# Queueing systems with heavy tails

# Queueing Systems with Heavy Tails

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Zwart, Albertus P.

Queueing Systems with Heavy Tails / by A.P. Zwart. - Eindhoven : Eindhoven University of Technology, 2001
Proefschrift. - ISBN 90-386-0891-8
NUGI 811
Subject headings : queuing theory / asymptotics
2000 Mathematics Subject Classification : 60K25, 60F10, 90B18, 90B22

Printed by Universiteitsdrukkerij Technische Universiteit Eindhoven

# Queueing Systems with Heavy Tails

## 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 dinsdag 11 september 2001 om 16.00 uur

door

Albertus Petrus Zwart

geboren te Hilversum

Dit proefschrift is goedgekeurd door de promotoren:

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

# Dankwoord (acknowledgments)

Met dit proefschrift sluit ik een speciale periode in mijn leven af. Ik maak daarom ook met veel plezier van de gelegenheid gebruik om enkele personen te bedanken.
Ik heb dit proefschrift geschreven onder de supervisie van Onno Boxma. Onno heeft me de afgelopen jaren op een voortreffelijke manier weten te begeleiden. Hij voelde feilloos aan wanneer ik wel, maar ook wanneer ik geen hulp nodig had. Daarnaast was Onno altijd bereid om naar problemen buiten het werk te luisteren. De resultaten uit Sectie 2.3 en Hoofdstuk 3 van dit proefschrift zijn gebaseerd op gezamenlijk onderzoek.
Ook aan mijn tweede promotor, Sem Borst, ben ik veel dank verschuldigd. In het bijzonder heeft een onschuldig ogend vraagje van Sem, ergens eind 1999, geleid tot een uiterst plezierige samenwerking, welke ik nooit zal vergeten. Hiervan zijn de resultaten terug te vinden in de hoofdstukken 6, 7 en 8 van dit proefschrift. Buiten dit onderzoek heeft Sem me ook diverse andere malen geholpen.
Ik ben verder NWO dankbaar voor de OIO positie die het mij heeft geboden. Daarnaast dank ik de Technische Universiteit Eindhoven en het Centrum voor Wiskunde en Informatica in Amsterdam voor de aan mij beschikbaar gestelde faciliteiten. Ik dank het onderzoeksinstituut BETA voor de financiële bijdrage aan mijn bezoek aan Columbia University.
Aan mijn collega's bewaar ik goede herinneringen, in het bijzonder de contacten tijdens congresbezoeken en de vele (thee-)pauzes. De meeste van deze herinneringen deel ik met mijn kamergenoten Sindo Núñez, Miranda van Uitert, Qing Deng en Tjark Vredeveld.
Tenslotte wil ik mijn familie en vrienden bedanken voor de geboden steun in de laatste jaren. Hierbij wil ik mijn paranimfen Nam Kyoo Boots en Daniël van Vuuren, mijn ouders, Peter, Margriet en Lonneke met name noemen. Zonder jullie had ik het niet gered.

Bert Zwart
juni 2001

# Contents

# Chapter 1

# Introduction

Queueing theory occupies a prominent role in the performance analysis of a wide range of systems in computer-communications, logistics, and manufacturing. One of the pillars of queueing theory is the fact that queues can often be modeled as continuous-time Markov chains, making extensive use of generalizations of the exponential distribution such as phase-type distributions. This class of distributions enables a tractable computation of various characteristics of the queueing model.

However, recent findings have shown that the statistical assumptions underlying this approach may not always be satisfied in practice. A crucial example is the empirical finding that traffic in communication networks can exhibit phenomena like *self-similarity* and *long-range dependence*. These phenomena are not present in queues in which all distributions are phase-type; it has been shown that heavy-tailed distributions are more appropriate. Similar observations have been made in insurance, a field that has given rise to quite similar models and problems as queueing. In risk theory, the claim size distribution is often not phase-type, but heavy-tailed. This monograph analyzes queueing systems with heavy-tailed input.

This first introductory chapter serves as further background to motivate the study of such queueing systems, and is organized as follows: In Section 1.1 we introduce long-range dependence, self-similarity, and heavy tails, and discuss the
relevance of these concepts in modeling communication networks. Section 1.2 reviews the standard queueing models, in particular the single-server queue and the fluid queue; both are key objects of study in this monograph. The first two sections are tied together in Section 1.3, where we argue that queueing models with heavy-tailed input are appropriate for incorporating the phenomena discussed in Section 1.1. Section 1.4 is concerned with the analysis of queues with heavy-tailed input, in particular with *large-deviations probabilities in the regime of large buffers.* Section 1.5 gives an overview of several other possible approaches. In particular, this section discusses other asymptotic approaches, non-asymptotic approaches, and other traffic models. A more detailed exposition of the contents of this monograph can be found in Section 1.6.

## 1.1 Long-range dependence, self-similarity and heavy tails

In this section we give a short introduction to the occurrence of self-similarity and long-range dependence in communication network traffic. More extensive treatments can be found in e.g. Park & Willinger [223] and Adler *et al.* [11].

### 1.1.1 Traffic measurements

In recent years, it has become possible to collect large amounts of high-quality measurement data on traffic in communication networks. Many of these data sets have been used to validate the traditional statistical assumptions made when analyzing such networks. These assumptions contain the premise that network traffic can be described by Markovian models. This implies that autocorrelations in network traffic decay exponentially fast. This kind of traffic behaves smoothly over long time scales.

It came as a shock when it was found that these traditional (Markovian) assumptions are *not* always satisfied. A careful statistical analysis in Leland *et al.* [181] showed that Ethernet LAN traffic at Bellcore exhibits properties like *self-similarity* and *long-range dependence* (LRD). In particular, this traffic behaves extremely bursty on a wide range of time scales.

This burstiness property is clearly illustrated by Figure 1.1 below (taken from [181]). The left part of this figure shows actual traces of Ethernet LAN traffic. Starting with a time unit of 100 seconds, each subsequent plot is obtained from the previous one by increasing the time resolution by a factor of 10 and by zooming in on a randomly chosen subinterval (a darker shaded area in the figure). The figure clearly shows that the observed traffic trace is bursty on all time scales. This is in stark contrast with traffic simulated from conventional traffic models. The right part of Figure 1.1 shows a trace obtained by simulating a conventional traffic model (based on exponential assumptions) with the same arrival intensity and average packet size. This traffic behaves smoothly on large time scales.

Further statistical analysis of the correlation structure of measured network traffic shows that its autocorrelation function decays extremely slowly. This property is closely related to the notion of long-range dependence.

The properties of long-range dependence and self-similarity are defined as follows.

#### Long-range dependence

Let $\mathcal{X} = \{X(t), t \geq 0\}$ be some strictly stationary stochastic process. Typically, $X(t)$ may be thought of as the rate of network traffic generated at time $t$. The cumulative amount of traffic up to time $t$ is given by $T(t) = \int_0^t X(u)\mathrm{d}u$. Define the autocorrelation

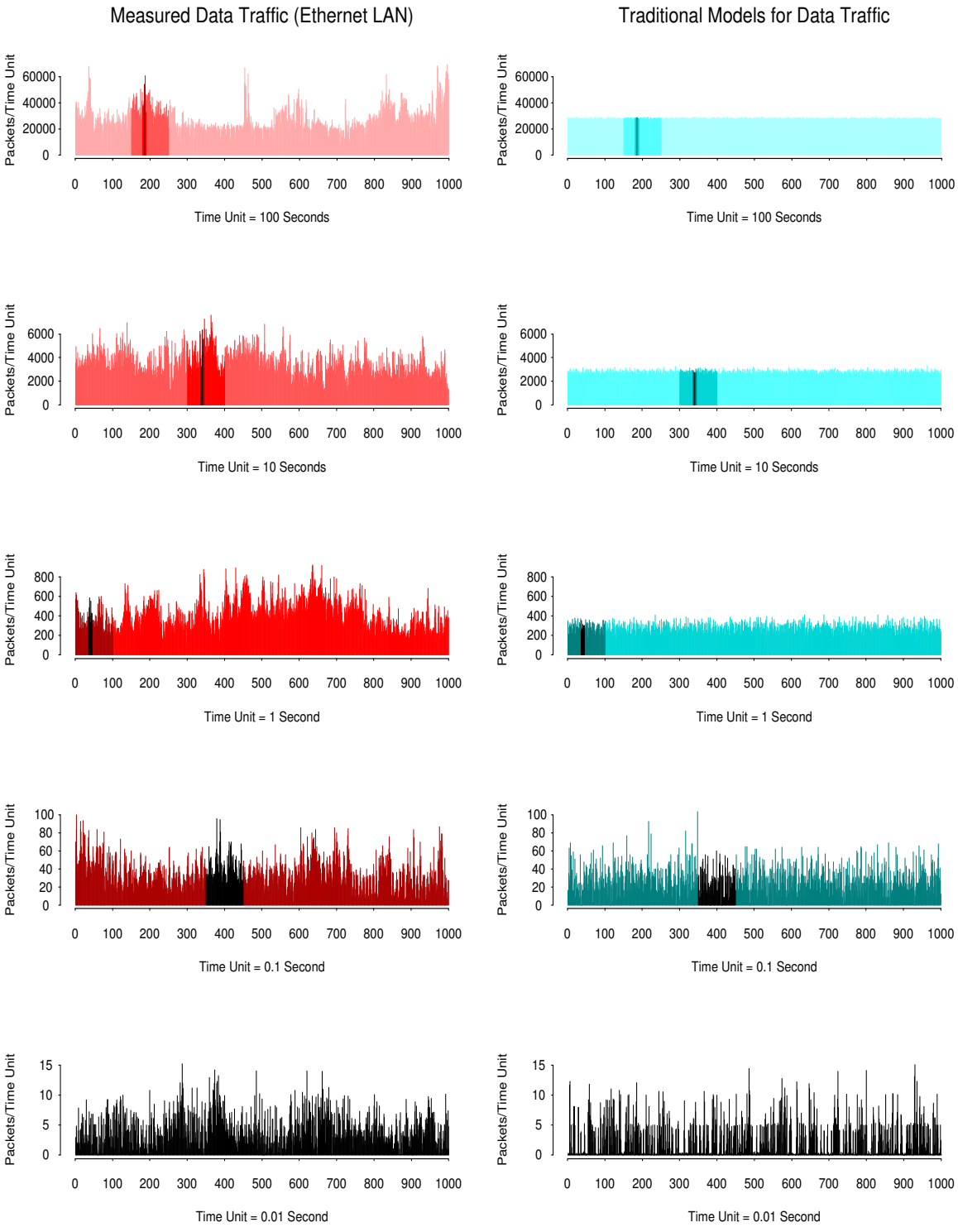


Figure 1.1: Traffic traces of Ethernet LAN traffic

function

$$c(t) = \mathbb{C}\text{ov}\{X(s), X(s+t)\}/\mathbb{V}\text{ar}\{X(s)\}.$$

The following definition is standard, see e.g. Cox [106].

**Definition 1.1.1** *$\mathcal{X}$ is short-range dependent if $\int_0^\infty |c(t)|\mathrm{d}t < \infty$. If $\int_0^\infty |c(t)|\mathrm{d}t = \infty$, then $\mathcal{X}$ is long-range dependent.*

There are other (strongly related) definitions of long-range dependence.
For example, cf. Beran [36], $\mathcal{X}$ is long-range dependent if the autocorrelation function $c(\cdot)$ shows a particular type of *power-law behavior*:

$$c(t) \sim c_0 t^{-\alpha}, \qquad 0 < \alpha < 1. \tag{1.1}$$

(With $f(x) \sim g(x)$ we mean $f(x)/g(x) \to 1$ as $x \to \infty$.) Traditional assumptions typically imply that $c(t)$ decreases negative-exponentially in $t$: $c(t) \sim c_0 \mathrm{e}^{-\gamma t}$. The above definition of long-range dependence is intimately related to the behavior of the variance of the cumulative traffic process $T(t)$. The identity

$$\mathbb{V}\text{ar}\{T(t)\} = 2\int_0^t \int_0^u c(v)\mathrm{d}v\mathrm{d}u, \tag{1.2}$$

shows that the variance of $T(t)$ behaves linear in $t$ if $\mathcal{X}$ is short-range dependent, and superlinear in $t$ if $\mathcal{X}$ is long-range dependent. In particular, if (1.1) holds, then

$$\mathbb{V}\text{ar}\{T(t)\} \sim c_1 t^{2-\alpha}, \tag{1.3}$$

where $c_1$ can be expressed in terms of $\alpha$ and $c_0$.

### Self-similarity

A second key property is self-similarity, which is defined as follows (see e.g. Samorodnitsky & Taqqu [246]).

**Definition 1.1.2** *A stochastic process $\mathcal{X} = \{X(t), t \geq 0\}$ is (strictly) self-similar with parameter $H$ if $\{X(t), t \geq 0\}$ and $\{\gamma^{-H} X(\gamma t), t \geq 0\}$ have the same finite-dimensional distributions for any $\gamma > 0$.*

Note that a self-similar process is non-stationary. The concept of self-similarity became popular due to the work of Mandelbrot, see e.g. [192, 193]. If $\mathcal{X}$ is self-similar with parameter $H$, then

$$\mathbb{V}\text{ar}\{X(t)\} = t^{2H}\mathbb{V}\text{ar}\{X(1)\}. \tag{1.4}$$

Any process $\mathcal{X}$ satisfying this property is called second-order self-similar. An even weaker form of self-similarity is *asymptotic second-order self-similarity* (which is also defined for discrete-time processes). $\mathcal{X}$ satisfies this property if a suitably centered and normalized version of $\{X(\gamma t), t \geq 0\}$ converges to a self-similar process when $\gamma \to \infty$. A formal definition may be found in Chapter 1 of [223].

The notions of self-similarity and long-range dependence are related in some examples, but not equivalent. For instance, Brownian motion is self-similar (with $H = \frac{1}{2}$) but not long-range dependent. Conversely, there are also long-range dependent processes which are not self-similar, see [223].
If $H > \frac{1}{2}$, then the definitions of asymptotic self-similarity and long-range dependence are equivalent, see Chapter 1 in [223]. For our purposes, it suffices to give an intuitive explanation: If $T(\cdot)$ is self-similar with Hurst parameter $H$, then

$$\mathbb{V}\mathrm{ar}\{T(t)\} \sim c_2 t^{2H}. \tag{1.5}$$

Hence, the variance of $T(t)$ behaves superlinear in $t$ if $H > \frac{1}{2}$, and the constants $H$ and $\alpha$ are related by the identity $H = (2 - \alpha)/2$. In view of this equivalence, we often refrain from mentioning self-similarity explicitly and just speak of long-range dependence (or even just use the acronym LRD).

### 1.1.2 Explaining long-range dependence via heavy tails

As mentioned earlier, there is now mounting statistical evidence that network traffic is self-similar and long-range dependent.
Besides the paper [181], other studies confirm these properties. See e.g. Willinger *et al.* [277] for traffic in Local-Area Networks, Paxson & Floyd [228] for traffic in Wide Area Networks and Beran *et al.* [37] for VBR video traffic. More references can be found in the recent monograph [223].
All these properties are examined at the *packet level*. A number of studies tried to explain these results by examining quantities related to network traffic at a much higher level of aggregation, particularly the *application level*. At this level, basic entities are file sizes, connection times, transmission times, etc.
Several studies at this level indicate that long-range dependence may be caused by *heavy-tailedness* of certain traffic characteristics. Crovella & Bestavros [108] show that file sizes and transmission times of files in the Internet are power-tailed with infinite variance: Let $Y$ be a generic file size or transmission time. Then, typically,

$$\mathbb{P}\{Y > t\} \sim c_3 t^{-\alpha}, \qquad 0 < \alpha < 2. \tag{1.6}$$

The infinite-variance property of various quantities in network traffic has been independently confirmed by a number of other studies, see e.g. Crovella *et al.* [109], Willinger

*et al.* [278], and references therein. Most of the above studies conclude that the mean of the above quantities is finite, but this may not always be the case. Resnick & Rootzén [237] statistically show that the *mean* of file sizes may also be infinite. Other characteristics of network traffic which are heavy-tailed include CPU times, idle times, peak rates, connection times and more; see again the monograph [223].

Heavy tails and LRD are intimately related. The canonical *On-Off* process for example (to be introduced in Subsection 1.2.2), is LRD if and only if (iff) the On- or Off-time has infinite variance, see Theorem 3.9 of Boxma & Dumas [67].

An important observation is that LRD may be due to heavy-tailedness of basic user-related characteristics (e.g., heavy-tailedness of files is a consequence of consumer demand), see Crovella *et al.* [109] for a discussion. Besides user behavior, there may still be other causes of LRD, such as traffic control mechanisms like the Transmission Control Protocol (TCP) used in the Internet, see Figueredo *et al.* [132].

## 1.2 Queueing models

Queues naturally arise in situations where there is competition for some "scarce resource". A typical example of a queue is the counter at the supermarket or the post-office, where customers are waiting until they receive their service. Congestion usually occurs because customer arrivals are random in nature. In addition, the time it takes to serve a customer is also often random. Apart from the counter-example above, queues also arise in situations where the basic entities are not customers, but packets at a link in a communication network, or jobs in a production system.

The queueing problems in this thesis are all motivated by problems in communication networks. Queueing theory has been quite a successful tool in the performance analysis of such networks. In fact, new results in queueing theory have often been inspired by new technological advances in computer-communications.

A classical example is the celebrated *Erlang loss model*, first studied by A.K. Erlang [127] in the beginning of the 20th century in the context of telephone networks. The *Erlang loss formula* has been and still is applied in a wide variety of problems. Another successful branch of queueing theory is the study of networks of queues motivated by computer-communication systems evolving in the 60's and 70's, which led to milestones like Baskett *et al.* [34] and Kelly [169].

Important monographs on queueing theory (and related subjects) include Asmussen [19], Cohen [97], Kleinrock [174], and Tijms [266]. A recent book focusing on the role of queueing theory in the performance analysis of computer-communication systems is Walrand & Varaiya [270].

In the following two subsections, we further elaborate on the queueing models studied in this thesis: (i) The single-server queue, and (ii) the fluid queue.

### 1.2.1 The single-server queue

The most elementary queueing model is the single-server queue. In this model, customers arrive at the queue one at a time. The time between the arrivals of two consecutive customers is called the interarrival time. A common assumption is that the sequence of interarrival times consists of independent and identically distributed (i.i.d.) random variables.

There is one server, which works at a constant speed $c$ whenever there are customers in the system. Similarly to the interarrival times, the service times of customers are usually assumed to form an i.i.d. sequence of random variables. Moreover, the sequences of interarrival times and service times are independent. After a customer has received its full service requirement, it leaves the system.

The above-described queueing model is usually called the $G/G/1$ queue. This notation was introduced by Kendall [170]. The first $G$ means that the interarrival time distribution may be of a general form; the second $G$ indicates the same for the service time distribution. If one wishes to stress the independence of interarrival times and service times, then one sometimes writes $GI/GI/1$ or $GI/G/1$. If the interarrival time distribution is exponential (i.e., if the arrival process is Poisson), then one speaks of the $M/G/1$ queue (with $M$ abbreviating 'memoryless' or 'Markovian'). Many extensions of this model exist. In Chapter 4 for example, we consider a queue in which the total work in the system is bounded by $K$; we shall refer to this system as the $G/G/1/K$ queue.

As mentioned before, the server works at speed $c$ as long as there is work in the system. This information is enough to describe the evolution of the total amount of unfinished work in the system (also called the buffer content or the workload). Other important performance measures are the number of customers in the queue and the waiting and sojourn times of customers. These processes are in addition governed by the *service discipline*. The most common service discipline is *First Come First Served*, abbreviated as FCFS. Other important service disciplines are *Last Come First Served* (LCFS) and *Processor Sharing* (PS). If the server operates according to the PS discipline, then it simultaneously serves all (say $n$) customers in the system at the same speed ($c/n$). PS queues are investigated in Chapter 3 of this thesis.

The single-server queue is a central model in applied probability. Problems in, for example, inventory and risk theory can often be reformulated as queueing problems (and vice versa). A key example is the equivalence between waiting-time probabilities in the $G/G/1$ queue with FCFS service and ruin probabilities in insurance risk models, see e.g. Asmussen [19, 29].

### 1.2.2 The fluid queue

Traffic in today's communication networks is heterogeneous in nature, not only consisting of voice traffic, but also of video and data. In addition, network traffic is inherently bursty

(as already stressed in Section 1.1). Traditional telephone networks are not flexible enough to cope with this burstiness and heterogeneity, as they assign a fixed amount of capacity (one channel) to each connection.

Hence, for these reasons, modern communication networks like ATM (Asynchronous Transfer Mode) and IP (Internet Protocol) networks operate in a more flexible way. Basic entities are not calls or connections, but *packets*.

Packet-switched networks can be studied on various time scales, see e.g. Hui [158] and Roberts *et al.* [241]. The burstiness of network traffic is explicitly modeled on the *burst scale*. On this time scale, traffic is modeled as a continuous fluid flow, thus neglecting the discrete nature of relatively small packets. In particular, a popular way of modeling bursty traffic is by means of an *On-Off source*. An On-Off source generates traffic at constant rate during On-periods, and no traffic during Off-periods. This has motivated the study of queueing models fed by a superposition of On-Off sources. From a queueing perspective, the main difference with ordinary queues is that work does not arrive instantaneously, but gradually over time.

The seminal paper which made the above fluid model *the* paradigm for modeling bursty traffic is Anick, Mitra, & Sondhi [16], where an explicit expression for the steady-state buffer content (workload) distribution is derived. The model considered in [16] was already studied earlier in a series of papers by Kosten [178, 179], Cohen [91, 99], and others.

The papers [16, 178] both consider a queue fed by the superposition of several homogeneous On-Off sources with exponentially distributed On- and Off-periods. Subsequent work extended the model in various directions, such as heterogeneous source characteristics, several source states to account for various activity levels, or activity periods with a general Markovian structure, see for instance Kosten [179], Mitra [210], and Stern & Elwalid [261].

The buffer content in fluid queues with generally distributed On- and Off-periods is studied by Cohen [91, 99]. Unfortunately, these papers make the assumption that a single On-period is sufficient for the buffer to fill. In general, the service rate of the queue is so large that several simultaneous On-periods are necessary for this to occur. Exact queueing analysis appears to be very hard in this general case.

More references on fluid queues can be found in the survey paper of Kulkarni [180], see also the thesis of Scheinhardt [250].

## 1.3 Long-range dependence and queues

In this section, we tie both previous sections together and incorporate LRD in a queueing model. We propose to model network traffic by one of the traffic processes described in the previous section. Indeed, from the queueing point of view, the most natural way to incorporate LRD in a traffic model is by simply allowing the input into a queue to have heavy-tailed characteristics. Other possible traffic models are briefly discussed in

Section 1.5.3.

In Section 1.1 we already mentioned that statistical analysis at the application level showed that various quantities, such as file sizes, have heavy-tailed distributions. In the queueing context, this naturally translates into heavy-tailed interarrival and/or service times in the single-server queue, or to heavy-tailed On- and/or Off-periods in the fluid queue. For the On-Off model with generic On-time $A$ and Off-time $U$, this implies LRD: Heath *et al.* [152] have shown that the autocovariance function satisfies $c(t) \sim c_4 t^{1-\min\{\alpha_A,\alpha_U\}}$ if $\mathbb{P}\{A > x\} \sim c_A t^{-\alpha_A}$ and $\mathbb{P}\{U > x\} \sim c_U t^{-\alpha_U}$. Hence, if $\alpha_A < 2$ or $\alpha_U < 2$, then the On-Off process is LRD. A related result can be found in Boxma & Dumas [67]. Other processes in (fluid) queues and queueing networks can also be LRD. Anantharam [14] shows that LRD input may propagate through a queueing network. Similar insights can be found in Boxma & Dumas [68], who consider the output process (busy period) of a fluid queue (see also Chapter 5 in this monograph). It is shown in [68] that the output of the fluid queue exhibits LRD if and only if the input process does, see also Chapter 5 of this thesis. Similar conclusions hold for finite-buffer systems, see Vamvakos & Anantharam [268]. A survey on the literature on (fluid) queues with heavy-tailed input (up to 1998) can be found in Boxma & Dumas [67].

## 1.4 Queueing systems with heavy tails

In the previous sections we gave an introduction to the occurrence of LRD network traffic, and described how this phenomenon can be attributed to heavy-tailedness of various quantities like file sizes. These observations motivate the analysis of queueing systems where some of the underlying variables (e.g. service times) are heavy-tailed. We are not only interested in heavy tails with infinite variance as in (1.6), but in any kind of tail which is heavier than a negative exponential; see Chapter 2 for convenient classes of heavy-tailed distributions. In the remainder of this monograph the analysis of queueing systems with heavy tails will play a central role.

The purpose of the present section is to elaborate upon our approach to analyze these queueing systems. Almost all results in this thesis are asymptotic expansions for tail probabilities in the large-buffer regime, as is described in Subsection 1.4.1. Further background and literature is provided in Subsection 1.4.2. Some limitations are discussed in Subsection 1.4.3.

### 1.4.1 The asymptotic approach

In general, there are many ways to analyze queueing systems. For example, Cohen & Boxma [98] make a distinction between the following approaches: (i) Exact analysis; (ii) numerical analysis; (iii) (asymptotic) approximations; (iv) experimental analysis and simulation.

This thesis is mainly concerned with asymptotics. Suppose that $X$ is some random variable in a queueing model, e.g. the waiting time in a single server queue or the workload in a fluid queue. *The central topic of this thesis is the development of asymptotic approximations for $\mathbb{P}\{X > x\}$ in the regime $x \to \infty$ for queues with heavy-tailed input.*
The above topic can be viewed as classical in queueing theory, but has been considered mostly for queueing systems with light-tailed input. Asymptotic results for queues with heavy tails are limited, especially for fluid queues (see also Chapter 2). Many theoretically challenging problems in this area are not well understood.
Besides the wish to tackle some of these problems, there are several reasons to consider asymptotics. An exact analysis may be impossible or may lead to cumbersome expressions. In this case, one needs to make some kind of approximation. This is especially the case when heavy-tailed distributions are involved: As mentioned before, this prohibits the use of phase-type distributions.
The reason for considering approximations in the regime $x \to \infty$ is motivated by Quality-of-Service requirements in communication networks. These typically include loss probabilities of the order $10^{-6}$ or less. Such small probabilities may be covered by the asymptotic regime $x \to \infty$. Another benefit of studying asymptotics is that they often lead to simple and important qualitative insights in how the event $\{X > x\}$ occurs.

### 1.4.2 Background on asymptotics

Asymptotic analysis has a rich tradition in queueing and insurance. Classical is the work of Cramèr [107] and Lundberg [191], see Asmussen [29] for a recent account from the insurance risk viewpoint. The literature on asymptotics in (fluid) queues is voluminous, see e.g. Asmussen [19, 29], Feller [131], Tijms [266], and more, e.g. the Ph.D theses of Mandjes [194] and Van Ommeren [218].
There are several types of asymptotics which can be considered. If $X$ is the waiting time or workload in some (fluid) queue, then the typical result is of the form

$$\mathbb{P}\{X > x\} \sim C\mathrm{e}^{-\theta f(x)}. \tag{4.1}$$

When the input is Markovian, one usually has $f(x) = x$, implying that the tail of $X$ is exponential. Most queueing papers assume $f(x)$ to be linear. Note that $X$ is power-tailed if $f(x) = \log x$. A thorough treatment of the case $f(x) = x$ was given by Asmussen [18] for waiting times in single-server queues and by the same author [20] for workloads in fluid queues. These papers also contain further references.
The above result gives a description of the *exact* tail asymptotics of $X$. In many cases, it is difficult to obtain the exact asymptotics and then one often considers *logarithmic* asymptotics. These have the form

$$\log \mathbb{P}\{X > x\} \sim -\theta f(x). \tag{4.2}$$

It is not surprising that logarithmic asymptotics (can be proven to) hold in considerably greater generality than exact asymptotics. In the single-server queue for example, it is not necessary to assume that the service times are independent for (4.2) to hold, see Glynn & Whitt [138]. This important paper considers the light-tailed case $f(x) = x$. Results for general $f(x)$ can be found in Duffield & O'Connell [117].
A third type of asymptotic result, which can be viewed as an intermediate case between exact and logarithmic asymptotics are *(asymptotic) bounds*, of the form

$$C_{-}\mathrm{e}^{-\theta f(x)} \leq \mathbb{P}\{X > x\} \leq C_{+}\mathrm{e}^{-\theta f(x)}. \tag{4.3}$$

Bounds are useful when it is difficult to prove exact asymptotics, or when the pre-factor $C$ is too difficult to compute. Bounds for light-tailed fluid models can be found in e.g. Gautam *et al.* [136], Palmowski & Rolski [220], and Palmowski [221]. Bounds for the heavy-tailed case can be found in Dumas & Simonian [120], Likhanov [184], and Likhanov & Mazumdar [185]. More exact asymptotics for the heavy-tailed regime can be found in the next chapter.

### 1.4.3 Limitations

We now discuss some practical as well as nearly philosophical issues which may arise in relation to the study of large-buffer asymptotics in queues with heavy tails.
In queues with phase-type service-time distributions, the accuracy of large-buffer asymptotics is usually good, since the speed of convergence of the asymptote to the true value is exponentially fast. This is not the case when heavy-tailed distributions are considered. Typically, the speed of convergence is linear, see Mikosch & Nagaev [207], but it can be even worse [205, 207]. We refer to Abate & Whitt [2] and Kalashnikov [166] for illustrative numerical examples. Thus, the asymptotic expansions as developed in this thesis should be handled with care. Furthermore, these asymptotics tend to *underestimate* the true value, see [2].
The justification of the regime underlying the asymptotic approximation is also of relevance in practice. In some cases, other asymptotic regimes (like the many-sources regime discussed below) may provide a more natural choice.
Another problem is the usual assumption that queues operate in steady state. When one wishes to view the steady-state distribution as an approximation of the transient distribution, one should realize that convergence of the transient distribution to steady state can be quite slow in the heavy-tailed case, see Asmussen & Teugels [23].

## 1.5 Other approaches

The previous two sections reduced the study of congestion and LRD network traffic to large-buffer asymptotics in queues with heavy tails. This section gives an overview of several possible different ways to analyze LRD in communication networks.

Since this thesis is about large-buffer asymptotics in queues with heavy tails, we divide this section into four parts, which are ordered in increasing level of aggregation. In the first subsection, we review some different asymptotic regimes (as opposed to large-buffer asymptotics). The second subsection does not consider asymptotics, but other approaches, like exact analysis, numerical analysis, and simulation; all in the context of queueing theory. Subsection 1.5.3 does not consider queues, but other traffic models for analyzing LRD in communication networks. Finally, the last subsection examines the practical relevance of LRD in these networks.

### 1.5.1 Other asymptotic regimes

The asymptotics studied in this thesis are usually referred to as *large-buffer asymptotics*, as they typically involve the build-up of a large buffer content. One may also consider various other types of asymptotic regimes which are of interest to queueing theory and performance analysis. Below, we mention a number of alternative asymptotic regimes, with a view towards heavy tails.

#### Many-sources asymptotics

Consider the fluid queue with capacity $c$, fed by $n$ identical On-Off sources. If a large number of sources are multiplexed (which often occurs in practice), then it is natural to consider what happens when the number of sources $n$ tends to infinity. To get a non-trivial limit, one needs to scale $c$ proportionally in $n$: $c = nc'$, with $c'$ the capacity per source. Thus, we consider a sequence of models. Let $V^{(n)}$ be the workload in the $n$-th model, fed by $n$ On-Off sources and with capacity $nc'$. Under certain regularity conditions the following result holds,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{V^{(n)} > nx\} = -I(x),$$

for some function $I(x)$ which is called the *loss curve*. The above limiting procedure was originally proposed in Weiss [271] and has been popularized and generalized since then, see e.g. Shwartz & Weiss [254], Botvitch & Duffield [62], Courcoubetis & Weber [105], Mandjes & Ridder [197], and Wischik [279].

In the context of queues with heavy tails, an important breakthrough was made by Likhanov & Mazumdar [183], who significantly relaxed the conditions under which the above asymptotics hold. As is shown in Mandjes & Borst [195], these conditions are satisfied by On-Off sources with heavy-tailed On-periods. Likhanov & Mazumdar also strengthen the above limiting result by obtaining the exact asymptotics.

A disadvantage is that $I(x)$ is not very explicit in general; it is the solution of some variational problem. Several authors have studied properties of the function $I(x)$. For the case of On-Off sources with heavy-tailed On-periods, see Mandjes & Borst [195] ($x \to \infty$), and Mandjes & Kim [196] ($x \downarrow 0$).

Important related work is that of Duffield [119] (see also Mandjes [198]), who considers a class of fluid queues with $M/G/\infty$ input (an infinite number of sources) and a limiting regime which is similar to the many-sources scaling. This type of input is also considered in this thesis, see Chapter 8.

**Heavy-traffic limits**

Consider the steady-state waiting-time distribution $W$ in the single-server queue with service speed 1. Let $\rho$ be the mean amount of work offered to the system per time unit. For convenience, index $W = W_\rho$. If $\rho \geq 1$, then the system is unstable and the steady-state waiting-time distribution does not exist. However, it is possible to consider the regime $\rho \uparrow 1$, by properly scaling the workload $W_\rho$. The typical result is: If the service-time distribution has finite variance, then $(1-\rho)W_\rho$ converges to a random variable with an exponential distribution, see e.g. Kingman [173], Borovkov [48], Iglehart [159], and Whitt [272].

The above result does not hold if the service time distribution is regularly varying (see Chapter 2) with infinite variance. This case was treated by Boxma & Cohen [71, 75], Cohen [100, 102], Furrer *et al.* [135], and Resnick & Samorodnitksy [238]. The typical result in these papers is that a properly scaled version of $W_\rho$ converges weakly to the supremum of a an infinite variance stable Lévy motion. which has the Mittag-Leffler distribution (cf. [135]) . A related result is that a suitably time-scaled and normalized version of the transient workload process converges (in a carefully chosen topology) to a Lévy process. It is important to note that the scaling is different from the finite-variance case. If the service-time distribution is Pareto with index $-\nu, 1 < \nu < 2$ (see Chapter 2), then the scaling factor is $(1-\rho)^{\frac{1}{\nu-1}}$.

If one considers the fluid queue with $n$ On-Off sources, then there are two possible scalings for the cumulative input process. A first option is to first scale time ($t \to \infty$) and then the number of sources ($n \to \infty$). In this case the limiting process is an infinite variance stable Lévy motion, see Taqqu *et al.* [264]. Another possible scaling is to first scale the number of sources ($n \to \infty$) and then time ($t \to \infty$). This scaling leads to *Fractional Brownian Motion* (FBM) and has already been known since Taqqu [263]. For On-Off sources, key references are Willinger *et al.* [277] and Brichet *et al.* [79]. These results establish a fundamental link between FBM and the On-Off model, and provide additional insight into the relation between the observed self-similarity and heavy-tailedness in network traffic.

The above two scalings lead to entirely different limiting processes. Mikosch *et al.* [206] consider the case where $n, t \to \infty$ simultaneously. It is also possible to combine the heavy-traffic limiting regimes with large-buffer or many-sources asymptotics, see e.g. Section 15 in Cohen [104], and Wischik [280].

Additional useful references on heavy traffic and heavy tails are Stegeman [259], and the monographs of Samorodnitsky & Taqqu [246] and Whitt [275].

### 1.5.2 Non-asymptotic approaches

Besides asymptotics, there are also several other approaches. These are partly motivated by the issues mentioned in Subsection 1.4.3, but are also interesting from a mathematical point of view.
Most of the approaches below are restricted to the waiting time $W$ in the $M/G/1$ queue. Since asymptotic expansions for $\mathbb{P}\{W > x\}$ may behave poorly for moderate values of $x$, it is worth looking for more explicit solutions and exploring other approaches. Below we will make a distinction between (i) Exact analysis; (ii) Bounds and multiterm asymptotic expansions; (iii) Numerical analysis; (iv) Rare-event simulation.

#### Exact analysis

Consider the stationary waiting time $W$ in the $M/G/1$ queue. An explicit expression for the transform of $W$ is well-known, but convenient expressions for the distribution of $W$ are in general restricted to phase-type service-time distributions.
An exception is provided by Boxma & Cohen [69], who found an explicit expression for $\mathbb{P}\{W > x\}$ for a particular choice of the service time $B$. This result has been generalized by Abate & Whitt [6]. Related results can be found in Gaver & Jacobs [137].

#### Multi-term expansions and bounds

To strike a balance between the difficulties in deriving exact expressions for $\mathbb{P}\{W > x\}$ and the limited accuracy of the (single-term) asymptotic expansions for $\mathbb{P}\{W > x\}$, several authors have tried to find multiterm expansions for $\mathbb{P}\{W > x\}$.
Willekens & Teugels [276] and Abate & Whitt [6] both obtain three-term expansions, using entirely different (probabilistic vs. transform) methods. In the latter paper, a convenient class of heavy-tailed distributions (Pareto mixtures of exponentials) is introduced, which have a tractable transform. Boxma & Cohen [69] introduce another convenient class of service time distributions with infinite variance and obtain a full series representation of the waiting-time distribution. Related results may be found in e.g. Borovkov & Borovkov [52].
All the results above are restricted to power-tailed service time distributions. Kalashnikov [166] and Kalashnikov & Tsitsiashvili [167] derive lower and upper bounds for the waiting-time distribution, which have the same asymptotic behavior. These bounds are valid for a large class of service time distributions (including Weibull distributions as defined in the next chapter).

#### Numerical analysis

There exist many tractable numerical algorithms for analyzing queueing systems. Most of these algorithms assume light-tailed (phase-type) distributions. Numerical analysis

of queues with heavy tails is still in its infancy; most algorithms are restricted to the single-server queue.
In a series of papers, Abate & Whitt [2, 7, 8] extend their transform inversion approach (cf. [1]) to heavy tails, and obtain tractable algorithms for the waiting-time distribution in the $M/G/1$ queue. Their results are based on manageable expressions for transforms of heavy-tailed distributions.
Another way to get computational results is to approximate a heavy-tailed distribution with a hyperexponential distribution (which can have an arbitrarily large variance), see e.g. Feldmann & Whitt [130] and Starobinski & Sidi [258]. Although this idea is generally applicable, it has only been tested for the $M/G/1$ waiting-time distribution. Even for this simple model, it seems difficult to give performance guarantees. Nevertheless, a related approach (using truncated power tail distributions) is proposed in Schwefel & Lipsky [248] and there applied to analyze the stationary buffer content distribution of a fluid queue with heavy-tailed On-Off sources.

#### Rare-event simulation

Rare-event simulation aims to provide reliable estimates of small tail probabilities in e.g. queueing and insurance risk models. A considerable body of theory exists for the light-tailed case, see e.g. Asmussen [29], Mandjes [194], and references there. The available literature in the heavy-tailed case is mainly concerned with ruin probabilities in insurance risk models, or equivalently, with waiting times in single-server queues.
Asmussen *et al.* [30] describe several algorithms for the $M/G/1$ queue which all heavily rely on the explicit random-sum representation of the $M/G/1$ waiting time. Boots & Shahabuddin [46] develop an efficient algorithm for the $GI/G/1$ queue for Weibullian service times. The results in [46] have been extended to a much wider class of risk models in [47].
The most popular technique in rare-event simulation is importance sampling. Unfortunately, the above studies show that severe problems arise when importance sampling is applied to queues with heavy tails.

### 1.5.3 Other traffic models

Besides the traditional traffic models in queueing theory, a number of other models have been proposed to model LRD network traffic. We give a brief overview of these traffic models; a more thorough treatment can be found in the references cited below.

