(t < B < y < \frac{B}{1-\rho})\lambda \mathrm{d}y \\
=\ & \lambda \int_{y=0}^{\infty} \mathbf{P}(B > t, y < \frac{\rho B}{1-\rho})\mathrm{d}y \\
=\ & \lambda \int_{y=0}^{\frac{\rho t}{1-\rho}} \mathbf{P}(B > t)\mathrm{d}y + \lambda \int_{y=\frac{\rho t}{1-\rho}}^{\infty} \mathbf{P}(B > \frac{(1-\rho)y}{\rho})\mathrm{d}y
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\lambda\rho t}{1-\rho}\mathbf{P}(B>t) + \frac{\lambda\rho}{1-\rho}\int_{z=t}^{\infty}\mathbf{P}(B>z)\mathrm{d}z \\
&\sim \nu\rho\mathbf{P}(V^{(1)}>t), \quad t\to\infty, \qquad (7.6.3)
\end{aligned}
$$

where the first equality follows since the integral only depends on the length of the interval for y.

2. $\frac{B}{1-\rho} < y < \frac{2B}{1-\rho} - \frac{t}{1-\rho}$. During this time interval, the workload at Q_2 is roughly the same as the sojourn time at Q_2. So, from (7.6.2), we may write

$$
\begin{aligned}
&\mathbf{P}(V^{(2)}>t, \text{ possibility 2 happens}) \\
\approx\ &\mathbf{P}(S^{(1)}>t) \sim \mathbf{P}(V^{(1)}>t), \quad t\to\infty. \qquad (7.6.4)
\end{aligned}
$$

Summing (7.6.3) and (7.6.4), we get the desired result:

$$
\mathbf{P}(V^{(2)}>t) \sim (\nu\rho+1)\mathbf{P}(V^{(1)}>t), \quad t\to\infty,
$$

which coincides with (7.5.5).

The total sojourn time, denoted by $S^{total} := S^{(1)} + S^{(2)}$ (note that $S^{(1)}$ and $S^{(2)}$ are dependent) is also an interesting quantity to study. In the following, we derive the tail behavior of the total sojourn time distribution $S^{total}(\cdot)$ by applying the above heuristic arguments.

Suppose we observe the system at time y ($y \geq 0$) and a customer arrives at Q_1. We claim that the total sojourn time S^{total} of this particular customer exceeds t, t large, because at time 0, a customer with a large service time B arrived at Q_1. The arrival rate is λ. Keeping Figure 7.1 in mind, there are two possibilities:

1. $0 < y < \frac{B}{1-\rho}$. In this case, $S^{(1)}$ is roughly the same as the workload at Q_1 and $S^{(2)}$ is roughly B. Thus, S^{total} is related to y as

$$
S^{total} = 2B - (1-\rho)y.
$$

We have

$$
\begin{aligned}
&\mathbf{P}(S^{total}>t, \text{ possibility 1 happens}) \\
\approx\ &\int_{y=0}^{\infty}\mathbf{P}(B>(1-\rho)y,\ B>\frac{(1-\rho)y}{2}+\frac{t}{2})\lambda\mathrm{d}y \\
=\ &\int_{y=0}^{\frac{t}{1-\rho}}\mathbf{P}(B>\frac{(1-\rho)y+t}{2})\lambda\mathrm{d}y + \int_{y=\frac{t}{1-\rho}}^{\infty}\mathbf{P}(B>(1-\rho)y)\lambda\mathrm{d}y
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2\rho}{1-\rho}\mathbf{P}(B^{res} > t/2) - \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t) \\
&\sim 2\mathbf{P}(S^{(1)} > t/2) - \mathbf{P}(S^{(1)} > t). \qquad (7.6.5)
\end{aligned}
$$

2. $\frac{B}{1-\rho} < y < \frac{2B}{1-\rho}$. In this case, $S^{(1)}$ is roughly 0 and $S^{(2)}$ is roughly the same as the workload at Q_2. By (7.6.2), we have

$$\mathbf{P}(S^{total} > t, \text{possibility 2 happens}) \approx \mathbf{P}(S^{(1)} > t). \qquad (7.6.6)$$

Summing (7.6.5) and (7.6.6) leads to

$$\mathbf{P}(S^{total} > t) \sim 2\mathbf{P}(S^{(1)} > t/2), \quad t \to \infty. \qquad (7.6.7)$$

It should be pointed out that Relation (7.6.7) is only a conjecture obtained from the above heuristic arguments. A more rigorous mathematical proof is needed. Boxma [18] gives the joint distribution of the waiting time distributions at the first and the second queue, which might be a starting point to prove (7.6.7). The above arguments can be extended to the tandem system consisting of k ($k \geq 2$) queues, where the service time is identical at each queue. We refrain from presenting the detailed analysis here.

7.7 The sojourn time distribution in heavy traffic

In [22], Boxma and Cohen have obtained a heavy-traffic limit theorem for the waiting time distribution in the $G/G/1$ queue (cf. Section 1.4), when the variance of the interarrival and/or the service time distribution is *infinite*. Exactly the same limit theorem holds at Q_1 for the sojourn time $S^{(1)}$ which is the sum of the waiting time and the (independent) service time. In the present section, we derive a heavy-traffic limit theorem for the sojourn time $S^{(2)}$ in the case of a regularly varying service time distribution of index $-\nu$, $\nu > 1$.