#### Chaotic maps

At the time when the performance analysis of systems with LRD input became popular, there was some activity in the application of non-linear dynamics, see Erramilli *et al.*

[128], and Pruthi [231]. Although chaotic maps allow for a concise description of traffic phenomena (see [67] for some examples), only limited progress has been made so far in the associated queueing analysis.

### Time series

Black-box (ARIMA) time series modeling may also be applied to model network traffic. There are two options to incorporate LRD or heavy tails. A first option is to allow the innovations of the ARIMA process to be heavy-tailed. An alternative is to consider Fractional ARIMA processes, see e.g. Brockwell & Davis [80].
These types of models have a strong tradition among statisticians, but their queueing analysis is still in its infancy, in spite of a recent paper of Mikosch & Samorodnitsky [208]. There are also statistically-oriented grounds for considering structural queueing models instead of black-box models, see e.g. the reply of Paxson & Willinger in Resnick [234].

### Fractional Brownian motion

Another way of modeling LRD network traffic is by Fractional Brownian Motion (FBM) or another Gaussian process exhibiting LRD. Such a process can then be used as input process in a fluid queue to study its performance. This approach was proposed by Norros [212, 213], and has been the subject of several investigations since then. For surveys, see Debiçki & Rolski [114] and Norros [214].
FBM may also be seen as an approximation of the traffic offered by the superposition of a large number of On-Off sources with heavy-tailed On-periods. This the canonical example of the intimate relationship between heavy tails and long-range dependence and provides a physical explanation of the detected statistical self-similarity. More on this limiting procedure can be found in Section 1.5.1.

### Multifractals

A key parameter in self-similar input traffic is the Hurst parameter $H$, which determines the behavior of correlations over various time scales, see (1.4). However, it was found that over short time scales (100 milliseconds and less) the behavior of network traffic may be more complex. To account for this behavior, it was proposed to model the second-order statistics of $\mathcal{X}$ as

$$\mathbb{E}\{[X(t+\tau)-X(t)]^2\} \sim \tau^{2h(t)},\ \tau \downarrow 0, t \text{ fixed}$$

(with $f(x)/g(x) \to 1, x \downarrow 0$, we mean $\lim_{x\downarrow 0} f(x)/g(x) = 1$). This is an extension of the standard self-similar case (1.4), in which case the equation above holds with $h(t) = H - 1$. The analysis of multifractional processes is a fascinating new research area. We confine ourself to mentioning the papers of Mannersalo *et al.* [199], Abry & Veitch [10],

and Riedi & Willinger [240]. These papers are mainly concerned with the statistical analysis of multifractional processes using wavelets; nothing seems to be known about the performance of queues with multifractional input.

### 1.5.4 The relevance of LRD in performance analysis

An important practical issue is the impact of LRD on network performance. Conclusions in the literature are mixed, and critically rely on the specific assumptions that are made. For large or infinite buffer sizes, several studies indicate that LRD input leads to severe performance degradation. Erramilli *et al.* [129] perform an experimental queueing analysis with existing traffic traces and find that LRD has a significant influence on queueing behavior. Resnick & Samorodnitsky [235] consider a $G/M/1$ queue with a dependent sequence of interarrival times. They demonstrate that dependence in the
interarrival times can lead to a heavy-tailed waiting-time distribution. (Note that the waiting-time distribution in the $GI/M/1$ queue is exponential, even if the interarrival time distribution is heavy-tailed!) Other studies, mostly concerning large-buffer asymptotics in queueing models with heavy-tailed input, give similar conclusions. The typical result is that heavy-tailed service times in single-server queues and heavy-tailed On-times in fluid queues lead to heavy-tailed waiting times and workloads. In particular, an infinite second moment for the service time in the single-server queue implies an infinite mean for the waiting time.
For moderate buffer sizes, the impact of LRD is not as pronounced, see Grossglauser & Bolot [148], Heyman & Lakshman [154], Mandjes & Kim [196], and Ryu & Elwalid [244]. In addition, flow control mechanisms play a critical role in preventing badly-behaved traffic from overwhelming the buffer content, see Arvidsson & Karlsson [17].
Besides the buffer size and the role played by feedback mechanisms, the performance impact of LRD also crucially depends on the service discipline. In Chapter 3 of this thesis it is shown for example that PS gives much better delay performance than FCFS. Borst *et al.* [54, 55, 56, 57, 58] obtain similar conclusions for a class of (fluid) queues operating under the Generalized Processor Sharing (GPS) policy. Another study on scheduling strategies and LRD is Anantharam [15].

## 1.6 Overview of the thesis

This section gives an overview of the results in this thesis.
In Chapter 2 we give an introduction to heavy-tailed distributions, and review some standard techniques and results for heavy-tailed queueing systems which are relevant in this thesis. We also provide supporting intuitive arguments. The type of intuition is illustrated with several examples. In particular, we provide an explanation of the waiting-time asymptotics of Boxma *et al.* [70] for an $M/G/2$ queue with heterogeneous servers.

In Chapter 3 we derive the sojourn-time asymptotics in the $M/G/1$ queue with the PS discipline. We approach this problem via the transform of the sojourn-time distribution, for which we obtain a novel expression. This chapter is based on Zwart [285] and Zwart & Boxma [287]. The main result we obtain is the following: We show that the tails of the service- and sojourn-time distribution are *equally heavy*. This result radically differs from the situation in the single-server queue with the FCFS discipline, where a heavy-tailed service-time distribution leads to a waiting-time distribution which is even heavier-tailed (see Chapter 2 for a precise result). The results in this chapter further suggest that a large sojourn time of a customer is not caused by other customers, but by its own large service time. This result continues to hold if other customers have an even heavier-tailed service time distribution. This shows that PS-based disciplines are more effective than FCFS in protecting individual customers, especially when service times are heavy-tailed.

Most models in this monograph assume an infinite buffer. An exception is made in Chapter 4, where we study a fluid queue with a finite buffer. This chapter is based on Zwart [286]. We are interested in the mean buffer content and the loss fraction as the buffer size grows large. We obtain several exact results for the stationary distribution of the fluid queue. In particular, we extend the well-known relationship of Kella & Whitt [168] between ordinary queues and fluid queues to finite-buffer systems. Furthermore, we show that the buffer content distributions of the finite- and infinite-buffer fluid queue are *proportional*. This proportionality result is then applied to obtain asymptotics for the loss fraction and mean buffer content. We show that these quantities are significantly influenced by the fact that the input is heavy-tailed. We also show that the output of the fluid queue is still long-range dependent, in spite of the fact that the buffer is finite.

Chapter 5, which is based on Zwart [289], investigates the tail behavior of the *busy-period distribution* in the single-server queue. We extend a result of De Meyer & Teugels [202]. A major (methodological) contribution of this chapter is the new method of proof; this method follows intuition quite closely. In particular, it is shown that a large busy period is caused by a large *cycle maximum*. Another important result is that the tail of the busy-period distribution is similar (up to a constant factor) to the tail of the service time distribution. As a by-product, we obtain asymptotic results for the $GI/G/1$ LCFS queue.

Chapters 6–8 are all devoted to fluid queues with infinite buffers, fed by multiple heavy-tailed On-Off sources. Chapter 6 treats a fluid queue fed by a superposition of light-tailed and heavy-tailed On-Off sources. The system under consideration has the special feature that the drift remains negative when all the heavy-tailed sources are On. Hence, in order to cause a large workload, the light-tailed sources need to deviate from their normal behavior as well. The main result in this chapter, which is based on Borst & Zwart [59], is derived by combining light-tailed and heavy-tailed large deviations. In particular, we show that the workload asymptotics are determined by the simultaneous occurrence of two events, which are entirely different in nature: The heavy-tailed sources are all simultaneously On for a long time, and the light-tailed input deviates from its mean by following a 'twisted'

distribution. The results in this chapter are improvements of previously obtained bounds by Dumas & Simonian [120].

In Chapter 7 (based on Zwart *et al.* [288]), we obtain the exact tail asymptotics for the workload distribution of the fluid queue fed by several heavy-tailed On-Off sources. The problem in this chapter has been studied by many authors, see e.g. [12, 65, 66, 161, 243]. The common assumption in these studies is that a single heavy-tailed On-period is sufficient for the buffer to fill. In practice, it is typically the case that the peak rate of an On-Off source is significantly smaller than the service capacity. Thus, several simultaneous On-periods are needed for a large buffer-content to build up. This constitutes an important open problem, and is solved in Chapter 7. So far, only asymptotic bounds were known in this general case, see Dumas & Simonian [120].

The main results of Chapter 7 can be described as follows. Under reasonably mild assumptions, we show that the workload is asymptotically equivalent to that in a reduced system. The reduced system consists of a 'dominant' subset of the sources, with the original service rate reduced by the mean rate of the other sources. It turns out that the 'dominant' subset may be found from a simple knapsack formulation. The corresponding set of sources may be interpreted as the most likely combination of sources to cause a persistent positive drift in the workload. The analysis of the reduced system involves a powerful probabilistic argument to characterize the most plausible scenario for the workload to reach a large level, and can be viewed as an extension of the analysis in Chapter 5.

Chapter 8 is related to Chapter 7, but now we study another class of input models, namely $M/G/\infty$ input (the number of active sessions is distributed as the number of customers in an $M/G/\infty$ queue). This class of input is more tractable than the superposition of On-Off sources, which makes it possible to give a more detailed analysis.

The contribution of this chapter is comparable to that of Chapter 7. Fluid queues with heavy-tailed $M/G/\infty$ input have been studied in many papers, see e.g. [65, 161, 164, 184, 185, 187, 224, 225, 226, 227, 239]. Like in Chapter 7, the exact asymptotics in these studies all rely on the assumption that a single long session is sufficient for the buffer to fill. Chapter 8 solves the important case where a large workload may be due to multiple long sessions. Besides obtaining the exact workload asymptotics in this system, we also determine the distribution of the most probable time to overflow. In addition, we derive asymptotic bounds for the transient workload distribution. This chapter is based on Borst & Zwart [60].

# Chapter 2

# Methodology

The first chapter of this thesis served as a general introduction to motivate the analysis of queueing systems with heavy tails. This chapter is concerned with the mathematical details involved in the study of such systems. In particular, the goals of this chapter are to

- give an introduction to heavy-tailed distributions;
- treat some basic asymptotic results for queueing systems with heavy tails;
- give the reader insight in the intuition behind these results;
- explain how one may use this intuition for constructing a proof.

Understanding the intuition behind the proofs is crucial, as it will appear many times in this thesis, sometimes in a complex form. In *light-tailed* situations, there is a well-established intuition regarding the occurrence of 'rare events', see e.g. Shwartz & Weiss [254] for a good discussion of the theory of large-deviations in light-tailed systems. Typically, the occurrence of a rare event can be explained by identifying a 'most probable scenario for the rare event to occur'. If several distributions in the queueing model are *heavy tailed* however, then the nature of such scenarios can become entirely different.
We illustrate large-deviations arguments for heavy-tailed phenomena by treating several queueing models. In particular, we discuss asymptotic results for the waiting-time distribution in an $M/G/2$ queue with heterogeneous servers. That discussion is based on Boxma *et al.* [70]. The chapter is concluded by presenting a framework which may be applied to use these heuristics in constructing a proof. This framework is used in Chapters 7 and 8.
The chapter is organized as follows. Section 2.1 gives an introduction to heavy tails. Basic queueing results can be found in Section 2.2; these results are also explained in an intuitive manner. The above-mentioned $M/G/2$ queue is treated in Section 2.3. Section 2.4 describes how one may strengthen intuition to formal proofs.

## 2.1 Heavy-tailed distributions

In this section we introduce some basic definitions and results concerning heavy-tailed distributions. The previous chapter motivates to consider distributions with infinite variance. However, we are not only interested in this class of distributions, but in all distributions for which the tail decreases slower than exponentially. We make this more precise by introducing several classes of heavy-tailed distributions.
Let $X, X_i, i \geq 1$, be independent non-negative random variables with common distribution function $F(x) = \mathbb{P}\{X \leq x\}$. Define $\bar{F}(x) = 1 - F(x)$. We make the following conventions.

**Definition 2.1.1** *$F$ is heavy tailed if, for all $\epsilon > 0$,*

$$\mathbb{E}\{\mathrm{e}^{\epsilon X}\} = \infty,$$

*or equivalently, if for all $\epsilon > 0$,*

$$\frac{\mathbb{P}\{X > x\}}{\mathrm{e}^{-\epsilon x}} \to \infty. \tag{1.1}$$

A major subclass of heavy-tailed distributions is the class of *long-tailed* distributions, defined by

**Definition 2.1.2** *$F$ is long tailed if for any fixed $y > 0$ and $x \to \infty$,*

$$\mathbb{P}\{X > x + y \mid X > x\} = \frac{\bar{F}(x+y)}{\bar{F}(x)} \to 1. \tag{1.2}$$

The class of long-tailed distributions is denoted by $\mathcal{L}$. We will often write $X \in \mathcal{L}$ instead of $F \in \mathcal{L}$. It can be shown that the convergence in (1.2) is uniform in $y$ on compact subintervals. The following lemma shows that $y$ can also be random:

**Lemma 2.1.1** *If $X \in \mathcal{L}$ and $Y$ is independent of $X$ and non-negative, then*

$$\frac{\mathbb{P}\{X - Y > x\}}{\mathbb{P}\{X > x\}} \to 1.$$

The defining property of $\mathcal{L}$ is appealing: If $X$ is long-tailed and $X > x$ for some large $x$, then it is likely that $X$ exceeds any larger value as well. If $X$ has finite mean, we define the excess random variable $X^r$ as a random variable with 'integrated-tail' distribution

$$\mathbb{P}\{X^r > x\} = \frac{1}{\mathbb{E}\{X\}} \int_x^\infty \mathbb{P}\{X > u\}\mathrm{d}u, \quad x \geq 0.$$

Another interesting property of long-tailed distributions is the following lemma.

**Lemma 2.1.2** *If $X \in \mathcal{L}$, then $X^r \in \mathcal{L}$ and*

$$\frac{\mathbb{P}\{X^r > x\}}{\mathbb{P}\{X > x\}} \to \infty.$$

Thus, if $X \in \mathcal{L}$, then the tail of $X^r$ is heavier than the tail of $X$. This result is in contrast with case in which $X$ is exponentially distributed. In this case the distributions of $X$ and $X^r$ coincide, implying equally heavy tails.
The following two subsections are concerned with the two most important subclasses of heavy-tailed distributions. We introduce subexponential distributions in Section 2.1.1. Section 2.1.2 treats regularly varying distributions.

### 2.1.1 Subexponentiality

In this subsection we review some standard (see e.g. [126]) results and definitions.
Let $F^{n*}$ be the $n$-fold convolution of $F$, i.e.,

$$F^{n*}(x) = \int_{u=0}^{x} F^{(n-1)*}(x-u)\mathrm{d}F(u).$$

The class of subexponential distribution functions, denoted by $\mathcal{S}$, is defined as follows.

**Definition 2.1.3** *$F$ is subexponential if*

$$\frac{\bar{F}^{2*}(x)}{\bar{F}(x)} = \frac{\mathbb{P}\{X_1 + X_2 > x\}}{\mathbb{P}\{X > x\}} \to 2, \qquad x \to \infty.$$

The definition of subexponentiality can be weakened: It has been shown in Embrechts & Goldie [124] that $F \in \mathcal{S}$ if

$$\mathbb{P}\{X_1 + \ldots + X_n > x\} \sim n\mathbb{P}\{X_1 > x\}$$

for *some* $n \geq 2$. If $F \in \mathcal{S}$, then this relation holds for all $n \geq 2$. A characterization of $\mathcal{S}$ which may be more appealing is the following.

**Definition 2.1.4** *$F$ is subexponential if, for some $n \geq 2$,*

$$\mathbb{P}\{X_1 + \ldots + X_n > x\} \sim \mathbb{P}\{\max\{X_1, \ldots, X_n\} > x\}.$$

Intuitively, subexponentiality means that large sums are most likely caused by a large value of a single summand; other summands do not make a significant contribution. This makes subexponentiality a commonly-used paradigm in insurance mathematics, especially in modeling catastrophes.
Subexponential distribution functions were introduced independently by Chistyakov [82] and Chover *et al.* [84]. In these references, the framework of subexponential distribution functions was used to derive asymptotic properties of branching processes, see also the textbook of Athreya & Ney [31]. One of the first papers that recognized the usefulness of the class of subexponential distributions is Teugels [265]. In the next section, we explain why subexponentiality plays a key role in queueing theory and insurance. Subexponentiality has also connections with other topics in probability theory, such as infinite divisibility, cf. [122].

Several well-known probability distributions are subexponential. Key examples are:

- *Pareto,*

$$\mathbb{P}\{X > x\} = \left(\frac{a}{a+x}\right)^{\nu}, \qquad a, \nu > 0.$$

- *Lognormal,*

$$\mathbb{P}\{X > x\} = \mathbb{P}\{\mathrm{e}^{\mu+\sigma U} > x\}, \qquad \mu \in \mathbb{R},\ \sigma > 0,$$

  with $U$ a standard-normal random variable.

- *Weibull,*

$$\mathbb{P}\{X > x\} = \mathrm{e}^{-ax^b}, \qquad a > 0, 0 < b < 1.$$

Subexponentiality of the above examples follows, for example, from a sufficient condition for membership of $\mathcal{S}$ of Pitman [229]. The Pareto case will be thoroughly studied in the next subsection.
We proceed by stating some results on subexponential distributions which are used in this thesis. For more extensive surveys (and proofs), we refer to Embrechts *et al.* [126], Goldie & Klüppelberg [139], Mikosch [205], and Sigman [256].
The next lemma is often useful in proofs, for example to justify the interchange of limits and sums. The lemma seems to be due to Kesten, see [31].

**Lemma 2.1.3** *Let $X_i \in \mathcal{S}, i \geq 1$. Then, for all $\epsilon > 0$ there exists a $K < \infty$ such that for all $x \geq 0, n \geq 1$,*

$$\mathbb{P}\{X_1 + \ldots + X_n > x\} \leq K(1+\epsilon)^n \mathbb{P}\{X_1 > x\}.$$

The class $\mathcal{S}$ is *not* closed under convolution, i.e., if $X$ and $Y$ are independent members of $\mathcal{S}$, then $X + Y$ is not necessarily in $\mathcal{S}$, see Leslie [182]. The next lemma gives some sufficient conditions for $X + Y$ to be subexponential.

**Lemma 2.1.4** *([122]) Let $X$ and $Y$ be independent. If $X \in \mathcal{S}$ and $\mathbb{P}\{Y > x\} \sim [K + \mathrm{o}(1)]\mathbb{P}\{X > x\}, K \geq 0$, then $X + Y \in \mathcal{S}$ and $\mathbb{P}\{X + Y > x\} \sim (1 + K)\mathbb{P}\{X > x\}$. Moreover, if $K > 0$, then also $Y \in \mathcal{S}$.*

There are many other properties of subexponential distributions. For example, if $X$ is subexponential and $Y$ is sufficiently well-behaved, then the product $XY$ is subexponential as well, see Cline & Samorodnitsky [87]. This property is used in this thesis in the special case that $X$ is regularly varying. Regular variation is the topic of the next subsection.

### 2.1.2 Regular variation

In the previous subsection we gave three examples of subexponential distributions: Pareto, lognormal, and Weibull. In this section, we study the class of *regularly varying distributions*. This class can be viewed as a generalization of the Pareto distribution. Regularly varying distributions are all subexponential (see Lemma 2.1.8 below). We note that Weibullian and lognormal distributions are not regularly varying.
Regular variation is a topic on its own, with applications in various fields, like complex analysis, number theory, and probability theory. Within probability theory, regular variation plays a key role in extreme-value theory, central limit theorems, branching processes, queueing theory, and more. An encyclopedic treatment of regular variation is Bingham *et al.* [44]. Other key references are De Haan [149], Resnick [232, 233], and Embrechts *et al.* [126].
This subsection is organized as follows. First, we define the class of regularly varying functions and give some general results which are used in this thesis. Next, we treat some basic properties of random variables which have a regularly varying distribution.

#### General results

All functions in this subsection are assumed to be measurable, non-negative and defined on $[x_0, \infty), x_0 > 0$.

**Definition 2.1.5** *$f$ is regularly varying of index $\alpha \in \mathbb{R}$ ($f \in \mathcal{R}_\alpha$), if for all $y > 0$,*

$$\frac{f(yx)}{f(x)} \to y^\alpha, \qquad x \to \infty.$$

*If $\alpha = 0$, then $f$ is called slowly varying.*

Slowly varying functions are usually denoted by $L$. Examples of slowly varying functions are constants and (iterated) logarithms. The class of all regularly varying distributions ($\cup_{\alpha\in\mathbb{R}}\mathcal{R}_\alpha$) is denoted by $\mathcal{R}$.
We now list some properties of regularly varying functions. All of these properties can be found in Bingham *et al.* [44], see also Feller [131]. The following basic property for slowly varying functions is often used without mention.

**Lemma 2.1.5** *Let $L$ be a slowly varying function. Then, for all $\epsilon > 0$, there exists a $T$ such that, if $x > T$,*

$$x^{-\epsilon} \leq L(x) \leq x^\epsilon. \tag{1.3}$$

The next lemma provides a useful bound for slowly varying functions, this bound is one instance of the *Potter bounds.*

**Lemma 2.1.6** *Let $L$ be a slowly varying function. Then, for any fixed $A > 1$, $\delta > 0$, there exists a finite constant $K$ such that for all $x, y > K$,*

$$\frac{L(y)}{L(x)} \leq A \max\{(y/x)^{\delta}, (y/x)^{-\delta}\}.$$

We now come to the first deep result of this subsection. The following lemma is part of Karamata's theorem, see Section 1.6 of [44], and shows that slowly varying functions are precisely those functions which can be treated like a constant in the (asymptotic) evaluation of integrals.

**Lemma 2.1.7** *Let $L$ be locally bounded in $\{x : x \geq T\}$. Let $\alpha > 1$. The following are equivalent:*

1. *$L$ is slowly varying,*
2. *$\int_x^{\infty} y^{-\alpha} L(y) \mathrm{d}y \sim \frac{1}{\alpha-1} x^{1-\alpha} L(x)$.*

Another beautiful result is *Karamata's Tauberian theorem*, see Theorems 1.7.1 and 5.2.4 in [44], which relates the asymptotic behavior of a regularly varying function at infinity to the behavior of its Laplace-Stieltjes transform (LST) near 0. The LST of a function $f$ is given by $\hat{f}(s) = \int_0^{\infty} \mathrm{e}^{-sx} \mathrm{d}f(x)$.

**Theorem 2.1.1** *Let $U$ be a non-decreasing and right-continuous function on $\mathbb{R}$ with $U(x) = 0$ for all $x < 0$. Let $\hat{U}$ be the LST of $U$. If $L$ varies slowly and $c \geq 0$ and $\alpha > 0$, the following are equivalent.*

$$U(x) \sim (c + \mathrm{o}(1)) x^{\alpha} L(x)/\Gamma(1+\alpha), \qquad x \to \infty, \tag{1.4}$$

$$\hat{U}(s) \sim (c + \mathrm{o}(1)) s^{-\alpha} L(1/s), \qquad s \downarrow 0. \tag{1.5}$$

*Conversely, if $U(x)/\hat{U}(1/x) \to \frac{1}{\Gamma(1+\alpha)}$, then $U \in \mathcal{R}_{\alpha}$ and (1.4), (1.5) hold for some slowly varying function $L$.*

The reverse part is a *Mercerian theorem*, see Chapter 5 of [44]. An encyclopedic treatment of Abelian and Tauberian theorems can be found in Chapter 4 of [44].

We now turn to random variables with regularly varying (tails of) distribution functions.

**Regularly varying distribution functions**

A non-negative random variable $X$ is called regularly varying of index $-\alpha$, if

$$\mathbb{P}\{X > x\} = \bar{F}(x) = L(x)x^{-\alpha}, \quad \alpha \geq 0,$$

with $L$ a slowly varying function. With a slight abuse of notation, we write $X \in \mathcal{R}_{-\alpha}$. We now state some basic properties of regularly varying distributions:

**Lemma 2.1.8** *Let $\mathbb{P}\{X > x\} = \bar{F}(x) = L(x)x^{-\alpha}$. Then,*

**(i)** *$X \in \mathcal{S}$.*

**(ii)** *$\mathbb{E}\{X^\theta\} < \infty$ if $\theta < \alpha$, $\mathbb{E}\{X^\theta\} = \infty$ if $\theta > \alpha$.*

**(iii)** *If $\alpha > 1$, then $X^r \in \mathcal{R}_{1-\alpha}$ and*

$$\mathbb{P}\{X^r > x\} \sim \frac{1}{(\alpha-1)\mathbb{E}\{X\}} L(x)x^{1-\alpha}.$$

**(iv)** *If $Y$ is non-negative and independent of $X$ such that $\mathbb{P}\{Y > x\} = L_2(x)x^{-\alpha_2}$, then $X + Y \in \mathcal{R}_{-\min\{\alpha,\alpha_2\}}$, and*

$$\mathbb{P}\{X + Y > x\} \sim \mathbb{P}\{X > x\} + \mathbb{P}\{Y > x\}.$$

**(v)** *If $Y$ is non-negative and independent of $X$ such that $\mathbb{E}\{Y^{\alpha+\epsilon}\} < \infty$ for some $\epsilon > 0$ then $XY \in \mathcal{R}_{-\alpha}$ and*

$$\mathbb{P}\{XY > x\} \sim \mathbb{E}\{Y^\alpha\}\mathbb{P}\{X > x\}.$$

**Proof**
Property (i) can be found in e.g. Feller [131]. (ii) follows from Lemma 2.1.5. (iii) follows from Karamata's theorem (Lemma 2.1.7). (iv) can be found in Feller [131], p. 271. Finally, property (v) is due to Breiman [78], see also [113, 123, 232]. □

Karamata's Tauberian theorem characterizes the asymptotic behavior of a regularly varying function in terms of its LST. This theorem is however not applicable to distribution functions (which are regularly varying with negative index). Fortunately, there exists another Tauberian theorem which is suitable for LST's of random variables. This theorem is due to Bingham & Doney [42], see also Theorem 8.1.6 in [44].
Let $\phi(s)$ be the LST of $F$. Suppose that $X$ has finite first $n$ moments $\mu_1, ..., \mu_n$ (and $\mu_0 = 1$). Define

$$\phi_n(s) := (-1)^{n+1} \left[ \phi(s) - \sum_{j=0}^{n} \mu_j \frac{(-s)^j}{j!} \right].$$

**Theorem 2.1.2** *Let $n < \nu < n+1, n \in \mathbb{N}, C \geq 0$. Then, the following are equivalent.*

$$\phi_n(s) \sim (C + \mathrm{o}(1))s^{\nu}L(1/s), \qquad s \downarrow 0, \qquad s \in \mathbb{R}, \tag{1.6}$$

$$1 - F(t) \sim (C + \mathrm{o}(1))\frac{(-1)^n}{\Gamma(1-\nu)}t^{-\nu}L(t), \qquad t \to \infty. \tag{1.7}$$

The case $C > 0$ is the Tauberian theorem as proven in [42]. The case $C = 0$ is Boxma & Dumas [68], Lemma 2.2. For the more complicated case when $\nu$ is integer, we refer to Theorem 8.1.6 and Chapter 3 of [44].

Bingham & Doney [42, 43] apply Theorem 2.1.2 to analyze asymptotic properties of – again – various branching processes. More applications of regular variation in probability theory can be found in Chapter 8 of [44].
Theorem 2.1.2 is a powerful tool for queueing theorists, as explicit expressions are available for the LST of many random variables occurring in queueing models, see e.g. Cohen [97]. We apply this theorem several times in this thesis, in particular in Chapter 3. Limitations of Theorem 2.1.2 are the restriction to non-integer $\nu$, and the fact that the LST is sometimes unavailable. For example, the LSTs of the stationary workload distributions in the fluid queues studied in Chapter 6–8 are all unavailable, except for some special cases.
There exist many extensions of regular variation, which may be found in [44]. An extension appearing in this thesis is *intermediate regular variation*, which has been introduced by Cline [86]. This class of distributions is characterized by property (1.8) below. The extension may look artificial, but sometimes its characterization is exactly the argument needed in proofs, and therefore the class which should be considered.

**Definition 2.1.6** *$X$ is of intermediate regular variation ($X \in \mathcal{IRV}$) if $X$ satisfies*

$$\lim_{\epsilon \downarrow 0} \liminf_{x \to \infty} \frac{\mathbb{P}\{X > (1+\epsilon)x\}}{\mathbb{P}\{X > x\}} = 1. \tag{1.8}$$

**Lemma 2.1.9** *([86]) If $X \in \mathcal{R}$, then $X \in \mathcal{IRV}$. If $X \in \mathcal{IRV}$, then $X \in \mathcal{S}$ and $X \in \mathcal{L}$ .*

We now proceed with some basic results for queueing models with heavy tails.

## 2.2 Asymptotics for some basic queueing models

### 2.2.1 The single-server queue

In this section, we give some results for the stationary waiting time in the single-server queue.

We start by introducing some notation. Assume that the server works at speed $c$. Let $T, T_i, i \geq 1$, be an i.i.d. sequence of interarrival times and let $B, B_i, i \geq 1$, be an i.i.d. sequence of service times. Define the arrival rate by $\lambda := 1/\mathbb{E}\{T\}$. It is well-known [189] that the stationary waiting time $W^c$ exists as a proper random variable when $\frac{\mathbb{E}\{B\}}{\mathbb{E}\{T\}} =: \rho < c$. Furthermore,

$$W^c \stackrel{d}{=} \sup_{n \geq 1} S_n^c,$$

with $S_n^c$ the random walk with step size $B_i - cT_i$, and '$\stackrel{d}{=}$' denoting equality in distribution. Hence, $W^c$ can be viewed as the supremum of a random walk with negative drift, so its distribution can be studied using Wiener-Hopf theory, cf. Asmussen [19], Cohen [97], and many others.

Computing asymptotics for $\mathbb{P}\{W^c > x\}$ when $x \to \infty$ is a universal problem in queueing theory. Crucial is the tail behavior of the service time distribution. The case of a regularly varying service-time distribution has been treated by Borovkov [50] and Cohen [89]. The following theorem covers the more general case of subexponential service times, and is originally due to Pakes [219] and Veraverbeke [269], see also Embrechts & Veraverbeke [125]. The version we state here is slightly more general and can be found in Korshunov [177].

**Theorem 2.2.1** *$B^r \in \mathcal{S}$ iff $W^c \in \mathcal{S}$ iff*

$$\mathbb{P}\{W^c > x\} \sim \frac{\rho}{c - \rho} \mathbb{P}\{B^r > x\}. \tag{2.1}$$

The above theorem shows why subexponentiality is such a convenient concept for queueing theory; it is precisely that class for which the asymptotics (2.1) hold: Subexponentiality of $B^r$ is not only sufficient but also necessary! This deep result was proved earlier in [219] for the $M/G/1$ case.

Theorem 2.2.1 is formulated in terms of the waiting time, but exactly the same result holds for the stationary workload $V^c$ (complementary results for queue lengths may be found in Asmussen *et al.* [27] and Foss & Korshunov [134]).

**Theorem 2.2.2** *If $B^r \in \mathcal{S}$ then*

$$\mathbb{P}\{V^c > x\} \sim \frac{\rho}{c - \rho} \mathbb{P}\{B^r > x\}. \tag{2.2}$$

**Proof**

Combine the identity $\mathbb{P}\{V^c > x\} = \frac{\rho}{c} \mathbb{P}\{W^c + B^r > x\}$ (see e.g. Asmussen [19] p. 189, or Cohen [97] p. 296) with the previous theorem and Lemma 2.1.4. □

In the $M/G/1$ case, the above theorems immediately follow from PASTA and the well-known Pollaczek-Khintchine formula for the waiting-time distribution:

$$\mathbb{P}\{W^c > x\} = (1-\rho/c)\sum_{n=1}^{\infty}(\rho/c)^n \mathbb{P}\{B_1^r + \ldots + B_n^r > x\}, \tag{2.3}$$

with $B_i^r, i \geq 1$, i.i.d. copies of $B^r$. Interchanging the summation and the limit in $\mathbb{P}\{W^c > x\}/\mathbb{P}\{B^r > x\}$ is justified by Lemma 2.1.3.

**Heuristics**

Theorem 2.2.2 can be explained in a heuristic manner. Suppose that we observe the system at time 0 and that $V^c > x$, for $x$ large. Assume for convenience that the arrival process is Poisson. Our claim is that $V^c$ is large because at some time $-t, t \geq 0$, a customer entered the system, which had a large service time $B$. At that time, the waiting time was O(1). After time $-t$, no exceptional things happen and the system simply drifts with rate $-(c-\rho)$. At time 0 the workload is then approximately $B-(c-\rho)t$. Hence, in order for $V^c$ to be larger than $x$, the large service time $B$ at time $-t$ needs to exceed $x+(c-\rho)t$. The intensity of occurrence of such an event is $\lambda\mathbb{P}\{B > x+(c-\rho)t\}$. Integrating over all $t$ we obtain

$$\mathbb{P}\{V^c > x\} \approx \int_{t=0}^{\infty} \lambda\mathbb{P}\{B > x+(c-\rho)t\}\mathrm{dt}.$$

This yields (2.2) after a straightforward computation.

The above heuristic argument essentially focuses on just one possible scenario for $V^c$ to get large. The fact that the corresponding probability coincides with that in (2.2) shows indirectly that the scenario is dominant, in the sense that the probability of all other possible scenarios is negligible.

### 2.2.2 The fluid queue

The previous subsection focused on asymptotics for the waiting-time distribution in the single-server queue. In this subsection, we review some results for the workload distribution in the fluid queue fed by single or multiple On-Off sources.

**A single on-off source**

Consider a fluid queue with capacity $c$, fed by a single On-Off source, indexed by $i$. As described in Chapter 1, an On-Off source alternates between On- and Off-periods. When the source is On (active), it sends input with rate $r_i > c$. Generic activity and silence (Off-) periods are denoted by $A_i$ and $U_i$. Let $p_i := \frac{\mathbb{E}\{A_i\}}{\mathbb{E}\{A_i\}+\mathbb{E}\{U_i\}}$ be the probability that the On-Off source is active (in steady state). The mean amount of input generated per unit

of time is denoted by $\rho_i$, $\rho_i = p_i r_i$. The stationary workload (buffer content) distribution of a fluid queue with capacity $c$, fed by source $i$, is denoted by $V_i^c$.
The following theorem, due to Jelenković & Lazar [161], yields the tail behavior of the workload distribution.

**Theorem 2.2.3** *If $A_i^r \in \mathcal{S}$, $\rho_i < c < r_i$, then*

$$\mathbb{P}\{V_i^c > x\} \sim (1-p_i)\frac{\rho_i}{c-\rho_i}\mathbb{P}\{A_i^r > \frac{x}{r_i - c}\}.$$

The proof of this theorem is simple: Kella & Whitt [168] express the distribution of $V_i^c$ in terms of the waiting-time distribution of a certain single-server queue. Theorem 2.2.3 then immediately follows by combining this result with Theorem 2.2.1. The relationship between fluid and single-server queues is also exploited in this thesis, see Chapter 4.

Theorem 2.2.3 can be explained in a similar manner as Theorem 2.2.2 above. In this case, the event $V_i^c > x$ is caused by a single long activity period of the On-Off source; we omit the details.

### Reduced-load equivalence

The fluid queue considered above is quite simple, as it only involves a single On-Off source. Several papers are concerned with the extension to multiple sources, see e.g. Boxma [65, 66], Rolski *et al.* [243], Jelenković & Lazar [161], and Agrawal *et al.* [12].
In this section we assume that the fluid queue is fed by two sources. Source 1 is an On-Off source with subexponential activity periods. Source 2 is some general 'well-behaved' source (e.g. a Markov-modulated fluid source; exact conditions may be found in [12]) with mean rate $\rho_2$. The stationary workload is denoted by $V_{\{1,2\}}^c$.
$A_1$ is called Weibullian with index $\alpha$ if $\mathbb{P}\{A_1 > x\} \sim c_1 \mathrm{e}^{-c_2 x^\alpha}$. The following result is derived in Agrawal *et al.* [12].

**Theorem 2.2.4 (Reduced-load equivalence)** *Suppose $r_1 + \rho_2 > c > \rho_1 + \rho_2$. If $A_1$ is of intermediate regular variation, lognormal, or Weibullian with index $0 < \alpha < \frac{1}{3}$, then*

$$\mathbb{P}\{V_{\{1,2\}}^c > x\} \sim \mathbb{P}\{V_1^{c-\rho_2} > x\}. \tag{2.4}$$

*If $A_1$ is Weibullian with index $\alpha \geq \frac{1}{2}$, then (2.4) does not hold.*

The most probable way for $V_{\{1,2\}}^c$ to get large is due to a single long activity period of source 1; source 2 shows no abnormal behavior and just uses its mean service requirement $\rho_2$. Hence, the only influence of source 2 in the above scenario is that it reduces the capacity of the fluid queue by its load from $c$ to $c - \rho_2$. This explains the term 'reduced-load equivalence'.

These heuristics may sound plausible, but do not always hold: If the tail of $A_1$ is not heavy enough, then it may be the case that source 2 behaves 'abnormally' as well. This case is not well understood yet. Another problem with the above theorem is the 'gap' between $\frac{1}{3}$ and $\frac{1}{2}$. Results in Agrawal *et al.* [12] and Zwart [290] suggest that (2.4) can be extended to Weibullian tails with index $< \frac{1}{2}$. The latter paper also contains results for the case of a Weibullian tail with index $\geq \frac{1}{2}$. Further evidence on the critical nature of the value $\frac{1}{2}$ may be found in Asmussen *et al.* [27] and Foss & Korshunov [134].
In this thesis, we do not consider Weibullian tails, and concentrate on tails of (intermediate) regular variation. (We make an exception to this rule in Chapter 4 though.) We do extend Theorem 2.2.4 in another way, namely by removing the condition $r_1 + \rho_2 > c$ (see Chapter 6) and by allowing multiple heavy-tailed On-Off sources (see Chapters 6, 7 and 8).
If the activity period distribution is of intermediate regular variation, then the proof (due to [161]) is quite straightforward. The asymptotic upper bound is established as follows. The system will behave less efficient when it is split in two parts: Serve source 1 with capacity $c - \rho_2 - \epsilon$ and serve source 2 with capacity $\rho_2 + \epsilon$. Then,

$$\begin{aligned}\mathbb{P}\{V_{1,2}^c > x\} &\leq \mathbb{P}\{V_1^{c-\rho_2-\epsilon} + V_2^{\rho_2+\epsilon} > x\} \\ &\sim \mathbb{P}\{V_1^{c-\rho_2-\epsilon} > x\} \\ &\sim (1-p_1)\frac{\rho_1}{c-\rho_1-\rho_2-\epsilon}\mathbb{P}\{A_1^r > \frac{x}{r_1-c+\rho_2+\epsilon}\}.\end{aligned}$$

The second step follows from the fact that $V_2^{\rho_2+\epsilon}$ is light-tailed, the third step follows from Theorem 2.2.3. The asymptotic upper bound then follows by letting $\epsilon \downarrow 0$. This is allowed because $A_1^r$ is of intermediate regular variation (which is one reason that sometimes intermediate regular variation may be the most convenient class to consider). The corresponding lower bound may be found by using a similar technique. We apply such bounding techniques in Chapters 6–8.

## 2.3 A multi-server queue

The asymptotic results considered so far all relied on a reduction to the waiting time in the single-server queue, or the workload in a fluid queue fed by a single On-Off source. In general, such a reduction may not be possible. Two prominent models where that is the case are the fluid queue with multiple heavy-tailed On-Off sources, considered in Chapters 6–8 of this thesis, and the multi-server queue. The latter model is the subject of the present section.
The tail behavior of the waiting-time distribution in the multi-server queue is unknown in the heavy-tailed case. Its characterization is one of the current challenging problems in queueing theory. Motivated by this, Boxma *et al.* [70] have studied a particular multi-server queue with heterogeneous servers, for which an exact (transform) analysis is

possible. It turns out that a (fairly complicated) expression for the LST of the waiting-time distribution can be obtained. In this section, we only state the final asymptotic results and concentrate on a heuristic interpretation. Depending on the traffic intensity, the system shows two qualitatively different overflow scenarios. This section presents heuristics in both cases; formal proofs may be found in [70].

The model under consideration can be described as follows. Customers arrive according to a Poisson process with rate $\lambda$. 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 involved. The service times at server 1 are exponentially distributed with rate $\mu$, and at server 2 they have a general distribution $B(x) := \mathbb{P}\{B \leq x\}$ with mean $\beta$. The steady-state queue length and waiting-time distributions exist if $\lambda < \mu + 1/\beta$. In the sequel, we assume this condition to hold. Denote the probability that server 2 is busy in steady state by $P_2$.

Denote the stationary waiting time by $W$. The tail behavior of the waiting-time distribution $W(t)$ is determined by two scenarios, which correspond to the two cases $\lambda < \mu$ and $\lambda > \mu$. In the first case, to be discussed in Subsection 2.3.1, the exponential server is capable of handling all incoming customers alone: The heavy-tailed server is not necessary for stability. If $\lambda > \mu$ (discussed in Subsection 2.3.2) on the other hand, then the second server is needed to ensure stability. We did not consider the delicate case $\lambda = \mu$.

Several studies contain related (partial) results for multi-server queues. Scheller-Wolf & Sigman [251], and Scheller-Wolf [252] obtain sufficient finite-moment conditions for the waiting time in the $GI/G/c$ queue. Asymptotic lower and upper bounds for $\mathbb{P}\{W > x\}$ in the $GI/G/c$ queue can be found in Foss & Korshunov [133] and Whitt [274]. These studies all indicate that the qualitative tail behavior of the waiting-time distribution crucially depends upon the value of the traffic intensity.

### 2.3.1 The case $\lambda > \mu$

In case $\lambda > \mu$, the exponential server alone cannot cope with all the traffic: The second, 'ill-behaved', server is necessary for stability of the system. This makes it plausible that the heavy-tailed service times at the second server result in a heavy-tailed waiting time. In fact, we have

**Theorem 2.3.1** *Suppose that $\lambda > \mu$ and that*

$$\mathbb{P}\{B > x\} \sim x^{-\nu} L(x), \tag{3.1}$$

*with $\nu \in (m, m+1), m \in \mathbb{N}$. Then*

$$\mathbb{P}\{W > x\} \sim \frac{P_2}{1 - \lambda\beta + \mu\beta} \mathbb{P}\{B^r > \frac{\lambda x}{\lambda - \mu}\}. \tag{3.2}$$

**Heuristics**

First, we make two preliminary observations:

1. The long-term fraction of customers served by server 2 equals $\frac{P_2}{\lambda\beta}$ (note that the mean number of customers handled by server 2 per time unit equals $\frac{P_2}{\beta}$).

2. If both servers are busy, 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

$$\frac{\lambda}{\mu}\frac{\beta\mu}{1+\beta\mu} + \lambda\beta\frac{1}{1+\beta\mu} - 2.$$

We now turn to the heuristic explanation of (3.2). Suppose a customer enters the system in steady state at time $\tau$ (say) and is served by server 2. This happens with probability $\frac{P_2}{\lambda\beta}$ (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 O(1). This means that all incoming customers will be allocated to server 1, implying that the workload at server 1 will increase linearly at rate $\rho - 1$ (with $\rho = \lambda/\mu$). As no work is allocated to the second server, the workload of server 2 decreases with at rate $-1$. This continues until both workloads are the same, which happens at time $\tau + B/\rho$, see Figure 2.1. After time $\tau + B/\rho$, the waiting time decreases at rate $1 - \frac{\lambda}{\mu+\beta^{-1}}$, by observation 2. Hence, at time $\tau + \frac{B}{(\mu-\lambda)\beta+1}$ the effect of the large customer entering the system at time 0 has vanished, see again Figure 2.1.

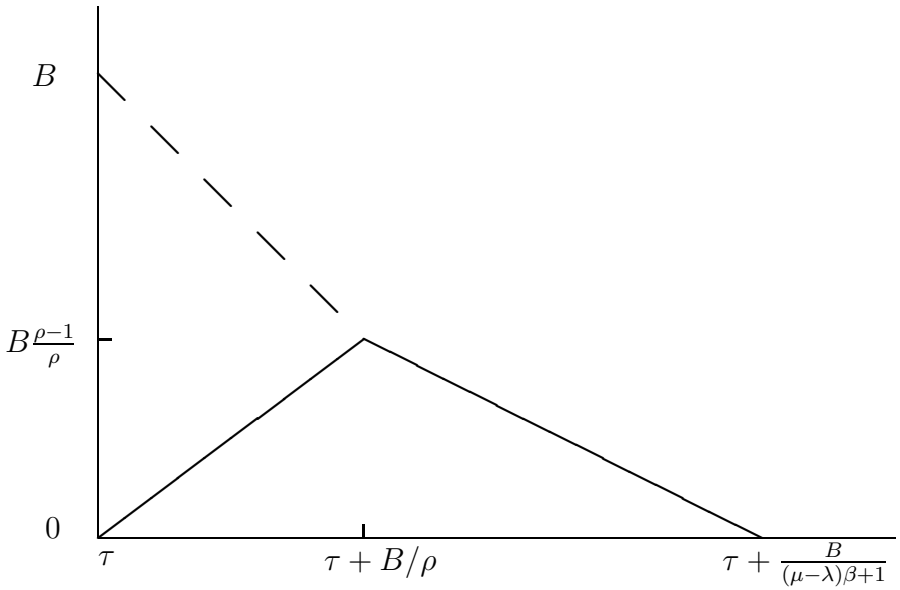


Figure 2.1: Evolution of the waiting time

Suppose that we observe the system at time 0 and that $W > x$, $x$ large. Our claim is that the waiting time is large because at time $\tau = -y$, a customer entered the system and went to server 2. This customer had a large service time $B$. Keeping Figure 2.1 in mind, there are two possible scenarios:

1. $y < B/\rho$. In this case, we are still in the first part of the excursion illustrated in Figure 2.1 (where all incoming customers are sent to the first server). In order to get $W > x$, we need $y > x/(\rho - 1)$.

2. $y > B/\rho$. Using observation 2 (in order to determine the drift after time $B/\rho$ in Figure 2.1), we obtain the condition

$$\begin{aligned} x &< B\frac{\rho-1}{\rho} - \frac{1+(\mu-\lambda)\beta}{1+\mu\beta}\left(y - \frac{B}{\rho}\right) \\ &= \frac{B}{1+\mu\beta} - \frac{1+(\mu-\lambda)\beta}{1+\mu\beta}y. \end{aligned}$$

Together with the condition $y > B/\rho$, this can be rewritten into

$$B > (1+\mu\beta)x + (1+(\mu-\lambda)\beta)y, \qquad y > \frac{x}{\rho-1}.$$

To summarize, the event $W > x$ occurs if at time $y > x/(\rho-1)$ a customer enters the system which is sent to server 2 and has a service time $B > (1+\mu\beta)x + (1+(\mu-\lambda)\beta)y$. By observation 1, the probability that the customer is sent to server 2 equals $\frac{P_2}{\lambda\beta}$. We conclude after a straightforward computation that

$$\begin{aligned} \mathbb{P}\{W > x\} &\approx \int_{\frac{x}{\rho-1}}^{\infty} \frac{P_2}{\lambda\beta}\mathbb{P}\{B > (1+\mu\beta)x + (1+(\mu-\lambda)\beta)y\}\lambda \mathrm{d}y \\ &= \frac{P_2}{1+(\mu-\lambda)\beta}\frac{1}{\beta}\int_{\frac{\rho x}{\rho-1}}^{\infty} \mathbb{P}\{B > z\}\mathrm{d}z, \end{aligned}$$

which is equal to (3.2).

### 2.3.2 The case $\lambda < \mu$

We now turn to the case $\lambda < \mu$. From an analytical (transform) perspective, this case is more intricate, as is explained in [70]. A more advanced Tauberian theorem is necessary in this case, in particular, the analysis in [70] relies on a theorem of Sutton [260]. Here, we ignore this and concentrate on heuristics.
Let $W_{M/M/1}$ be the steady state waiting time in an $M/M/1$ queue with arrival rate $\lambda$ and service rate $\mu$. The precise conditions in the following theorem can be found in [70].

**Theorem 2.3.2** *Suppose that $\lambda < \mu$ and that $\mathbb{P}\{B > x\} = L(x)x^{-\nu}$, with $\nu$ non-integer. Then*

$$\mathbb{P}\{W > x\} \sim P_2\mathbb{P}\{B^r > \frac{\mu x}{\mu - \lambda}\}\mathbb{P}\{W_{M/M/1} > x\}. \tag{3.3}$$

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 [254] or Anantharam [13]). Hence, the amount of time it takes until $W_{M/M/1} > x$ (given that this event occurs) is equal to $\lambda x/(\mu - \lambda)$.
In order for the deviant behavior of the $M/M/1$ queue to take place, server 2 needs to be occupied (which has probability $P_2$) and the past service time $B^p$ of the customer must be larger than $\lambda x/(\mu - \lambda)$. Finally, the residual service time $B^r$ of the customer at server 2 must be larger than $x$. Standard renewal theory (see e.g. [97], p. 113) gives

$$\mathbb{P}\{B^p > \frac{\lambda x}{\mu - \lambda}, B^r > x\} = \mathbb{P}\{B^r > \frac{\mu x}{\mu - \lambda}\}.$$

Combining all these observations yields (3.3). The above interpretation shows an interesting feature of this model: A waiting time becomes very large by the simultaneous occurrence of two events: A very long waiting time at the exponential server ($M/M/1$ large deviations) and one large service time of the heavy-tailed server. Another interesting point is that the nature of these two events is qualitatively different: The latter is heavy-tailed, the former light-tailed. A similar phenomenon can occur in fluid queues, see Chapter 6.

## 2.4 How to make heuristics precise

In the previous two subsections we sketched some heuristic ideas for the single-server queue and a particular $M/G/2$ queue. It turned out that these heuristic arguments provide the correct answer, although their formal proofs rely on different (e.g. transform) methods; they do not use the heuristic ideas at all.
The goal of this section is to give an outline of how to use these heuristic arguments in a formal proof. As an underlying vehicle, we use the workload process in the $M/G/1$ queue. The outline of this section may be viewed as a warming-up for Chapters 7 and 8, where the most probable overflow scenarios are much more difficult to identify.
Starting point is the following representation for the stationary waiting-time distribution in the $M/G/1$ queue with server speed $c$:

$$V^c \stackrel{d}{=} \sup_{t \geq 0}\{A(0,t) - ct\}. \tag{4.1}$$

Here $A(0,t)$ is the cumulative amount of traffic offered to the system between time 0 and time $t$. This type of representation (in terms of a supremum of some process) holds quite generally, in particular for fluid queues (see Chapters 7 and 8). In the $M/G/1$ case considered here, $A(0,t)$ is simply a compound Poisson process with rate $\lambda$ and generic jump size $B$.
In terms of the representation (4.1), the heuristics for the single-server queue can be rephrased as follows. Given $V^c > x$, the process $A(0,t)$ shows average behavior up to time $y$, and simply jumps to a level $> x$ at time $y$. At time $y$, we approximately have $A(0,y) - cy \approx -(c-\rho)y$, so the jump size should be at least $x + (c-\rho)y$.
The probability of a jump of at least $x + (c-\rho)y$ at any time $y \geq 0$ is given by

$$\int_0^\infty \lambda \mathbb{P}\{B > x + (c-\rho)y\} \mathrm{d}y.$$

A straightforward computation then gives the desired value $\frac{\rho}{c-\rho}\mathbb{P}\{B^r > x\}$.

### 2.4.1 Lower bound: Use the law of large numbers

It is not very difficult to use the above heuristics to get a lower bound. The above scenario of a single large jump is a sufficient condition for the event $V^c > x$ to occur. The only thing that needs to be shown is that at the same time $t$, $A(0,t)$ is not much smaller than $\rho t$, but this follows from the law of large numbers.
More formally, we have that for any $\delta, \epsilon > 0$ there exists a finite $t_{\delta,\epsilon}$ such that $\mathbb{P}\{A(0,t) > (\rho-\epsilon)t\} > 1-\delta$ if $t \geq t_{\delta,\epsilon}$. Thus, for any $\delta, \epsilon > 0$,

$$\begin{aligned}
& \mathbb{P}\{V^c > x\} \\
\geq & \int_{t=t_{\delta,\epsilon}}^\infty \lambda \mathbb{P}\{B > x + (c-\rho)t + \epsilon t\} \mathbb{P}\{A(0,t) > (\rho-\epsilon)t\} \mathrm{d}t \\
\geq & (1-\delta) \int_{t=t_{\delta,\epsilon}}^\infty \lambda \mathbb{P}\{B > x + (c-\rho)t + \epsilon t\} \mathrm{d}t \\
= & (1-\delta) \frac{\rho}{c-\rho+\epsilon} \mathbb{P}\{B^r > x + t_{\delta,\epsilon}\} \\
\sim & (1-\delta) \frac{\rho}{c-\rho+\epsilon} \mathbb{P}\{B^r > x\}.
\end{aligned}$$

The second step follows from the law of large numbers; the fourth step from the fact that $B^r \in \mathcal{L}$. The desired lower bound now follows by letting $\epsilon, \delta \downarrow 0$.

### 2.4.2 Upper bound (I): Isolate large jumps

As we saw above, it is not difficult to get a lower bound. Obtaining the corresponding upper bound is a much more demanding task. The most difficult part is giving (and proving) a formal version of the statement "overflow happens as a consequence of a single

big jump". Another difficulty is that one needs to show that other overflow scenarios do not contribute to the asymptotics of the probability under consideration.
In order to prove formal statements, we need to introduce the notion of a large service time. Given that we want to estimate $\mathbb{P}\{V^c > x\}$ for $x \to \infty$, we call a service time $B$ 'large' if $B > \epsilon x$. We need to control the effect of jumps that are smaller than $\epsilon x$. This can be achieved through the following extremely useful lemma, which is due to Resnick & Samorodnitsky [236].

**Lemma 2.4.1** *Let $S_n = X_1 + \ldots + X_n$ be a random walk with i.i.d. step sizes such that $\mathbb{E}\{X_1\} < 0$ and $\mathbb{E}\{(X_1^+)^p\} < \infty$ for some $p > 1$. Then, for any $\alpha < \infty$, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\alpha}$ such that for $\epsilon \in (0, \epsilon^*]$,*

$$\mathbb{P}\{S_n > x | X_j \leq \epsilon x, j = 1, \ldots, n\} \leq \phi(x),$$

*for all $n$ and all $x$.*

In the next subsection, we describe how to apply this result.

**Remark 2.4.1**
Exact asymptotics in the above setting, for both $S_n$ and $\sup_n S_n$ and a regularly varying right tail of $X_1$, have been computed by Jelenković [163]. Note that if $X_j$ can be represented as the difference of two non-negative independent random variables $X_j^1$ and $X_j^2$, then the lemma remains valid if the $X_j$'s are replaced by $X_j^1$.

### 2.4.3 Upper bound (II): Eliminate unlikely scenarios

We now give an outline of how to eliminate all scenarios that are unlikely, emphasizing the main steps. In view of the lower bound, we may neglect all scenarios whose probabilities can be bounded by a regularly varying function of arbitrarily negative index. This is where Lemma 2.4.1 is brought into action.
The scheme presented below is applied to fluid queues in Chapters 7 and 8. A similar scheme can be extracted from a recent study of Resnick & Samorodnitsky [239], who consider a fluid queue with $M/G/\infty$ input.
With $\mathcal{N}(Mx, \epsilon x)$, we denote the number of large jumps before time $Mx$.

- *Overflow occurs in linear time.*
  When considering the process $A(0,s) - cs$, it is convenient to ignore extremely large $s$. This is possible when $\sup_{s\geq 0}\{A(0,s) - cs\} \approx \sup_{s\in[0,Mx]}\{A(0,s) - cs\}$. Formally, one needs to show that

$$\lim_{M\to\infty} \liminf_{x\to\infty} \frac{\mathbb{P}\{\sup_{s\in[0,Mx]}\{A(0,s) - cs\} > x\}}{\mathbb{P}\{\sup_{s\geq 0}\{A(0,s) - cs\} > x\}} = 1. \tag{4.2}$$

- *There is* at least *one large jump in* $[0, Mx]$.
  In this step, one needs to evaluate the probability that $\{\sup_{s\in[0,Mx]}\{A(0,s) - cs\} > x\}$ occurs, while all service times are smaller than $\epsilon x$. This can be made sufficiently small by invoking Lemma 2.4.1.

- *There is* at most *one large jump in* $[0, Mx]$.
  Here one needs to compute the asymptotic behavior of the probability that at least two large jumps occur before time $Mx$. It is not difficult to show that this probability is regularly varying of index $1 - 2\nu > 1 - \nu$ (use the fact that $\mathcal{N}(Mx, \epsilon x)$ has a Poisson distribution). Thus, it can be neglected.

- *The process* $A(0,s) - cs$ *must reach level* $(1-\delta)x$ *when the large jump occurs.*
  This must be shown for every $\delta > 0$. If $A(0,s) - cs$ does not reach level $(1-\delta)x$ at the time of the large jump, then it needs to increase at least $\delta x$ more, by making small jumps only. This event has negligible probability, which follows from another application of Lemma 2.4.1.

Let $\tau(\epsilon x) > 0$ be the time of the first large jump.
The above steps reduce the problem of evaluating $\mathbb{P}\{V^c > x\}$ to the (asymptotic) computation of

$$\mathbb{P}\{A(0, \tau(\epsilon x)) - c\tau(\epsilon x) > (1-\delta)x\}.$$

This calculation is lengthy, but quite straightforward.

Of course, the machinery presented here is unnecessarily heavy for the $M/G/1$ queue (for which much simpler proofs exist, see Section 2.2.1). Nevertheless, it gives detailed insights in the overflow behavior of this system. Another advantage of the method discussed in this section is that it is applicable to more complex models, as will be demonstrated in Chapters 7 and 8.

# Chapter 3

# Sojourn-time asymptotics in the $M/G/1$ PS queue

## 3.1 Introduction

Processor Sharing (PS) queues first became popular by the work of Kleinrock [174], and were originally proposed to analyze the performance of time sharing disciplines in computer systems. Nowadays, PS has also become relevant in modeling (elastic) traffic in communication networks, as is observed in e.g. Núñez-Queija [215] and Roberts [242]. In the PS discipline, the service capacity is always equally shared among all customers present. Thus, if there are $n$ customers present, then each one receives a fraction $\frac{1}{n}$ of the service capacity. From a probabilistic perspective, PS queues are interesting in view of their connections with branching processes, see e.g. Yashkov [281] and Grishechkin [146]. An extensive overview on PS queues can be found in the surveys of Yashkov [282, 283].
This chapter contains various new results for the steady-state sojourn-time distribution of the $M/G/1$ PS queue. In particular, we present explicit asymptotic expansions for the tail of the sojourn-time distribution in case of a regularly varying service-time distribution. The main result of this chapter, Theorem 3.4.1, states that the sojourn-time distribution is regularly varying of index $-\nu$ (with $\nu > 1$ and non-integer) iff the service-time distribution satisfies the same property.
Theorem 3.4.1 reveals a crucial property of PS: It shows that the tail of the service-time and sojourn-time distribution are *equally heavy-tailed*. This is in stark contrast with the $GI/G/1$ FCFS queue. In this case, a result of Cohen [89] (see Theorem 2.2.1) implies that the waiting-time distribution is regularly varying of index $1-\nu$ iff the service-time distribution is regularly varying of index $-\nu$. This implies that if the latter is the case, also the sojourn time is regularly varying of index $1-\nu$. Thus, the tail of the sojourn-time distribution is even fatter than the tail of the service-time distribution. This is due to the FCFS discipline, in which short jobs can be held up by long jobs. Theorem 3.4.1 implies that PS is more effective in handling heavy-tailed service times: Short jobs can overtake

long jobs, so the influence of long jobs on the sojourn time of short jobs is limited.

These nice properties of PS are further exemplified by Theorem 3.4.2, which generalizes Theorem 3.4.1 to the case of several customer classes. Theorem 3.4.2 implies that the tail behavior of the sojourn-time distribution is not heavier than the tail of the service-time distribution, *even if service-time distributions of other customer classes are heavier-tailed*. The additional insight offered by Theorem 3.4.2 is the following: If a job has a long sojourn time, this is due to the fact that *its own* service time is long, so the delay is *not* caused by extremely long service times of other jobs.

In this respect, PS differs from LCFS (with pre-emption). The sojourn time of an arbitrary customer in the LCFS case has (up to a constant) the same tail behavior as in the PS case, but this tail behavior is the same for all types of customers (see e.g. Boxma & Cohen [75] and Chapter 5 of this thesis).

The above-mentioned theorems are proven by means of an application of the Tauberian theorem of Bingham & Doney [42] (stated in Chapter 2 as Theorem 2.1.2), and the expression for the LST of the sojourn-time distribution given by Ott [217]. As a first step, we rewrite the expression for the LST of the sojourn-time distribution. Known expressions for the LST of the sojourn time (see also Schassberger [247], and Yashkov [281]) all contain contour integrals which are inversion formulas of Laplace transforms. We show how to get rid of these contour integrals and thus to obtain a more explicit formula. Using this result, we show how the moments of the sojourn time can be calculated recursively and prove that the $k$-th moment of the sojourn time is finite iff the $k$-th moment of the service time is finite.

Apart from the tail behavior of the sojourn-time distribution, we also study some properties of the sojourn time in heavy traffic. It turns out that, in contrast to the FCFS case (discussed in Chapter 1), it is not necessary to make a distinction between the cases of finite and infinite variance. We give a new proof of a heavy-traffic theorem due to Sengupta [249] & Yashkov [284], and prove similar statements for the moments of the sojourn time in heavy traffic. When the service time has a Pareto distribution, it is possible to give an explicit formula for the heavy traffic limiting distribution. More generally, we show that the heavy traffic limiting distribution is regularly varying of index $-\nu$ if the service-time distribution is regularly varying of index $-\nu$, $\nu > 1$.

This chapter is organized as follows. Preliminary results are given in Section 3.2. In Section 3.3, we derive a new expression for the LST of the sojourn-time distribution and study the moments of the sojourn time. Section 3.4 establishes the above-mentioned asymptotic results for the tail behavior of the sojourn-time distribution, and also contains some additional upper bounds. The proofs of Theorems 3.4.1 and 3.4.2 are given in Section 3.5. The heavy-traffic analysis is performed in Section 3.6. Section 3.7 contains some concluding remarks.

## 3.2 Preliminaries

This section contains some preliminary results for the multi-class $M/G/1$ PS queue. Since we want to study one type of customer in isolation, it suffices to consider only two streams (indexed by $i = 1, 2$), the second stream possibly being the aggregate of several arrival streams.

Customers of type $i$ enter the system according to a Poisson process with rate $\lambda_i > 0$. The service time of a customer of type $i$ is denoted by $B_i$, with distribution function $B_i(x)$, $B_i(0+) = 0$. The moments (if finite) and LST's of these service times are given by $\beta_{i,k}, k \geq 1$, (with $\beta_{i,1} > 0$) and $\beta_i(s)$, respectively. The traffic load offered by class $i$ is given by $\rho_i := \lambda_i \beta_{i,1}$. We also consider the aggregate interarrival and service times. For this purpose, we define $\rho := \rho_1 + \rho_2$, $\lambda := \lambda_1 + \lambda_2$, and

$$\begin{aligned} B(x) &:= \frac{\lambda_1}{\lambda} B_1(x) + \frac{\lambda_2}{\lambda} B_2(x), \qquad x \geq 0, \\ \beta_k &:= \frac{\lambda_1}{\lambda} \beta_{1,k} + \frac{\lambda_2}{\lambda} \beta_{2,k}, \qquad k \geq 1, \\ \beta(s) &:= \frac{\lambda_1}{\lambda} \beta_1(s) + \frac{\lambda_2}{\lambda} \beta_2(s), \qquad \text{Re } s \geq 0. \end{aligned}$$

We denote a random variable with distribution function $B(.)$ by $B$ and assume that the system is stable, i.e. $\rho < 1$. (The server is assumed to work at unit speed.) The distribution of the excess service time $B^r$ (see Chapter 2) and its LST are given by $B^r(.)$ and

$$\beta^r(s) = \int_0^\infty \mathrm{e}^{-st} \mathrm{d}B^r(t) = \frac{1 - \beta(s)}{\beta_1 s}, \qquad \text{Re } s \geq 0.$$

A similar definition holds for $B_i^r(t)$ and $\beta_i^r(s)$.

We are now in a position to describe the queue-length and sojourn-time distributions. A well-known result, due to Sakata *et al.* [245] (see also Kelly [169]), is that the steady-state distribution $(P_n)_{n \geq 0}$ of the number of customers in the system is geometric, and only depends on the service-time distribution through its mean:

$$P_n = (1 - \rho)\rho^n.$$

In the multi-class case, we have for the steady-state distribution $(P_{i,j})_{i,j \geq 0}$ of the number of customers of type 1 and 2, cf. Baskett *et al.* [34], Cohen [95],

$$P_{i,j} = (1 - \rho) \begin{pmatrix} i + j \\ j \end{pmatrix} \rho_1^i \rho_2^j.$$

The sojourn time of a customer (the time that a customer spends in the system) of type $i$ is denoted by $R_i$ with LST $r_i(s)$. Of special interest is the conditional sojourn time $R(\tau)$, defined as the sojourn time of a customer having processing time (service requirement) $\tau$.

It is not difficult to see that this random variable has the same distribution for all types of customers, so we can omit the subscripts. Let $r(s,\tau)$ be the LST of $R(\tau)$. Obviously, we have the identity

$$r_i(s) = \int_0^\infty r(s,\tau)\mathrm{d}B_i(\tau), \qquad i = 1,2. \tag{2.1}$$

The sojourn time of an arbitrary customer is denoted by $R$, and has LST

$$r(s) = \int_0^\infty r(s,\tau)\mathrm{d}B(\tau).$$

Contrasting with the simple product form of the queue-length distribution, the distribution of the sojourn time has a fairly complex form. Yashkov [281] has derived an expression for $r(s,\tau)$ by writing the sojourn time as a functional on a branching process. Using the structure of the branching process, Yashkov found (and solved) a system of differential equations determining $r(s,\tau)$. The analysis in [281] has been extended by Ott [217]. Different approaches are followed in Van den Berg [38] and Schassberger [247].

The expression for $r(s,\tau)$ derived in [217] is the most suitable one for our purposes. It is given by (see also [217], p. 367–368)

$$r(s,\tau) = \frac{1-\rho}{(1-\rho)H_1(s,\tau) + sH_2(s,\tau)}, \tag{2.2}$$

where the functions $H_1$ and $H_2$ are given by,

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}H_1(s,\tau) = \frac{x-\lambda(1-\beta(x))}{x-s-\lambda(1-\beta(x))}, \qquad \mathrm{Re}\ x > 0, \tag{2.3}$$

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}H_2(s,\tau) = \frac{\rho x-\lambda(1-\beta(x))}{x(x-s-\lambda(1-\beta(x)))}, \qquad \mathrm{Re}\ x > 0. \tag{2.4}$$

Denote the $k$-th moment of $R(\tau)$ by $\bar{r}_k(\tau)$. The first moment of $R(\tau)$ is given by, cf. [174], p. 168:

$$\bar{r}_1(\tau) = \frac{\tau}{1-\rho}. \tag{2.5}$$

Note that $\bar{r}_1(\tau)$ is linear in $\tau$. An immediate consequence of (2.5) (or of the expression for $(P_n)_{n\geq 0}$ and Little's formula) is that the first moment of the sojourn time $\mathbb{E}\{R\}$ is finite and equals $\frac{\beta_1}{1-\rho}$ if $\beta_1 < \infty$. Similar statements hold for $R_i, i = 1,2$. In Section 3.3, we will show that a similar result holds for higher moments of the sojourn time. This property contrasts with the FCFS service discipline, where finiteness of the mean sojourn time requires $\beta_2 < \infty$. We come back to this in Section 3.4.

## 3.3 Properties of the conditional sojourn-time distribution

The goal of this section is to provide a novel expression for $r(s,\tau)$ that is suitable to analyze the tail behavior of the sojourn-time distribution in the next section. In particular, we show that $r(s,\tau)^{-1}$ can be written as a power series in $s$. It turns out that the expression contains the LST of the waiting-time distribution $W(.)$ in the M/G/1 FCFS queue, which is given by the Pollaczek-Khintchine formula, i.e.

$$\omega(s) := \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}W(x) = \frac{1-\rho}{1-\rho\beta^r(s)}. \tag{3.1}$$

It can easily be shown by inversion of $\omega(s)^k$ that, for $k \geq 1$ and $x \geq 0$,

$$W^{k*}(x) = (1-\rho)^k \sum_{n=0}^{\infty} \left( \begin{array}{c} n+k-1 \\ k-1 \end{array} \right) \rho^n B^{r,n*}(x). \tag{3.2}$$

We introduce some definitions before the main result of this section is presented. Define the coefficients $\alpha_k(\tau)$, with $k \geq 0$ and $\tau \geq 0$, by $\alpha_0(\tau) := 1$, $\alpha_1(\tau) := \frac{\tau}{1-\rho}$, and for $k \geq 2$,

$$\alpha_k(\tau) := \frac{k}{(1-\rho)^k} \int_{x=0}^{\tau} (\tau - x)^{k-1} W^{(k-1)*}(x)\mathrm{d}x. \tag{3.3}$$

Obviously we can write

$$\alpha_k(\tau) = \left(\frac{\tau}{1-\rho}\right)^k - \delta_k(\tau), \tag{3.4}$$

with $\delta_0(\tau) = \delta_1(\tau) := 0$, and

$$\delta_k(\tau) := \frac{k}{(1-\rho)^k} \int_0^{\tau} (\tau - x)^{k-1} (1 - W^{(k-1)*}(x))\mathrm{d}x, \qquad k = 2, 3, .... \tag{3.5}$$

The next theorem expresses $r(s,\tau)^{-1}$ as a power series in $s$ with coefficients $\frac{\alpha_k(\tau)}{k!}$.

**Theorem 3.3.1** *For* Re $s \geq 0, \tau \geq 0$ :

$$r(s,\tau) = \left[ \sum_{k=0}^{\infty} \frac{s^k}{k!} \alpha_k(\tau) \right]^{-1}. \tag{3.6}$$

This theorem is proven below by analyzing the LST of $r(s,\tau)^{-1}$. It is also possible to prove Theorem 3.3.1 without using transforms, starting from Formula (5.2) in [282]. However, this proof is rather lengthy and therefore omitted. Instead, we give a short proof of Theorem 3.3.1 with the aid of the following lemma.

**Lemma 3.3.1** *For* Re $s \geq 0$ *and* Re $x > 0$ :

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}r(s,\tau)^{-1} = 1 + \frac{1}{1-\rho}\frac{s}{x}\frac{1}{1-\frac{1}{1-\rho}\frac{s}{x}\omega(x)}. \tag{3.7}$$

**Proof**
By (2.2)–(2.4) and (3.1) we have for Re $x > 0$:

$$\begin{aligned}
\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}r(s,\tau)^{-1} &= \frac{x-\lambda(1-\beta(x))}{x-s-\lambda(1-\beta(x))} + \frac{s}{1-\rho}\frac{\rho x-\lambda(1-\beta(x))}{x(x-s-\lambda(1-\beta(x)))} \\
&= 1 + \frac{1}{1-\rho}\frac{s-s\lambda(1-\beta(x))/x}{x-s-\lambda(1-\beta(x))} \\
&= 1 + \frac{1}{1-\rho}\frac{s}{x}\frac{1-\rho\beta_r(x)}{1-\rho\beta_r(x)-\frac{s}{x}} \\
&= 1 + \frac{1}{1-\rho}\frac{s}{x}\frac{1}{1-\frac{1}{1-\rho}\frac{s}{x}\omega(x)},
\end{aligned}$$

which proves the lemma.

□

**Proof of Theorem 3.3.1**
It is sufficient to show that the LST of the power series in the denominator of the right-hand side of (3.6) has the same LST as $r(s,\tau)^{-1}$ for Re $x > |s| + \lambda$. Using the expression for $\omega(s)$, it is not difficult to show that $\left|\frac{s\omega(x)}{(1-\rho)x}\right| < 1$ if Re $x > |s| + \lambda$. Hence, we have by Lemma 3.3.1 that

$$\begin{aligned}
\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}r(s,\tau)^{-1} &= 1 + \frac{1}{1-\rho}\frac{s}{x}\frac{1}{1-\frac{1}{1-\rho}\frac{s}{x}\omega(x)} \\
&= 1 + \sum_{k=1}^\infty \left(\frac{1}{1-\rho}\frac{s}{x}\right)^k \omega(x)^{k-1}.
\end{aligned} \tag{3.8}$$

On the other hand, we have for $k \geq 1$, cf. (3.3),

$$\int_0^\infty \mathrm{e}^{-x\tau}\mathrm{d}\alpha_k(\tau) = \frac{1}{x^k}\frac{k!}{(1-\rho)^k}\omega(x)^{k-1},$$

which completes the proof. □

As a first application of Theorem 3.3.1 we show how the moments $\bar{r}_k(\tau)$ can be found recursively. Note that all $\bar{r}_k(\tau)$ exist and are equal to $(-1)^k \left(\frac{\partial^k}{\partial s^k} r\right)(0,\tau)$, since Theorem 3.1 implies that $r(s,\tau)$ is analytic in $s=0$. From Theorem 3.3.1 we obtain the identity

$$r(s,\tau)\sum_{n=0}^{\infty}\frac{s^n}{n!}\alpha_n(\tau)=1.$$

Differentiating both sides $k$ times w.r.t. $s$ and putting $s=0$, we obtain the following result (with $\bar{r}_0(\tau):=1$).

**Corollary 3.3.1** *For $k\geq 1$ and $\tau\geq 0$,*

$$\bar{r}_k(\tau)=-\sum_{j=1}^{k}\binom{k}{j}\bar{r}_{k-j}(\tau)\alpha_j(\tau)(-1)^j. \tag{3.9}$$

In particular, the variance of $R(\tau)$ is given by

$$\mathbb{V}\mathrm{ar}\{R(\tau)\}=\delta_2(\tau), \qquad \tau\geq 0. \tag{3.10}$$

This result is also obtained in Yashkov [281].

**Remark 3.3.1**
Besides being a tool in the proof of Theorem 3.3.1, Lemma 3.3.1 is also useful for the determination of a tractable expression for $r(s,\tau)$. For example, if the service time is exponentially distributed with parameter $\mu$, it is possible to invert the right hand side of (3.7) by partial fraction expansion, which yields the following expression for $r(s,\tau)$:

$$r(s,\tau)=\left[\frac{s}{1-\rho}\frac{\mu+\lambda-x_1(s)}{x_1(s)x_0(s)}\mathrm{e}^{x_1(s)\tau}-\frac{s}{1-\rho}\frac{\mu+\lambda-x_2(s)}{x_2(s)x_0(s)}\mathrm{e}^{x_2(s)\tau}-\frac{2\rho}{1-\rho}\right]^{-1},$$

with $x_0(s)=x_1(s)-x_2(s)$ and

$$x_1(s)=\frac{1}{2}\left[s+\lambda-\mu+\sqrt{(s+\lambda-\mu)^2+4\mu s}\right],$$

$$x_2(s)=\frac{1}{2}\left[s+\lambda-\mu-\sqrt{(s+\lambda-\mu)^2+4\mu s}\right].$$

The LST of the sojourn-time distribution in the $M/M/1$ PS queue was derived earlier by Coffman *et al.* [88]. Agreement with the result in [88] can be established by noting that $x_1(s)=\lambda\pi(s)-\mu$, where $\pi(s)$ is the LST of the busy period distribution in the M/M/1 queue. We omit the details.
More generally, the right hand side of (3.7) can be inverted when the LST of the service-time distribution is a rational function, since then also $\omega(s)$ and the right hand side of (3.7) are rational functions.

**Remark 3.3.2**
If $\beta_2 < \infty$, then we have the following two-term asymptotic expansion for $\bar{r}_k(\tau)$.

$$\bar{r}_k(\tau) = \bar{r}_1^k(\tau) + \frac{\beta_2}{2\beta_1}\frac{\rho}{1-\rho}\frac{k(k-1)}{(1-\rho)^k}\tau^{k-1} + \mathrm{o}(\tau^{k-1}), \qquad \tau \to \infty.$$

This result can be derived by analyzing the behavior of $\delta_k(\tau)$ for $\tau \to \infty$ by means of its LST and Karamata's Tauberian theorem (see Theorem 2.1.1). Then, use Corollary 3.3.1 and induction. We omit the details. The case $k = 2$ is similar to a result in [281].
If $1 - B(x) = x^{-\nu}L(x)$, $1 < \nu < 2$, then the results are different. We give the asymptotic expansion for $k = 2$. With $L$ we denote a slowly varying function, see Chapter 2.

$$\bar{r}_2(\tau) - \bar{r}_1^2(\tau) = \mathbb{V}\mathrm{ar}\{R(\tau)\} = \delta_2(\tau) \sim \frac{B(2, 2-\nu)}{(1-\rho)^3}\frac{2\lambda}{\nu - 1}\tau^{3-\nu}L(\tau), \tag{3.11}$$

where $B(.,.)$ is the Beta-function. This result can be derived from Equation (3.5), Karamata's theorem (see Lemma 2.1.7) and the asymptotics for $1 - W(x)$ which follow from Theorem 2.2.1.

Using Corollary 3.3.1, it is not difficult to prove the following corollary, which states that the $k$-th moment of the sojourn time $R$ is finite iff the $k$-th moment of the service time $B$ is finite. A similar result holds for $R_i, i = 1, 2$.

**Corollary 3.3.2** *For integer $k \geq 1$,*

$$\mathbb{E}\{R^k\} < \infty \qquad \Leftrightarrow \qquad \beta_k < \infty.$$

**Proof**
Since $R \geq B$ for any particular customer, '$\Rightarrow$' is trivial. To prove '$\Leftarrow$', fix $k \geq 1$ and write

$$\mathbb{E}\{R^k\} = \int_0^\infty \bar{r}_k(\tau)\mathrm{d}B(\tau). \tag{3.12}$$

Note that, cf. (3.3), for $j \geq 1$,

$$\alpha_j(\tau) \leq \frac{\tau^j}{(1-\rho)^j}.$$

From this and Corollary 3.3.1, it is easily shown that

$$\bar{r}_k(\tau) \leq \frac{C_k}{(1-\rho)^k}\tau^k, \tag{3.13}$$

with $C_0 = 1$ and

$$C_k = \sum_{j=0}^{k-1} \binom{k}{j} C_j, \qquad k \geq 1. \tag{3.14}$$

The proof now follows from (3.12)–(3.14). □

Corollary 3.3.2 indicates that the tail behavior of the service-time distribution and the sojourn-time distribution is similar. In the next section, we will study this relation in the case that the service-time distribution or the sojourn-time distribution has a regularly varying tail of index $-\nu$.