Theorem 7.7.1 *For the stable tandem queue with Poisson input process and identical service times at both queues, and with the service time distribution satisfying the condition of Theorem 7.4.2, i.e.,*

$$\mathbf{P}(B > t) \sim t^{-\nu}L(t), \quad t \to \infty, \qquad (7.7.1)$$

where $\nu > 1$, the contracted sojourn time $\Delta(\rho)S^{(2)}$ converges in distribution for $\rho \uparrow 1$. The limit distribution function $H(t)$ is given by:

$$H(t) = \frac{\exp(-t^{1-\nu})}{1+\nu t^{1-\nu}}, \quad t > 0,$$

and the coefficient of contraction $\Delta(\rho)$ is the unique root (for the uniqueness concept, we refer to Lemma 3.5.1) of the following equation:

$$x^{\nu-1}L(1/x) = \frac{(\nu-1)(1-\rho)}{\lambda}, \tag{7.7.2}$$

with the property that $\Delta(\rho) \downarrow 0$ for $\rho \uparrow 1$.

Proof. Let $\Delta(\rho)$ be the solution to Equation (7.7.2) with the property $\Delta(\rho) \downarrow 0$ for $\rho \uparrow 1$. As proved in Lemma 3.5.1, the solution $\Delta(\rho)$ with such a property exists and is unique. Using Theorem 1.6.1 in [12], it follows from (7.4.4) that for $t > 0$,

$$\lambda \int_{t/\delta}^{\infty} (1 - M(x))\mathrm{d}x \sim \frac{\lambda}{1-\rho}\frac{1}{\nu-1}\frac{t^{1-\nu}}{\delta^{1-\nu}}L(t/\delta), \quad \delta \downarrow 0,$$

which in combination with the definition of $\Delta(\rho)$ yields

$$\lim_{\rho\uparrow 1} \lambda \int_{t/\Delta(\rho)}^{\infty} (1 - M(x))\mathrm{d}x = t^{1-\nu}.$$

Thus by (7.2.5), for $t > 0$,

$$\lim_{\rho\uparrow 1} Y(t/\Delta(\rho)) = \lim_{\rho\uparrow 1} \exp\left(-\lambda \int_{t/\Delta(\rho)}^{\infty} (1 - M(x))\mathrm{d}x\right) = \exp(-t^{1-\nu}). \tag{7.7.3}$$

By (7.7.1), it is easy to get

$$\lim_{\rho\uparrow 1} B(t/\Delta(\rho)) = 1. \tag{7.7.4}$$

Applying Theorem 1.6.5 in [12], (7.4.3) implies that, for $t > 0$,

$$\int_{t/\delta}^{\infty} x\mathrm{d}B(x) \sim \frac{\nu}{\nu-1}\frac{t^{1-\nu}}{\delta^{1-\nu}}L(t/\delta), \quad \delta \downarrow 0,$$

which further implies that

$$\lim_{\rho\uparrow 1} \frac{\lambda}{1-\rho} \int_{t/\Delta(\rho)}^{\infty} x\mathrm{d}B(x) = \nu t^{1-\nu}. \tag{7.7.5}$$

Since (7.3.9) and (7.7.5) imply that $\lim_{\rho\uparrow 1} F(t/\Delta(\rho)) = 0$ where $F(\cdot)$ is given by (7.3.5), it follows from (7.2.4) and (7.3.6) that, for $t > 0$,

$$\lim_{\rho\uparrow 1} X(t/\Delta(\rho)) = \frac{1}{1+\nu t^{1-\nu}}. \tag{7.7.6}$$

By (7.2.3) and (7.2.6), we can rewrite $S^{(2)}(t)$ as

$$S^{(2)}(t) = X(t)Y(t)B(t). \tag{7.7.7}$$

Combining (7.7.3), (7.7.4), (7.7.6) and (7.7.7) leads to

$$\lim_{\rho\uparrow 1} S^{(2)}(t/\Delta(\rho)) = \frac{\exp(-t^{1-\nu})}{1+\nu t^{1-\nu}},$$

which finally implies that, for $t > 0$,

$$\lim_{\rho\uparrow 1} \mathbf{P}(\Delta(\rho)S^{(2)} \le t) = \lim_{\rho\uparrow 1} S^{(2)}(t/\Delta(\rho)) = \frac{\exp(-t^{1-\nu})}{1+\nu t^{1-\nu}}.$$

□

Remark 7.7.1 The limiting distribution function $H(t)$ is easily seen to have a regularly varying tail of the same index as the tail of $S^{(2)}(t)$. It is interesting to observe that $\exp(-t^{1-\nu})$, $t > 0$, is a Weibull distribution, cf. Feller [52] p. 52.

Remark 7.7.2 The above heavy-traffic limit theorem may be used to provide an approximation for $S^{(2)}(t)$; for such an approach to the ordinary $M/G/1$ queue and the $M/G/1$ queue with priority classes, see [21] and Section 3.8, respectively.

Remark 7.7.3 In case both the service time and interarrival time distributions have a finite second moment, Kingman [68] derives a standard heavy-traffic limit theorem for the stationary waiting time W in the $G/G/1$ queue. In our tandem model, if $\nu > 2$, a similar limit theorem holds for the sojourn time $S^{(1)}$ at Q_1, i.e.,

$$\lim_{\rho\uparrow 1} \mathbf{P}(\zeta(\rho)S^{(1)} \le t) = 1 - e^{-t}, \quad t \ge 0,$$

with $\zeta(\rho) := 2\lambda(1-\rho)/[1+\lambda^2(\mathrm{E}B^2 - \beta^2)]$ (cf. [68]).

Remark 7.7.4 Taking $L(x) = 1$ in (7.7.2), one can immediately get $\Delta(\rho) = (\frac{\nu-1}{\lambda})^{\frac{1}{\nu-1}}(1-\rho)^{\frac{1}{\nu-1}}$. In fact, it is not surprising that when $\nu > 2$, the contraction coefficient $\Delta(\rho)$ of $S^{(2)}$ is much larger than the above contraction coefficient $\zeta(\rho)$ of $S^{(1)}$ for $\rho \uparrow 1$. As has been shown in [18], if the third moment β_3 of the service time is finite, then

$$\lim_{\rho\uparrow 1} \frac{\mathrm{E}S^{(2)}}{\mathrm{E}S^{(1)}} = 0, \qquad \lim_{\rho\uparrow 1} \frac{\mathrm{Var}(S^{(2)})}{\mathrm{Var}(S^{(1)})} = 0.$$

In fact, applying the technique used in [18] to derive the above limit results, one can show that if the $(n+1)$th moment β_{n+1} is finite $(n \geq 2)$, then

$$\mathrm{E}[(S^{(2)})^n] \leq \frac{c}{(1-\rho)^{\frac{2n+1}{n+1}}},$$

for some positive constant c.

Bibliography

[1] J. Abate, G.L. Choudhury and W. Whitt (1994). *Waiting-time tail probabilities in queues with long-tail service-time distributions.* Queueing Systems **16**, 311-338.

[2] J. Abate and W. Whitt (1992). *The Fourier-series method for inverting transforms of probability distributions.* Queueing Systems **10**, 5-88.

[3] J. Abate and W. Whitt (1997). *Asymptotics for M/G/1 low-priority waiting-time tail probabilities.* Queueing Systems **25**, 173-233.

[4] R. Agrawal, A.M. Makowski and Ph. Nain (1999). *On a reduced load equivalence for fluid queues under subexponentiality.* Queueing Systems **33**, 5-41.

[5] E. Altman, P. Konstantopoulos and Z. Liu (1992). *Stability, monotonicity and invariant quantities in general polling systems.* Queueing Systems **11**, 35-57.

[6] V. Anantharam (1996). *Networks of queues with long-range dependent traffic streams.* In: Stochastic Networks - Stability and Rare Events. Eds. P. Glasserman, K. Sigman and D.D. Yao. Springer-Verlag, Berlin, 237-256.

[7] S. Asmussen, C. Klüppelberg and K. Sigman (1999). *Sampling at subexponential times, with queueing applications.* Stochastic Processes and their Applications **79**, 265-286.

[8] K.B. Athreya and P.E. Ney (1972). Branching Processes. Springer-Verlag, Berlin.

[9] F. Baccelli, S. Schlegel and V. Schmidt (1999). *Asymptotics of stochastic networks with subexponential service times.* Queueing Systems **33**, 205-232.

[10] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger (1995). *Long-range dependence in variable-bit-rate video traffic.* IEEE Transactions on Communications **43**, 1566-1579.

[11] D. Bertsimas and G. Mourtzinou (1999). *Decomposition results for general polling systems and their applications.* Queueing Systems **31**, 295-316.

[12] N.H. Bingham, C.M. Goldie and J.L. Teugels (1987). Regular Variation. Cambridge University Press, Cambridge.

[13] S.C. Borst (1994). Polling Systems. Ph.D. thesis, Tilburg University.

[14] S.C. Borst and O.J. Boxma (1997). *Polling models with and without switchover times.* Operations Research **45**, 536-543.

[15] S.C. Borst, O.J. Boxma and P.R. Jelenković (1999). *Generalized processor sharing with long-tailed traffic sources.* In: Teletraffic Engineering in a Competitive World, Proceedings of ITC-16. Eds. P. Key and D. Smith. North-Holland, Amsterdam, 345-354.

[16] S.C. Borst, O.J. Boxma and M.J.G. van Uitert (2001). *Two coupled queues with heterogeneous traffic.* To appear in the proceedings of ITC-17.

[17] O.J. Boxma (1978). *On the longest service time in a busy period of the M/G/1 queue.* Stochastic Processes and their Applications **8**, 93-100.

[18] O.J. Boxma (1979). *On a tandem queueing model with identical service times at both counters, I, II.* Advances in Applied Probability **11**, 616-643; 644-659.

[19] O.J. Boxma (1980). *The longest service time in a busy period.* Zeitschrift für Operations Research **24**, 235-242.

[20] O.J. Boxma (1989). *Workloads and waiting times in single-server systems with multiple customer classes.* Queueing Systems **5**, 185-214.

[21] O.J. Boxma and J.W. Cohen (1998). *The M/G/1 queue with heavy-tailed service time distribution.* IEEE Journal on Selected Areas in Communications **16**, 349-363.

[22] O.J. Boxma and J.W. Cohen (1999). *Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions.* Queueing Systems **33**, 177-204.

[23] O.J. Boxma and J.W. Cohen (2000). *The single server queue: heavy tails and heavy traffic.* In: Self-Similar Network Traffic and Performance Evaluation. Eds. K. Park and W. Willinger. John Wiley & Sons, New York, 143-170.

[24] O.J. Boxma, J.W. Cohen and Q. Deng (1999). *Heavy-traffic analysis of the M/G/1 queue with priority classes.* In: Teletraffic Engineering in a Competitive World, Proceedings of ITC-16. Eds. P. Key and D. Smith. North-Holland, Amsterdam, 1157-1167.

[25] O.J. Boxma and Q. Deng (2000). *Asymptotic behaviour of the tandem queueing system with identical service times at both queues.* Mathematical Methods of Operations Research **52**, 307-323.

[26] O.J. Boxma, Q. Deng and J.A.C. Resing (2000). *Polling systems with regularly varying service and/or switchover times.* Advances in Performance Analysis **3**, 71-107.

[27] O.J. Boxma, Q. Deng and A.P. Zwart (2001). *Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers.* To appear in Queueing Systems.