## 3.4 Main asymptotic results

In this section we present the main results of this chapter. Section 3.4.1 treats the tail behavior of the sojourn time of an arbitrary customer. Section 3.4.2 presents extensions to the multi-class case. Several complementing bounds for the sojourn-time distribution are derived in Section 3.4.3.

### 3.4.1 The single-class case

In this subsection we present the first main result of this chapter, and establish an asymptotic equivalence between the tails of the service-time distribution and the sojourn-time distribution.

**Theorem 3.4.1** *Let $\nu > 1$, $\nu$ not an integer. The following are equivalent.*

$$(i) \qquad \mathbb{P}\{B > x\} \sim x^{-\nu} L(x), \tag{4.1}$$

$$(ii) \qquad \mathbb{P}\{R > x\} \sim (1-\rho)^{-\nu} x^{-\nu} L(x). \tag{4.2}$$

*Both imply*

$$\mathbb{P}\{R > x\} \sim \mathbb{P}\{B > (1-\rho)x\}. \tag{4.3}$$

The proof of Theorem 3.4.1 is deferred to Section 3.5.

Theorem 3.4.1 illuminates a crucial property of Processor Sharing. We explain this property by a comparison with the FCFS discipline. Theorem 2.2.1 implies that the sojourn-time distribution in the $GI/G/1$ FCFS queue is regularly varying of index $1-\nu$ iff the service-time distribution is regularly varying of index $-\nu$, $\nu > 1$, a result originally due to Cohen [89]. Thus, a heavy-tailed service time leads to an even heavier-tailed sojourn time.
Theorem 3.4.1 shows that this is not the case in the $M/G/1$ PS queue: The sojourn time is as heavy as the tail of the service time. This reveals a crucial property of PS: Long

service times have a much smaller effect on the delay of other customers than in the case of FCFS.

In addition, Theorem 3.4.1 provides insight in the most likely way that the sojourn time of a customer becomes large. In particular, Equation (4.3) can be explained as follows. When a tagged customer is in the system for a long time, the distribution of the total number of customers is approximately equal to the steady-state distribution of the number of customers in a PS queue with one permanent customer. This model is a special case of the $M/G/1$ *generalized* processor sharing queue, as studied by Cohen [95]. Using the results obtained in [95], it is possible to show that the mean service rate in steady state for the tagged (permanent) customer equals $1-\rho$. (Which is no surprise, since the non-permanent customers require mean service rate $\rho$.) Hence, if a tagged customer has been in the system for $x$ time periods, with $x$ large, one would expect that the amount of attained service is roughly equal to $x(1-\rho)$.

It must be emphasized that the above heuristics do not apply in general. For example, (4.1) is not true if the service time is exponentially distributed, as can be shown from the expression for $\mathbb{P}\{R>x\}$ in the $M/M/1$ PS queue given by Morrison [211]. An explanation for this is that, when the service time distribution is exponential, the tagged customer does not stay in the system long enough to reach the equilibrium situation sketched above.

**Remark 3.4.1**

Define the *delay time* $R_d$ of a customer entering the system in steady state as the sojourn time minus the size of the service request. The conditional delay time $R_d(\tau)$ is given by, cf. [282],

$$R_d(\tau) = R(\tau) - \tau. \tag{4.4}$$

The LST's of $R_d$ and $R_d(\tau)$ are denoted by $r_d(s)$ and $r_d(s,\tau)$. Note that

$$\mathbb{E}\{R_d(\tau)\} = \frac{\rho\tau}{1-\rho}, \tag{4.5}$$

$$r_d(s,\tau) = \mathrm{e}^{s\tau} r(s,\tau). \tag{4.6}$$

One can show that the $k$-th moment of $R_d$ is finite iff the $k$-th moment of the service time is finite. If the latter is the case, then this follows from Corollary 3.3.2 and the fact that $R_d \le R$ for any particular customer. If the former holds, then use Jensen's inequality and $\rho > 0$:

$$\infty > \mathbb{E}\{R_d^k\} = \int_0^\infty \mathbb{E}\{R_d(\tau)^k\}\mathrm{d}B(\tau) \ge \left(\frac{\rho}{1-\rho}\right)^k \int_0^\infty \tau^k \mathrm{d}B(\tau) = \left(\frac{\rho}{1-\rho}\right)^k \beta_k.$$

If the service-time distribution is regularly varying of index $-\nu$, $1<\nu<2$, it is possible to show from (4.3)–(4.5), following a similar analysis as in the proof of Theorem 3.4.1 in the next section, that for $x\to\infty$,

$$\mathbb{P}\{R_d > x\} \sim \mathbb{P}\{\frac{\rho}{1-\rho} B > x\}.$$

### 3.4.2 The multi-class case

In this subsection we present asymptotic results for the class-$i$ sojourn-time tail $\mathbb{P}\{R_i > x\}$. Compared with the single-class case, we go one step further and show that the tail of the sojourn-time distribution is as heavy as that of the service-time distribution, even if another customer class possesses a service-time distribution with a heavier tail.

**Theorem 3.4.2** *If there exists a $\mu > 1$ such that $\mathbb{E}\{B^\mu\} < \infty$, then the following are equivalent for non-integer $\nu > 1$,*

$$(i) \qquad \mathbb{P}\{B_1 > x\} \sim x^{-\nu} L(x), \tag{4.7}$$

$$(ii) \qquad \mathbb{P}\{R_1 > x\} \sim (1-\rho)^{-\nu} x^{-\nu} L(x). \tag{4.8}$$

*Both imply*

$$\mathbb{P}\{R_1 > x\} \sim \mathbb{P}\{B_1 > (1-\rho)x\}. \tag{4.9}$$

The condition $\mathbb{E}\{B^\mu\} < \infty$ in Theorem 3.4.2 is made for technical reasons (for which we refer to the proof in the next section); it is weak enough for all practical purposes. In particular, Theorem 3.4.2 provides explicit asymptotics for the following case. Suppose that we have $N$ types of customers, with service times $B_i$ and stationary sojourn times $R_i$. We immediately obtain the following result (choose $\mu \in (1, \min_i \nu_i)$ in Theorem 3.4.2).

**Corollary 3.4.1** *For $i = 1, ..., N$, and non-integer $\nu_i > 1$, $\mathbb{P}\{B_i > x\}$ is regularly varying of index $-\nu_i$ if and only if $\mathbb{P}\{R_i > x\}$ is regularly varying of index $-\nu_i$. Both imply that*

$$\mathbb{P}\{R_i > x\} \sim \mathbb{P}\{B_i > (1-\rho)x\}, \qquad i = 1, \ldots, N. \tag{4.10}$$

To appreciate the implications of Theorem 3.4.2 and Corollary 3.4.1, we compare the multi-class $M/G/1$ PS queue with other service disciplines. Suppose we have a stable $M/G/1$ queue with $N$ types of customers and suppose that the service time of customers of type $i$ is regularly varying of non-integer index $-\nu_i$, with $1 < \nu_1 < \nu_2 < \cdots < \nu_N$. Note that the service time of an arbitrary customer is regularly varying of index $-\nu_1$. We are interested in the tail behavior of the sojourn-time distribution of a customer under the service disciplines FCFS, LCFS and PS.

For a customer of type $i$, the following holds. In the FCFS case, the tail of the customers of type 1 dominates all other types, which leads to a regularly varying sojourn-time distribution of index $1 - \nu_1$ for all types. The index is increased by 1 since an arbitrary customer has to wait with positive probability for a residual service-time period of a customer of type 1. In Anantharam [15] it has been shown that this is the case for all non-preemptive service disciplines where at most one customer is being served at the same time.

The situation under the LCFS pre-emptive regime is slightly better; in this case the sojourn time of an arbitrary customer is regularly varying of index $-\nu_1$, see Boxma & Cohen [75] and Chapter 5 of this thesis. However, the customers of type 1 still dominate the sojourn time of a customer of type $i$. With positive probability, a customer of type 1 enters the system when a customer of type $i$ is being served, so customers of type 1 dominate the tail of the sojourn-time distribution of type $i$.
Theorem 3.4.2 and Corollary 3.4.1 show that under the PS regime, the tail of the sojourn-time distribution of a customer of type $i$ is *not* dominated by a heavier tail of a customer of another type, so that in this case, the sojourn-time distribution is regularly varying of index $-\nu_i$.

**Remark 3.4.2**
Theorem 3.4.2 and Corollary 3.4.1 show that the tail behavior of customer class $i$ is (in case of regular variation) the same as in the $M/G/1$ PS queue where (only) customers of class $i$ enter and where the server works at speed $\rho_i/\rho$.

### 3.4.3 Bounds

The results presented so far all rely on regular variation assumptions. In this section, we derive some upper bounds for the tails of $R(\tau)$ and $R_1$, without assuming regular variation. We believe that these bounds provide some additional insight, but caution that they might be rather crude.
The first result can be proven along the same lines as Theorem 3.4.2.

**Proposition 3.4.1** *If there exists a $\mu > 1$ such that $\mathbb{E}\{B^\mu\} < \infty$, then the following are equivalent,*

$$\mathbb{P}\{R_i > x\} = \mathrm{o}(x^{-\alpha}),\ \forall \alpha > 0, \tag{4.11}$$

$$\mathbb{P}\{B_i > x\} = \mathrm{o}(x^{-\alpha}),\ \forall \alpha > 0. \tag{4.12}$$

In words: The sojourn time distribution is lighter than any power tail iff the service time distribution satisfies the same property. In particular, this result remains true if the service-time distribution of another customer class is heavy-tailed. We can even go a step further: Conditional upon its service requirement, the sojourn time $R(\tau)$ is *always* light-tailed. Formally, we have

**Proposition 3.4.2** *For each $\gamma > \mathrm{e} - 1$ and $\tau > 0$,*

$$\mathbb{P}\{R(\tau) > x\} = \mathrm{o}(\mathrm{e}^{-\frac{1-\rho}{\gamma\tau}x}),$$

*as $x \to \infty$.*

**Proof**
Using Corollary 3.3.1, it can easily be shown that the moments $\bar{r}_k(\tau)$ satisfy the inequality

$$\bar{r}_k(\tau) \leq k! \left( \frac{(\mathrm{e}-1)\tau}{1-\rho} \right)^k. \tag{4.13}$$

This implies that $r(s,\tau)$ can be extended to Re $s \geq -\frac{1-\rho}{\gamma\tau}$, $\gamma > \mathrm{e}-1$. □

This qualitative result supports the conjecture that a large sojourn time is not due to excessive behavior of other customers, but does not imply that $R_1$ is always light-tailed whenever $B_1$ is. For example, when $B_1$ has an exponential distribution, we only get a Weibullian upper bound for $\mathbb{P}\{R_1 > x\}$.

**Proposition 3.4.3** *If $B_1$ is exponentially distributed with rate $\mu$, then*

$$\limsup_{x\to\infty} \frac{\log \mathbb{P}\{R_1 > x\}}{\sqrt{x}} \leq -\sqrt{\frac{1-\rho}{(\mathrm{e}-1)\mu}}. \tag{4.14}$$

**Proof**
From Proposition 3.4.2, we conclude that for each $\gamma > \mathrm{e}-1$, there exists a constant $C$ such that

$$\mathbb{P}\{R(\tau) > x\} \leq C\mathrm{e}^{-\frac{1-\rho}{\gamma\tau}x}.$$

Thus,

$$\begin{aligned}
\mathbb{P}\{R_1 > x\} &= \int_0^\infty \mathbb{P}\{R(\tau) > x\}\mu\mathrm{e}^{-\mu\tau}\mathrm{d}\tau \\
&\leq C\mu \int_0^\infty \mathrm{e}^{-\mu\tau-\frac{1-\rho}{\gamma\tau}x}\mathrm{d}\tau \\
&= C\mu\sqrt{x}\int_{t=0}^\infty \mathrm{e}^{-(\mu t-\frac{1-\rho}{\gamma t})\sqrt{x}}\mathrm{d}t.
\end{aligned}$$

The transformation $t = \tau/\sqrt{x}$ in the last step has paved the way for applying the Laplace method, see for example p. 80 in Olver [216]. Invoking Theorem 7.1 of [216], we conclude that, for some constant $C_2$,

$$\int_{t=0}^\infty \mathrm{e}^{-(\mu t-\frac{1-\rho}{\gamma t})\sqrt{x}}\mathrm{d}t \sim C_2 x^{-\frac{1}{4}}\mathrm{e}^{-\sqrt{\frac{1-\rho}{\gamma\mu}x}}. \tag{4.15}$$

This holds for each $\gamma > \mathrm{e}-1$, yielding (4.14). □

## 3.5 Proof of Theorems 3.4.1 and 3.4.2

In this section we give a proof of the theorems in the previous section. We concentrate on the most general multi-class case (Theorem 3.4.2), from which the single-class case (Theorem 3.4.1) easily follows. The proof makes use of the Tauberian Theorem 2.1.2 and the Expression (3.6) for the LST of the sojourn time distribution.
Before we give a proof of Theorem 3.4.2, we make some preparations in the following three lemmas.

**Lemma 3.5.1** *If $\mathbb{E}\{B^\mu\} < \infty$ for some $\mu > 1$, then there exists a $\delta > 0$ such that for every $n \geq 1$,*

$$\bar{r}_n(\tau) - \bar{r}_1^n(\tau) = \mathrm{o}(\tau^{n-\delta}), \qquad \tau \to \infty, \tag{5.1}$$

$$\sqrt{\mathbb{V}\mathrm{ar}\{R^n(\tau)\}} = \mathrm{o}(\tau^{n-\delta}), \qquad \tau \to \infty. \tag{5.2}$$

**Proof**
From (3.1) and (3.5) we obtain for $k \geq 2$,

$$\int_0^\infty \mathrm{e}^{-s\tau} \mathrm{d}\delta_k(\tau) = \frac{1}{s^k} \frac{k!}{(1-\rho)^k} \left(1 - \omega^{k-1}(s)\right). \tag{5.3}$$

Since $\mathbb{E}\{B^\mu\} < \infty$ for a $\mu > 1$, it follows that, cf. p. 199 in [188],

$$\beta(s) = 1 - \beta_1 s + \mathrm{O}(|s^\mu|), \qquad s \downarrow 0. \tag{5.4}$$

This implies, using (3.1), for $\delta \in (0, \mu - 1)$ and $k \geq 2$,

$$\omega^{k-1}(s) = 1 - \mathrm{o}(s^\delta), \qquad s \downarrow 0. \tag{5.5}$$

Hence, from (5.3) it follows for $k \geq 2$,

$$\int_0^\infty \mathrm{e}^{-s\tau} \mathrm{d}\delta_k(\tau) = \mathrm{o}(s^{\delta-k}), \qquad s \downarrow 0. \tag{5.6}$$

Since the function $\delta_k(\tau)$ is non-decreasing in $\tau$, it follows from Karamata's Tauberian theorem (Theorem 2.1.1) that for $k \geq 2$,

$$\delta_k(\tau) = \mathrm{o}(\tau^{k-\delta}), \qquad \tau \to \infty. \tag{5.7}$$

Equation (5.1) now follows by an inductive argument using (3.3), Corollary 3.3.1 and (5.7). To prove (5.2), we have by using (5.1) for both $\bar{r}_{2n}(\tau)$ and $\bar{r}_n(\tau)$,

$$\mathbb{V}\mathrm{ar}\{R^n(\tau)\} = \bar{r}_{2n}(\tau) - \bar{r}_n^2(\tau) = \mathrm{o}(\tau^{2n-\delta}), \quad \tau \to \infty,$$

which proves (5.2) with $\delta$ replaced by $\frac{1}{2}\delta$. □

Define

$$f(s,\tau) := r(s,\tau) - \mathrm{e}^{-\frac{s\tau}{1-\rho}}. \tag{5.8}$$

We have the following useful lemma, controlling the behavior of $f(s,\tau)$ for small $s$ and $\tau$ not too large (i.e., $\tau \leq K/s$ with $K$ some large constant).

**Lemma 3.5.2** *If $\mathbb{E}\{B^\mu\} < \infty$ for a $\mu > 1$, then, for $\gamma \in (0,1)$, $\gamma < \mu - 1$,*

$$f(s,\tau) = \mathrm{o}(s^\gamma), \qquad \tau = \mathrm{O}\left(1/s\right), \qquad s \downarrow 0.$$

**Proof**
Without loss of generality, it can be assumed that $\mu < 2$. From Theorem 3.3.1 and (3.3) we get

$$f(s,\tau) = \frac{\mathrm{e}^{-\frac{2s\tau}{1-\rho}} \sum_{k=2}^{\infty} \frac{s^k}{k!}\delta_k(\tau)}{1 - \mathrm{e}^{-\frac{s\tau}{1-\rho}} \sum_{k=2}^{\infty} \frac{s^k}{k!}\delta_k(\tau)}. \tag{5.9}$$

It follows immediately from (4.9), using $\delta_k(\tau) \leq \frac{\tau^k}{(1-\rho)^k}$, that for real $s \geq 0$,

$$f(s,\tau) \leq \frac{1}{1+\frac{s\tau}{1-\rho}} \mathrm{e}^{-\frac{s\tau}{1-\rho}} \sum_{k=2}^{\infty} \frac{s^k}{k!}\delta_k(\tau) \leq \mathrm{e}^{-\frac{s\tau}{1-\rho}} \sum_{k=2}^{\infty} \frac{s^k}{k!}\delta_k(\tau). \tag{5.10}$$

We now derive an upper bound for $\delta_k(\tau)$. In view of (3.5), we need an upper bound for $1 - W^{(k-1)*}(x)$. From (5.5) with $k = 2$ and Theorem 2.1.2 with $C = 0$, we obtain for $\epsilon \in (0, \mu - 1)$

$$1 - W(x) = \mathrm{o}(x^{-\epsilon}), \qquad x \to \infty. \tag{5.11}$$

Let $(W_i)_{i\geq 1}$ be an i.i.d. sequence with distribution function $W(x)$. Then, we have

$$\begin{aligned} 1 - W^{(k-1)*}(x) &= \mathbb{P}\{W_1 + \cdots + W_{k-1} > x\} \leq \mathbb{P}\{\cup_{i=1}^{k-1}\{W_i > \frac{x}{k-1}\}\} \\ &\leq (k-1)\mathbb{P}\{W_1 > \frac{x}{k-1}\}. \end{aligned}$$

Combining this with (5.11) we get for $x \to \infty$,

$$1 - W^{(k-1)*}(x) \leq (k-1)^2 \mathrm{o}(x^{-\epsilon}), \tag{5.12}$$

where $\mathrm{o}(x^{-\epsilon})$ is independent of $k \geq 2$. This implies, for Re $s \geq 0$,

$$\delta_k(\tau) \leq k(k-1)^2 \left(\frac{\tau}{1-\rho}\right)^k \mathrm{o}(\tau^{-\epsilon}), \tag{5.13}$$

where $\tau \to \infty$. Since $\tau = \mathrm{O}(1/s), s \to 0$, it follows that

$$\begin{aligned} f(s,\tau) &\leq \mathrm{e}^{-\frac{s\tau}{1-\rho}} \sum_{k=2}^{\infty} \frac{s^k}{k!}(k-1)^2 k \left(\frac{\tau}{1-\rho}\right)^k \mathrm{o}(\tau^{-\epsilon}) \\ &\leq \mathrm{o}(\tau^{-\epsilon})\mathrm{e}^{-\frac{s\tau}{1-\rho}} \left(\frac{s\tau}{1-\rho}\right)^2 \sum_{k=0}^{\infty} \frac{k+1}{k!} \left(\frac{s\tau}{1-\rho}\right)^k \\ &= \mathrm{o}(\tau^{-\epsilon}) \left(\frac{s\tau}{1-\rho}\right)^2 \left[1 + \left(\frac{s\tau}{1-\rho}\right)\right] \\ &= \mathrm{o}(\tau^{-\epsilon}) \left(\frac{s\tau}{1-\rho}\right)^2 (1 + \mathrm{O}(1)) = \mathrm{o}(s^{\epsilon}). \end{aligned}$$

Since this result applies for all $\epsilon \in (0, \mu - 1)$, the lemma is proven. □

The $n$-th derivative of $f(s,\tau)$ with respect to $s$ is defined by

$$f^{(n)}(s,\tau) := \frac{\partial^n}{\partial s^n} f(s,\tau). \tag{5.14}$$

The following upper bound for $f^{(n+1)}(s,\tau)$ will be useful.

**Lemma 3.5.3** *For $n \geq 1$, $s \geq 0$, $\tau \geq 0$,*

$$|f^{(n+1)}(s,\tau)| \leq \mathrm{e}^{-s\tau}\sqrt{\mathbb{V}\mathrm{ar}\{R^{n+1}(\tau)\}} + r(s,\tau)\left(\bar{r}_{n+1}(\tau) - \bar{r}_1^{n+1}(\tau)\right) + \left(\frac{\tau}{1-\rho}\right)^{n+1} f(s,\tau).$$

**Proof**
Using the probabilistic interpretation $f(s,\tau) = \mathbb{E}\{\mathrm{e}^{-sR(\tau)}\} - \mathrm{e}^{-\frac{s\tau}{1-\rho}}$, we obtain

$$\begin{aligned} |f^{(n+1)}(s,\tau)| &= \left|\mathbb{E}\{R^{n+1}(\tau)\mathrm{e}^{-sR(\tau)}\} - \left(\frac{\tau}{1-\rho}\right)^{n+1} \mathrm{e}^{-\frac{s\tau}{1-\rho}}\right| \\ &\leq \left|\mathbb{E}\{R^{n+1}(\tau)\mathrm{e}^{-sR(\tau)}\} - \mathbb{E}\{R^{n+1}(\tau)\}\mathbb{E}\{\mathrm{e}^{-sR(\tau)}\}\right| \\ &+ \left|\mathbb{E}\{R^{n+1}(\tau)\}\mathbb{E}\{\mathrm{e}^{-sR(\tau)}\} - \left(\frac{\tau}{1-\rho}\right)^{n+1} \mathbb{E}\{\mathrm{e}^{-sR(\tau)}\}\right| \\ &+ \left|\left(\frac{\tau}{1-\rho}\right)^{n+1} \mathbb{E}\{\mathrm{e}^{-sR(\tau)}\} - \left(\frac{\tau}{1-\rho}\right)^{n+1} \mathrm{e}^{-\frac{s\tau}{1-\rho}}\right| \\ &= |\mathbb{C}\mathrm{ov}\{R^{n+1}(\tau), \mathrm{e}^{-sR(\tau)}\}| + r(s,\tau)\left(\bar{r}_{n+1}(\tau) - \left(\frac{\tau}{1-\rho}\right)^{n+1}\right) \\ &+ \left(\frac{\tau}{1-\rho}\right)^{n+1} f(s,\tau). \end{aligned}$$

The first term can be bounded by using the inequality of Cauchy-Schwarz and noting that

$$\mathbb{V}\mathrm{ar}\{\mathrm{e}^{-sR(\tau)}\} \leq \mathbb{E}\{\mathrm{e}^{-2sR(\tau)}\} \leq \mathrm{e}^{-2s\tau},$$

since $R(\tau) \geq \tau$.

□

**Proof of Theorem 3.4.2**

Recall that $\mathbb{E}\{B^\mu\} < \infty$ for some $\mu > 1$. Let $\nu \in (n, n+1)$. Without loss of generality, we can assume that $\mu < \min(\nu, 2)$. By Theorem 2.1.2, it suffices to show that (i) or (ii) in Theorem 3.4.2 implies that for real $s \downarrow 0$,

$$r_1(s) - \beta_1\left(\frac{s}{1-\rho}\right) - \sum_{k=0}^{n} \frac{(-s)^k}{k!}\left(\mathbb{E}\{R_1^k\} - \frac{\beta_{1,k}}{(1-\rho)^k}\right) = \mathrm{o}(s^\nu L(1/s)). \tag{5.15}$$

Write

$$r_1(s) - \beta_1\left(\frac{s}{1-\rho}\right) - \sum_{k=0}^{n} \frac{(-s)^k}{k!}\left(\mathbb{E}\{R_1^k\} - \frac{\beta_{1,k}}{(1-\rho)^k}\right) = \int_0^\infty f_n(s,\tau)\mathrm{d}B_1(\tau),$$

with $f_n(s,\tau)$ the residual term of the $n$-term Taylor expansion of $f(s,\tau)$ in $s=0$, i.e.

$$f_n(s,\tau) = f(s,\tau) - \sum_{k=0}^{n} \frac{s^k}{k!} f^{(k)}(0,\tau). \tag{5.16}$$

Since $f(s,\tau)$ is analytic in $s=0$, we can apply Taylor's theorem, which gives, for $s$ in a neighborhood of 0,

$$|f_n(s,\tau)| = \left|\int_0^s \frac{(s-u)^n}{n!} f^{(n+1)}(u,\tau)\mathrm{d}u\right| \leq s^n \int_0^s \left|f^{(n+1)}(u,\tau)\right| \mathrm{d}u. \tag{5.17}$$

Using Lemma 3.5.3 and (5.17) we obtain

$$\begin{aligned}
&\int_0^\infty |f_n(s,\tau)|\,\mathrm{d}B_1(\tau) \\
\leq\ & s^n \int_0^\infty \sqrt{\mathbb{V}\mathrm{ar}\{R^{n+1}(\tau)\}} \int_0^s \mathrm{e}^{-u\tau}\mathrm{d}u\mathrm{d}B_1(\tau) \\
+\ & s^n \int_0^\infty \left(\bar{r}_{n+1}(\tau) - \bar{r}_1^{n+1}(\tau)\right) \int_0^s r(u,\tau)\mathrm{d}u\mathrm{d}B_1(\tau) \\
+\ & s^n \int_0^\infty \bar{r}_1^{n+1}(\tau) \int_0^s f(u,\tau)\mathrm{d}u\mathrm{d}B_1(\tau) =: I + II + III.
\end{aligned}$$

This implies that the proof of Theorem 3.4.2 is complete once we have shown that all three integrals (I, II and III) on the right hand side are of $o(s^\nu L(1/s))$ for $s \downarrow 0$. Thus, the remainder of the proof is split up in three parts (I,II, and III). Suppose that (i) in Theorem 3.4.2 holds with equality (which is no restriction, since the class of regularly varying distributions is closed under tail equivalence).

**Part I**

It is convenient to split the integral in two parts. Using Part (ii) of Lemma 3.5.1 for $\tau$ large we get for a $\delta > 0$ and a finite constant $M$:

$$\begin{aligned} & s^n \int_0^\infty \sqrt{\mathbb{V}\mathrm{ar}\{R^{n+1}(\tau)\}} \int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \mathrm{d}B_1(\tau) \qquad (5.18) \\ \leq\ & Ms^n \int_0^{s^{-1}} \tau^{n+1-\delta} \int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \mathrm{d}B_1(\tau) + Ms^n \int_{s^{-1}}^\infty \tau^{n+1-\delta} \int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \mathrm{d}B_1(\tau). \end{aligned}$$

The first part of (5.18) can be bounded by using $\mathrm{e}^{-u\tau} \leq 1$ and $\tau \leq s^{-1}$:

$$\begin{aligned} & Ms^n \int_0^{s^{-1}} \tau^{n+1-\delta} \int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \mathrm{d}B_1(\tau) \leq Ms^{n+1} \int_0^{s^{-1}} \tau^{n+1-\delta} \mathrm{d}B_1(\tau) \\ \leq\ & Ms^{\nu+\frac{1}{2}\delta} \int_0^{s^{-1}} \tau^{\nu-\frac{1}{2}\delta} \mathrm{d}B_1(\tau) \leq M\mathbb{E}\{B_1^{\nu-\frac{1}{2}\delta}\} s^{\nu+\frac{1}{2}\delta}. \end{aligned}$$

To bound the second integral in the right-hand side of (5.18), use $\int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \leq \frac{1}{\tau}$ and apply partial integration:

$$\begin{aligned} & Ms^n \int_{s^{-1}}^\infty \tau^{n+1-\delta} \int_0^s \mathrm{e}^{-u\tau} \mathrm{d}u \mathrm{d}B_1(\tau) \\ \leq\ & -Ms^n \int_{s^{-1}}^\infty \tau^{n-\delta} \mathrm{d}(1 - B_1(\tau)) \\ =\ & M(1 - B_1(1/s))s^\delta + M(n-\delta)s^n \int_{s^{-1}}^\infty \tau^{n-1-\delta}(1 - B_1(\tau))\mathrm{d}\tau \\ =\ & Ms^{\nu+\delta} L(1/s) + M(n-\delta)s^n \int_{s^{-1}}^\infty \tau^{n-1-\delta-\nu} L(\tau)\mathrm{d}\tau. \end{aligned}$$

It follows from Karamata's theorem (Lemma 2.1.7) that the expression in the right-hand side behaves proportionally to $s^{\nu+\delta}L(1/s)$ for real $s \downarrow 0$, which completes the proof of

Part I.

**Part II**

Identical to Part I, using Part (i) of Lemma 3.5.1, and $r(s,\tau) \leq \mathrm{e}^{-s\tau}$.

**Part III**

We split the leftmost integral in III again up in two parts, namely $0 \leq \tau \leq T/s$ for some finite $T$, and $\tau \geq T/s$. Using Lemma 3.5.2 and a similar calculation as in the first part of the proof of Part I, we can conclude that the first integral is $\mathrm{o}(s^{\nu+\epsilon})$, for an $\epsilon > 0$ and $s \downarrow 0$. This result holds for each finite $T$. Bounding the second integral is more difficult than in Part I and II. It follows from the first inequality in (5.10) and $\delta_k(\tau) \leq \left(\frac{\tau}{1-\rho}\right)^k$, $k \geq 2$, that $f(s,\tau) \leq 1/(1+\frac{s\tau}{1-\rho})$. This implies that

$$\int_0^s f(u,\tau)\mathrm{d}u \leq \int_0^s \frac{1}{1+\frac{u\tau}{1-\rho}}\mathrm{d}u = \frac{1-\rho}{\tau}\ln\left(1+\frac{s\tau}{1-\rho}\right).$$

Note that for every $\gamma > 0$ we have for large $T$ that $\ln\left(1+\frac{s\tau}{1-\rho}\right) \leq (s\tau)^\gamma$ if $\tau \geq T/s$. The second integral can now be handled by using partial integration, with $\gamma \in (0, \nu-n)$:

$$\begin{aligned}
& s^n \int_{T/s}^{\infty} \left(\frac{\tau}{1-\rho}\right)^{n+1} \int_0^s f(u,\tau)\mathrm{d}u\mathrm{d}B_1(\tau) \\
\leq\ & s^{n+\gamma} \int_{T/s}^{\infty} \tau^{n+\gamma}\mathrm{d}B_1(\tau) \\
=\ & s^{n+\gamma}(T/s)^{n+\gamma}(1-B_1(T/s)) + (n+\gamma)s^{n+\gamma}\int_{T/s}^{\infty}(1-B_1(\tau))\tau^{n+\gamma-1}\mathrm{d}\tau \\
=\ & s^\nu L(1/s)T^{n+\gamma-\nu} + (n+\gamma)s^{n+\gamma}\int_{T/s}^{\infty}\tau^{n+\gamma-1-\nu}L(\tau)\mathrm{d}\tau \\
\sim\ & s^\nu L(1/s)T^{n+\gamma-\nu}\left(1+\frac{n+\gamma}{\nu-n-\gamma}\right), \qquad s \downarrow 0. \qquad (5.19)
\end{aligned}$$

The last result holds for for every $T > 0$ by another application of Karamata's theorem. Part III is then completed by choosing $T$ arbitrarily large.

We conclude that (5.15) holds, which implies (ii) of Theorem 3.4.2 by invoking Theorem 2.1.2. If Part (ii) of Theorem 3.4.2 holds with equality, the proofs of I, II, and III (and hence that of (i) of Theorem 3.4.2) follow similarly, using the stochastic dominance $1-B_1(\tau) \leq \mathbb{P}\{R_1 > \tau\} = (1-\rho)^{-\nu}\tau^{-\nu}L(\tau)$; we omit the details.

## 3.6 Heavy traffic and heavy tails

In this section we give a new proof of a heavy-traffic theorem (due to [249, 284]) based on Theorem 3.4.1. We will show that the 'contracted' moments of the sojourn times converge to the moments of the limiting distribution. Finally, we give both explicit and asymptotic results for the sojourn-time distribution in heavy traffic when the service-time distribution has a regularly varying tail.

### 3.6.1 General results

We present a new proof for the following result, see [249, 284].

**Theorem 3.6.1** *If $\beta_1 < \infty$, then*

$$\lim_{\rho\uparrow 1} v(s(1-\rho),\tau) = \frac{1}{1+s\tau}, \qquad \text{Re } s \geq 0, \quad \tau \geq 0, \tag{6.1}$$

$$\lim_{\rho\uparrow 1} \mathbb{P}\{(1-\rho)R(\tau) \leq x\} = 1 - \mathrm{e}^{-\frac{x}{\tau}}, \qquad x \geq 0, \quad \tau \geq 0. \tag{6.2}$$

A heavy-traffic theorem for the $GI/G/1$ PS queue is also known, see Grishechkin [147]. Note that it is only required that the first moment of the service-time distribution is finite, which is not the case in the FCFS service discipline, as discussed in Chapter 1.

**Proof**
Note that (6.1) and (6.2) are equivalent. Since

$$r(s(1-\rho),\tau) = \left[1 + s\tau + \sum_{k=2}^{\infty} \frac{s^k}{k!}(1-\rho)^k \alpha_k(\tau)\right]^{-1},$$

it suffices to show that, for $k \geq 2$,

$$\lim_{\rho\uparrow 1}(1-\rho)^k \alpha_k(\tau) = 0. \tag{6.3}$$

This follows immediately from (3.3) and the fact that $\lim_{\rho\uparrow 1} W(x) = 0$ for $x \geq 0$. Indeed, when $\beta_2 < \infty$ this follows from the standard heavy-traffic limit, which can be found in e.g. Cohen [97], p. 597. If $\beta_2 = \infty$, then it must hold that $B^r(x) < 1$. Hence, since $B^{r,n*}(x) \leq B^n(x)$,

$$W(x) = (1-\rho)\sum_{n=0}^{\infty} \rho^n B^{r,n*}(x) \leq \frac{1-\rho}{1-\rho B^r(x)} \to 0,$$

when $\rho \to 1$. □

Since $r(s(1-\rho),\tau) \leq 1$, we have by dominated convergence and Theorem 3.6.1 the following heavy-traffic limit for the unconditional sojourn-time distribution.

**Corollary 3.6.1** *For* Re $s \geq 0$*,*

$$\lim_{\rho \uparrow 1} r(s(1-\rho)) = \int_0^\infty \frac{1}{1+s\tau} \mathrm{d}B(\tau), \tag{6.4}$$

*and*

$$\lim_{\rho \uparrow 1} r(s(1-\rho)) = \int_0^\infty \mathrm{e}^{-x} \beta(sx) \mathrm{d}x. \tag{6.5}$$

**Proof**
Equation (6.4) follows from Theorem 3.6.1 and bounded convergence. Equation (6.5) follows easily from (6.4) since

$$\int_0^\infty \frac{1}{1+s\tau} \mathrm{d}B(\tau) = \int_0^\infty \int_0^\infty \mathrm{e}^{-x-s\tau x} \mathrm{d}x \mathrm{d}B(\tau) = \int_0^\infty \mathrm{e}^{-x} \beta(sx) \mathrm{d}x.$$

□

This result has also been obtained by Sengupta [249]. Note that (6.5) is the LST of a random variable $Y := XB$, where $B$ is equal to the service time and $X$ is exponentially distributed with mean 1 and independent of $B$. A similar interpretation is given in [249], where it serves as a basis for approximations for the sojourn-time distribution in the $GI/G/1$ PS queue.
The above results remain of course true in the multi-class case. In this case, $(1-\rho)R_i$ converges weakly to $XB_i$. Hence, in heavy traffic, the sojourn-time distribution of a customer is completely determined by its own service-time distribution. To state it differently: PS provides perfect isolation between customer classes in heavy traffic.

We now turn to convergence of the moments of the sojourn time in heavy traffic. It will be shown that the moments of the contracted sojourn time converge to the corresponding moments of the heavy-traffic limiting distribution. Instead of using arguments concerning uniform integrability, cf. [41] p. 338, after which Theorem 3.6.2 below readily follows from (6.2), we follow another approach by using Corollary 3.3.1.

**Theorem 3.6.2** *If* $\beta_1 < \infty$*, then*

$$\lim_{\rho \uparrow 1} \mathbb{E}\{((1-\rho)R(\tau))^k\} = k!\tau^k, \qquad \tau \geq 0, \quad k \geq 1.$$

**Proof**
We apply induction w.r.t. $k$. Fix $\tau \geq 0$. By (2.4), the result holds for $k = 1$. Suppose the result is true for $k = 1, \ldots, n$, $n > 1$. By Corollary 3.3.1 we have

$$(1-\rho)^{n+1}\bar{r}_{n+1}(\tau) = -\sum_{j=1}^{n+1} \left( \begin{array}{c} n+1 \\ j \end{array} \right) (1-\rho)^{n+1-j}\bar{r}_{n+1-j}(\tau)(1-\rho)^j \alpha_j(\tau)(-1)^j. \quad (6.6)$$

The result follows after some simple calculations for $k = n+1$ by the induction hypothesis, (6.3), and $(1-\rho)\alpha_1(\tau) \equiv \tau$. □

A similar result holds for the unconditional moments of the sojourn time, whenever they exist.

**Corollary 3.6.2** *If $\beta_k < \infty$, $k \geq 1$, then*

$$\lim_{\rho \uparrow 1} \mathbb{E}\{((1-\rho)R)^k\} = k!\beta_k.$$

**Proof.** The same idea is used as in the proof of Corollary 3.3.2. Write

$$\mathbb{E}\{((1-\rho)R)^k\} = \int_0^\infty (1-\rho)^k \bar{r}_k(\tau) \mathrm{d}B(\tau).$$

Note that, cf. (3.3),

$$(1-\rho)^k \alpha_k(\tau) \leq \tau^k. \quad (6.7)$$

From (6.6), (6.7) and by induction w.r.t. $k$, it is trivially seen that $(1-\rho)^k r_k(\tau) \leq C_k \tau^k$, with $C_0 = 1$ and

$$C_k = \sum_{j=0}^{k-1} \left( \begin{array}{c} k \\ j \end{array} \right) C_j, \qquad k \geq 1.$$

The result follows by dominated convergence and Theorem 3.6.2.

□

Van den Berg [38] (Chapter 4) has proven Theorem 3.6.2 and Corollary 3.6.2 in the case $k = 2$. Numerical results in [38] indicate that the heavy-traffic approximation for the second moment of the sojourn time performs well.

**Remark 3.6.1**
Abate & Whitt [5] perform a heavy-traffic analysis for the waiting time in the $M/G/1$ LCFS system. They prove a heavy-traffic theorem for the moments of the waiting time

under additional assumptions to meet uniform integrability conditions. The latter concept can also be applied in our case without making any additional assumptions. Note that in our case the $k$-th moment of the heavy-traffic limiting distribution is equal to $k!\beta_k$ if $\beta_k < \infty$.