[28] O.J. Boxma and V. Dumas (1998). *Fluid queues with heavy-tailed activity period distributions.* Computer Communications **21**, 1509-1529.

[29] O.J. Boxma and V. Dumas (1998). *The busy period in the fluid queue.* Performance Evaluation Review **26**, 100-110.

[30] O.J. Boxma and W.P. Groenendijk (1987). *Pseudo-conservation laws in cyclic service systems.* Journal of Applied Probability **24**, 949-964.

[31] O.J. Boxma and I.A. Kurkova (2001). *The M/G/1 queue with two service speeds.* To appear in Advances in Applied Probability **33**, No. 2.

[32] V.P. Chistyakov (1964). *A theorem on sums of independent, positive random variables and its applications to branching processes.* Theory of Probability and its Applications **9**, 640-648.

[33] G.L. Choudhury and W. Whitt (1996). *Computing distributions and moments in polling models by numerical transform inversion.* Performance Evaluation **25**, 267-292.

[34] J.W. Cohen (1973). *Some results on regular variation for distributions in queueing and fluctuation theory.* Journal of Applied Probability **10**, 343-353.

[35] J.W. Cohen (1974). *Superimposed renewal processes and storage with gradual input.* Stochastic Processes and their Applications **2**, 31-58.

[36] J.W. Cohen (1982). The Single Server Queue. North-Holland, Amsterdam; 2nd revised edition.

[37] R.B. Cooper (1972). Introduction to Queueing Theory. Macmillan, London.

[38] R.B. Cooper (1990). *Queueing theory.* In: Stochastic Models, Handbooks in Operations Research and Management Science **2**. Eds. D.P. Heyman and M.B. Sobel. North-Holland, Amsterdam, 469-518.

[39] D.R. Cox (1962). Renewal Theory. Methuen, London.

[40] D.R. Cox (1984). *Long-range dependence: a review.* In: Statistics: An Appraisal. Eds. H.A. David and H.T. David. Iowa State University Press, 55-74.

[41] T. Daniëls (1999). Asymptotic Behaviour of Queueing Systems. Ph.D. Thesis, Universiteit Antwerpen.

[42] Q. Deng (2001). *The two-queue $E/1$-L polling model with regularly varying service and/or switchover times.* In preparation.

[43] G. Doetsch (1950). Handbuch der Laplace Transformation, Vol. I, II, III. Birkhäuser Verlag, Basel.

[44] G. Doetsch (1974). Introduction to the Theory and Applications of the Laplace Transformation. Springer-Verlag, New York.

[45] N.G. Duffield (1997). *Exponents for the tail of distributions in some polling models.* Queueing Systems **26**, 105-119.

[46] N.G. Duffield (1998). *Queueing at large resources driven by long-tailed $M/G/\infty$-modulated processes.* Queueing Systems **28**, 245-266.

[47] N.G. Duffield and N. O'Connell (1995). *Large deviations and overflow probabilities for the general single server queue, with applications.* Mathematical Proceedings of the Cambridge Philosophical Society **118**, 363-374.

[48] V. Dumas and A. Simonian (2000). *Asymptotic bounds for the fluid queue fed by sub-exponential On/Off sources.* Advances in Applied Probability **32**, 244-255.

[49] M. Eisenberg (1972). *Queues with periodic service and changeover times.* Operations Research **20**, 440-451.

[50] P. Embrechts, C. Klüppelberg and T. Mikosch (1997). Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Heidelberg.

[51] P. Embrechts, C.M. Goldie and N. Veraverbeke (1979). *Subexponentiality and infinite divisibility.* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **49**, 335-347.

[52] W. Feller (1970). An Introduction to Probability Theory and its Applications, Vol. II. John Wiley & Sons, New York.

[53] S.G. Foss and D.A. Korshunov (1999). *On waiting time distribution in GI/GI/2 queue system with heavy tailed service times.* Unpublished manuscript.

[54] C. Fricker and M.R. Jaibi (1994). *Monotonicity and stability of periodic polling models.* Queueing Systems **15**, 211-238.

[55] S.W. Fuhrmann and R.B. Cooper (1985). *Stochastic decompositions in the M/G/1 queue with generalized vacations.* Operations Research **33**, 1117-1129.

[56] B.V. Gnedenko and V.Yu. Korolev (1996). Random Summation. CRC Press, Boca Raton.

[57] C.M. Goldie and C. Klüppelberg (1998). *Subexponential distributions.* In: A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions. Eds. R. Adler, R. Feldman and M.S. Taqqu. Birkhäuser, Boston.

[58] J. Grandell (1997). Mixed Poisson Processes. Chapman & Hall, London.

[59] W.P. Groenendijk (1990). Conservation Laws in Polling Systems. Ph.D. Thesis, University of Utrecht.

[60] R. Haji and G.F. Newell (1971). *A relation between stationary queue length and waiting time distributions.* Journal of Applied Probability **8**, 617-620.

[61] J.M. Harrison (1973). *A limit theorem for priority queues in heavy traffic.* Journal of Applied Probability **10**, 907-912.

[62] P. Heidelberger and S.S. Lavenberg (1984). *Computer performance evaluation methodology.* IEEE Transactions on Computers **C-33**, 1195-1220.

[63] T. Huang and K. Sigman (1999). *Delay asymptotics for tandem, split & match, and other feedforward queues with heavy tailed service.* Queueing Systems **33**, 233-259.

[64] O.C. Ibe (1990). *Analysis of polling systems with mixed service disciplines.* Communications in Statistics, Stochastic Models **6**, 667-689.

[65] J. Keilson and L.D. Servi (1990). *The distributional form of Little's law and the Fuhrmann-Cooper decomposition.* Operations Research Letters **9**, 239-247.

[66] D.G. Kendall (1953). *Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain.* Annals of Mathematical Statistics **24**, 338-354.

[67] J. Kiefer and J. Wolfowitz (1956). *On the characteristics of the general queueing process with applications to a random walk.* Annals of Mathematical Statistics **27**, 147-161.

[68] J.F.C. Kingman (1965). *The heavy traffic approximation in the theory of queues.* In: Proceedings of the Symposium on Congestion Theory. Eds. W.L. Smith and W.E. Wilkinson. The University of North Carolina Press, Chapel Hill, 137-159.

[69] J.F.C. Kingman and S.J. Taylor (1966). Introduction to Measure and Probability. Cambridge University Press, Cambridge.

[70] L. Kleinrock (1964). Communication Nets: Stochastic Message Flow and Delay. McGraw-Hill, New York.

[71] L. Kleinrock (1975), (1976). Queueing Systems, Vol. I, II. John Wiley & Sons, New York.

[72] C. Klüppelberg (1988). *Subexponential distributions and integrated tails.* Journal of Applied Probability **25**, 132-141.

[73] C. Knessl, B.J. Matkowsky, Z. Schuss and C. Tier (1990). *An integral equation approach to the M/G/2 queue.* Operations Research **38**, 506-518.

[74] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson (1994). *On the self-similar nature of Ethernet traffic* (extended version). IEEE/ACM Transactions on Networking **2**, 1-15.

[75] N. Likhanov, B. Tsybakov and N. Georganas (1995). *Analysis of an ATM buffer with self-similar ("fractal") input traffic.* In: Proceedings of IEEE INFOCOM '95, 985-992.

[76] A.V. Makarichev (1984). *Analysis of a tandem queueing system with identical service times at both counters for various service disciplines.* In: Proceedings of 3rd ITC Seminar (Moscow, June 1984), 298-301.

[77] B.B. Mandelbrot (1982). The Fractal Geometry of Nature. W.H. Freeman, New York.

[78] B.B. Mandelbrot and J.W. Van Ness (1968). *Fractional Brownian motions, fractional noises and applications.* SIAM Review **10**, 422-437.

[79] M.R.H. Mandjes and S.C. Borst (2000). *Overflow behavior in queues with many long-tailed inputs.* Advances in Applied Probability **32**, 1150-1167.

[80] L. Massoulié and A. Simonian (1999). *Large buffer asymptotics for the queue with fractional Brownian input.* Journal of Applied Probability **36**, 894-906.

[81] A. De Meyer and J.L. Teugels (1980). *On the asymptotic behavior of the distributions of the busy period and service time in M/G/1.* Journal of Applied Probability **17**, 802-813.

[82] O. Narayan (1998). *Exact asymptotic queue length distribution for fractional Brownian traffic.* Advances in Performance Analysis **1**, 39-64.

[83] I. Norros (1994). *A storage model with self-similar input.* Queueing Systems **16**, 387-396.

[84] I. Norros (1995). *On the use of fractional Brownian motion in the theory of connectionless networks.* IEEE Journal on Selected Areas in Communications **13**, 953-962.

[85] A.G. Pakes (1975). *On the tails of waiting-time distributions.* Journal of Applied Probability **12**, 555-564.

[86] K. Park and W. Willinger (2000). *Self-similar network traffic: an overview.* In: Self-Similar Network Traffic and Performance Evaluation. Eds. K. Park and W. Willinger. John Wiley & Sons, New York, 1-38.

[87] K. Park and W. Willinger, eds. (2000). Self-Similar Network Traffic and Performance Evaluation. John Wiley & Sons, New York.

[88] M. Parulekar and A.M. Makowski (1996). *Tail probabilities for a multiplexer with self-similar traffic.* In: Proceedings of IEEE INFOCOM '96, 1452-1459.

[89] V. Paxson and S. Floyd (1995). *Wide area traffic: the failure of Poisson modeling.* IEEE/ACM Transactions on Networking **3**, 226-244.

[90] J.A.C. Resing (1993). *Polling systems and multitype branching processes.* Queueing Systems **13**, 409-426.

[91] S. Resnick and G. Samorodnitsky (1999). *Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues.* Report no. 1242, Department of ORIE, Cornell University; to appear in Bernoulli.

[92] S. Resnick and G. Samorodnitsky (2000). *A heavy traffic approximation for workload processes with heavy tailed service requirements.* Management Science **46**, 1236-1248.

[93] J.W. Roberts, U. Mocci and J. Virtamo (1996). Broadband Network Traffic - Final Report of Action COST 242. Springer-Verlag, Berlin.

[94] G. Samorodnitsky and M.S. Taqqu (1994). Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance. Chapman & Hall, London.

[95] A. Scheller-Wolf (2000). *Further delay moment results for FIFO multiserver queues.* Queueing Systems **34**, 387-400.

[96] A. Scheller-Wolf and K. Sigman (1997). *Delay moments for FIFO GI/GI/s queues.* Queueing Systems **25**, 77-95.

[97] E. Seneta (1981). *Non-negative Matrices and Markov Chains.* Springer-Verlag, New York; 2nd edition.

[98] A. Shwartz and A. Weiss (1995). Large Deviations for Performance Analysis. Chapman & Hall, London.

[99] K. Sigman (1999). *Appendix: A primer on heavy-tailed distributions.* Queueing Systems **33**, 261-275.

[100] M.M. Srinivasan, S.-C. Niu and R.B. Cooper (1995). *Relating polling models with nonzero and zero switchover times.* Queueing Systems **19**, 149-168.

[101] W.G.L. Sutton (1934). *The asymptotic expansion of a function whose operational equivalent is known.* Journal of the London Mathematical Society **9**, 131-137.