### 3.6.2 An explicit expression for the limiting distribution

Let $R_{HT}$ be a random variable with a distribution equal to the heavy-traffic limit, i.e.

$$\mathbb{P}\{R_{HT} \leq x\} := \lim_{\rho \to 1} \mathbb{P}\{(1-\rho)R \leq x\}.$$

If the service time has a Pareto distribution, given by

$$1 - B(\tau) = \left(\frac{r-1}{r}\right)^r \tau^{-r}, \qquad \tau \geq \frac{r-1}{r}, \tag{6.8}$$

($B(\tau) = 0$ otherwise), then an explicit expression for $\mathbb{P}\{R_{HT} \leq x\}$ can be found if $r$ is integer-valued and a multi-term asymptotic expansion is available for $\mathbb{P}\{R_{HT} \leq x\}$ if $r$ is non-integer. A similar result holds if we consider finite mixtures of (6.8).
To show this, we exploit results of Abate & Whitt [2]. They define the class of Pareto Mixtures of Exponentials (PME) as follows. A distribution function $F$ is a PME if

$$1 - F(x) = \int_0^\infty \mathrm{e}^{-\frac{x}{\tau}} \mathrm{d}B(\tau), \qquad x \geq 0, \tag{6.9}$$

with $B(.)$ given by (6.8). From this definition and (6.2) we can conclude that the heavy-traffic limiting distribution is a PME if the service-time distribution is Pareto. We get, cf. [2],

$$\begin{aligned}
\mathbb{P}\{R_{HT} > x\} &= \int_0^\infty \mathrm{e}^{-\frac{x}{y}} \mathrm{d}B(y) \\
&= \int_{\frac{r-1}{r}}^\infty \mathrm{e}^{-\frac{x}{y}} r \left(\frac{r-1}{r}\right)^r y^{-r-1} \mathrm{d}y \\
&= r\left(\frac{r-1}{r}\right)^r \int_0^{\frac{r}{r-1}} \mathrm{e}^{-yx} y^{r-1} \mathrm{d}y.
\end{aligned}$$

This expression is (up to a multiplicative constant) equal to the incomplete Gamma function. Applications of well-known results for the incomplete Gamma function (see Abramovitz and Stegun [9], (4.2.55) and §6.5) give the following results. For integer $r \geq 2$ we have,

$$\mathbb{P}\{R_{HT} > x\} = \left(\frac{r-1}{r}\right)^r \frac{r!}{x^r} \left[1 - \mathrm{e}^{-\frac{rx}{r-1}} \sum_{k=0}^{r-1} \frac{1}{(r-1-k)!} \left(\frac{xr}{r-1}\right)^{r-1-k}\right]. \tag{6.10}$$

And, for non-integer $r > 1$:

$$\mathbb{P}\{R_{HT} > x\} =$$

$$\left(\frac{r-1}{r}\right)^r \frac{r}{x^r}\left[\Gamma(r) - \left(\frac{rx}{r-1}\right)^{r-1} \mathrm{e}^{-\frac{rx}{r-1}}\left[1 + \frac{r-1}{\frac{rx}{r-1}} + \frac{(r-1)(r-2)}{\left(\frac{rx}{r-1}\right)^2} + \cdots\right]\right]. \quad (6.11)$$

It is not difficult to obtain an explicit expression for $\mathbb{P}\{R_{HT} > x\}$ when the service-time distribution is a mixture of Pareto distributions as given in (6.8). In that case, the distribution of $R_{HT}$ is a mixture of PME's, which readily leads to an extension of (6.10) and (6.11).
Both (6.10) and (6.11) indicate that a one-term asymptotic expansion for the heavy-traffic limiting distribution will behave quite accurately since the residual terms decrease exponentially fast if $x \to \infty$ (cf. the observation in [2] p. 321). Another interesting observation is that the one-term expansion for $\mathbb{P}\{R_{HT} > x\}$ behaves like $\Gamma(r+1)\mathbb{P}\{B > x\}$, $x \to \infty$. In the next subsection, we will show that this property still holds if we only assume that the service-time distribution is regularly varying.

### 3.6.3 Tail behavior

In this subsection we study the behavior of $\mathbb{P}\{R_{HT} > x\}$ for $x$ large in the case that the service-time distribution is regularly varying. In particular, it will be shown that the heavy-traffic approximation

$$\mathbb{P}\{R > x\} \approx \mathbb{P}\{R_{HT} > (1-\rho)x\}$$

for the sojourn time *over*estimates the true sojourn-time distribution for large $x$.

**Theorem 3.6.3** *If* $1 - B(x) = x^{-\nu}L(x)$ *with* $\nu > 1$, *then*

$$\mathbb{P}\{R_{HT} > x\} \sim \Gamma(\nu+1)\mathbb{P}\{B > x\}.$$

**Proof**
Since $R_{HT} \stackrel{d}{=} YB$, with $Y$ exponentially distributed with mean 1 and $B$ the service time independent of $Y$, Theorem 6.3 immediately follows from Proposition 3 in [78], which is stated only for $0 < \nu < 1$, but can easily be extended to $\nu > 0$ (see also [113, 123, 232]). □

**Remark 3.6.2**
Using a result of Cline & Samorodnitsky [87], we can conclude that $R_{HT}$ is subexponential when $B$ is subexponential. However, it is not possible to give a general characterization of the tail behavior of $R_{HT}$, see [87] (and also Schmidli [253]) for a discussion.

**Remark 3.6.3**
It is possible to get more refined asymptotics for $\mathbb{P}\{R_{HT} > x\}$. Suppose $1 - B(x)$ is given by

$$1 - B(x) = \sum_{i=1}^{N} p_i x^{-\nu_i} + \mathrm{o}(x^{-\nu_N}), \qquad x \to \infty, \tag{6.12}$$

with $1 < \nu_1 < \cdots < \nu_N$, and $p_i > 0$. Applying (6.10), (6.11), and Theorem 3.6.3, we get

$$\mathbb{P}\{R_{HT} > x\} = \sum_{i=1}^{N} p_i \Gamma(\nu_i + 1) x^{-\nu_i} + \mathrm{o}(x^{-\nu_N}), \qquad x \to \infty. \tag{6.13}$$

**Remark 3.6.4**
By Theorems 3.6.3 and 3.4.1, we have the following interesting result:

$$\lim_{x\to\infty} \lim_{\rho\to 1} \frac{\mathbb{P}\{(1-\rho)R > x\}}{\mathbb{P}\{B > x\}} = \Gamma(\nu + 1) > 1 = \lim_{\rho\to 1} \lim_{x\to\infty} \frac{\mathbb{P}\{(1-\rho)R > x\}}{\mathbb{P}\{B > x\}}. \tag{6.14}$$

Hence, the heavy-traffic approximation for $\mathbb{P}\{R > x\}$ overestimates the true value when $x$ is large. This indicates that the approximations for the waiting-time distribution associated with Theorems 3.4.1 and 3.6.3 will behave differently.

## 3.7 Concluding remarks

In this chapter, we have investigated asymptotic properties of the sojourn time distribution in the $M/G/1$ PS queue. Our main results (Theorems 3.4.1 and 3.4.2) show that the sojourn-time distribution and the service-time distribution are equally heavy-tailed. More precisely, in the case of regular variation, we have $\mathbb{P}\{R_1 > x\} \sim \mathbb{P}\{B_1 > (1-\rho)x\}$. The 'if' part of Theorem 3.4.1 has recently been generalized by Núñez-Queija [215] (Theorem 5.2.3) to the class of intermediately regularly varying distributions. Chapter 5 of [215] also contains similar results for related service disciplines, such as foreground-background PS, shortest remaining processing time first, and PS with a varying service rate. The extension of Theorem 3.4.1 to the more general class of subexponential distributions remains a challenging open problem.
The results in this chapter confirm the viewpoint in Kleinrock [174] that PS is a 'fair' service discipline. PS offers protection for short jobs against the long ones, and for well-behaved customer classes against odd-behaved ones. Similar insights for the class of *Generalized Processor Sharing* queues have recently been obtained in a series of papers by Borst *et al.* [54, 55, 56, 57, 58].

# Chapter 4

# A fluid queue with a finite buffer

## 4.1 Introduction

Most of the early papers on fluid queues with heavy-tailed input assume an infinite buffer size, see e.g. Boxma [65, 66], Jelenković & Lazar [161], and Rolski *et al.* [243]. The results in [65, 66, 243] are based on a distributional equivalence between the stationary buffer content in the fluid queue and the waiting time in the $GI/G/1$ queue. A systematic treatment of this equivalence has been developed by Kella & Whitt [168], and is applied in Boxma & Dumas [67] to derive asymptotics for fluid queues from Theorem 2.2.1; see also the discussion accompanying Theorem 2.2.3.

Besides the above relation between queues with instantaneous and gradual input, there is sometimes also a relation between queues with finite and infinite buffer size. A classical example is the equivalence between the stationary waiting-time distributions of the $M/G/1$ queue with finite and infinite buffer size. These distributions are *proportional*, see Equation (3.6) below. This proportionality relation has already been known since Takács [262].

In this chapter, we analyze a fluid queue with finite buffer size $K$. We provide useful relations between this model and other models, like the fluid queue with infinite buffer size, and the single-server queue with finite capacity $K$ (the finite dam). In particular, we extend the results of Kella & Whitt [168] to the finite buffer case, and we prove that the stationary buffer-content distributions in the fluid queue with finite and infinite buffer size are proportional, see Theorem 4.5.2 below. These results are then applied to investigate the influence of heavy-tailed input characteristics on performance measures like the loss fraction and the mean buffer content.

The asymptotic expansions that are derived for these performance measures indicate that heavy-tailed input characteristics can have a significant influence on the performance of the fluid queue. In particular, loss fractions decay less than exponentially fast to zero when the buffer size gets large. This implies that very large buffers are needed to guarantee a small loss fraction, which differs from the case where Cramèr-type conditions are satisfied.

In the latter case, the loss fraction is known to behave negative exponentially as function of the buffer size. Another performance measure which is influenced by heavy-tailed input is the mean buffer content. When the activity periods of the On-Off sources have infinite second moments, the mean buffer content may behave like a (positive) power of the buffer size when the latter gets large. Complementing results have been obtained by Heath *et al.* [151, 153], Resnick & Samorodnitsky [236], who investigate the (asymptotic behavior of the) expected time to buffer overflow.

This chapter is organized as follows. In Section 4.2, we introduce the fluid model and indicate its relation to the finite dam with instantaneous input. We present some new results for the latter model in Sections 4.3 and 4.4. The main results for the fluid model can be found in Section 4.5. The results obtained in Sections 4.3–4.5 are applied in Section 4.6, where the fluid queue fed by a number of On-Off sources is discussed. Section 4.7 treats the case of overloaded queues. An alternative proof of Theorem 4.5.2 can be found in the Appendix.

## 4.2 Preliminaries

In this section we describe the dynamics of the fluid model introduced by Kella & Whitt [168], and extend this description to a fluid queue with a finite buffer. There are four elements governing the dynamics of the fluid model: Two collections of random variables $\{A_k : k \geq 1\}$ and $\{U_k : k \geq 1\}$, and two collections of stochastic processes $\{\{B_k(t) : t \geq 0\} : k \geq 1\}$, and $\{\{T_k(t) : t \geq 0\} : k \geq 1\}$, both classes having right-continuous sample paths with left limits. In the terminology of [168], $A_k$ and $U_k$ can be interpreted as successive down- and up-times respectively, a terminology motivated by queues with service interruptions.

Fluid in the buffer *increases* according to $\{B_k(t) : t \geq 0\}$ during the $k$-th *downtime* (of the server), and fluid in the buffer *decreases* by the stochastic process $\{T_k(t) : t \geq 0\}$ during the $k$-th *uptime*. Therefore we use a different terminology, which is motivated by fluid queues: We call $A_i$ an activity period (of a global fluid source) and $U_i$ a silence period. Define

$$\tau_k = A_1 + U_1 + \cdots + A_k + U_k, \qquad k \geq 1, \tag{2.1}$$

and $\tau_0 = 0$. The buffer content of the fluid queue with infinite buffer size at time $t$ is denoted by $V(t)$, and is given by, cf. [168],

$$V(\tau_{k+1}-) = \max\{V(\tau_k-) + B_{k+1}(A_{k+1}-) - T_{k+1}(U_{k+1}-), 0\}, \tag{2.2}$$

$$V(t) = V(\tau_k-) + B_{k+1}(t-\tau_k-), \qquad \tau_k \leq t < \tau_k + A_{k+1},$$

$$V(t) = \max\{V(\tau_k-) + B_{k+1}(A_{k+1}-) - T_{k+1}(t-\tau_k-A_{k+1}), 0\},$$

$$\tau_k + A_{k+1} \leq t < \tau_{k+1}. \tag{2.3}$$

In this chapter we assume that the *main independence assumption* stated in [168] holds, i.e. $\{(A_k, U_k, \{B_k(t) : t \geq 0\}, \{T_k(t) : t \geq 0\}) : k \geq 1\}$ is an i.i.d. sequence. Moreover, it is assumed that the moments $\mathbb{E}\{A_1\}$, $\mathbb{E}\{U_1\}$, $\mathbb{E}\{B_1(A_1-)\}$, and $\mathbb{E}\{T_1(U_1-)\}$ are finite. Then, under the condition $\mathbb{E}\{B_1(A_1-)\}/\mathbb{E}\{T_1(U_1-)\} < 1$, it is shown in [168] that $V(\tau_k-)$ converges in distribution to a random variable (tentatively denoted by) $W$ as $k \to \infty$. Moreover, when $A_1$, $U_1$, and $A_1 + U_1$ are non-lattice, the buffer-content process $V(t)$ converges in distribution to a random variable $V$.

It is obvious that $W$ corresponds to the waiting-time of the $G/G/1$ queue with service times $B_1(A_1-)$ and interarrival times $T_1(U_1-)$. One of the main contributions of Kella & Whitt [168] is to relate the distribution of $V$ to the stationary waiting-time distribution in a $G/G/1$ queue, see Theorems 4–6 in [168].

Next, we introduce the fluid model with finite buffer size $K > 0$. For each $K$, the buffer content $V^K(t)$ at time $t$ can be described by $V^K(0) = V^K(\tau_0) = 0$, and

$$V^K(\tau_{k+1}-) = \max\{\min\{V^K(\tau_k-) + B_{k+1}(A_{k+1}-), K\} - T_{k+1}(U_{k+1}-), 0\}, \tag{2.4}$$

$$V^K(t) = \min\{V^K(\tau_k-) + B_{k+1}(t - \tau_k-), K\}, \qquad \tau_k \leq t < \tau_k + A_{k+1},$$

$$V^K(t) = \max\{\min\{V^K(\tau_k-) + B_{k+1}(A_{k+1}-), K\} - T_{k+1}(t - \tau_k - A_{k+1}), 0\},$$

$$\tau_k + A_{k+1} \leq t < \tau_{k+1}. \tag{2.5}$$

The dynamics of the finite-buffer model are the same as those of the infinite-buffer model, except that when the buffer content reaches level $K$, all the excess amount of fluid offered to the buffer during the remaining activity period will be lost.

It is easily shown that $V^K(\tau_{k+1}-)$ can be identified with the waiting time of the $(k+1)$-st customer in the $G/G/1$ queue with finite capacity $K$, in which the interarrival times and service times are distributed as $T_1(U_1-)$ and $B_1(A_1-)$. Under the condition $\mathbb{P}\{T_1(U_1-) = B_1(A_1-)\} < 1$, it is shown in Section III.5.3 of [97] that $V^K(\tau_{k+1}-)$ converges in distribution to a limiting random variable $W^K$ as $k \to \infty$.

We wish to extend the results of [168] to finite buffer queues; we will establish a relationship between the stationary distribution of the finite buffer model and the stationary distribution of the $G/G/1$ queue with a buffer having finite capacity $K$. The latter model is also known as the finite dam, see Chapter III.5 in [97].

Define the environment indicator process by

$$I(t) = I_{\{\tau_k \leq t < \tau_k + A_{k+1} \text{ for some } k \geq 1\}},$$

so $I(t) = 1$ if the global fluid source is active at time $t$. The amount of time the global fluid source is active (resp. silent) up to time $t$, $t \geq 0$, is defined by

$$C_a(t) = \int_0^t I(x)\mathrm{d}x, \tag{2.6}$$

$$C_s(t) = t - C_a(t). \tag{2.7}$$

The inverse processes of $C_a$ and $C_s$ are defined by

$$C_a^{-1}(t) = \inf_{x\geq 0}\{C_a(x) > t\}, \tag{2.8}$$

$$C_s^{-1}(t) = \inf_{x\geq 0}\{C_s(x) > t\}. \tag{2.9}$$

Since the random variables $A_i$ and $U_i$ are finite a.s., we may assume that $C_s(t) \to \infty$ if $t \to \infty$ everywhere. So the following processes are well-defined,

$$V_a^K(t) = V^K(C_a^{-1}(t)), \qquad t \geq 0, \tag{2.10}$$

$$V_s^K(t) = V^K(C_s^{-1}(t)), \qquad t \geq 0. \tag{2.11}$$

Note that $V_s^K(U_1) = V^K(A_1+U_1+A_2)$. We similarly define $V_a(t)$ and $V_s(t)$ for the infinite buffer model. Cf. [168], we define the r.v. $B_1(A_1^r)$ (which is non-trivial since $B_1$ and $A_1$ are dependent in general) by

$$\begin{aligned}\mathbb{P}\{B_1(A_1^r) > x\} &= \frac{1}{\mathbb{E}\{A_1\}}\mathbb{E}\{\int_0^{A_1} 1_{\{B_1(t)>x\}}\mathrm{d}t\} \\ &= \int_0^\infty \mathbb{P}\{B_1(t) > x \mid A_1 > t\}\mathrm{d}\mathbb{P}\{A_1^r \leq t\}.\end{aligned} \tag{2.12}$$

We are now ready to give the main result of this section, which can be viewed as an extension of Theorem 4 in [168]. Note that no assumptions on the traffic load are needed, since the state space is bounded.

**Theorem 4.2.1** *Suppose that the main independence assumption holds, that $A_1$, $U_1$, and $A_1 + U_1$ are non-lattice, and that $\mathbb{P}\{T_1(U_1-) = B_1(A_1-)\} < 1$. Then there exist r.v.'s $V_s^K$, $V_a^K$, $V^K$, and $I$ such that, when $t \to \infty$,*

1. $V_a^K(t) \Rightarrow V_a^K \stackrel{d}{=} \min\{W^K + B_1(A_1^r), K\}$,
2. $V_s^K(t) \Rightarrow V_s^K$,
3. $[V^K(t), I(t)] \Rightarrow [V^K, I]$.

*Here $B_1(A_1^r)$ is independent of $W^K$, $V_s^K \stackrel{d}{=} (V^K \mid I = 0)$, $V_a^K \stackrel{d}{=} (V^K \mid I = 1)$, and*

$$\mathbb{P}\{V^K > x\} = (1-p)\mathbb{P}\{V_s^K > x\} + p\mathbb{P}\{V_a^K > x\}, \tag{2.13}$$

*where*

$$p = \mathbb{P}\{I = 1\} = \frac{\mathbb{E}\{A_1\}}{\mathbb{E}\{A_1\} + \mathbb{E}\{U_1\}}. \tag{2.14}$$

**Proof**
The proof is almost identical to the proof of Theorem 4 in [168]. The processes $[V^K(t), I(t)]$, $V_a^K(t)$ and $V_s^K(t)$ are all regenerative with the exit times of state $[0, 0]$, resp. 0 (i.e. the end of idle periods) as regeneration points. The regeneration cycles are non-lattice when $U_1$, $A_1$, and $U_1 + A_1$ are non-lattice, due to the main independence assumption. Since all state spaces are finite, it is trivially seen that all regeneration cycles have finite means. The convergence of the processes $[V^K(t), I(t)]$, $V_a^K(t)$ and $V_s^K(t)$ now follows by the results on pp. 125–127 of [19].
By a result of Green [143], we can study the sequences of activity periods and silence periods separately. This gives the relationship between the limiting distributions of $[V^K, I]$, $V_a^K$, and $V_s^K$, and the characterization of the distribution of $V_a^K$. □

**Remark 4.2.1**
The non-lattice conditions can be omitted if $U_1$ is exponentially distributed and independent of $A_1$ and $\{B_1(t)\}$. In Theorem 4.2.1, the condition on $U_1 + A_1$ is imposed since $U_1$ and $A_1$ are allowed to be dependent.

If the outflow from the buffer is constant during silence periods, then it is also possible to specify the limiting distribution $V_s^K$.

**Theorem 4.2.2** *Suppose the assumptions stated in Theorem 4.2.1 hold and that*

$$T_1(t) \equiv t.$$

*Then* $\{V_s^K(t), t \geq 0\}$ *is distributed as the workload process in the finite dam with capacity* $K$*, interarrival times* $U_1$*, and service times* $B_1(A_1-)$*.*

**Proof**
Similar to the proof of Theorem 2 in [168]. Both processes have reflecting barriers in the origin and $K$, decrease linearly at rate 1, and have jumps of size $B_{k+1}(A_{k+1}-)$ at times $U_1 + \cdots + U_k$. □

One can apply Theorems 4.2.1 and 4.2.2 to compute (characteristics of) the distribution of $V^K$ when the steady-state distribution for the $G/G/1$ finite dam is sufficiently tractable, which is the case for the $M/G/1$ and $G/M/1$ finite dams, see [97, 209]. In Section 4.5, we further specify the distribution of $V^K$ by using Theorems 4.2.1 and 4.2.2. Both theorems indicate a clear relationship between the fluid model with gradual input and the $G/G/1$ finite dam with instantaneous input; we will study the latter model in the next two sections.

In the remainder of the chapter, we assume that the buffer content declines linearly during silence periods, i.e., we assume that $T_1(t) \equiv t$. In this case, the fluid model can process one unit of fluid per unit of time. The amount of fluid offered to the system per unit of time, given by $\rho$, equals (with $\mathbb{E}\{U_1\} = 1/\lambda$)

$$\rho = \frac{\mathbb{E}\{B_1(A_1-)\} + \mathbb{E}\{A_1\}}{\lambda^{-1} + \mathbb{E}\{A_1\}}.$$

## 4.3 The stationary distribution of the finite dam

The distribution of the random variable $W^K$ in the previous section corresponds to the stationary waiting-time distribution in the finite dam having capacity $K$. The relation between the models with gradual and instantaneous input will turn out to be useful in the rest of the chapter. In this section, we give some new results for the finite dam. In particular, we give a relationship between the virtual and actual waiting time which is very similar to the relationship in the infinite-buffer case. The latter is well-known, see the references below.

First, we introduce some notation in the traditional queueing setting. Customers arrive at a single-server queue (which is initially empty) with interarrival times $T_n, n \geq 1$. These customers have service times $B_n, n \geq 1$. It is assumed that the interarrival times and service times are all independent of each other and have the same distributions as random variables $T$ and $B$, respectively. The means of $T$ and $B$ are denoted by $\lambda^{-1}$ and $\beta$, respectively. The distribution function of the service time is denoted by $B(.)$. The traffic load $\hat{\rho}$ is given by $\hat{\rho} := \lambda\beta$ and is assumed to be strictly positive.

The waiting time of the $n$-th customer is given by $W_n^K$. When $W_n^K + B_n$ exceeds $K$, a quantity of $W_n^K + B_n - K$ is lost (so we consider partial overflow). Hence, $W_n^K$ is given by $W_0^K = 0$ and (see e.g. Chapter III.5 in [97]),

$$W_{n+1}^K = \max\{\min\{W_n^K + B_n, K\} - A_{n+1}, 0\}. \tag{3.1}$$

Denote the stationary waiting-time by $W^K$ (cf. Section 4.2, with $B_n \equiv B_n(A_n-)$ and $T_n \equiv U_n$). We also consider the amount of work present in the system at time $t$, given by $V_q^K(t)$; its stationary distribution is denoted by $V_q^K$. Finally, the long-run fraction of work lost is defined by $L_{q,K}$.

### 4.3.1 General results

The loss fraction $L_{q,K}$ can be obtained by a simple renewal argument:

$$L_{q,K} = \frac{\mathbb{E}\{\max\{W^K + B - K, 0\}\}}{\mathbb{E}\{B\}} = \mathbb{P}\{W^K + B^r > K\}. \tag{3.2}$$

The second equality, which is quite useful for further analysis as is shown below, can be obtained by partial integration. For the virtual waiting time $V_q$ and the actual waiting

time $W$ in the $GI/G/1$ queue (with $\hat{\rho} < 1$), it is well-known that (see e.g. Asmussen [19] p. 189, and Cohen [93], [97] p. 296)

$$V_q \mid V_q > 0 \stackrel{d}{=} W + B^r. \tag{3.3}$$

The following result is very similar to (3.3) and appears to be new.

**Theorem 4.3.1** *For all $\hat{\rho} > 0$ and $0 < K < \infty$,*

$$V_q^K \mid V_q^K > 0 \stackrel{d}{=} (W^K + B^r) \mid W^K + B^r \leq K, \tag{3.4}$$

$$\mathbb{P}\{V_q^K > x\} = \hat{\rho}\, \mathbb{P}\{x < W^K + B^r \leq K\}. \tag{3.5}$$

**Proof**

The results can be obtained in a similar way as for the infinite-buffer queue, namely by a level-crossing argument, see [93]. Following the same lines as in [93], we obtain for almost every $0 < v < K$,

$$\frac{\mathrm{d}}{\mathrm{d}v}\mathbb{P}\{V^K < v\} = \hat{\rho} \int_0^v \frac{1 - \mathbb{P}\{B < v - u\}}{\beta} \mathrm{d}\mathbb{P}\{W^K < u\}.$$

Hence, for $0 < x < K$,

$$\begin{aligned} \mathbb{P}\{V_q^K < x\} &= \mathbb{P}\{V^K = 0\} + \hat{\rho} \int_0^x \int_0^{x-u} \frac{\mathbb{P}\{B > w\}}{\beta} \mathrm{d}w \mathrm{d}\mathbb{P}\{W^K < u\} \\ &= \mathbb{P}\{V_q^K = 0\} + \hat{\rho}\mathbb{P}\{W^K + B^r < x\}. \end{aligned}$$

By Little's law for a busy server (see e.g. Example 4.3 in Whitt [273]) and (3.2), we have

$$\mathbb{P}\{V_q^K = 0\} = 1 - \hat{\rho}(1 - L_{q,K}) = 1 - \hat{\rho}\mathbb{P}\{W^K + B^r \leq K\}.$$

Hence, for $0 < x < K$,

$$\begin{aligned} \mathbb{P}\{V_q^K < x\} &= 1 - \hat{\rho}(\mathbb{P}\{W^K + B^r \leq K\} - \mathbb{P}\{W^K + B^r < x\}) \\ &= 1 - \hat{\rho}\mathbb{P}\{x \leq W^K + B^r \leq K\}. \end{aligned}$$

This expression is also valid for $x \downarrow 0$ and $x = K$, which yields (3.5), since $B^r$ has a continuous distribution. It is easily shown from (3.5) that

$$\mathbb{P}\{0 < V_q^K \leq x\} = \hat{\rho}\mathbb{P}\{W^K + B^r \leq x\}.$$

Hence,

$$\begin{aligned} \mathbb{P}\{V_q^K \leq x \mid V_q^K > 0\} &= \frac{\mathbb{P}\{0 < V_q^K \leq x\}}{\mathbb{P}\{V_q^K > 0\}} = \frac{\hat{\rho}\mathbb{P}\{W^K + B^r \leq x\}}{\hat{\rho}\mathbb{P}\{W^K + B^r \leq K\}} \\ &= \mathbb{P}\{W^K + B^r \leq x \mid W^K + B^r \leq K\}. \end{aligned}$$

This proves (3.4).

□

### 4.3.2 Exponentially distributed interarrival times

If the interarrival times are exponentially distributed, then the following *proportionality relation* holds, see Takács [262], Cohen [92, 97], Hooghiemstra [156], and many others:

$$\mathbb{P}\{W^K \leq x\} = \frac{\mathbb{P}\{W \leq x\}}{\mathbb{P}\{W \leq K\}}, \tag{3.6}$$

for $0 \leq x \leq K$. Proportionality relations like (3.6) have been applied in a number of studies to determine loss probabilities, see e.g. Daley [110], Stanford [257], Gouweleeuw [140, 141], Boots & Tijms [45] and references therein. The main idea applied in these studies is to combine the proportionality result with Little's formula for a busy server (see e.g. Example 4.3 in Whitt [273]). Applying the latter together with PASTA to the finite- and infinite-buffer queue, we obtain for $0 < \hat{\rho} < 1$ and $\hat{\rho} > 0$ respectively,

$$\mathbb{P}\{W = 0\} = 1 - \hat{\rho}, \tag{3.7}$$

$$\mathbb{P}\{W^K = 0\} = 1 - \hat{\rho}(1 - L_{q,K}). \tag{3.8}$$

Using the proportionality relation

$$\frac{\mathbb{P}\{W^K = 0\}}{\mathbb{P}\{W = 0\}} = \frac{\mathbb{P}\{W^K \leq K\}}{\mathbb{P}\{W \leq K\}} = \frac{1}{\mathbb{P}\{W \leq K\}},$$

we obtain from (3.7) and (3.8), for $0 < \hat{\rho} < 1$,

$$L_{q,K} = \frac{1 - \hat{\rho}}{\hat{\rho}} \left( \frac{1}{\mathbb{P}\{W \leq K\}} - 1 \right) = \frac{1 - \hat{\rho}}{\hat{\rho}} \frac{\mathbb{P}\{W > K\}}{\mathbb{P}\{W \leq K\}}. \tag{3.9}$$

**Remark 4.3.1**
By PASTA, we have that $V_q^K \stackrel{d}{=} W^K$. Using this and the proportionality relation, it is also possible to derive (3.9) from (3.2) and (3.5).

**Remark 4.3.2**
Another performance measure is the probability that the work offered by a customer (entering the system in its stationary regime) cannot be completely accepted; denote this probability by $P_{q,K}$. For the $GI/G/1$ finite dam, we have

$$P_{q,K} = \mathbb{P}\{W^K + B > K\}. \tag{3.10}$$

(Note that $P_{q,K} = L_{q,K}$ in the $GI/M/1$ finite dam.) When $\hat{\rho} < 1$, we have the following remarkable relation for the $M/G/1$ finite dam. It follows from the proportionality relation that

$$P_{q,K} = \frac{\mathbb{P}\{W + B > K\} - \mathbb{P}\{W > K\}}{\mathbb{P}\{W \leq K\}}. \tag{3.11}$$

But this quantity can be identified with $\mathbb{P}\{C_{\max} > K\}$, where $C_{\max}$ is the maximal content in the infinite dam during a busy cycle, see e.g. Section 3.3 in [92] or [97], p. 297, and p. 618, so we conclude that

$$P_{q,K} = \mathbb{P}\{C_{\max} > K\} = \frac{1}{\lambda} \frac{\frac{\mathrm{d}}{\mathrm{d}K}\mathbb{P}\{W \leq K\}}{\mathbb{P}\{W \leq K\}}. \tag{3.12}$$

## 4.4 Asymptotic results for the finite dam

For the case $\hat{\rho} < 1$, we are interested in the asymptotic behavior of $L_{q,K}$ when $K \to \infty$, in particular when the service-time distribution is subexponential. In the case of exponentially distributed silence periods, it is possible to apply Theorem 2.2.1 for the single-server queue with infinite buffer size.

**Theorem 4.4.1** *If $\hat{\rho} < 1$, $B^r$ is subexponential, and if the interarrival times are exponentially distributed, then*

$$L_{q,K} \sim \mathbb{P}\{B^r > K\}, \qquad K \to \infty. \tag{4.1}$$

**Proof**
Using Theorem 2.2.1 we have, if $B^r$ is subexponential,

$$\mathbb{P}\{W > x\} \sim \frac{\hat{\rho}}{1-\hat{\rho}}\mathbb{P}\{B^r > x\}, \qquad x \to \infty. \tag{4.2}$$

Theorem 4.4.1 now follows directly from (3.9) and (4.2). □

When the interarrival times have a general distribution, the proportionality relation does not hold, so it is not possible to apply results for the infinite dam directly. However, it is still possible to extend Theorem 4.4.1 to the case of generally distributed interarrival times. This is established in the following theorem, under the additional assumption that the service-time distribution is regularly varying.

**Theorem 4.4.2** *For generally distributed interarrival times and $\hat{\rho} < 1$, (4.1) holds if the service-time distribution is regularly varying of index $-\nu$, $\nu > 1$.*

**Proof**
Note that $L_{q,K} \geq \mathbb{P}\{B^r > K\}$, so it suffices to show that

$$\limsup_{K\to\infty} \frac{\mathbb{P}\{W^K + B^r > K\}}{\mathbb{P}\{B^r > K\}} \leq 1. \tag{4.3}$$

Let $\phi(K)$ be a function such that $\phi(K) \to \infty$ and $\phi(K)/K \to 0$ if $K \to \infty$, and let $\epsilon > 0$. Write

$$\mathbb{P}\{W^K + B^r > K\} = P_{1,K} + P_{2,K} + P_{3,K}, \tag{4.4}$$

with

$$P_{1,K} = \mathbb{P}\{W^K + B^r > K; W^K \leq \epsilon K\}, \tag{4.5}$$
$$P_{2,K} = \mathbb{P}\{W^K + B^r > K; \epsilon K < W^K \leq K - \phi(K)\}, \tag{4.6}$$
$$P_{3,K} = \mathbb{P}\{W^K + B^r > K; W^K > K - \phi(K)\}. \tag{4.7}$$

Since $P_{1,K} \leq \mathbb{P}\{B^r > (1-\epsilon)K\}$ and since $B^r$ is regularly varying of index $1-\nu$, we have

$$\limsup_{K\to\infty} \frac{P_{1,K}}{\mathbb{P}\{B^r > K\}} \leq \left(\frac{1}{1-\epsilon}\right)^{\nu-1}, \qquad \forall \epsilon > 0. \tag{4.8}$$

We can bound $P_{2,K}$ using that $W^K$ is stochastically dominated by $W$:

$$P_{2,K} \leq \mathbb{P}\{B^r \geq \phi(K)\}\mathbb{P}\{W^K > \epsilon K\} \leq \mathbb{P}\{B^r \geq \phi(K)\}\mathbb{P}\{W > \epsilon K\}.$$

Using (4.2) for the $GI/G/1$ queue and the fact that $B^r$ is regularly varying we obtain for each $\epsilon > 0$,

$$\lim_{K\to\infty} \frac{\mathbb{P}\{W > \epsilon K\}}{\mathbb{P}\{B^r > K\}} = \frac{\hat{\rho}}{1-\hat{\rho}}\epsilon^{1-\nu},$$

which implies, since $\phi(K) \to \infty$ if $K \to \infty$,

$$\limsup_{K\to\infty} \frac{P_{2,K}}{\mathbb{P}\{B^r > K\}} = 0, \qquad \forall \epsilon > 0. \tag{4.9}$$

Finally, we deal with the last term. Note that

$$P_{3,K} \leq \mathbb{P}\{W^K \geq K - \phi(K)\}. \tag{4.10}$$

We make some additional definitions. Define the random walk $(S_n)_{n\geq 0}$ by $S_0 = 0$, and for $n \geq 1$,

$$S_n = \sum_{i=1}^{n} (B_i - T_i). \tag{4.11}$$

Note that this random walk has negative drift $\beta - \lambda^{-1}$. We also define the sequence of random variables $(\bar{W}_n^K)_{n\geq 0}$ by $\bar{W}_0^K = 0$, and $\bar{W}_{n+1}^K = \min\{\max\{\bar{W}_n^K + B_n - T_{n+1}, 0\}, K\}$. Denote the stationary solution of this recursion by $\bar{W}^K$. From the construction of both $W^K$ and $\bar{W}^K$ it is clear that $\mathbb{P}\{W^K \geq x\} \leq \mathbb{P}\{\bar{W}^K \geq x\}$, $0 \leq x \leq K$. Hence,

$$P_{3,K} \leq \mathbb{P}\{\bar{W}^K > K - \phi(K)\}. \tag{4.12}$$

We now use a representation of the distribution of $\bar{W}^K$ in terms of an absorption probability of the random walk $(S_n)$, which seems to be due to Lindley [186], see also Loynes [190]. Define the stopping times

$$\tau(K) = \inf\{n : S_n \geq K - \phi(K)\}, \qquad \tau'(K) = \inf\{n : S_n \leq -\phi(K)\}.$$

Then, from [186, 190],

$$\mathbb{P}\{\bar{W}^K > K - \phi(K)\} = \mathbb{P}\{\tau(K) < \tau'(K)\}. \tag{4.13}$$

Rewriting this yields

$$\mathbb{P}\{\bar{W}^K > K - \phi(K)\} = \mathbb{P}\{S_1, ..., S_{\tau(K)-1} > -\phi(K) \mid \tau(K) < \infty\}\mathbb{P}\{\tau(K) < \infty\}.$$

Since $\sup_n S_n$ can be identified with $W$, and $\tau(K) < \infty$ iff $\sup_n S_n > K - \phi(K)$, this equals

$$\mathbb{P}\{S_1, ..., S_{\tau(K)-1} > -\phi(K) \mid \tau(K) < \infty\}\mathbb{P}\{W > K - \phi(K)\}.$$

Using $\phi(K)/K \to 0$, we have by (4.2) that

$$\mathbb{P}\{W > K - \phi(K)\}/\mathbb{P}\{B^r > K\} \to \frac{\hat{\rho}}{1-\hat{\rho}}, \qquad K \to \infty.$$

Thus, we can conclude that $P_{3,K} = \mathrm{o}(\mathbb{P}\{B^r > K\})$ if

$$\mathbb{P}\{S_{\tau(K)-1} > -\phi(K) \mid \tau(K) < \infty\} \to 0, \qquad K \to \infty. \tag{4.14}$$

For this we use a theorem of Asmussen & Klüppelberg, see Theorem 1.1 in [21]. This result provides the following conditional limit theorem for $S_{\tau(K)-1}$ (which is the last value of $S_n$ before making a jump to level $K-\phi(K)$). Define $a(u) = \int_u^\infty (1-B(z))\mathrm{d}z/(1-B(u))$. Then,

$$\lim_{K\to\infty} \mathbb{P}\{-S_{\tau(K)-1}/a(K) > x \mid \tau(K) < \infty\} = (1 + x/(\nu-1))^{1-\nu}, \qquad x \geq 0. \tag{4.15}$$

Note that $a(K) \sim K/(\nu-1)$ if $K \to \infty$, so $\phi(K)/a(K) \to 0$ if $K \to \infty$. Hence,

$$\mathbb{P}\{S_{\tau(K)-1} > -\phi(K) \mid \tau(K) < \infty\} = \mathbb{P}\{-S_{\tau(K)-1}/a(K) < \phi(K)/a(K) \mid \tau(K) < \infty\} \to 0.$$

This proves (4.14). Hence, we have for each $\epsilon > 0$ that

$$\limsup_{K\to\infty} L_{q,K}/\mathbb{P}\{B^r > K\} \leq \left(\frac{1}{1-\epsilon}\right)^{\nu-1}, \tag{4.16}$$

which implies (4.3) by letting $\epsilon \to 0$. □

In Section 4.6, we apply the above results to obtain asymptotics for the loss fraction and mean buffer-content in the fluid queue.

## 4.5 The stationary distribution of the fluid queue

In this section, we study the distribution of the steady-state buffer-content $V^K$ in the fluid queue. Under certain assumptions, we express the distribution of $V^K$ completely in terms of $W^K$, thereby extending the results in [168] to the finite-buffer case. In a special case, it is also possible to express the distribution of $V^K$ in terms of $V$, by showing that the two probability measures are proportional.