[102] H. Takagi (1990). *Queueing analysis of polling models: an update.* In: Stochastic Analysis of Computer and Communication Systems. Ed. H. Takagi. Elsevier, Amsterdam, 267-318.

[103] H. Takagi (1997). *Queueing analysis of polling models: Progress in 1990-1994.* In: Frontiers in Queueing. Ed. J.H. Dshalalow. CRC Press, Boca Raton, 119-146.

[104] H. Takagi, T. Takine and O.J. Boxma (1992). *Distribution of the workload in multiclass queueing systems with server vacations.* Naval Research Logistics **39**, 41-52.

[105] L. Takács (1967). Combinatorial Methods in the Theory of Stochastic Processes. John Wiley & Sons, New York.

[106] M.S. Taqqu (1986). *A bibliographical guide to self-similar process and long-range dependence.* In: Dependence in Probability and Statistics. Eds. E. Eberlein and M.S. Taqqu. Birkhäuser, Boston, 137-162.

[107] E.C. Titchmarsh (1952). The Theory of Functions. Oxford University Press, London.

[108] B. Tsybakov and N.D. Georganas (1998). *Self-similar traffic and upper bounds to buffer overflow in an ATM queue.* Performance Evaluation **36**, 57-80.

[109] O.P. Vinogradov (1984). *On the distribution of sojourn time in the tandem system with identical service times.* In: Proceedings of 3rd ITC Seminar (Moscow, June 1984), 449-450.

[110] J. Walrand and P. Varaiya (1996). High-Performance Communication Networks. Morgan Kaufmann, San Francisco.

[111] W. Whitt (1971). *Weak convergence theorems for priority queues: preemptive-resume discipline.* Journal of Applied Probability **8**, 74-94.

[112] W. Whitt (2000). *The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution.* Queueing Systems **36**, 71-87.

[113] W. Willinger, M.S. Taqqu, W.E. Leland and D.V. Wilson (1995). *Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements.* Statistical Science **10**, 67-85.

[114] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson (1997). *Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level.* IEEE/ACM Transactions on Networking **5**, 71-86.

[115] A.P. Zwart (1999). *Sojourn times in a multiclass processor sharing queue.* In: Teletraffic Engineering in a Competitive World, Proceedings of ITC-16. Eds. P. Key and D. Smith. North-Holland, Amsterdam, 335-344.

[116] A.P. Zwart and O.J. Boxma (2000). *Sojourn time asymptotics in the M/G/1 processor sharing queue.* Queueing Systems **35**, 141-166.

[117] A.P. Zwart (2001). *Tail asymptotics for the busy period in the GI/G/1 queue.* To appear in Mathematics of Operations Research.

Summary

Queueing theory plays an important role in the design of telecommunication networks. Simple models, like fluid queues or classical single-server queues, can often be used to obtain insightful results, e.g., to predict the global traffic behavior. Traditional queueing models typically assume that the interarrival and service times have finite variance (e.g., exponential or Erlang distribution). As a result, the aggregate traffic that is offered by a collection of sources behaves like white noise. Recently, it has become clear that delay and buffer content distributions in modern communication networks often do not exhibit such a behavior like white noise. Many studies on traffic measurements from a variety of communication networks, have shown a striking difference between actual network traffic and assumptions in traditional theoretical traffic models. That is, actual network traffic is often *self-similar* or *long-range dependent* in nature. In other words, the traffic looks statistically the same over a wide range of time scales, from milliseconds to minutes and even hours. This conclusion is supported by statistical analysis of numerous high-quality Ethernet and Internet traffic measurements.

It has been shown that queueing models with regularly varying service times, with an index of regular variation between -2 and -1, may be very useful in modelling modern network traffic. This thesis is devoted to the performance analysis of several fundamental classes of queueing models, with the special feature of regularly varying service times. Tail probabilities (of the waiting time or workload) receive special attention.

More specifically, we study in detail the following four queueing models: (i) the $M/G/1$ queue with priority classes, (ii) the tandem queueing system with Poisson input processes and identical service times at both queues, (iii) the cyclic polling system with Poisson input processes, and (iv) the $M/G/2$ queue with heterogeneous servers. For these models, we assume that at least one of the service times has a regularly varying (sometimes heavy-tailed) distribution. By analyzing the asymptotic behavior of the corresponding Laplace-Stieltjes Transforms (LSTs) in the neighborhood of the rightmost singularity and ap-

plying the Tauberian Theorem, we find, for models (i), (ii) and (iii), the cyclic polling system), that the service time with the largest tail probability governs the tail behavior of the waiting time and workload distributions. For the multi-server queue (the $M/G/2$ queue with heterogeneous servers), the waiting time tail behavior depends not only on the service time tail behavior, but also on the total traffic load. We also developed intuitive arguments which illustrate how large waiting time (or workload) occurs.

We now briefly discuss the chapters in this thesis. Chapter 1 provides the background and motivation for this thesis, and presents some basic knowledge of queueing theory and the performance analysis of computer-communication networks. Some relevant work on the topic of queues with heavy tails is also discussed.

Chapter 2 is devoted to the basic properties of heavy-tailed distributions. Special attention is paid to regularly varying distributions.

We study the $M/G/1$ queue with two priority classes in Chapter 3. The service times of the high- and/or low-priority customers are assumed to be regularly varying of index $-\nu$ $(1 < \nu < 2)$. Based on an expression for the LST of the low-priority waiting time distribution, we establish relations between the tail behavior of the waiting time distribution of the low-priority customers and that of the service time distributions. Furthermore, we derive a heavy-traffic limit theorem the waiting time distribution of the low-priority customers when the total traffic load $\rho \uparrow 1$.