**Theorem 4.5.1** *For $\rho > 0$ and $0 \leq x < K$,*

$$\mathbb{P}\{V^K > x\} = p\mathbb{P}\{W^K + B_1(A_1^r) > x\} + (1-p)\hat{\rho}\mathbb{P}\{K \geq W^K + B_1^r(A_1) > x\}. \quad (5.1)$$

*In particular, if the silence periods are exponentially distributed, then*

$$\mathbb{P}\{V^K > x\} = p\mathbb{P}\{W^K + B_1(A_1^r) > x\} + (1-p)\mathbb{P}\{W^K > x\}, \quad (5.2)$$

*with $p$ given by (2.14).*

**Proof**
In view of Theorem 4.2.1, we only need to specify the distribution of $V_s^K$. By Theorem 4.2.2, we have $V_s^K \stackrel{d}{=} V_q^K$. The first part of the theorem now follows from Theorem 4.3.1 and the second part can be obtained using PASTA. □

In the case that $U_1$ has an exponential distribution and $\rho < 1$, one can establish the following relation between the distributions of $V^K$, $V$, and $W$. Recall that $W$ can be identified with the waiting-time distribution of the $GI/G/1$ queue with service time $B_1(A_1-)$ and interarrival time $U_1$.

**Theorem 4.5.2** *If $U_1$ is exponentially distributed and if $\rho < 1$, then, for $0 \leq x < K$,*

$$\mathbb{P}\{V^K \leq x\} = \frac{\mathbb{P}\{V \leq x\}}{\mathbb{P}\{W \leq K\}}. \quad (5.3)$$

**Proof**
Use the second part of the previous theorem, the proportionality relation (3.6), and

$$\mathbb{P}\{V \leq x\} = p\mathbb{P}\{W + B_1(A_1^r) \leq x\} + (1-p)\mathbb{P}\{W \leq x\}, \quad (5.4)$$

cf. [168]. □

Note that (5.3) is not valid for $x = K$ and note the appearance of the term $\mathbb{P}\{W \leq K\}$ (and not $\mathbb{P}\{V \leq K\}$) in (5.3). An implication of this is that the probability that the buffer

is full ($\mathbb{P}\{V^K = K\}$) is strictly positive. This is not the case when input is instantaneous, cf. (3.6).
In the proof of Theorem 4.5.2, we used the relation between the models with gradual and instantaneous input (Theorem 4.5.1 and (5.4)), and the proportionality relation (3.6) between the two models with instantaneous input. It is also possible to prove Theorem 4.5.2 directly (without using connections with models with instantaneous input) by a regenerative argument, for which we refer to the appendix.
In the remainder of this section, we will study two important performance measures: The long-run fraction of fluid lost, denoted by $L_K$, and the mean buffer content.

**Theorem 4.5.3** *For all $\rho > 0$,*

$$L_K = \frac{\mathbb{E}\{B_1(A_1-)\}}{\mathbb{E}\{A_1\} + \mathbb{E}\{B_1(A_1-)\}} L_{q,K}, \tag{5.5}$$

*where $L_{q,K} = \mathbb{P}\{W^K + B_1^r(A_1-) > K\}$.*

**Proof**
We can establish a relation between the fluid model and the finite dam in the following manner. Suppose that both models are fed by the same input process. The amount of fluid lost during the $k$-th activity (and silence) period in the fluid model is identical to the work lost of the $k$-th customer in the finite buffer queue. However, the amount of fluid offered during the $k$-th activity period is $B_k(A_k-) + A_k$, whereas the amount of work offered by the $k$-th customer equals $B_k(A_k-)$. The result now follows by the renewal reward theorem, see e.g. Tijms [266]. □

Finally, we investigate the mean buffer content $\mathbb{E}\{V^K\}$. We restrict ourself to the case $\rho < 1$ and exponentially distributed silence periods.

**Theorem 4.5.4** *Under the conditions of Theorem 4.5.2,*

$$\mathbb{E}\{V^K\} = \frac{1}{\mathbb{P}\{W \le K\}} \int_0^K \mathbb{P}\{V > x\}\mathrm{d}x - \frac{K\mathbb{P}\{W > K\}}{\mathbb{P}\{W \le K\}}.$$

**Proof**
Use the representation $\mathbb{E}\{V^K\} = \int_0^{K^-} \mathbb{P}\{V^K > x\}\mathrm{d}x$, and the identity

$$\mathbb{P}\{V^K > x\} = \frac{\mathbb{P}\{V > x\} - \mathbb{P}\{W > K\}}{\mathbb{P}\{W \le K\}},$$

which follows easily from Theorem 4.5.2. □

**Remark 4.5.1**
Using the proportionality relations (3.6) and (5.3), it is possible to formulate heavy traffic limit theorems for $W^K$ and $V^K$, based on heavy-traffic limits for the $M/G/1$ queue.
Suppose that silence periods are exponentially distributed, that $\rho < 1$ (hence $\hat{\rho} < 1$), and that a function $\Delta(\hat{\rho})$ exists such that $\Delta(\hat{\rho})W$ converges in distribution to a random variable $W_{HT}$ if $\hat{\rho} \to 1$. This holds quite generally, see Section 1.4.2 for references.
Under these assumptions, we can formulate a heavy-traffic limit for $W^K$ by letting $\hat{\rho} \to 1$ and $K \to \infty$ such that $K\Delta(\hat{\rho}) = c$ for some constant $c$. Using (3.6), it is not difficult to see that, if $\hat{\rho} \to 1$ and $\Delta(\hat{\rho})K \equiv c$,

$$\mathbb{P}\{\Delta(\hat{\rho})W^K \leq x\} \to \frac{\mathbb{P}\{W_{HT} \leq x\}}{\mathbb{P}\{W_{HT} \leq c\}}, \tag{5.6}$$

for $0 \leq x \leq c$. By (5.4), $\Delta(\hat{\rho})V$ converges to the same heavy-traffic limit as $\Delta(\hat{\rho})W$. Hence, $\Delta(\hat{\rho})V^K$ has the same heavy-traffic limit as $\Delta(\hat{\rho})W^K$ using (5.2) or (5.3).

For a similar result for the $G/G/1$ queue with uniformly bounded actual waiting time (Chapter III.4 in [97]), see Kennedy [171] and references therein. More results can be found in the monograph of Whitt [275].

## 4.6 Asymptotic results for the fluid queue

In this section, we apply the results derived in the previous sections to obtain asymptotic expansions for various performance measures, in particular the loss fraction and the mean buffer content. We concentrate on the case where $A_1$ and $B_1(A_1-)$ have a subexponential tail.
The general case will be treated in Subsection 4.6.1. In Subsection 4.6.2 we study the simplest possible fluid model, namely the case of a single On-Off source. The last Subsection 4.6.3 treats the case of multiple On-Off sources.

### 4.6.1 General input

We start with the case of general input, where we assume that $B_1(A_1-)$ has a subexponential distribution. The following result follows immediately from Theorem 4.5.3 and the results in Section 4.4.

**Theorem 4.6.1** *Under the conditions of Theorem 4.4.1 or 4.4.2,*

$$L_K \sim \frac{\mathbb{E}\{B_1(A_1-)\}}{\mathbb{E}\{A_1\} + \mathbb{E}\{B_1(A_1-)\}} \mathbb{P}\{B_1^r(A_1-) > K\}, \qquad K \to \infty. \tag{6.1}$$

Asymptotics for the mean buffer content are more difficult to obtain in general. Such a result would involve the tail behavior of $B_1(A_1^r)$ (cf. Theorem 3.15 in [66] and (2.12)), for which no results are available.

### 4.6.2 A simple On-Off source

Suppose that the fluid queue is fed by a single On-Off source. When the source is active, it sends input at rate $r > 1$ during a period of $A_1$. Off-periods are exponentially distributed with parameter $\lambda$. In terms of the model in the previous sections, this implies that $B_1(t) \equiv (r-1)t$. In the terminology of [168], this is the linear fluid model with random disruptions, with the additional assumption that the idle periods are exponentially distributed.
We first derive the asymptotics for the loss fraction.

**Proposition 4.6.1** *If the distribution of $A_1^r$ is subexponential and if the Off-periods are exponentially distributed, then, for $\rho < 1$ and $K \to \infty$,*

$$L_K \sim \frac{r-1}{r}\mathbb{P}\{A_1^r > \frac{K}{r-1}\}. \tag{6.2}$$

*If the off-periods are generally distributed and $\mathbb{P}\{A_1 > x\} = L(x)x^{-\nu}$, $\nu > 1$, then*

$$L_K \sim \frac{(r-1)^\nu}{r(\nu-1)\mathbb{E}\{A_1\}}L(K)K^{1-\nu}, \qquad K \to \infty. \tag{6.3}$$

**Proof**
Equation (6.2) follows immediately from Theorem 4.6.1 (or alternatively, use Theorem 4.5.3, (3.9), and Theorem 2.2.1). Equation (6.3) follows from Theorem 4.6.1, Theorem 2.2.3, and Karamata's theorem (Lemma 2.1.7). □

**Remark 4.6.1**
Awater ([32], p. 131) has suggested the following approximation for the fraction of fluid lost,

$$L_{K,app} = \frac{(1-\rho)\mathbb{P}\{V > K\}}{1-\rho\mathbb{P}\{V > K\}}.$$

Numerical experiments in [32] that $L_{K,app}$ can be a good approximation for $L_K$. Variants of $L_{K,app}$ have been shown to be exact in various other cases like the loss probability of a customer in the $M^X/G/1/B$ queue (see [140]) and the $M/M/c$ queue with impatient customers (see [45]).
If we evaluate the performance of $L_{K,app}$ in the simplest possible case $B_1(t) \equiv (r-1)t$, then it is easily shown from Proposition 4.6.1 and (6.4) that the asymptotic behavior of $L_{K,app}$ is not entirely correct: Under the conditions of Proposition 4.6.1, $L_K/L_{K,app}$ converges to a constant which is positive and finite, but not equal to one. The same conclusion can be drawn if the activity periods are exponentially distributed.

We now turn to the mean buffer content, where we restrict ourself to the (important) special case of activity periods with infinite second moments (corresponding to long-range dependent input, see Chapter 1). It is also assumed that the silence periods are exponentially distributed.

**Proposition 4.6.2** *If* $\mathbb{P}\{A_1 > x\} = L(x)x^{-\nu}$, $1 < \nu < 2$ *and if the conditions in Theorem 4.5.2 hold, then*

$$\mathbb{E}\{V^K\} \sim \frac{\rho}{1-\rho}\frac{(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}\{A_1\}}\left[\frac{1-p}{2-\nu} - \frac{r-1}{r}\right]L(K)K^{2-\nu}, \tag{6.4}$$

*if* $K \to \infty$.

**Proof**
We will obtain an asymptotic expansion for both terms in the formula for $\mathbb{E}\{V^K\}$ given in Theorem 4.5.4:

$$\mathbb{E}\{V^K\} = \frac{1}{\mathbb{P}\{W \le K\}}\int_0^K \mathbb{P}\{V > x\}\mathrm{d}x - \frac{K\mathbb{P}\{W > K\}}{\mathbb{P}\{W \le K\}}. \tag{6.5}$$

For the second term we have, by (6.3) and the identity $\frac{\hat{\rho}}{1-\hat{\rho}} = \frac{\rho}{1-\rho}\frac{r-1}{r}$,

$$\frac{K\mathbb{P}\{W > K\}}{\mathbb{P}\{W \le K\}} \sim \frac{\rho}{1-\rho}\frac{(r-1)^{\nu}}{r(\nu-1)\mathbb{E}\{A_1\}}L(K)K^{2-\nu}, \tag{6.6}$$

if $K \to \infty$. The tail behavior for $V$ follows straightforwardly from that of $A_1^r$, which follows by applying Karamata's theorem (Lemma 2.1.7). This gives for $x \to \infty$,

$$\mathbb{P}\{V > x\} \sim (1-p)\frac{\rho}{1-\rho}\frac{(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}\{A_1\}}L(x)x^{1-\nu}. \tag{6.7}$$

Applying Karamata's theorem once more to the first term in the right-hand side of (6.5), we get

$$\int_0^K \mathbb{P}\{V > x\}\mathrm{d}x \sim \frac{(1-p)(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}\{A_1\}}\frac{\rho}{1-\rho}\frac{1}{2-\nu}K^{2-\nu}L(K). \tag{6.8}$$

The proof follows by combining (6.6) and (6.8), thereby noting that the constant in (6.8) is larger than the constant appearing in (6.6). This follows from $\rho = \lambda(r-1)\mathbb{E}\{A_1\} < 1$ and $1 < \nu < 2$, which implies $(1-p)/(2-\nu) > 1-p > (r-1)/r$. □

Loosely speaking, the mean buffer content behaves like a positive power of the buffer size in case of long-range dependent input. This shows once more that the impact of long-range dependence on the performance of fluid queues can be quite substantial – even if buffers are finite.

**Remark 4.6.2**
For the model with a single On-Off source it is also possible to obtain multi-term asymptotic expansions or even explicit results for the loss fraction. The classes of (heavy-tailed)

service-time distributions introduced in Boxma & Cohen [69] and Abate & Whitt [6] lead to explicit results for the waiting-time distribution in the $M/G/1$ queue. These results may also be used to obtain more refined asymptotics and explicit results for the mean buffer content.

### 4.6.3 A superposition of $N$ On-Off sources

The characteristics of this model can be described as follows. When source $i, 1 \leq i \leq N$, is On, it transmits fluid at rate $r_i \geq 1$ during a generic activity period $A_{1i}$ having mean $\alpha_i$. The silence periods $U_{1i}$ are exponentially distributed with parameter $\lambda_i$.
We have to make the restrictive assumption $r_i \geq 1$ in order to apply the framework developed in the previous sections: The general activity period $A_1$ now corresponds to the period where at least one on-off source is on. During this period, the buffer content is non-decreasing (its increments are the same as that of $B_1(.)$).
The stationary probability of silence $p_i^s$ equals $1/(1+\alpha_i\lambda_i)$, the mean offered load per unit of time offered by source $i$ is denoted by $\rho_i$ and equals $r_i\frac{\lambda_i\alpha_i}{1+\lambda_i\alpha_i}$. Note that in our setting, $\rho = \rho_1 + \cdots + \rho_N$, $\lambda = \lambda_1 + \cdots + \lambda_N$, and $p^s = \prod_i p_i^s$. Using this, it is not difficult to calculate $\mathbb{E}\{A_1\}$ and $\mathbb{E}\{B_1(A_1-)\}$. The following result is part of Theorem 4.6 in [67], see also Theorem 2.2.4 in this thesis.

**Lemma 4.6.1** *Assume that the activity periods of the sources 2,...,N are exponentially distributed and assmue that*

$$\mathbb{P}\{A_{11} > x\} = L(x)x^{-\nu},$$

*for $\nu > 1$. Assume that $\rho < 1$ and define $c = 1 - \sum_{i=2}^{N} \rho_i$. Then, the following asymptotics hold for $x \to \infty$:*

$$\mathbb{P}\{W > x\} \sim \frac{\lambda_1(r_1 - c)\alpha_1}{c - \lambda_1(r_1 - c)\alpha_1}\mathbb{P}\{(r_1 - c)A_{11}^r > x\}, \tag{6.9}$$

$$\mathbb{P}\{V > x\} \sim p_1^s\frac{\rho_1}{c - \rho_1}\mathbb{P}\{(r_1 - c)A_{11}^r > x\}. \tag{6.10}$$

Lemma 4.6.1 leads to the following results for the loss fraction and the mean buffer content in the finite buffer case.

**Theorem 4.6.2** *Assume that the conditions stated in Lemma 4.6.1 are satisfied. Then, for $K \to \infty$,*

$$L_K \sim M\mathbb{P}\{(r_1 - c)A_{11}^r > K\}, \tag{6.11}$$

*where $M$ is given by*

$$M = \frac{1-\rho}{\rho}\frac{\lambda_1(r_1 - c)\alpha_1}{c - \lambda_1(r_1 - c)\alpha_1}.$$

*If $1 < \nu < 2$, the mean buffer content satisfies for $K \to \infty$,*

$$\mathbb{E}\{V^K\} \sim \frac{(r-c)^{\nu-1}}{(\nu-1)\alpha_1}\left[p_1^s \frac{\rho_1}{1-\rho_1}\frac{1}{2-\nu} - \frac{\rho}{1-\rho}M\right] L(K)K^{2-\nu}. \tag{6.12}$$

**Proof**
The first part follows easily from Lemma 4.6.1, Theorem 4.3.1, and Theorem 4.5.3, or alternatively, from Theorem 4.6.1 and the tail behavior of $B_1(A_1-)$, which is given in Theorem 4.6 of [66]. The proof of the second part follows the same lines as the proof of Proposition 4.6.2 and is therefore omitted. □

The asymptotics for $L_K$ have recently been extended by Jelenković & Momčilović [165] (see also Jelenković [162]) to the case of multiple heavy-tailed On-Off sources. The methods in [165] do not require the assumption $r_i \geq 1$, see Chapter 7 for more discussion.

## 4.7 Overloaded queues

In this section we consider (for completeness) the case when the traffic load is at least 1, i.e., when $\rho \geq 1$ (equivalently $\hat{\rho} \geq 1$). If the silence periods are exponentially distributed, then it is possible to use the results for the $M/G/1$ queue with finite capacity $K$ given in Section III.5 of [97]. For this model we develop asymptotic expansions for the loss fraction, which can easily be applied to the fluid model by means of Theorem 4.5.3. Starting point of our analysis is the following expression for $\mathbb{P}\{W^K = 0\}$, given on p. 535 of [97], which is, just as (3.3), valid for all $\hat{\rho} > 0$.

$$\mathbb{P}\{W^K = 0\} = \left[\frac{1}{2\pi \mathrm{i}} \int_{s=-\mathrm{i}\infty+\epsilon}^{\mathrm{i}\infty+\epsilon} \frac{\mathrm{e}^{sK}}{s-\lambda+\lambda\beta(s)}\mathrm{d}s\right]^{-1}, \tag{7.1}$$

where $\beta(s)$ is the LST of the service time $B$ (which will equal $B_1(A_1-)$ when applied to the fluid model). $\epsilon$ must be chosen such that all zeroes of $s-\lambda+\lambda\beta(s)$ have real part smaller than $\epsilon$. If $\hat{\rho} \leq 1$, any $\epsilon > 0$ suffices. Note that the LST of $\mathbb{P}\{W^K = 0\}^{-1}$ with respect to $K$ is given by, for Re $s > \epsilon$,

$$\int_0^\infty \mathrm{e}^{-sK}\mathrm{d}\left[\mathbb{P}\{W^K = 0\}^{-1}\right] = \frac{s}{s-\lambda+\lambda\beta(s)}. \tag{7.2}$$

We now apply Equation (7.2) to derive asymptotic expressions for the loss probability when $\rho \geq 1$. Define $\beta_k$ as the $k$-th moment of the service time $B$. We first consider the case $\rho = 1$ (and hence $\hat{\rho} = 1$).

**Proposition 4.7.1** *Let* $\hat{\rho} = 1$.

1. *If* $\beta_2 < \infty$, *then*

$$L_{q,K} \sim \frac{\beta_2}{2\beta_1}\frac{1}{K}, \qquad K \to \infty.$$

2. *If* $\mathbb{P}\{B > x\} = L(x)x^{-\nu}$, $1 < \nu < 2$, *then*

$$L_{q,K} \sim \frac{1}{\beta_1}\frac{\pi}{\sin(\pi(\nu-1))}L(K)K^{1-\nu}, \qquad K \to \infty.$$

**Proof**
Both assertions will be proven by the use of Tauberian theorems. Since $\hat{\rho} = 1$, (3.8) reduces to

$$L_{q,K} = \mathbb{P}\{W^K = 0\}. \tag{7.3}$$

First, we prove Part 1. Since $\beta_2 < \infty$, we have

$$\beta(s) = 1 - \beta_1 s + \frac{1}{2}\beta_2 s^2 + \mathrm{o}(s^2), \qquad s \downarrow 0. \tag{7.4}$$

Inserting (7.4) in (7.2) yields

$$\int_0^\infty \mathrm{e}^{-sK}\mathrm{d}\left[\mathbb{P}\{W^K = 0\}^{-1}\right] = \frac{2\beta_1}{\beta_2 s} + \mathrm{o}(1/s), \qquad s \downarrow 0,$$

which gives Part 1 of Proposition 4.7.1 by using Theorem 2.1.1, and Theorem 4.3.1.
We now turn to Part 2. If $\mathbb{P}\{B > x\} = L(x)x^{-\nu}$, $1 < \nu < 2$, $\beta(s)$ satisfies (in view of Theorem 2.1.2)

$$\beta(s) - 1 + \beta_1 s \sim -\Gamma(1-\nu)s^\nu L(1/s), \qquad s \downarrow 0. \tag{7.5}$$

This gives, since $\hat{\rho} = 1$,

$$\int_0^\infty \mathrm{e}^{-sK}\mathrm{d}\left[\mathbb{P}\{W^K = 0\}^{-1}\right] \sim \frac{\beta_1}{-\Gamma(1-\nu)}s^{1-\nu}/L(1/s), \qquad s \downarrow 0. \tag{7.6}$$

Applying Theorem 2.1.1, we get for $K \to \infty$,

$$\mathbb{P}\{W^K = 0\}^{-1} \sim \frac{\beta_1}{-\Gamma(\nu)\Gamma(1-\nu)}K^{\nu-1}/L(K). \tag{7.7}$$

Part 2 now follows from $\Gamma(\nu)\Gamma(1-\nu) = \pi/\sin(\pi\nu)$ and $\sin(a) = -\sin(a-\pi)$.

□

**Remark 4.7.1**
We refrain from discussing the case where $\mathbb{P}\{B > x\} = L(x)x^{-2}$ (and $\beta_2 = \infty$). The Tauberian theorems are now much more delicate, see e.g. Theorem 8.1.6 in [44].

**Remark 4.7.2**
Although the asymptotic formula for the loss probability in case $\hat{\rho} < 1$ (given in Theorems 4.4.1 and 4.4.2) is independent of $\hat{\rho}$, it is not valid for $\hat{\rho} = 1$, as Proposition 4.7.1 shows. However, note that the asymptotic behavior of the loss probability in the heavy-tailed (infinite-variance) case is the same for $\hat{\rho} < 1$ and $\hat{\rho} = 1$, apart from a constant. Since $\sin x < x$ for $x > 0$, $\pi/\sin(\pi(\nu-1)) > 1/(\nu-1)$, so the constant in the asymptotic approximation for $L_{q,K}$ is strictly larger for $\hat{\rho} = 1$ than for $\hat{\rho} < 1$.

When $\hat{\rho} > 1$, it follows immediately that $\mathbb{P}\{W^K = 0\} \to 0$ as $K \to \infty$, which gives

$$L_{q,K} \to 1 - \frac{1}{\hat{\rho}}. \tag{7.8}$$

Using a result of Cohen [94], it is easy to derive the rate of convergence.

**Proposition 4.7.2** *If $\hat{\rho} > 1$, then we have for the $M/G/1$ queue with finite capacity $K$,*

$$L_{q,K} - \frac{\hat{\rho}-1}{\hat{\rho}} \sim -\delta\widetilde{\beta}'(\delta)\mathrm{e}^{-\delta K}, \qquad K \to \infty, \tag{7.9}$$

*where $\widetilde{\beta}(s) = \frac{1-\beta(s)}{\beta_1 s}$, $\widetilde{\beta}'(s)$ is the derivative of $\widetilde{\beta}(s)$, and $\delta$ is the unique positive real solution of*

$$\hat{\rho}\widetilde{\beta}(s) = 1. \tag{7.10}$$

**Proof**
Follows immediately from (3.8) and Part (iii) of Theorem 2.3 in [94]. □

# Appendix

## 4.A An alternative proof of Theorem 4.5.2

In this section we give an alternative proof of the proportionality result Theorem 4.5.2, which we believe is of independent interest. It is an extension of the proof of Hooghiemstra [156] for the proportionality result (3.6) for the $M/G/1$ finite dam. We start with two preliminary observations.

1. Let $C$ and $C^K$ be the length of a busy cycle for the infinite-buffer model and the model with finite buffer $K$, respectively. Then, the distributions of $V$ and $V^K$ are given by, cf. [19, 92]:

$$\mathbb{P}\{V \leq x\} = \frac{1}{\mathbb{E}\{C\}} \mathbb{E}\{\int_0^C 1_{\{V(t) \leq x\}} \mathrm{d}t\},$$

$$\mathbb{P}\{V^K \leq x\} = \frac{1}{\mathbb{E}\{C^K\}} \mathbb{E}\{\int_0^{C^K} 1_{\{V^K(t) \leq x\}} \mathrm{d}t\}.$$

2. Let $x < K < \infty$ and suppose that a downcrossing at level $x$ occurs for the process $V^K(t)$ for some $t$, so that the environment process $I(t) = 0$. Then, since $U_1$ is exponentially distributed, the time that elapses until $I(t)$ reaches 1 is distributed as $U_1$, due to the memoryless property of the silence periods.

We now construct a stochastic process $\widehat{V}^K(t)$ directly from $V(t)$. Consider an arbitrary sample path of $V(t)$, e.g. the sample path in Figure 4.1.

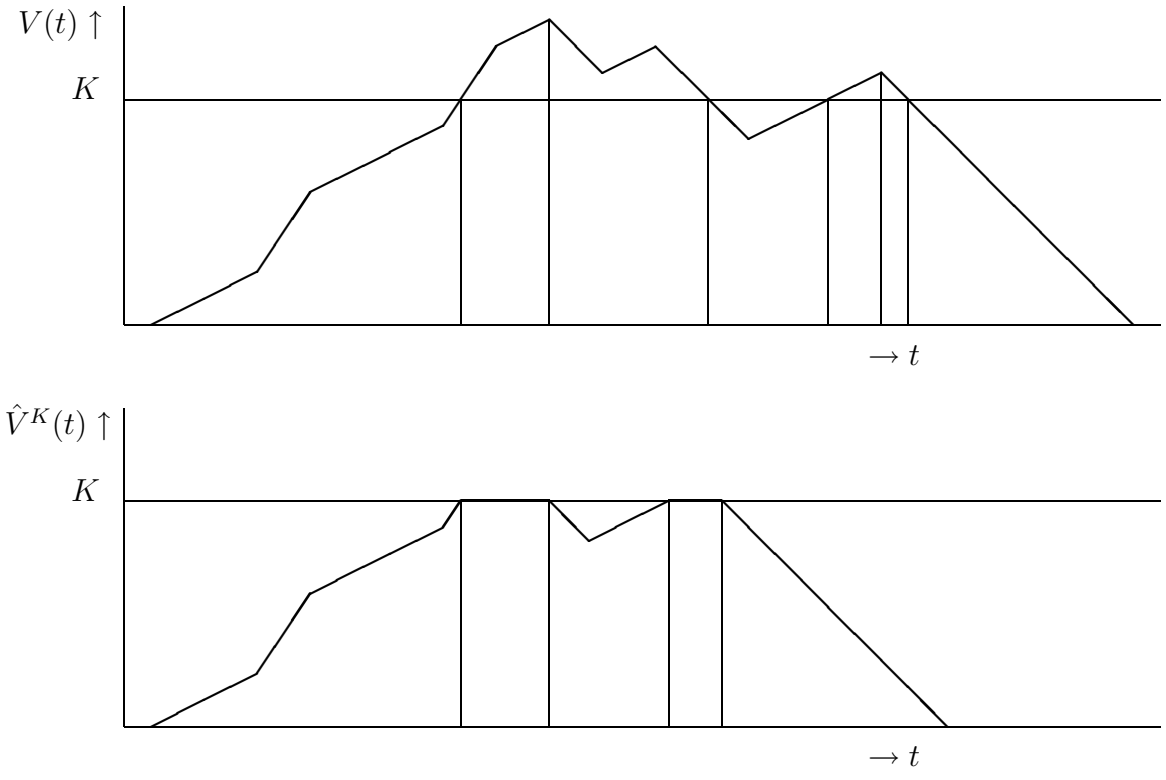


Figure 4.1: Construction of a sample path of $\widehat{V}^K(t)$ from $V(t)$.

The corresponding sample path for $\widehat{V}^K(t)$ is constructed as follows. The parts of the sample path of $V(t)$ below level $K$ remain unchanged. Consider the parts of the sample path of $V(t)$ between an upcrossing and a consecutive downcrossing of level $K$. Each of these parts can be divided into two sub-parts. The first sub-part is defined as the remaining activity period and the second sub-part as the remainder of the part. Now delete the second sub-part and truncate the first sub-part to level $K$, cf. Figure 4.1.
Since the silence periods in the infinite-buffer model are exponentially distributed, the same holds for the silence periods in the process $\{\widehat{V}^K(t) : t \geq 0\}$, by Observation 2. It follows immediately from the construction of $\{\widehat{V}^K(t)\}$ that the durations of the activity periods in $\{\widehat{V}^K(t)\}$ have the same distribution as $A_1$, and are all independent and independent of the silence periods. Finally, the trajectories during activity periods kan be chosen identically (in distribution) and independent from each other according to the stochastic process $\{B_1(t) : t \geq 0\}$. Hence, the dynamics of $\{\widehat{V}^K(t)\}$ satisfy the same dynamics as the process $\{V^K(t)\}$ as defined by Equations (2.4) and (2.5). This proves that $\{\widehat{V}^K(t) : t \geq 0\}$ has the same law as $\{V^K(t) : t \geq 0\}$.
To simplify the notation, we now define the process $V^K(t)$ as

$$V^K(t) := \widehat{V}^K(t), \qquad t \geq 0.$$

It follows immediately from the construction of $V^K(t)$ from $V(t)$ that the number of downcrossings from level $x \leq K$ is the same for their respective sample paths, which implies that the number of downcrossings at level $x \leq K$ of the process $V^K(t)$ during a busy cycle has the same distribution as that of $V(t)$ for any $K \in (0, \infty)$.
A second implication of the construction carried out is that

$$\int_0^{C^K} 1_{\{V^K(t) \leq x\}} \mathrm{d}t = \int_0^{C} 1_{\{V(t) \leq x\}} \mathrm{d}t,$$

for $0 \leq x < K$. This implies the proportionality between the stationary distributions of $V^K(t)$ and $V(t)$, by Observation 1: Define $\gamma := \frac{\mathbb{E}\{C\}}{\mathbb{E}\{C^K\}}$.

We relate $\gamma$ to the loss fraction $L_K$ by using variants of Little's formula, see also Section 4.3. The amount of work brought into the system per unit of time equals $\rho$ in the infinite-buffer model and $\rho(1 - L_K)$ in the finite-buffer model. Hence, we have by Little's formula that

$$\mathbb{P}\{V = 0\} = 1 - \rho,$$

and, for $K \geq 0$,

$$\mathbb{P}\{V^K = 0\} = 1 - \rho(1 - L_K).$$

Consequently,

$$\gamma = \frac{1 - \rho(1 - L_K)}{1 - \rho}.$$

A straightforward computation (use (3.9) and Theorem 4.5.3) shows that $\gamma = 1/\mathbb{P}\{W \leq K\}$.

# Chapter 5

# Busy-period asymptotics in single-server queues

## 5.1 Introduction

The $GI/G/1$ queue with heavy-tailed service-time distribution has been the subject of many studies. Most of them focus on the tail of the waiting-time distribution, see e.g. the list of references accompanying Theorem 2.2.1. The subject of investigation in the present chapter is the tail behavior of the *busy-period* distribution. Besides its intrinsic interest, the (tail behavior of the) busy period has applications to various other problems, like (networks of) fluid queues (Boxma [68]), Generalized Processor Sharing (Borst *et al.* [57, 58]), priority queues (Abate & Whitt [4]), and convergence rates in queueing and ruin problems (Asmussen & Teugels [23]). Another motivation for studying the busy-period distribution, is that it coincides with the sojourn-time distribution of a customer in the $GI/G/1$ queue with the LCFS service discipline with pre-emption.

The tail behavior of the busy-period distribution in the $M/G/1$ queue has been studied earlier in [4] under Cramèr-type assumptions. Two main references for the heavy-tailed case are De Meyer & Teugels [202], where the case of regularly varying service times is treated, and Asmussen *et al.* [27]. In the latter paper, it is shown that the result proven in [202] cannot be true for the entire class of subexponential distributions, thereby giving a negative answer to a conjecture posed in [202, 23]. Although the approaches in [202] and [27] are quite different, they are both based on the special branching structure (see e.g. Cohen [97], Section II.5) of the busy period in the $M/G/1$ queue, which heavily depends on the Poisson nature of the arrival stream. In this chapter, we look at the busy period from a different perspective: We show that the occurrence of a large busy period is related to the occurrence of a large *cycle maximum*, and then exploit asymptotic results for the latter random variable. These are known for the $GI/G/1$ queue with subexponential service times, see Asmussen [24].

The main result in this chapter (Theorem 5.3.1) is an extension of the result in [202], where

the $M/G/1$ queue is considered. Our main result is valid for renewal arrival streams. The assumption on the service-time distribution is weakened as well (to intermediate regular variation, see Remark 5.3.1). The general subexponential case left unanswered in [27] is not solved here. However, we give a partial result by showing an asymptotic lower bound, which coincides with the exact tail behavior under the conditions of Theorem 5.3.1. Finally, we give some counter-intuitive results for the busy period in the null-recurrent $M/G/1$ queue: It is shown (by analytic methods) that a heavier tail of the service time distribution can give rise to a *lighter* tail of the busy-period distribution.

We note that, in addition to the references mentioned above, several related results can be found in the random walk literature, see e.g. Doney [116], Bertoin & Doney [39, 40], and Baltrunas [35].

The chapter is organized as follows. In Section 5.2, we state some preliminary results and give a very short proof of the logarithmic asymptotics. The main result is proven in Section 5.3. This section also contains some additional remarks on the extension of Theorem 5.3.1 to other classes of (heavy-tailed) distributions. The null-recurrent case is treated in Section 5.4.

## 5.2 Preliminaries

### 5.2.1 The GI/G/1 queue

Suppose that the first customer enters an empty system at time 0. The service time of customer $i$ is denoted by $B_i$ and the time between the arrivals of customers $i$ and $i+1$ is denoted by $T_i$. It is assumed that $T_i, i \geq 1$, and $B_i, i \geq 1$, are i.i.d. sequences and that both sequences are independent of each other. The traffic load $\rho$ equals $\mathbb{E}\{B\}/\mathbb{E}\{T\}$ (with $T \stackrel{d}{=} T_1$ and $B \stackrel{d}{=} B_1$). Unless specified otherwise, it is assumed that $\rho < 1$.

Let $V(t)$ be the amount of work in the system at time $t$. The busy period $P$ is then defined as

$$P := \inf\{t > 0 : V(t) = 0\}.$$

The number of customers served during the busy period will be denoted by $N$. If $\rho < 1$, the process $\{V(t), t \geq 0\}$, is positive recurrent and the means of $P$ and $N$ are finite. We note that (using Wald's lemma) $\mathbb{E}\{P\} = \mathbb{E}\{B\}\mathbb{E}\{N\}$. An expression for $\mathbb{E}\{N\}$ is given on p. 279 of [97] and we repeat it here for later use. Define $S_n = \sum_{i=1}^{n}(B_i - A_i)$. Then,

$$\mathbb{E}\{N\} = \exp\{\sum_{n=1}^{\infty} \frac{1}{n}\mathbb{P}\{S_n > 0\}\}. \tag{2.1}$$

### 5.2.2 The cycle maximum

A random variable which will play a crucial role in the next sections is the cycle maximum $C_{\max}$, given by

$$C_{\max} := \sup\{V(t), 0 \le t \le P\}.$$

Asmussen [24] has obtained the tail behavior of the maximum *waiting time* $W_{\max}$ during a busy cycle for subexponential $B$. In particular,

$$\mathbb{P}\{W_{\max} > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > x\}.$$

Heath *et al.* [151] have shown that subexponentiality of $B$ implies $\mathbb{P}\{C_{\max} > x\} \sim \mathbb{P}\{W_{\max} > x\}$, see Corollary 2.2 in [151]. Together, these results imply that, if $B$ has a subexponential distribution,

$$\mathbb{P}\{C_{\max} > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > x\}. \tag{2.2}$$

We also need the first passage time of level $x$, so we define

$$\tau(x) := \inf\{t \ge 0 : V(t) \ge x\}.$$

Note that $C_{\max} \ge x$ iff $\tau(x) < P$.

### 5.2.3 An upper bound and crude asymptotics

As a preliminary result, we give a qualitative upper bound for the tail of $P$, which shows that $\mathbb{P}\{P > x\} = \mathrm{O}(\mathbb{P}\{B > x\})$. This result readily follows from general upper bounds for the distribution tails of stopping times which are given in Borovkov [52]. With $L(.)$, we denote a slowly varying function.

**Proposition 5.2.1** *If $\mathbb{P}\{B > x\} = L(x)x^{-\nu}$, there exists a finite constant $C$ such that*

$$\mathbb{P}\{P > x\} \le CL(x)x^{-\nu}.$$

**Proof**
Theorem 43.3 of [52] guarantees the existence of a constant $C_1$ such that $\mathbb{P}\{N > x\} \le C_1 L(x)x^{-\nu}$. Next, use the representation $P = B_1 + \cdots + B_N$ and apply Theorem 42.2 of [52], noting that $N$ is a stopping time w.r.t. the filtration generated by $(A_n, B_n)_{n\ge1}$. □

Together with the trivial lower bound $\mathbb{P}\{P > x\} \ge \mathbb{P}\{B > x\}$, Proposition 5.2.1 implies (if $\mathbb{P}\{B > x\}$ is regularly varying of index $-\nu$):

$$\lim_{x\to\infty} \frac{\log \mathbb{P}\{P > x\}}{\log x} = -\nu. \tag{2.3}$$

In view of the generality of the results in [52], we expect this result to be true more generally (e.g. by relaxing independence assumptions), but we will not pursue this here, since we are primarily interested in the exact asymptotics.

## 5.3 Main result

In this section we prove the following theorem.

**Theorem 5.3.1** *If the service-time distribution is regularly varying of index $-\nu$, $\nu > 1$, then,*

$$\mathbb{P}\{P > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > x(1-\rho)\}. \tag{3.1}$$

This result is a generalization of [202], where it was assumed that the interarrival times are exponential. Note that, in that case, $\mathbb{E}\{N\} = \frac{1}{1-\rho}$.
Before giving a proof of Theorem 3.1 we provide some heuristic arguments. When the busy period is large, there is probably a large cycle maximum within that busy period. In view of the results and arguments in the works of Asmussen (see [24, 26]), this is most likely due to one early large service time. After this early large service time, things go back to normal and the workload goes to zero with negative rate $-(1-\rho)$. Hence, if $C_{\max}$ is large, then one would expect that

$$P \approx \frac{C_{\max}}{1-\rho}.$$

Together with the tail behavior (2.2) of $C_{\max}$, this yields (3.1).
A strongly related performance measure is the *longest service time in a busy period*, denoted by $B_{\max}$. Results of Boxma [63, 64] imply that $\mathbb{P}\{B_{\max} > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > x\}$ for *any* service-time distribution (see Asmussen [24] for an alternative proof). Thus, a large busy period and a large cycle maximum are both caused by a single large service time in the beginning of the busy period.
Theorem 5.3.1 will be proven by providing lower and upper bounds, which asymptotically coincide. The derivation of these bounds is strongly related to the framework of Section 2.4. The lower bound formally shows that a large cycle maximum is sufficient for a large busy period to occur. In particular, the lower bound follows from the law of large numbers for renewal processes. The proof of the upper bound is more involved. In particular, we need the truncation Lemma 2.4.1.