Chapters 4 and 5 are devoted to the cyclic polling system with Poisson arrival processes. In Chapter 4, we study a two-queue model with exhaustive service at one queue and 1-limited service at the other queue. Note that this model reduces to the $M/G/1$ queue with two priority classes of Chapter 3 if there is no switchover time. For the case in which there are switchover times and the service times have an infinite variance, we derive a heavy-traffic limit theorem for the waiting time at the second queue. Finally we numerically test the approximation of the waiting time distribution at the second queue suggested by the heavy-traffic limit theorem.

In Chapter 5, we study the cyclic polling system with gated or exhaustive service at each queue. It is assumed that the service time distribution with the heaviest tail behavior has a regularly varying tail of index $-\nu$ $(\nu > 1)$. Based on an explicit expression for the LST of the waiting time distributions, we prove that the waiting time distribution at each queue is regularly varying of index $1 - \nu$.

Chapter 6 is devoted to the $M/G/2$ queue with one exponential server and one general server. Using the supplementary variable technique, we establish

a set of differential equations satisfying some boundary condition. In the case that the LST of the service time distribution at the general server is rational, we can explicitly solve the differential equations and thus the LST of the steady-state waiting time distribution follows. In the case that the service time at the general server has a regularly varying tail, we derive the tail behavior of the waiting time by using analytic methods. Furthermore, we provide intuitive arguments for the waiting time tail behavior.

In Chapter 7, we turn to the tandem queueing system with identical service times at both queues. We focus on the steady-state sojourn time and workload at the second queue. Starting from explicit expressions for the distributions of the sojourn time and workload at the second queue, we relate the tail behavior of these distributions to the tail behavior of the (residual) service time distribution. As a by-product, we prove that both the sojourn time distribution and the workload distribution at the second queue are regularly varying of index $1-\nu$, if the service time distribution is regularly varying of index $-\nu$ ($\nu > 1$), which coincides with the results we obtain by using intuitive arguments. Furthermore, in the latter case, we derive a heavy-traffic limit theorem for the sojourn time at the second queue when the traffic load $\rho \uparrow 1$.

Samenvatting (Summary)

Wachtrijtheorie speelt een belangrijke rol bij het ontwerp van telecommunicatienetwerken. Eenvoudige modellen, zoals vloeistofmodellen of klassieke wachtrijsystemen met één bediende, kunnen dikwijls worden gebruikt om inzicht te verkrijgen in het globale gedrag van communicatienetwerken en voorspellingen te doen omtrent de kwaliteit van hun dienstverlening. In de klassieke wachtrijmodellen wordt doorgaans verondersteld dat de aankomst- en bedieningstijden stochastische variabelen zijn met een eindige variantie. Dikwijls worden zij door een verdeling met een exponentiële staart gerepresenteerd, zoals de exponentiële verdeling of de Erlang verdeling. Het is in recente studies echter duidelijk geworden, dat zulke veronderstellingen niet altijd opgaan voor het verkeer in moderne communicatienetwerken. Diverse studies betreffende verkeersmetingen aan een groot scala aan communicatienetwerken hebben opvallende verschillen aan het licht gebracht tussen echt netwerkverkeer en verkeer dat is gemodelleerd met traditionele verdelingen met exponentiële staarten. Echt netwerkverkeer is dikwijls *self-similar* of *long-range dependent*. Met andere woorden, het verkeer vertoont hetzelfde patroon over een groot aantal tijdschalen, van milliseconden tot minuten en zelfs uren. Deze conclusie wordt ondersteund door statistische analyses van talrijke metingen aan verkeer in lokale Ethernet netwerken, Internet verkeer, enz.

Wachtrijsystemen met regulier variërende bedieningstijden, met een index van variatie tussen -2 en -1, zijn onlangs nuttig gebleken bij de modellering en analyse van modern communicatieverkeer. Dit proefschrift is daarom gewijd aan de bestudering van verscheidene belangrijke klassen van wachtrijsystemen onder de aanname dat de bedieningstijden een regulier variërende verdeling hebben. We zijn in het bijzonder geinteresseerd in staartkansen van wachttijden en werklast.

We bestuderen de volgende vier wachtrijsystemen in detail: (i) de $M/G/1$ wachtrij met prioriteiten, (ii) twee wachtrijen in serie, met een Poisson aankomstproces bij de eerste wachtrij en identieke bedieningstijden van klanten bij de beide wachtrijen; (iii) het cyclische *polling* systeem met Poisson aankom-

stprocessen; en (iv) de $M/G/2$ wachtrij met heterogene bedienden. Bij al deze modellen veronderstellen we dat minstens één der bedieningstijdverdelingen zwaarstaartig is (doorgaans: regulier variërend). Bestudering van het asymptotisch gedrag van de Laplace-Stieltjes getransformeerden (LST) van wachttijd- en werklastverdelingen in de omgeving van de meest rechtse singulariteit, en toepassing van een Tauberstelling, stelt ons in staat de volgende conclusie te trekken betreffende de wachtrijsystemen (i), (ii) en (iii): de bedieningstijdverdeling met de zwaarste staart bepaalt het staartgedrag van de wachttijd- en werklastverdelingen. In de $M/G/2$ wachtrij blijkt het staartgedrag van de wachttijdverdeling echter ook af te hangen van de grootte van het totale verkeersaanbod. De afleidingen van deze resultaten worden gecomplementeerd door intuitieve redeneringen die inzicht verschaffen in de meest waarschijnlijke wijze waarop grote wachttijden of werklasten op kunnen treden.

We bespreken nu de globale inhoud van elk der zeven hoofdstukken van het proefschrift. Hoofdstuk 1 verklaart het belang van de bestudering van wachtrijmodellen met zware (regulier variërende) staarten. Ook bevat dit hoofdstuk basiskennis over wachtrijen en over de prestatie-analyse van computercommunicatienetwerken. Tevens wordt aandacht geschonken aan de literatuur betreffende wachtrijsystemen met zwaarstaartige verdelingen.

Hoofdstuk 2 is gewijd aan zwaarstaartige kansverdelingen. We richten ons daarbij vooral op de klasse van regulier variërende verdelingen.

In hoofdstuk 3 bestuderen we de $M/G/1$ wachtrij met twee prioriteitsklassen. We veronderstellen dat de bedieningstijdverdelingen van de hoge en/of lage prioriteitsklanten regulier variërend zijn met index $-\nu$ $(1 < \nu < 2)$. Uitgaande van een uitdrukking voor de LST van de wachttijdverdeling van de klanten met lage prioriteit, verkrijgen we relaties tussen het staartgedrag van de wachttijdverdeling van deze klanten, en het staartgedrag van de beide bedieningstijdverdelingen. Bovendien leiden we een limietstelling af voor de wachttijdverdeling van de klanten met lage prioriteit, voor het geval dat de belasting van het systeem een kritische grens nadert ('*heavy traffic*').

De hoofdstukken 4 en 5 zijn gewijd aan cyclische *polling* systemen met Poisson aankomstprocessen. In hoofdstuk 4 bestuderen we een systeem met twee wachtrijen. De bedieningsdiscipline bij rij 1 is 1-*limited* (de bediende bedient ten hoogste één klant per bezoek), en de bedieningsdiscipline bij rij 2 is *exhaustive* oftewel uitputtend (de bediende bedient de wachtrij tot deze leeg is). Merk op dat dit model reduceert tot de $M/G/1$ wachtrij met twee prioriteitsklassen uit hoofdstuk 3 als er geen omschakeltijden tussen de rijen zijn. Onder de veronderstellingen dat er *wel* omschakeltijden zijn, en dat de bedieningstijden een oneindige variantie hebben, leiden we een *heavy-traffic* limietstelling af

voor de wachttijdverdeling bij rij 1. Deze limietstelling suggereert een benadering voor de betreffende wachttijdverdeling. Deze benadering wordt numeriek getest.

In hoofdstuk 5 bestuderen we het cyclische *polling* model voor het geval de bedieningsdiscipline bij elke wachtrij *exhaustive* is of *gated* (de bediende bedient bij een bezoek aan de wachtrij precies de klanten die hij bij zijn aankomst aantreft). We veronderstellen dat de bedieningstijdverdeling met de zwaarste staart regulier variërend is, met index $-\nu$ ($\nu > 1$). Uitgaande van een expliciete uitdrukking voor de LSTs van de wachttijdverdelingen, bewijzen we dat de wachttijdverdelingen bij alle rijen regulier variërend zijn met index $1 - \nu$.

Hoofdstuk 6 is gewijd aan de $M/G/2$ wachtrij met heterogene bedienden: bedieningstijden bij bediende 1 zijn negatief exponentieel verdeeld, terwijl bedieningstijden bij bediende 2 een – voorlopig – niet nader gespecificeerde verdeling hebben. Door gebruik te maken van de techniek van de supplementaire variabele, leiden we een stelsel differentiaalvergelijkingen met randvoorwaarden af, waaruit de LST van de wachttijdverdeling kan worden bepaald als de LST van de bedieningstijdverdeling bij bediende 2 een rationele functie is. Langs analytische weg bepalen we het staartgedrag van de wachttijdverdeling, voor het geval de betreffende bedieningstijdverdeling regulier variërend is. Ook geven we een intuitieve verklaring voor dit staartgedrag.

In hoofdstuk 7 wordt een wachtrijsysteem bestudeerd dat bestaat uit twee wachtrijen in serie, met identieke bedieningstijden van de klanten bij de beide rijen. Uitgaande van expliciete uitdrukkingen voor de verdelingen van de verblijftijd en de hoeveelheid werk bij de tweede rij, relateren we het staartgedrag van deze verdelingen aan het staartgedrag van de (residuele) bedieningstijdverdeling. Voor het geval dat de bedieningstijdverdeling regulier variërend is met index $-\nu$, leidt dit tot de conclusie dat de verdelingen van de verblijftijd en de hoeveelheid werk bij de tweede rij regulier variërend zijn met index $1-\nu$. We geven een intuitieve verklaring voor dit staartgedrag. Tenslotte bewijzen we een limietstelling voor de verblijftijdverdeling in de tweede rij voor het geval dat de belasting van het systeem een kritische grens nadert.

Curriculum vitae

Qing Deng was born in Jiangxi, China on 20 January 1976. She studied fundamental mathematics in the Department of Mathematics at Beijing Normal University in China for her Bachelor's degree during 1992-1996. In September 1996, she joined the master class in the field of Stochastic Operations Research at MRI (Mathematical Research Institute) of the Universities of Groningen, Nijmegen, Twente and Utrecht in the Netherlands. In June 1997, she got her master class diploma after completing her master thesis 'A problem about symmetrization of random variables' (under the supervision of prof.dr. F. den Hollander). From September 1997 till August 1998, she was a Ph.D. student in the Department of Econometrics and the CentER for Economic Research at Tilburg University, studying queueing systems with regular variation under the supervision of prof.dr.ir. O.J. Boxma and dr. J.P.C. Blanc. Since September 1998, she has been a Ph.D. student with the same research project under the supervision of prof.dr.ir. O.J. Boxma and prof.dr.ir. S.C. Borst in the Department of Mathematics and Computer Science at Eindhoven University of Technology.