### 5.3.1 Lower bound

**Proposition 5.3.1** *Assmue that the service-time distribution is regularly varying. Then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{P > x\}}{\mathbb{E}\{N\}\mathbb{P}\{B > x(1-\rho)\}} \geq 1. \tag{3.2}$$

**Proof**
For all $\epsilon > 0$, we have, when $x$ is large enough,

$$\begin{aligned}\mathbb{P}\{P > x\} &\geq \mathbb{P}\{P > x, C_{\max} > x(1-\rho+\epsilon)\} \\ &= \mathbb{P}\{P > x \mid \tau(x(1-\rho+\epsilon)) < P\}\mathbb{P}\{C_{\max} > x(1-\rho+\epsilon)\} \\ &\geq \mathbb{P}\{P - \tau(x(1-\rho+\epsilon)) > x \mid \tau(x(1-\rho+\epsilon)) < P\} \times \\ &\qquad \mathbb{P}\{C_{\max} > x(1-\rho+\epsilon)\}. \end{aligned} \tag{3.3}$$

Note that

$$\begin{aligned}&\mathbb{P}\{P - \tau(x(1-\rho+\epsilon)) > x \mid \tau(x(1-\rho+\epsilon)) < P\} = \\ &\int_{x(1-\rho+\epsilon)}^{\infty} \mathbb{P}\{P - \tau(x(1-\rho+\epsilon)) > x \mid \tau(x(1-\rho+\epsilon)) < P, V(\tau(x(1-\rho+\epsilon))) = y\} \\ &\qquad \mathrm{d}\mathbb{P}\{V(\tau(x(1-\rho+\epsilon))) \leq y \mid \tau(x(1-\rho+\epsilon)) < P\}.\end{aligned} \tag{3.4}$$

The probability inside the integral equals

$$\begin{aligned}&\mathbb{P}\{P - \tau(x(1-\rho+\epsilon)) > x \mid V(\tau(x(1-\rho+\epsilon))) = y\} \\ =\ &\mathbb{P}\{V(t) > 0; 0 \leq t \leq x \mid V(0) = y\}.\end{aligned}$$

Note that the right-hand side is increasing in $y$. Taking $y$ as small as possible (i.e., take $y = x(1-\rho+\epsilon)$) and combining this with (3.3) and (3.4), we obtain

$$\mathbb{P}\{P > x\} \geq \mathbb{P}\{V(s) > 0; 0 \leq s \leq x \mid V(0) = x(1-\rho+\epsilon)\}\mathbb{P}\{C_{\max} > x(1-\rho+\epsilon)\}.$$

The right-hand side is lower bounded by

$$(1-\delta)\mathbb{P}\{C_{\max} > x(1-\rho+\epsilon)\}$$

for each $\delta, \epsilon > 0$ and $x$ large enough. This follows from the strong law of large numbers for renewal processes (see also the proof of Proposition 4.2 in [26]). After dividing $\mathbb{P}\{P > x\}$ by $\mathbb{P}\{C_{\max} > x(1-\rho)\}$, the result follows using (2.2), letting $x \to \infty$, and then letting $\epsilon, \delta \downarrow 0$. □

### 5.3.2 Upper bound

**Proposition 5.3.2** *Assume that the service-time distribution is regularly varying of index $-\nu, \nu > 1$. Then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{P > x\}}{\mathbb{E}\{N\}\mathbb{P}\{B > x(1-\rho)\}} \leq 1. \tag{3.5}$$

**Proof**

First, we note that it suffices to prove the proposition for the case that the interarrival times are bounded by a finite constant $M$. If this is not the case, then truncate all interarrival times (this does not decrease the length of the busy period and does not violate stability as long as $M$ is chosen large enough). It is clear that the workload $\rho_M$ of the modified system converges to $\rho$, since $\mathbb{E}\{\min(T, M)\} \to \mathbb{E}\{T\}$, when $M \to \infty$. We also have to show that the expected number of customers served during a busy period in the modified system converges to $\mathbb{E}\{N\}$ when $M \to \infty$.

We give a simple proof of this result (since we found no direct reference), using the expression for $\mathbb{E}\{N\}$ given in Section 5.2.1. When the interarrival times are truncated at $M$, then the expected number of customers served in a busy period is given by

$$\exp\{\sum_{n=1}^{\infty} \frac{1}{n} \mathbb{P}\{\sum_{i=1}^{n} B_i > \sum_{i=1}^{n} \min(T_i, M)\}\}.$$

Since the main sum in the exponent is decreasing in $M$, and finite when $M$ is large enough, the desired result follows from a straightforward application of the dominated convergence theorem. Finally, the result follows easily since

$$\lim_{M\to\infty} \lim_{x\to\infty} \frac{\mathbb{P}\{B > x(1-\rho)\}}{\mathbb{P}\{B > x(1-\rho_M)\}} = 1, \tag{3.6}$$

because the tail of $B$ is regularly varying.

Henceforth, it will be assumed that the interarrival times are bounded by $M$. For $\delta > 0$, we get

$$\begin{aligned} \mathbb{P}\{P > x\} &= \mathbb{P}\{P > x, C_{\max} > x(1-\rho-\delta)\} + \mathbb{P}\{P > x, C_{\max} \le x(1-\rho-\delta)\} \\ &\le \mathbb{P}\{C_{\max} > x(1-\rho-\delta)\} + \mathbb{P}\{P > x, C_{\max} \le x(1-\rho-\delta)\}. \end{aligned}$$

Hence, using (2.2), it suffices to show that, for all $\delta > 0$,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{P > x, C_{\max} \le x(1-\rho-\delta)\}}{\mathbb{P}\{C_{\max} > x(1-\rho)\}} = 0. \tag{3.7}$$

Let $\epsilon > 0$. Write

$$\begin{aligned} & \mathbb{P}\{P > x, C_{\max} \le x(1-\rho-\delta)\} \\ = \; & \mathbb{P}\{P > x, C_{\max} \le \epsilon x\} + \mathbb{P}\{P > x, \epsilon x < C_{\max} \le x(1-\rho-\delta)\} \\ = \; & I + II. \end{aligned}$$

We start with the first term. Recall that the service and interarrival times of customer $i$ are given by $B_i$ and $T_i$. Under $C_{\max} \le \epsilon x$, it must hold that $X_i := B_i - T_i \le \epsilon x$ for $i = 1, \ldots, N$. Hence,

$$\mathbb{P}\{P > x, C_{\max} \le \epsilon x\} \le \mathbb{P}\{P > x, B_i - T_i \le \epsilon x; i = 1, \ldots, N\}. \tag{3.8}$$

Denote the integer part of $a$ by $[a]$. Since the interarrival times are bounded by $M$, the number of customers who have entered the system between time 0 and $x$ is at least $[x/M]$ (and in particular $N \geq [x/M]$). It follows easily that

$$\begin{aligned}
& \mathbb{P}\{P > x, X_i \leq \epsilon x; i = 1, \ldots, N\} \\
\leq\ & \mathbb{P}\{ \sum_{i=1}^{[x/M]-1} X_i > 0, X_i \leq \epsilon x; i = 1, \ldots, [x/M] - 1\} \\
\leq\ & \mathbb{P}\{ \sum_{i=1}^{[x/M]-1} X_i > 0 \mid X_i \leq \epsilon x; i = 1, \ldots, [x/M] - 1\} \\
=\ & \mathbb{P}\{ \sum_{i=1}^{[x/M]} (X_i - \frac{1}{2}\mathbb{E}\{X_1\}) > \frac{1}{2}\mathbb{E}\{X_1\}(1 - [x/M]) \mid X_i \leq \epsilon x; i =\leq [x/M] - 1\}.
\end{aligned}$$

We now apply the truncation Lemma 2.4.1 of Resnick & Samorodnitsky. This lemma guarantees that, for $\epsilon$ small enough, the above probability can be upper bounded by $\phi(\frac{1}{2}(\mathbb{E}\{-X_1\})[x/M]) = \mathrm{o}(\mathbb{P}\{B > x\})$. This completes the estimation of Term I. We now turn to Term II.

If we condition on $C_{\max} > \epsilon x$, then we obtain

$$II = \mathbb{P}\{P > x, C_{\max} \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\}\mathbb{P}\{C_{\max} > \epsilon x\}. \tag{3.9}$$

Since $\mathbb{P}\{C_{\max} > \epsilon x\} = \mathrm{O}(\mathbb{P}\{B > x\})$, it suffices to show that

$$\mathbb{P}\{P > x, C_{\max} \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\} \to 0, \quad x \to \infty. \tag{3.10}$$

Observe that $V(\tau(\epsilon x)) \leq C_{\max}$ when $C_{\max} > \epsilon x$. Hence, we can bound (3.10) by

$$\mathbb{P}\{P > x, V(\tau(\epsilon x)) \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\}.$$

Choose $\gamma > 0$ such that $(1-\rho-\delta)/(1-\gamma) < 1-\rho$, i.e. choose $\gamma < \delta/(1-\rho)$. If $P > \tau(\epsilon x)$, then $P > x$ implies that either $\tau(\epsilon x) > \gamma x$ or $P - \tau(\epsilon x) > (1 - \gamma)x$. Hence,

$$\begin{aligned}
& \mathbb{P}\{P > x, V(\tau(\epsilon x)) \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\} \\
\leq\ & \mathbb{P}\{\tau(\epsilon x) > \gamma x, V(\tau(\epsilon x)) \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\} \\
+\ & \mathbb{P}\{P - \tau(\epsilon x) > (1 - \gamma)x, V(\tau(\epsilon x)) \leq x(1 - \rho - \delta) \mid C_{\max} > \epsilon x\} \\
=\ & IIa + IIb.
\end{aligned}$$

We start with $IIb$. Using a similar argument as in the proof of the lower bound in the previous subsection, we get

$$IIb \leq \mathbb{P}\{V(s) > 0; 0 \leq s \leq (1 - \gamma)x \mid V(0) = x(1 - \rho - \delta)\}, \tag{3.11}$$

which converges to zero by the law of large numbers.

Note that

$$IIa = \frac{\mathbb{P}\{V(\tau(\epsilon x)) \leq x(1-\rho-\delta), \gamma x < \tau(\epsilon x) < P\}}{\mathbb{P}\{\tau(\epsilon x) < P\}} \leq \frac{\mathbb{P}\{\gamma x < \tau(\epsilon x) < P\}}{\mathbb{P}\{\tau(\epsilon x) < P\}}.$$

To deal with $IIa$, we need to prove that

$$\mathbb{P}\{\gamma x < \tau(\epsilon x) < P\} = \mathrm{o}(\mathbb{P}\{B > x\}). \tag{3.12}$$

Let $x_0$ be large, but not larger than $\epsilon x$. Then

$$\begin{aligned} & \mathbb{P}\{\gamma x < \tau(\epsilon x) < P\} \\ = \; & \mathbb{P}\{\gamma x < \tau(\epsilon x) < P, \tau(\epsilon x) > \tau(x_0)\} \\ + \; & \mathbb{P}\{\gamma x < \tau(\epsilon x) < P, \tau(\epsilon x) = \tau(x_0)\} \\ = \; & IIa1 + IIa2. \end{aligned}$$

A useful fact is Lemma 4.4 from Asmussen & Möller [25], which implies

$$\lim_{x_0 \to \infty} \limsup_{x \to \infty} \frac{\mathbb{P}\{\tau(x_0) < \tau(x) < P\}}{\mathbb{P}\{B > x\}} = 0. \tag{3.13}$$

This controls term $IIa1$. The second term can be bounded as follows:

$$\mathbb{P}\{\gamma x < \tau(\epsilon x) < P, \tau(\epsilon x) = \tau(x_0)\} \leq \mathbb{P}\{P > \gamma x, \tau(x_0) > \gamma x\}.$$

The probability on the right hand side equals

$$\mathbb{P}\{0 < V(s) < x_0; 0 \leq s \leq \gamma x\}.$$

Since the interarrival times are bounded by $M$, at least $[\gamma x/M]$ customers must have entered the system by time $\gamma x$. All these customers have a service time which is at most $x_0$. Hence,

$$\mathbb{P}\{0 < V(s) < x_0; 0 \leq s \leq \gamma x\} \leq \mathbb{P}\{B_i < x_0; i = 1, \ldots, [\gamma x/M]\} = \mathbb{P}\{B < x_0\}^{[\gamma x/M]},$$

which decays exponentially fast in $x$. The proof of the theorem now follows by first letting $x \to \infty$, then $x_0 \to \infty$, then $\gamma \downarrow 0$ and finally $\delta, \epsilon \downarrow 0$ (and eventually $M \to \infty$).

□

**Remark 5.3.1**
Theorem 5.3.1 remains true if the service-time distribution is of intermediate regular variation. The only non-trivial change in the proof is the treatment of Term I. Under the assumption $\mathbb{E}\{B^\theta\} < \infty$ for some $\theta > 1$, we can use the fact [86] that $C^{-1}x^{-\gamma_1} \leq \mathbb{P}\{B > x\} \leq Cx^{-\gamma_2}$, where $\gamma_1 \geq \gamma_2 > 1$ and $C > 1$. This makes Lemma 2.4.1 applicable.

**Remark 5.3.2**
The proof of the lower bound can be extended to the case where $B$ is subexponential and $(\log \mathbb{P}\{B > x\})/\sqrt{x} \to \infty$: Instead of the law of large numbers, apply the central limit theorem, starting with the inequality $\mathbb{P}\{P > x\} \geq \mathbb{P}\{P > x, C_{\max} \geq x(1-\rho) + K\sqrt{x}\}$, with $K$ large. Note that (3.1) does not hold if $\log \mathbb{P}\{B > x\} = \mathrm{o}(\sqrt{x})$, cf. [27]. Baltrunas [35] considers a related random walk problem. The results in [35] can be used to obtain exact asymptotics for $\mathbb{P}\{N > x\}$ for the case $(\log \mathbb{P}\{B > x\})/\sqrt{x} \to \infty$ and some additional regularity conditions.

**Remark 5.3.3**
The relationship between the busy period and the cycle maximum as provided by (the proof of) Theorem 5.3.1 is entirely different when the service times are light-tailed. Results in Hooghiemstra [155] and Cohen & Hooghiemstra [96] indicate that a (light-tailed) busy period of size $x$ implies a cycle maximum of $\mathrm{O}(\sqrt{x})$.

## 5.4 On the critical case

In this final section we derive the tail behavior of the busy-period distribution in the case that the workload $\rho = 1$. It will be assumed that the interarrival time distribution is exponential, so let the arrival process be a Poisson process with intensity $\lambda$. It is well-known that the distribution of $P$ is still proper but has infinite mean when $\rho = 1$. It is difficult to develop intuition for this boundary case $\rho = 1$. In that respect, the –surprising– result of the present section may be helpful.
Our method of proof requires Laplace-transform techniques. Let $\pi(s)$ be the LST of $P$ and let $\beta(s)$ be the LST of the service-time distribution. It is well-known that $\pi(s)$ is the unique solution (when $\rho \leq 1$) of

$$\pi(s) = \beta(s + \lambda - \lambda\pi(s)), \qquad \text{Re } s \geq 0,$$

with $|\pi(s)| \leq 1$. Define $\beta_1 = \mathbb{E}\{B\}$.

**Theorem 5.4.1** *If $\beta_2 = \mathbb{E}\{B^2\} < \infty$, then*

$$\mathbb{P}\{P > x\} \sim \frac{1}{\lambda}\sqrt{\frac{\beta_1}{2\pi\beta_2}}x^{-\frac{1}{2}}. \tag{4.1}$$

*If $\mathbb{P}\{B > x\} \sim Cx^{-\nu}, C > 0, 1 < \nu < 2$, then*

$$\mathbb{P}\{P > x\} \sim \frac{1}{C\Gamma(1-1/\nu)}\left(\frac{\beta_1}{-\Gamma(1-\nu)}\right)^{\frac{1}{\nu}} x^{-\frac{1}{\nu}}, \tag{4.2}$$

*with $\Gamma(.)$ being the Gamma function.*

Hence, we can conclude in the case $\rho = 1$ that the heavier the tail of the service time distribution, the *lighter* the tail of the busy-period distribution.

**Proof**
First, assume that $\beta_2 < \infty$. Then we can write for $s \downarrow 0$,

$$\beta(s) = 1 - \beta_1 s + \frac{1}{2}\beta_2 s^2 + o(s^2).$$

Combining this with the functional equation for $\pi(s)$, we obtain after using $\lambda = 1/\beta_1$, $s = o(1 - \pi(s))$ and $1 - \pi(s) = o(1)$,

$$1 - \pi(s) \sim \frac{1}{\lambda}\sqrt{\frac{2\beta_1}{\beta_2}}\sqrt{s}, \qquad s \downarrow 0.$$

The result now follows from Theorem 2.1.2, noting that $-\Gamma(-\frac{1}{2}) = 2\sqrt{\pi}$.
Next we assume that $\mathbb{P}\{B > x\} \sim Cx^{-\nu}, C > 0, 1 < \nu < 2$. According to Theorem 2.1.2 we can write in this case

$$\beta(s) = 1 - \beta_1 s + (-\Gamma(1-\nu))Cs^\nu + o(s^\nu), \qquad s \downarrow 0.$$

After some tedious but straightforward computations we obtain

$$1 - \pi(s) \sim \frac{1}{C}\left(\frac{\beta_1}{-\Gamma(1-\nu)}\right)^{\frac{1}{\nu}} s^{\frac{1}{\nu}}, \qquad s \downarrow 0.$$

The result now follows from yet another application of Theorem 2.1.2.

□

# Chapter 6

# The fluid queue I: Reduced-peak

## 6.1 Introduction

The central subject of investigation of the present and the next chapter is the fluid queue fed by a finite number of On-Off sources with heavy-tailed On- and/or Off-periods, and possibly some additional light-tailed input. Both chapters focus on the asymptotic behavior of the workload distribution. In particular, we extend the results for the fluid queue considered in Subsection 2.2.2 to the case of multiple heavy-tailed On-Off sources. Furthermore, we remove the assumptions on the peak rates imposed by previous studies, see again Subsection 2.2.2.

It turns out that the workload asymptotics crucially depend on whether or not activity of heavy-tailed sources alone is sufficient for severe congestion to arise. First results in this realm are asymptotic bounds obtained by Dumas & Simonian [120]. These bounds show a sharp dichotomy in the qualitative tail behavior of the workload, depending on whether the mean rate of the light-tailed input plus the aggregate peak rate of the heavy-tailed sources exceeds the link rate (service capacity) or not. In case the link rate is smaller, the workload distribution has heavy-tailed characteristics, whereas the link rate being larger results in light-tailed characteristics.

The asymptotic bounds in [120] as well as results of Agrawal *et al.* [12] (see also Section 2.2 of this thesis) indicate that in the former case one can often identify a 'dominant' heavy-tailed source or a set of such sources. As far as tail behavior is concerned, all other sources can be accounted for by subtracting their aggregate traffic intensity from the service capacity. This may formally be phrased in terms of a 'reduced-load equivalence', implying that the workload is asymptotically equivalent to that in a *reduced* system. The reduced system consists only of the set of dominant sources, served at the link rate reduced by the mean rate of all other sources. This suggests that the most likely way for overflow to occur is for the sources in the dominant subset to experience extremely long On-periods, while all other sources show roughly average behavior. These phenomena are studied in great detail in Chapter 7.

In the present chapter, we focus on the opposite case where the peak rate of the heavy-tailed sources plus the mean rate of the light-tailed sources is *smaller* than the link rate. Thus, the overflow scenario described above cannot occur, and now the light-tailed sources too must deviate from their 'normal' behavior in order for the queue to grow. Our results will show in detail how a conjunction of extreme activity of the light-tailed and heavy-tailed sources, both in their own characteristic ways, results in a large queue building up.

We will find that the workload distribution is asymptotically equivalent to that in a somewhat 'dual' reduced system, multiplied with a certain pre-factor. The reduced system now consists of only the *light*-tailed sources, served at the link rate reduced by the *peak* rate of the *heavy*-tailed sources, hence the phrase 'reduced-peak equivalence'. The pre-factor represents the probability that the heavy-tailed sources have sent at their peak rate for more than a certain amount of time. This amount of time may be interpreted as the 'time to overflow' for the light-tailed sources in the reduced system. This suggests that the most likely way for overflow to occur is for the light-tailed sources to show temporarily similar 'abnormal' behavior as is the typical cause of overflow in the reduced system. During that time period, the heavy-tailed sources constantly send at their peak rate. Loosely stated, the heavy-tailed sources must send at their peak rate long enough for the light-tailed sources to be able to cause overflow. The subtle combination of light-tailed and heavy-tailed large deviations is similar to that for an $M/G/2$ queue with heterogeneous servers as described in Section 2.3 of this thesis.

The remainder of the chapter is organized as follows. In Section 6.2 we present a detailed model description and give an important preliminary result. We determine the exact asymptotics of the workload distribution in Section 6.3. The 'reduced-peak equivalence' involves some new results for light-tailed input, which may be of independent interest. In Section 6.4 we show that our assumptions regarding the light-tailed input are satisfied for two important traffic scenarios: (i) Markov-modulated fluid input; (ii) instantaneous input.

## 6.2 Model description

We first present a detailed model description. We consider $N$ traffic sources sharing a link of unit rate. Denote by $A_i(s,t)$ the amount of traffic generated by source $i$ during the time interval $(s,t]$. We assume that the process $A_i(s,t)$ has stationary increments. Let $\mathcal{I} = \{1,\ldots,N\}$ index the sources. For any $E \subseteq \mathcal{I}$, denote by $A_E(s,t) := \sum_{i\in E} A_i(s,t)$ the aggregate amount of traffic generated by the sources $i \in E$ during $(s,t]$. In particular, $A(s,t) := A_{\mathcal{I}}(s,t)$ is the total amount of traffic generated during $(s,t]$.

Denote by $\rho_i$ the traffic intensity of source $i$ (as will be defined in detail below). For any $E \subseteq \mathcal{I}$, define $\rho_E := \sum_{i\in E} \rho_i$ as the aggregate traffic intensity of the sources $i \in E$.

For any $c \geq 0$, $E \subseteq \mathcal{I}$, define $V_E^c(t) := \sup_{0 \leq s \leq t} \{A_E(s,t) - c(t-s)\}$ as the workload at time $t$ in a queue of capacity $c$ fed by the sources $i \in E$ (assuming $V_E^c(0) = 0$). For $c > \rho_E$, let $V_E^c$ be a random variable with the limiting distribution of $V_E^c(t)$ for $t \to \infty$ (assuming it exists). In particular, $V(t) := V_{\mathcal{I}}^1(t)$ is the total workload at time $t$, and $V$ is a random variable with the limiting distribution of $V(t)$ for $t \to \infty$. Note that

$$V \stackrel{d}{=} \sup_{t \geq 0} \{A(-t,0) - t\}.$$

We now describe the traffic scenario that we consider. We assume that the sources may be partitioned into two sets; $\mathcal{I}_1$ is the set of 'light-tailed' sources; $\mathcal{I}_2$ is the set of 'heavy-tailed' sources. For the sources in $\mathcal{I}_1$, we make the (weak) assumption that the input process $A_{\mathcal{I}_1}(s,t)$ satisfies a large-deviations principle. In particular, we follow Glynn & Whitt [138] and assume the following:

**Assumption 6.2.1** *There exist positive constants $\theta^* = \theta^*(c)$ and $\epsilon^*$ such that*

$$t^{-1} \log \mathbb{E}\{\exp\{\theta(A_{\mathcal{I}_1}(0,t) - ct)\}\} \to \phi_c(\theta)$$

*as $t \to \infty$, for $|\theta - \theta^*| \leq \epsilon^*$, such that $\phi_c(\theta^*) = 0$, $\phi_c'(\theta^*) > 0$, and*

$$\mathbb{E}\{\exp\{\theta^* A_{\mathcal{I}_1}(0,t)\}\} < \infty$$

*for all $t > 0$.*

Assumption 6.2.1 and a stability condition yield the following large-deviations estimate (cf. Theorem 4 of [138]):

$$\lim_{x \to \infty} x^{-1} \log \mathbb{P}\{V_{\mathcal{I}_1}^c > x\} = -\theta^*. \tag{2.1}$$

For a more elaborate discussion on Assumption 6.2.1 and its connections with classical large-deviations theory, we refer to [138] and references therein.

For the sources in $\mathcal{I}_2$, we assume that each source $i$ generates traffic according to a semi-Markov process on a finite state space $\{1, \ldots, n_i\}$. Note that this process is a generalization of the On-Off source to multiple activity states. While source $i$ is in state $j$, it generates traffic at rate $r_{ij}$, with $r_{i1} > r_{i2} > \ldots > r_{in_i} = 0$. Define $r_i \equiv r_{i1}$ as the peak rate of source $i$. For any $E \subseteq \mathcal{I}_2$, denote by $r_E := \sum_{i \in E} r_i$ the aggregate peak rate of the sources $i \in E$. The time that source $i$ stays in state $j$ before jumping to another state has some general distribution $A_{ij}(\cdot)$ with finite mean $\alpha_{ij}$. The state transitions are governed by some irreducible Markov chain (we assume self-transitions are not possible). The fraction of time that source $i$ spends in state $j$ is denoted by $p_{ij}$, with $p_i \equiv p_{i1}$ the fraction of time that source $i$ sends at its peak rate. Note that

$$\rho_i = \sum_{j=1}^{n_i} p_{ij} r_{ij} = \sum_{j=1}^{n_i - 1} p_{ij} r_{ij}.$$

An important special case is $n_i = 2$. In this case, source $i$ behaves as an On-Off source, and we have $p_i = \frac{\alpha_{i1}}{\alpha_{i1}+\alpha_{i2}}$ as the fraction of On-time and $\rho_i = p_i r_i$.

Let $A_i$ be a random variable with distribution $A_i(\cdot) \equiv A_{i1}(\cdot)$, i.e., the amount of time that source $i$ stays in state 1 (sends at its peak rate). Denote by $A_i^r(\cdot)$ the distribution of the residual lifetime of $A_i$, i.e., $A_i^r(x) := \frac{1}{\mathbb{E}\{A_i\}} \int_0^x (1 - A_i(y))\mathrm{d}y$. Let $A_i^r$ be a random variable with distribution $A_i^r(\cdot)$.
We now give an important preliminary result, which (besides of independent interest) will be used in establishing our main theorem in the next sections. In the special case of On-Off sources, the result is due to Jelenković & Lazar [161], see Theorem 2.2.3.

**Theorem 6.2.1** *If $A_i(\cdot) \in \mathcal{L}$, $A_i^r(\cdot) \in \mathcal{S}$, $\rho_i < c$, and $r_{i2} < c < r_i = r_{i1}$, then*

$$\mathbb{P}\{V_i^c > x\} \sim p_i \frac{r_i - \rho_i}{c - \rho_i} \mathbb{P}\{A_i^r > \frac{x}{r_i - c}\}.$$

**Proof**
The condition $r_{i2} < c < r_i = r_{i1}$ ensures that the workload process falls within the framework of Kella & Whitt [168], see also Chapter 4. In particular, the stationary distribution has the following representation:

$$\mathbb{P}\{V_i^c > x\} = p_i \mathbb{P}\{W_i^c + (r_i - c)A_i^r > x\} + (1 - p_i)\mathbb{P}\{W_i^c + (r_i - c)A_i - T_i(U_i^r) > x\}.$$

The exact form of $T_i(U_i^r)$ is not relevant for our purposes. The random variable $W_i^c$ represents the waiting time in a $GI/G/1$ queue of capacity 1 with service times $(r_i - c)A_i$. The interarrival times are equal to the decrease in the workload during the time that source $i$ spends in states $\{2, \ldots, n_i\}$ between two successive visits to state 1. Like in Chapter 4, we denote such a decrease (with a slight abuse of notation) by $T_i(U_i)$, and the corresponding time interval by $U_i$. Note that $T_i(U_i) \equiv cU_i$ in case $n_i = 2$.
From Theorem 2.2.1, we have

$$\mathbb{P}\{W_i^c > x\} \sim \frac{\tilde{\rho}_i}{1 - \tilde{\rho}_i} \mathbb{P}\{(r_i - c)A_i^r > x\},$$

with $\tilde{\rho}_i = \frac{(r_i - c)\mathbb{E}\{A_i\}}{\mathbb{E}\{T_i(U_i)\}}$. Using standard properties of long-tailed and subexponential distribution functions, we obtain

$$\mathbb{P}\{V_i^c > x\} \sim \left( \frac{\tilde{\rho}_i}{1 - \tilde{\rho}_i} + p_i \right) \mathbb{P}\{(r_i - c)A_i^r > x\}.$$

The statement now follows after a straightforward computation, using the expression for $\tilde{\rho}_i$ and the identities

$$p_i = \frac{\mathbb{E}\{A_i\}}{\mathbb{E}\{A_i\} + \mathbb{E}\{U_{i1}\}}, \qquad \rho_i = p_i r_i + (1 - p_i)\left( c - \frac{\mathbb{E}\{T_i(U_i)\}}{\mathbb{E}\{U_i\}} \right).$$

□

## 6.3 Asymptotic analysis

In this section we analyze the tail behavior of the workload distribution $\mathbb{P}\{V > x\}$. As mentioned in Section 6.1, asymptotic bounds in Dumas & Simonian [120] show a sharp dichotomy in the qualitative tail behavior, depending on the value of $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2}$ (i.e. the mean rate of the light-tailed sources plus the peak rate of the heavy-tailed sources) relative to the link rate. In case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$, the workload distribution has heavy-tailed characteristics (to be treated in Chapter 7), whereas $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$ implies light-tailed characteristics. In this section we determine the exact asymptotics of $\mathbb{P}\{V > x\}$ in the latter case. Results for the boundary case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} = 1$ can be found in Zwart [290].

To put the main result of this chapter in perspective, we first provide a heuristic derivation of the tail behavior of $\mathbb{P}\{V > x\}$ in the case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$. Large-deviations theory suggests that, given that a 'rare event' occurs, with overwhelming probability 'it happens in the most likely way'. In the asymptotic regime considered here ('large buffers'), the most likely way usually consists of a linear build-up of the workload, due to temporary instability of the system. In case of heavy-tailed distributions, the temporary instability typically arises from a 'minimal set' of potential causes. The minimal set corresponds to the minimal *number* of causes when these are homogeneous in nature. In general however, when the potential causes have heterogeneous characteristics, not only the number of them matters, but also their relative likelihood, and their relative contribution to the occurrence of the rare event under consideration.
Translated to our situation, temporary instability is most likely caused by a 'minimal set' of sources generating an extreme amount of traffic, while all other sources show roughly average behavior. These considerations give rise to the following characterization of the tail behavior of $\mathbb{P}\{V > x\}$:

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_{S^*}^{c_{S^*}} > x\},$$

with $S^*$ representing the 'minimal set', and $c_{S^*} := 1 - \rho_{\mathcal{I}\setminus S^*}$ the service rate subtracted by the aggregate traffic intensity of all other sources. In the next chapter, we will provide a proof of the above equivalence relation, and determine the asymptotic behavior of $\mathbb{P}\{V > x\}$ as $x \to \infty$.

We now turn to the case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$. Before formulating our main theorem, we first provide a heuristic derivation of the tail behavior of $\mathbb{P}\{V > x\}$. The overflow scenario described above for the case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$ cannot occur, and now the light-tailed sources too must deviate from their 'normal' behavior in order for the queue to grow. Specifically, large-deviations results suggest that the light-tailed sources must behave as if their aggregate traffic intensity is temporarily increased from $\rho_{\mathcal{I}_1}$ to, say, $\hat{\rho}_{\mathcal{I}_1}$. During that time period, all heavy-tailed sources constantly send at their peak rate, leaving capacity

$1 - r_{\mathcal{I}_2}$ for the sources in $\mathcal{I}_1$. (Notice that, for a given workload level to be reached, any alternative behavior of the sources in $\mathcal{I}_2$ would have to be compensated for by the sources in $\mathcal{I}_1$ showing even greater anomalous activity.)
To summarize, our claim is as follows: a large workload level $x$ occurs as a consequence of two rare events:

1. The sources in $\mathcal{I}_1$ show similar 'abnormal' behavior as is the typical cause of overflow when served in isolation, thus behaving as if their aggregate traffic intensity is increased from $\rho_{\mathcal{I}_1}$ to $\hat{\rho}_{\mathcal{I}_1}$ for a period of time $x/(\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1)$.

2. During that time period, all sources in $\mathcal{I}_2$ constantly send at their peak rate, leaving capacity $1 - r_{\mathcal{I}_2}$ for the sources in $\mathcal{I}_1$.

These considerations lead to the following asymptotic characterization of $\mathbb{P}\{V > x\}$:

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\} \prod_{j \in \mathcal{I}_2} p_j \mathbb{P}\{A_j^r > \frac{x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}\}. \tag{3.1}$$

Thus, the workload distribution is asymptotically equivalent to that in a reduced system, but now multiplied with a pre-factor. The reduced system consists of only the *light*-tailed sources, served at the link rate reduced by the *peak* rate of the *heavy*-tailed sources. The pre-factor essentially represents the probability that the heavy-tailed sources have sent at their peak rate long enough for the light-tailed sources to be able to cause overflow. The conjunction of light-tailed and heavy-tailed large deviations is reminiscent of that for the $M/G/2$ queue with heterogeneous servers as described in Section 2.3.

To prove (3.1), we now give two preliminary results, which may be of independent interest. The proofs are given in Appendices 6.A and 6.B.

**Proposition 6.3.1** *If Assumption 6.2.1 holds, then, for any $\alpha > 0$,*

$$\liminf_{x \to \infty} \frac{\mathbb{P}\{V_{\mathcal{I}_1}^c(\frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1} - c}) > x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}} = 1, \tag{3.2}$$

*where $\hat{\rho}_{\mathcal{I}_1} := \phi_c'(\theta^*(c)) + c$.*

**Proposition 6.3.2** *If Assumption 6.2.1 holds, then*

$$\theta^*(c + \epsilon) = \theta^*(c) + \epsilon \frac{\theta^*(c)}{\hat{\rho}_{\mathcal{I}_1} - c} + \mathrm{o}(\epsilon), \qquad \epsilon \downarrow 0,$$

*where $\hat{\rho}_{\mathcal{I}_1}$ is the same as in Proposition 6.3.1.*
*In particular, for any $\alpha > 0$, there exists an $\epsilon_\alpha > 0$ such that*

$$\limsup_{x \to \infty} \frac{x^\beta \mathbb{P}\{V_{\mathcal{I}_1}^{c+\epsilon} > \frac{\hat{\rho}_{\mathcal{I}_1} - c - \epsilon(1-\alpha)}{\hat{\rho}_{\mathcal{I}_1} - c} x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}} = 0 \tag{3.3}$$

*for all $\epsilon \in (0, \epsilon_\alpha)$ and $\beta > 0$.*

The first proposition is related to the folk theorem that a large workload level in the large-buffer regime is due to a temporary change in the traffic intensity from $\rho_{\mathcal{I}_1}$ to $\hat{\rho}_{\mathcal{I}_1}$. The second proposition will be used to show that the two rare events mentioned above are the only contributing factors to the tail distribution of the workload.

We now state our main theorem. We note that the result actually holds for any light-tailed input process for which (3.2), (3.3) are satisfied.

**Theorem 6.3.1** *(Reduced-peak equivalence)*
*Suppose that the input process $A_{\mathcal{I}_1}(s,t)$ satisfies Assumption 6.2.1. If $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$ and $A_j^r(\cdot) \in \mathcal{IRV}$ for all $j \in \mathcal{I}_2$, then*

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\} \prod_{j\in\mathcal{I}_2} p_j \mathbb{P}\{A_j^r > \frac{x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}\}.$$

**Proof**
The proof consists of the derivation of a lower bound and an upper bound which asymptotically coincide.
We start with the lower bound. For any $\alpha > 0$, we have

$$\begin{aligned}
\mathbb{P}\{V > x\} &= \mathbb{P}\{\sup_{t\geq 0}\{A(-t,0) - t\} > x\} \\
&\geq \mathbb{P}\{\sup_{0\leq t\leq \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}} \{A(-t,0) - t\} > x\} \\
&\geq \mathbb{P}\{\sup_{0\leq t\leq \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}} \{A_{\mathcal{I}_1}(-t,0) - (1-r_{\mathcal{I}_2})t\} > x, \\
&\qquad A_{\mathcal{I}_2}(-u,0) \geq r_{\mathcal{I}_2} u \text{ for all } u \in \left[0, \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}\right]\} \\
&= \mathbb{P}\{\sup_{0\leq t\leq \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}} \{A_{\mathcal{I}_1}(-t,0) - (1-r_{\mathcal{I}_2})t\} > x, \\
&\qquad A_j(-u,0) \geq r_j u \text{ for all } u \in \left[0, \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}\right], j \in \mathcal{I}_2\} \\
&= \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}(\frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}) > x\} \prod_{j\in\mathcal{I}_2} p_j \mathbb{P}\{A_j^r > \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1}\}.
\end{aligned}$$

Thus,

$$\frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\} \prod_{j\in\mathcal{I}_2} p_j \mathbb{P}\{A_j^r > \frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \geq$$

$$\frac{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}(\frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}) > x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\}} \prod_{j\in\mathcal{I}_2} \frac{\mathbb{P}\{A_j^r > \frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}{\mathbb{P}\{A_j^r > \frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}.$$

Using Proposition 6.3.1, we find that

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V>x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2} p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \geq \prod_{j\in\mathcal{I}_2} \liminf_{x\to\infty} \frac{\mathbb{P}\{A_j^r>\frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}{\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} =$$

$$\prod_{j\in\mathcal{I}_2} \liminf_{x\to\infty} \frac{\mathbb{P}\{A_j^r>(1+\alpha)x\}}{\mathbb{P}\{A_j^r>x\}}.$$

Letting $\alpha\downarrow 0$ and using the fact that $A_j^r(\cdot)\in\mathcal{IRV}$ for all $j\in\mathcal{I}_2$ then completes the proof of the lower bound.

We now turn to the upper bound. Notice that $V$ is stochastically smaller (in fact sample path wise) than $V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}$ as well as $V_{\mathcal{I}_1\cup\{j\}}^{1-r_{\mathcal{I}_2}+r_j}$ for all $j\in\mathcal{I}_2$. The latter random variable can be dominated by $V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}+V_j^{r_j-\epsilon}$. Hence,

$$\begin{aligned}
\mathbb{P}\{V>x\} &\leq \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x, V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}+V_j^{r_j-\epsilon}>x \text{ for all } j\in\mathcal{I}_2\}\\
&\leq \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x, V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}>(1-\delta)x \text{ or } V_j^{r_j-\epsilon}>\delta x \text{ for all } j\in\mathcal{I}_2\}\\
&\leq \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}>(1-\delta)x \text{ or } V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x, V_j^{r_j-\epsilon}>\delta x \text{ for all } j\in\mathcal{I}_2\}\\
&\leq \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}>(1-\delta)x\}+\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2}\mathbb{P}\{V_j^{r_j-\epsilon}>\delta x\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{\mathbb{P}\{V>x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2} p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}\\
\leq\ &\frac{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}>(1-\delta)x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2} p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}\\
+\ &\prod_{j\in\mathcal{I}_2}\frac{\mathbb{P}\{V_j^{r_j-\epsilon}>\delta x\}}{p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}.
\end{aligned}$$

Now take $\delta=\frac{\epsilon(1-\alpha)}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}$.

$$\begin{aligned}
&\frac{\mathbb{P}\{V>x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2} p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}\\
\leq\ &\frac{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon}>\frac{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1-\epsilon(1-\alpha)}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}}>x\}\prod_{j\in\mathcal{I}_2} p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}\\
+\ &\prod_{j\in\mathcal{I}_2}\frac{\mathbb{P}\{V_j^{r_j-\epsilon}>\frac{\epsilon(1-\alpha)}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}x\}}{p_j\mathbb{P}\{A_j^r>\frac{x}{\hat{\rho}_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}.
\end{aligned}$$

Because $A_j^r(\cdot) \in \mathcal{IRV}$ for all $j \in \mathcal{I}_2$, there exists a $\beta$ such that

$$\lim_{x\to\infty} x^\beta \prod_{j\in\mathcal{I}_2} \mathbb{P}\{A_j^r > x\} = \infty. \tag{3.4}$$

Using Theorem 6.2.1 we get

$$\begin{aligned}
& \limsup_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\} \prod_{j\in\mathcal{I}_2} p_j \mathbb{P}\{A_j^r > \frac{x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \\
\leq\ & \frac{1}{\prod_{j\in\mathcal{I}_2} p_j} \prod_{j\in\mathcal{I}_2} \limsup_{x\to\infty} \frac{\mathbb{P}\{A_j^r > x\}}{\mathbb{P}\{A_j^r > \frac{x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \limsup_{x\to\infty} \frac{1}{x^\beta \prod_{j\in\mathcal{I}_2} \mathbb{P}\{A_j^r > x\}} \times \\
& \limsup_{x\to\infty} \frac{x^\beta \mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}+\epsilon} > \frac{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1-\epsilon(1-\alpha)}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1} x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^{1-r_{\mathcal{I}_2}} > x\}} \\
+\ & \prod_{j\in\mathcal{I}_2} \limsup_{x\to\infty} \frac{\mathbb{P}\{V_j^{r_j-\epsilon} > \frac{\epsilon(1-\alpha)}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1} x\}}{p_j \mathbb{P}\{A_j^r > \frac{x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}.
\end{aligned}$$

The first term is seen to converge to zero by using the fact that $A_j^r(\cdot) \in \mathcal{IRV}$ for all $j \in \mathcal{I}_2$, Equation (3.4), and Proposition 6.3.2. The second term equals, by Theorem 6.2.1,

$$\begin{aligned}
& \prod_{j\in\mathcal{I}_2} \limsup_{x\to\infty} \frac{p_j \frac{r_j-\rho_j}{r_j-\epsilon-\rho_j} \mathbb{P}\{A_j^r > \frac{(1-\alpha)x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}{p_j \mathbb{P}\{A_j^r > \frac{x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \\
=\ & \prod_{j\in\mathcal{I}_2} \frac{r_j-\rho_j}{r_j-\epsilon-\rho_j} \limsup_{x\to\infty} \frac{\mathbb{P}\{A_j^r > \frac{(1-\alpha)x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}}{\mathbb{P}\{A_j^r > \frac{x}{\hat\rho_{\mathcal{I}_1}+r_{\mathcal{I}_2}-1}\}} \\
=\ & \prod_{j\in\mathcal{I}_2} \frac{r_j-\rho_j}{r_j-\epsilon-\rho_j} \limsup_{x\to\infty} \frac{\mathbb{P}\{A_j^r > (1-\alpha)x\}}{\mathbb{P}\{A_j^r > x\}}.
\end{aligned}$$

Letting $\epsilon \downarrow 0$ and then $\alpha \downarrow 0$ and using the fact that $A_j^r(\cdot) \in \mathcal{IRV}$ for all $j \in \mathcal{I}_2$ then completes the proof of the upper bound. □

## 6.4 Examples

We now apply Theorem 6.3.1 to obtain a complete characterization of the tail behavior of the workload distribution $\mathbb{P}\{V > x\}$ for two important traffic scenarios for the light-tailed sources: (i) Markov-modulated fluid input; (ii) instantaneous input.

### 6.4.1 Markov-modulated fluid input

In this subsection we check that Assumption 6.2.1 is satisfied in case the light-tailed sources are Markovian On-Off sources. We follow Asmussen [20], and assume that the input process $A_{\mathcal{I}_1}(s,t)$ can be represented as follows. Let $\mathbf{J}(t)$ be an irreducible continuous-time Markov process on a finite state space $\mathcal{J}$ with $Q$-matrix $\Lambda$. $\mathbf{J}(t)$ converges in distribution to the random variable $\mathbf{J}$. If $\mathbf{J}(t) = j$, then $A_{\mathcal{I}_1}(t, t + \mathrm{d}t) = r_j \mathrm{d}t$. Thus,

$$A_{\mathcal{I}_1}(s,t) = \int_s^t r_{\mathbf{J}(u)} \mathrm{d}u.$$

We introduce some additional notation, following [20]. Define the matrix polynomial

$$K_c(s) = \Lambda + s(R - cI),$$

where $R$ is a diagonal matrix with elements $r_j$, and $I$ is the identity matrix. $K_c(s)$ has a simple and unique eigenvalue with maximal real part. Denote this eigenvalue by $\kappa_c(s)$. A simple computation shows that $\phi_c(s) = \kappa_c(s)$. From [20] we know that the equation $\kappa_c(s) = 0$ has a unique solution $\theta^*(c) > 0$ and that all other conditions of Assumption 6.2.1 are satisfied as well.
In this special case, the *exact* asymptotics of $\mathbb{P}\{V^c_{\mathcal{I}_1} > x\}$ are available: Corollary 4.9 in [20] yields

$$\mathbb{P}\{V^c_{\mathcal{I}_1} > x\} \sim D^c \mathrm{e}^{-\theta^*(c)x}. \tag{4.1}$$

An explicit, but quite elaborate expression for the pre-factor $D^c$ may be found in [20].

Together, Theorem 6.3.1 and Equation (4.1) provide a complete characterization of the tail behavior of $\mathbb{P}\{V > x\}$. As mentioned above, the pre-factor $D^c$ is quite complicated in general. However, that is not the case when the input process $A_{\mathcal{I}_1}(s,t)$ is the superposition of several statistically identical On-Off sources with exponentially distributed On- and Off-periods, see Anick, Mitra & Sondhi [16].

**Example 6.4.1** As an illustrating example, consider the following special case of two On-Off sources. Source 1 has exponentially distributed On- and Off-periods with parameters $\mu$ and $\lambda$, respectively. While On, the source generates traffic at rate $r_1$, so that $\rho_1 = \lambda r_1/(\lambda + \mu)$. Source 2 has On-periods which are regularly varying of index $-\nu < -1$, i.e., $\mathbb{P}\{A_2 > x\} = L(x)x^{-\nu}$, with $L(\cdot)$ a slowly varying function. Thus $\mathbb{P}\{A^r_2 > x\} \sim 1/((\nu-1)\mathbb{E}\{A_2\})L(x)x^{1-\nu}$. Some calculations show that for $\rho_1 < c$,

$$\begin{aligned} \mathbb{P}\{V^c_1 > x\} &= \frac{\lambda}{\lambda+\mu}\frac{r_1}{c}\exp\left\{-\left(\frac{\mu}{r_1-c} - \frac{\lambda}{c}\right)x\right\}, \\ \hat{\rho}_1 &= \frac{\mu}{\mu + \lambda\left(\frac{r_1-c}{c}\right)^2} r_1. \end{aligned}$$

Taking $c = 1 - r_2$, Theorem 6.3.1 implies that for $\rho_1 + r_2 < 1$ (see also Chapter 7),

$$\begin{aligned} & \mathbb{P}\{V > x\} \\ \sim \; & \frac{\lambda}{\lambda+\mu}\frac{r_1}{1-r_2}\frac{p_2}{(\nu-1)\mathbb{E}\{A_2\}}L(x)\left(\frac{x}{\hat{\rho}_1 + r_2 - 1}\right)^{1-\nu} \times \\ & \exp\left\{-\left(\frac{\lambda}{1-r_2} - \frac{\mu}{r_1+r_2-1}\right)x\right\}. \end{aligned}$$

In contrast, reduced-load equivalence (see Theorem 2.2.4 and Chapter 7), combined with Theorem 6.2.1, gives for $\rho_1 + r_2 > 1 > \rho_1 + \rho_2$,

$$\begin{aligned} \mathbb{P}\{V > x\} \; & \sim \; \mathbb{P}\{V_2^{1-\rho_1} > x\} \\ & \sim \; \frac{r_2 - \rho_2}{1-\rho_1-\rho_2}\frac{p_2}{(\nu-1)\mathbb{E}\{A_2\}}L(x)\left(\frac{x}{\rho_1+r_2-1}\right)^{1-\nu}. \end{aligned} \tag{4.2}$$

### 6.4.2 Instantaneous input

In this subsection we assume that the input process of the light-tailed sources is that of a $GI/G/1$ queue. Observe that in terms of total workload, the model may equivalently be viewed as a $GI/G/1$ queue with several service speeds (depending on which of the heavy-tailed sources are active). The assumption $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$ implies that the queue is stable, even when served at the lowest possible speed $1 - r_{\mathcal{I}_2}$. We refer to Boxma & Kurkova [73] for related results.

Instead of showing the validity of Assumption 6.2.1, we take a more direct approach and use results from Asmussen [18] to show that (3.2), (3.3) hold (which is sufficient for Theorem 6.3.1 to hold). If one wishes to stay within the general large-deviations framework, one should invoke additional regularity conditions, in particular Equations (1.23)–(1.26) in [138].

We assume i.i.d. interarrival times $T_n$ and i.i.d. service times $B_n$, $n = 1, 2, \ldots$ . We follow [18], and impose the following two technical conditions:

1. The distribution of $B_1 - cT_1$ is non-lattice,
2. There exists a $\theta^*(c) > 0$ such that $\mathbb{E}\{\mathrm{e}^{\theta^*(c)(B_1 - cT_1)}\} = 1$ and $\mathbb{E}\{|B_1 - cT_1|\,\mathrm{e}^{\theta^*(c)(B_1-cT_1)}\} < \infty$.

Let $\alpha(\cdot)$, $\beta(\cdot)$ be the LSTs of $\mathbf{T}_1$, $\mathbf{B}_1$, respectively. We define the 'twisted' (also called associated, cf. [18]) random variables $\hat{T}$ and $\hat{B}$ through their transforms

$$\mathbb{E}\{\mathrm{e}^{-s\hat{T}}\} = \hat{\alpha}(s) = \frac{\alpha(s + c\theta^*(c))}{\alpha(c\theta^*(c))}, \qquad \mathbb{E}\{\mathrm{e}^{-s\hat{B}}\} = \hat{\beta}(s) = \frac{\beta(s - \theta^*(c))}{\beta(-\theta^*(c))}.$$

Like in the previous subsection, it is possible to refine the logarithmic asymptotics in (2.1): The exact asymptotic behavior of $\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}$ is given by

$$\mathbb{P}\{V_{\mathcal{I}_1}^c > x\} \sim D^c \mathrm{e}^{-\theta^*(c)x}. \tag{4.3}$$

An expression for the pre-factor $D^c$ is specified on page 158 of [18].

We now show that (3.2), (3.3) are satisfied with the definition $\hat{\rho}_{\mathcal{I}_1} := \frac{\mathbb{E}\{\hat{B}\}}{\mathbb{E}\{\hat{T}\}}$. Equation (3.2) is a direct consequence of Theorem 6.2 in [18]. To check (3.3), we compute the derivative of $\theta^*(c)$ using the implicit function theorem. A straightforward computation yields

$$\frac{\mathrm{d}}{\mathrm{d}c}\theta^*(c) = \frac{\theta^*(c)}{\hat{\rho}_{\mathcal{I}_1} - c}.$$

This yields (3.3), see Appendix 6.B.
Together, Theorem 6.3.1 and Equation (4.3) determine the exact asymptotics of $\mathbb{P}\{V > x\}$.

**Example 6.4.2** To illustrate our results, consider the following example with two sources. The traffic model of source 1 is that of an $M/M/1$ queue with arrival rate $\lambda$ and service rate $\mu$, so that $\rho_1 = \lambda/\mu$. Source 2 is an On-Off source with regularly varying On-periods of index $-\nu < -1$, i.e., $\mathbb{P}\{A_2 > x\} = L(x)x^{-\nu}$, with $L(x)$ a slowly-varying function. As mentioned in the beginning of the subsection, in terms of total workload, the model may be viewed as an $M/M/1$ queue with two service speeds, $c_1 = 1$ and $c_2 = 1 - r_2$, regulated by the activity of source 2. For the ordinary $M/M/1$ queue we have, for any $c > \rho_1$,

$$\begin{aligned} \mathbb{P}\{V_1^c > x\} &= \frac{\lambda}{c\mu}\mathrm{e}^{-(\mu - \frac{\lambda}{c})x}, \\ \hat{\rho}_1 &= \frac{c\mu}{\lambda}. \end{aligned}$$

Taking $c = c_2 - 1 - r_2$, Theorem 6.3.1 yields for $\rho_1 < c_2$,

$$\mathbb{P}\{V > x\} \sim \frac{\lambda}{c_2\mu}\frac{p_2}{(\nu-1)\mathbb{E}\{A_2\}}L(x)\left(\frac{x}{\hat{\rho}_1 - c_2}\right)^{1-\nu}\mathrm{e}^{-(\mu - \frac{\lambda}{c_2})x}.$$

For $\rho_1 > c_2$, the tail behavior is identical to that when source 1 is an On-Off source with mean rate $\rho_1$ as given in Equation (4.2).

# Appendix

## 6.A Proof of Proposition 6.3.1

**Proposition 6.3.1** *If Assumption 6.2.1 is satisfied, then, for any $\alpha > 0$,*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_{\mathcal{I}_1}^c(\frac{(1+\alpha)x}{\hat{\rho}_{\mathcal{I}_1}-c}) > x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}} = 1,$$

*where $\hat{\rho}_{\mathcal{I}_1} := \phi_c'(\theta^*(c)) + c$.*

**Proof**
Note that $V_{\mathcal{I}_1}^c(t)$ has the same distribution as $\sup_{0\le s\le t}\{\bar{A}_{\mathcal{I}_1}(0,s) - cs\}$, with $\bar{A}_{\mathcal{I}_1}$ the time-reversed version of $A_{\mathcal{I}_1}$, i.e., $\bar{A}_{\mathcal{I}_1}(s,t) = A_{\mathcal{I}_1}(-t,-s)$. Define $\tau(x) := \inf\{\bar{A}_{\mathcal{I}_1}(0,t) - ct \ge x\}$. For integer $i$ and $n$, we define $X_i := \bar{A}_{\mathcal{I}_1}(i,i+1) - c$ and $S_n := X_1 + \ldots + X_n = \bar{A}_{\mathcal{I}_1}(0,n) - cn$. Following [138], we define the 'twisted' probability measures $\mathbb{P}_n^*\{\cdot\}$ by

$$\mathbb{P}_n^*\{\mathrm{d}x_1,\ldots,dx_n\} := \mathrm{e}^{\theta^*\sum_{i=1}^n x_i - \phi_n(\theta^*)}\mathbb{P}\{\mathrm{d}x_1,\ldots,\mathrm{d}x_n\},$$

where $\phi_n(\theta) = \log\mathbb{E}\{\exp\{\theta S_n\}\}$. Note that $\mathbb{P}_n^*$ and $\hat{\rho}_{\mathcal{I}_1}$ are independent of the system capacity $c$. To prove the proposition, we use similar arguments as in the proof of Theorem 2 of [138]. It suffices to show that

$$\mathbb{P}\{\infty > \tau(x) > x(1+\alpha)/(\hat{\rho}_{\mathcal{I}_1}-c)\} = \mathrm{o}(\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}), \qquad x\to\infty.$$

Define $m(x) = [x(1+\alpha)/(\hat{\rho}_{\mathcal{I}_1}-c)]$ (with $[y]$ the entier of $y$), and write

$$\begin{aligned}\mathbb{P}\{\infty > \tau(x) > x(1+\alpha)/(\hat{\rho}_{\mathcal{I}_1}-c)\} &\le \sum_{j=m(x)}^{\infty}\mathbb{P}\{j-1\le\tau(x)<j\}\\ &\le \sum_{j=m(x)}^{\infty}\mathbb{P}\{S_{j-1}\le x, S_j > x-c\}.\end{aligned}$$

We need some auxiliary results which are also stated in the proof of Theorem 4 in [138]. The following bounds are valid for some $\eta < 1$ when $x$ and $j$ are large enough:

$$\phi_j(\theta^*) < -\frac{1}{2}j\log\eta,$$

$$\mathbb{P}_j^*\{S_{j-1}\le x\} \le \eta^j, \qquad j\ge m(x).$$

Both bounds rely on Theorem 7 of [138], which basically shows that the speed of convergence of $S_n/n$ is exponentially fast under $\mathbb{P}_n^*\{\cdot\}$. The first bound is Equation (2.6)

in [138], while the second bound is derived on page 147 of [138]. From these bounds, we obtain

$$\begin{aligned} & \mathbb{P}\{S_{j-1} \leq x, S_j > x - c\} \\ = \; & \mathrm{E}_j^*\{\exp\{-\theta^* S_j + \phi_j(\theta^*)\}; S_{j-1} \leq x; S_j > x - c\} \\ \leq \; & \mathrm{e}^{\theta^* c}\mathrm{e}^{-\theta^* x}\mathrm{e}^{\phi_j(\theta^*)}\mathbb{P}_j^*\{S_{j-1} \leq x\} \\ \leq \; & \mathrm{e}^{\theta^* c}\mathrm{e}^{-\theta^* x}(\sqrt{\eta})^j. \end{aligned}$$

Combining all results, we obtain, for some finite constant $C$,

$$\mathbb{P}\{\infty > \tau(x) > x(1+\alpha)/(\hat{\rho}_{\mathcal{I}_1} - c)\} \leq C\mathrm{e}^{-\theta^* x}(\sqrt{\eta})^{m(x)},$$

which is negligible compared to $\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}$ according to Equation (2.1).

□

# 6.B Proof of Proposition 6.3.2

**Proposition 6.3.2** *If Assumption 6.2.1 is satisfied, then*

$$\theta^*(c+\epsilon) = \theta^*(c) + \epsilon\frac{\theta^*(c)}{\hat{\rho}_{\mathcal{I}_1} - c} + \mathrm{o}(\epsilon), \qquad \epsilon \downarrow 0,$$

*where $\hat{\rho}_{\mathcal{I}_1}$ is the same as in Proposition 6.3.1.*
*In particular, for any $\alpha > 0$, there exists an $\epsilon_\alpha > 0$ such that*

$$\limsup_{x\to\infty} \frac{x^\beta \mathbb{P}\{V_{\mathcal{I}_1}^{c+\epsilon} > \frac{\hat{\rho}_{\mathcal{I}_1} - c - \epsilon(1-\alpha)}{\hat{\rho}_{\mathcal{I}_1} - c}x\}}{\mathbb{P}\{V_{\mathcal{I}_1}^c > x\}} = 0$$

*for all $\epsilon \in (0, \epsilon_\alpha)$ and $\beta > 0$.*

**Proof**
First, we show that

$$\theta^*(c+\epsilon) = \theta^*(c) + \epsilon\frac{\theta^*(c)}{\hat{\rho}_{\mathcal{I}_1} - c} + \mathrm{o}(\epsilon), \qquad \epsilon \downarrow 0.$$

Using Taylor's theorem, it suffices to compute the derivative of $\theta^*(c)$ w.r.t. $c$. Note that

$$\phi_{c+\epsilon}(\theta) = \phi_c(\theta) - \epsilon\theta.$$

Combining this with the implicit function theorem, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}c}\theta^*(c) = -\frac{\frac{\mathrm{d}}{\mathrm{d}c}\phi_c(s)|_{s=\theta^*(c)}}{\frac{\mathrm{d}}{\mathrm{d}s}\phi_c(s)|_{s=\theta^*(c)}} = \frac{\theta^*(c)}{\phi_c'(\theta^*(c))}.$$

Finally, we prove the second part of Proposition 6.3.2. We may write

$$\mathbb{P}\{V^c_{\mathcal{I}_1} > x\} = f_c(x)\mathrm{e}^{-\theta^*(c)x},$$

with $f_c(x) = \mathrm{o}(\mathrm{e}^{\delta x})$ and $1/f_c(x) = \mathrm{o}(\mathrm{e}^{\delta x})$ for any $\delta > 0$ and $x \to \infty$. We obtain, for any fixed $\alpha$ and $\epsilon$ small enough,

$$\begin{aligned}
& \limsup_{x\to\infty} \frac{x^\beta \mathbb{P}\{V^{c+\epsilon}_{\mathcal{I}_1} > \frac{\hat\rho_{\mathcal{I}_1} - c - \epsilon(1-\alpha)}{\hat\rho_{\mathcal{I}_1} - c} x\}}{\mathbb{P}\{V^c_{\mathcal{I}_1} > x\}} \\
= & \limsup_{x\to\infty} \frac{f_{c+\epsilon}(x)}{f_c(x)} \frac{x^\beta \mathrm{e}^{-\theta^*(c+\epsilon)\frac{\hat\rho_{\mathcal{I}_1} - c - \epsilon(1-\alpha)}{\hat\rho_{\mathcal{I}_1} - c} x}}{\mathrm{e}^{-\theta^*(c)x}} \\
= & \limsup_{x\to\infty} \frac{f_{c+\epsilon}(x)}{f_c(x)} \frac{x^\beta \mathrm{e}^{-\theta^*(c)x(1+\frac{\epsilon}{\hat\rho_{\mathcal{I}_1} - c} + \mathrm{o}(\epsilon))(1-\frac{\epsilon(1-\alpha)}{\hat\rho_{\mathcal{I}_1} - c})}}{\mathrm{e}^{-\theta^*(c)x}} \\
= & \limsup_{x\to\infty} \frac{f_{c+\epsilon}(x)}{f_c(x)} x^\beta \mathrm{e}^{-\theta^*(c)x(\frac{\epsilon\alpha}{\hat\rho_{\mathcal{I}_1} - c} + \mathrm{o}(\epsilon))} = 0.
\end{aligned}$$

□

# Chapter 7

# The fluid queue II: Reduced-load

## 7.1 Introduction

In this chapter we revisit the fluid queue introduced in the previous chapter. As already mentioned there, asymptotic bounds in Dumas & Simonian [120] show a sharp dichotomy in the qualitative tail behavior of the workload distribution, depending on whether the mean rate of the light-tailed sources plus the peak rate of the heavy-tailed sources exceeds the link rate or not. In case the link rate is larger, the workload distribution has light-tailed characteristics (see Chapter 6), whereas the link rate being smaller results in heavy-tailed characteristics. The latter case will be studied in the present chapter.

The bounds in [120] indicate that one can usually identify a 'dominant' set, which is a minimal set of sources that can cause a positive drift in the buffer. As far as bounds is concerned, all other sources can essentially be accounted for by subtracting their aggregate mean rate from the link rate. Exact asymptotics however, have remained elusive for all but a few special cases. Results of Agrawal *et al.* [12] show that the dominance principle described above in fact extends to the exact asymptotics in the case of a *single* dominant source. This may be expressed in terms of a 'reduced-load equivalence', implying that the workload is asymptotically equivalent to that in a reduced system. The reduced system consists only of the dominant source, with the link rate subtracted by the aggregate mean rate of all other sources, see Subsection 2.2.2 for a more elaborate discussion and further references. This extends results of Boxma [65], Jelenković & Lazar [161], and Rolski *et al.* [243] for multiplexing a single (intermediately) regularly varying source with several exponential sources.

In the present chapter we determine the exact asymptotics for the case where *several* On-Off sources must be active for the buffer to fill (under the assumption that the distribution of the On-periods is regularly varying). From a practical perspective, this case appears particularly relevant, as the peak rate of a single source is usually substantially smaller than the link rate. However, the rather subtle interaction of several sources that is involved in filling the buffer drastically complicates the analysis. We start with extending the

reduced-load equivalence to the case of a reduced system consisting of several sources, using sample-path arguments. We then build on a qualitative understanding of the large-deviations behavior to obtain the exact asymptotics for the reduced system. A stylized version of our approach for the $M/G/1$ queue can be found in Section 2.4.
The remainder of the chapter is organized as follows. In Section 7.2, we present a detailed model description. In Section 7.3, we give a broad overview of the main results of the chapter, and describe how the dominant set may be determined from a simple knapsack formulation. Section 7.4 gives some preliminary results. The reduced-load equivalence result is established in Section 7.5. Section 7.6 develops the detailed probabilistic arguments involved in deriving the tail asymptotics for the reduced system. In Section 7.7, we discuss the relationship between the asymptotic regime considered here ('large buffers') and a many-sources regime.

## 7.2 Preliminaries

The model under consideration is similar to the model introduced in Chapter 6. Again, we assume that the sources may be partitioned into two sets: $\mathcal{I}_1$ is the set of 'light-tailed' sources; $\mathcal{I}_2$ is the set of 'heavy-tailed' sources. The precise assumptions on these sets are somewhat different from those in the previous chapter. For the sources $i \in \mathcal{I}_1$ we make the following assumption.

**Assumption 7.2.1** *For any $c > \rho_{\mathcal{I}_1}$, $\mu > 0$,*

$$\lim_{x\to\infty} x^{\mu}\mathbb{P}\{V_{\mathcal{I}_1}^{c} > x\} = 0.$$

The above assumption is quite weak; it is satisfied by the light-tailed input considered in the previous chapter. However, (superpositions of) On-Off sources of which the activity period has a Weibull distribution satisfy Assumption 7.2.1 too. Instantaneous renewal input of which the tail of the jump sizes (bursts) is lighter than any power tail is covered by Assumption 7.2.1 as well.

We assume that the sources in $\mathcal{I}_2$ generate traffic according to independent On-Off processes (which is a stronger assumption than made in the previous chapter, where we considered semi-Markov sources). The Off-periods of source $i$ are generally distributed with mean $1/\lambda_i$. The On-periods $A_i$ have a heavy-tailed distribution $A_i(\cdot)$ with mean $\alpha_i < \infty$. While On, source $i$ produces traffic at constant rate $r_i$, so the mean burst size is $\alpha_i r_i$. The fraction of time that source $i$ is On is

$$p_i = \frac{\alpha_i}{1/\lambda_i + \alpha_i} = \frac{\lambda_i\alpha_i}{1 + \lambda_i\alpha_i}.$$

Thus the traffic intensity of source $i$ is

$$\rho_i := p_i r_i = \frac{\lambda_i \alpha_i r_i}{1 + \lambda_i \alpha_i}.$$

For each source $i \in \mathcal{I}_2$, we assume that the On-period distribution is regularly varying of index $-\nu_i$, i.e., $A_i(\cdot) \in \mathcal{R}_{-\nu_i}$ for some $\nu_i > 1$.

We now give a convenient representation for the stationary workload $V_E^c$, with $E \subseteq \mathcal{I}_2$ an arbitrary set of heavy-tailed On-Off sources. We start from the definition $V_E^c(t) := \sup_{0 \le s \le t} \{A_E(s,t) - c(t-s)\}$ (assuming $V_E^c(0) = 0$), see also Chapter 6. Since the process $A_E(\cdot,\cdot)$ has stationary and reversible increments, we have

$$\sup_{0 \le s \le t} \{A_E(s,t) - c(t-s)\} \stackrel{d}{=} \sup_{0 \le s \le t} \{A_E(0,s) - cs\}.$$

In the sequel, we simply use the latter expression as the *definition* of $V_E^c(t)$. Accordingly, for $c > \rho_E$, the stationary workload as $t \to \infty$ may be represented as

$$V_E^c := \sup_{t \ge 0} \{A_E(0,t) - ct\}.$$

Recall (see Chapter 6) that $V(t) := V_{\mathcal{I}}^1(t)$ is the total workload at time $t$, and $V$ is a random variable with the limiting distribution of $V(t)$ for $t \to \infty$ (like in Chapter 6, we assume that $\rho_{\mathcal{I}} = \rho < 1$).

Explicit constructions of $A_i(0,t)$ (satisfying the stationarity condition) may be found in Dumas & Simonian [120] and Heath *et al.* [152]. For completeness, we review the construction in [152] which will be extensively used in Section 7.6.

Let $\{A_{im}, m \ge 0\}$ be a sequence of i.i.d. random variables representing On-periods of source $i$. Similarly, let $\{U_{im}, m \ge 1\}$ be Off-periods. Define three additional random variables $A_{i0}^r$, $U_{i0}^r$, and $I_i$ such that $A_{i0}^r \stackrel{d}{=} A_i^r$, $U_{i0}^r \stackrel{d}{=} U_i^r$, and

$$\mathbb{P}\{I_i = 1\} = \frac{\mathbb{E}\{A_{i1}\}}{\mathbb{E}\{A_{i1}\} + \mathbb{E}\{U_{i1}\}} = 1 - \mathbb{P}\{I_i = 0\}.$$

Note that $I_i = 1$ corresponds to source $i$ being On (in stationarity).

To obtain a stationary alternating renewal process, we define the delay random variable $D_{i0}$ by

$$D_{i0} = I_i A_{i0}^r + (1 - I_i)(U_{i0}^r + A_{i0}).$$

Then the delayed renewal sequence

$$\{Z_{in}, n \ge 0\} = \{D_{i0}, D_{i0} + \sum_{m=1}^{n} (U_{im} + A_{im}), n \ge 1\}$$

is stationary.

Next, we define the process $\{I_i(t), t \geq 0\}$ as follows. $I_i(t)$ is the indicator of the event that source $i$ is On at time $t$. Formally, we have

$$I_i(t) = I_i 1_{\{t<A^r_{i0}\}} + (1 - I_i)1_{\{U^r_{i0} \leq t < U^r_{i0}+A_{i0}\}} + \sum_{n=0}^{\infty} 1_{\{Z_{in}+U_{i,n+1} \leq t < Z_{i,n+1}\}}.$$

The On-Off process $\{J_i(t), t \geq 0\}$ is strictly stationary, see Theorem 2.1 of [152]. The process $\{A_i(0,t), t \geq 0\}$ is defined by

$$A_i(0,t) := r_i \int_0^t I_i(u)\mathrm{d}u.$$

Finally, note that the number of elapsed Off-periods during $[0,t]$ which started after time 0 is given by

$$N^A_i(t) := \max\{n : Z_{i,n-1} + U_{in} \leq t\}. \tag{2.1}$$

We conclude this section by introducing two notational conventions. With $f(x) \overset{<}{\sim} g(x)$ we denote $\limsup_{x\to\infty} f(x)/g(x) \leq 1$. Similarly, $f(x) \overset{>}{\sim} g(x)$ denotes $\liminf_{x\to\infty} f(x)/g(x) \geq 1$.

## 7.3 Overview of the results

We now give a broad overview of the main results of the chapter. As mentioned in Section 7.1, asymptotic bounds in Dumas & Simonian [120] show a sharp dichotomy in the qualitative behavior of $\mathbb{P}\{V > x\}$, depending on the value of $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2}$ (i.e. the mean rate of the light-tailed sources plus the peak rate of the heavy-tailed sources) relative to the service rate. In case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$, the workload has light-tailed characteristics, whereas $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$ implies heavy-tailed characteristics. In the present chapter we determine the exact asymptotics of $\mathbb{P}\{V > x\}$ in the latter case; see Section 6.3 for some intuitive arguments.

Before we state our main result, we first introduce some helpful notions. For any subset $S \subseteq \mathcal{I}_2$, define $c_S := 1 - \rho_{\mathcal{I}\setminus S}$ as the service rate subtracted by the aggregate traffic intensity of all other sources $j \in \mathcal{I} \setminus S$. Observe that the stability condition implies $\rho_S < c_S$ for any $S \subseteq \mathcal{I}_2$.

For any subset $S \subseteq \mathcal{I}_2$, denote by $r_S := \sum_{j\in S} r_j$ the aggregate peak rate of the sources $j \in S$. Define $d_S := r_S - c_S = r_S + \rho_{\mathcal{I}\setminus S} - 1$ as the net input rate (i.e. the drift) when all sources in $S$ are On and all other sources show average behavior.

A set $S \subseteq \mathcal{I}_2$ is called (strictly) *critical* if $d_S \geq (>)0$, i.e., if

$$r_S + \rho_{\mathcal{I}\setminus S} \geq (>)1.$$

Thus, when all sources in a (strictly) critical set are On, the workload has a (strictly) positive drift. A critical set $S$ is termed *minimally-critical* if no proper subset of $S$ is critical, i.e., $d_S < \min_{j\in S}\{r_j - \rho_j\}$.
For any subset $S \subseteq \mathcal{I}_2$, denote $\mu_S := \sum_{j\in S}(\nu_j - 1)$. A strictly critical set $S \subseteq \mathcal{I}_2$ is said to be (weakly) *dominant* if $\mu_S < (\leq)\mu_U$ for any other critical set $U \subseteq \mathcal{I}_2$. Observe that for a set $S \subseteq \mathcal{I}_2$ to be dominant, it must be minimally-critical (because otherwise the defining property would be violated for any critical subset $U \subset S$).

The quantity $\mu_S$ may be interpreted as a measure for the 'cost' associated with a temporary drift $d_S$: the probability of all sources in $S$ being On for a time of the order $x$ in steady state is roughly equal to $x^{-\mu_S}$. Thus, a set $S$ is (weakly) dominant if the sources in $S$ being On causes the drift to be positive in the cheapest possible way.
In case of light-tailed distributions, the cost minimization is usually not so simple; one then also needs to consider how long a certain positive drift must be maintained in order for a given workload level $x$ to be reached. This issue does not arise in case of regularly varying On periods, since $\mathbb{P}\{A_i^r > ax\}$ is of the same order of magnitude (up to a constant) as $\mathbb{P}\{A_i^r > x\}$ for any constant $a > 1$. This implies that the value of the temporary drift is not relevant as long as it is positive.

### 7.3.1 Tail behavior of the workload distribution

We now state our main theorem.

**Theorem 7.3.1** *(Reduced-load equivalence)*
*Suppose the set of sources $S^* \subseteq \mathcal{I}_2$ is dominant. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in \mathcal{I}_2$, then*

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}, \tag{3.1}$$

*with*

$$\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\} \sim \left(\prod_{j\in S^*} p_j\right) \sum_{\mathcal{J}_0 \subseteq S^*} P_{\mathcal{J}_0}(x), \tag{3.2}$$

*where $P_{\mathcal{J}_0}(x)$ is given by (with $\mathcal{J}_1 = S^* \setminus \mathcal{J}_0$, and $d_{S^*} = r_{S^*} - c_{S^*}$ as defined earlier)*

$$\begin{aligned} P_{\mathcal{J}_0}(x) \;=\; & \frac{1}{\prod_{i\in\mathcal{J}_1}\mathbb{E}\{A_i\}} \int_{y_i\in(0,\infty), i\in\mathcal{J}_1} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{d_{S^*}A_i > \sum_{j\in\mathcal{J}_1} y_j(r_j-\rho_j) - d_{S^*}y_i + x\} \\ & \prod_{i\in\mathcal{J}_0} \mathbb{P}\{d_{S^*}A_i^r > \sum_{j\in\mathcal{J}_1} y_j(r_j-\rho_j) + x\} \prod_{i\in\mathcal{J}_1} \mathrm{d}y_i. \end{aligned} \tag{3.3}$$

*In particular, $\mathbb{P}\{V > x\}$ and $P_{\mathcal{J}_0}(x)$ are regularly varying of index $-\mu_{S^*} = -\sum_{j\in S^*}(\nu_j - 1)$.*

The proof of the above theorem may be found in Subsection 7.5.1 (Equation (3.1)) and Section 7.6 (Equations (3.2) and (3.3) and the regular variation property).

Note that in case the reduced system consists of just a single source, i.e., $S^* = \{i^*\}$, the tail asymptotics follow directly from Theorem 2.2.3. This is in fact the reduced-load equivalence established in Agrawal *et al.* [12] (under somewhat weaker distributional assumptions), see also Section 2.2.2. Note that in this case the right hand side of (3.2) takes the form $p_{i^*}[P_\emptyset(x) + P_{i^*}(x)]$, with

$$P_{i^*}(x) = \mathbb{P}\{A^r_{i^*} > \frac{x}{r_{i^*} - c_{i^*}}\},$$

and (after a straightforward calculation)

$$P_\emptyset(x) = \frac{r_{i^*} - c_{i^*}}{c_{i^*} - \rho_{i^*}} \mathbb{P}\{A^r_{i^*} > \frac{x}{r_{i^*} - c_{i^*}}\},$$

so that

$$p_{i^*}[P_\emptyset(x) + P_{i^*}(x)] = (1 - p_{i^*}) \frac{\rho_{i^*}}{c_{i^*} - \rho_{i^*}} \mathbb{P}\{A^r_{i^*} > \frac{x}{r_{i^*} - c_{i^*}}\},$$

which is consistent with Theorem 2.2.3.
In case the reduced system consists of several sources, the tail asymptotics cannot be obtained from known results. In fact, the analysis of the reduced system then poses a major challenge because of the rather subtle mechanics involved in reaching a large workload level. By definition though, the reduced system has the special feature that all sources must be On for the drift in the workload to be positive, i.e., $r_{S^*} - \min_{j \in S^*}\{r_j - \rho_j\} < c_{S^*} < r_{S^*}$. In Section 7.6 we determine the exact asymptotics for systems satisfying this property, yielding the integral expression given in Theorem 7.3.1.

### 7.3.2 Knapsack formulation for determining a dominant set

We now describe how a dominant set may be determined from a simple knapsack formulation. Recall that the On-period distributions of the sources $i \in \mathcal{I}_2$ are regularly varying of index $-\nu_i$.
For a strictly critical set $S \subseteq \mathcal{I}_2$ to be dominant, it must necessarily solve the optimization problem

$$\begin{array}{ll} \min\limits_{S \subseteq \mathcal{I}_2} & \sum\limits_{j \in S} (\nu_j - 1) \\ \text{sub} & \sum\limits_{j \in S} r_j + \sum\limits_{j \in \mathcal{I}_2 \setminus S} \rho_j > 1 - \rho_{\mathcal{I}_1}. \end{array}$$

Note that the constraint is equivalent to $d_S > 0$. If we define $\theta_i := r_i - \rho_i$ for all $i \in \mathcal{I}_2$, then the above problem may be expressed in the standard knapsack form as

$$\begin{aligned} \max_{U \subseteq \mathcal{I}_2} \quad & \sum_{j \in U} (\nu_j - 1) \\ \text{sub} \quad & \sum_{j \in U} \theta_j \leq \rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1 - \epsilon, \end{aligned}$$

with $U = \mathcal{I}_2 \setminus S$ and $\epsilon$ some small positive number. The above problem may not always have a unique solution. In case it does, the corresponding set $S$ is dominant, except for the case when some set $T$ exists which is critical but not strictly critical (i.e. $r_T + \rho_{I \setminus T} = 1$), with $\mu_T \leq \mu_S$ (see the definition of a dominant set). Although intriguing, this 'critical case' is not further considered in the present chapter. In this case, the temporary drift may be *zero* for a long period of time during the path to overflow. Partial results for this case have been obtained in [290].
In case the knapsack problem has several solutions, the corresponding sets are weakly dominant (except for the critical case again). The next theorem extends the reduced-load equivalence to the case of weakly dominant sets.

**Theorem 7.3.2** *(Generalized reduced-load equivalence; weakly dominant sets)*
*Let $\Upsilon \subseteq 2^{\mathcal{I}_2}$ be the collection of all weakly dominant sets. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, $S \in \Upsilon$, then*

$$\mathbb{P}\{V > x\} \sim \sum_{S \in \Upsilon} \mathbb{P}\{V_S^{cs} > x\}, \tag{3.4}$$

*with $\mathbb{P}\{V_S^{cs} > x\}$ as in (3.2), (3.3).*

### 7.3.3 Homogeneous On-Off sources

We briefly consider the case of homogeneous On-Off sources as an important special case with weakly dominant sets. Assume that the sources $i \in \mathcal{I}_2$ have identical characteristics. With some minor abuse of notation, let $A(\cdot) := A_i(\cdot)$, $\nu := \nu_i$, $\rho := \rho_i$, $r := r_i$, $p := p_i$. Define $N^* := \arg\min\{N : Nr + (|\mathcal{I}_2| - N)\rho > 1 - \rho_{\mathcal{I}_1}\}$. (Observe that the assumption $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$ ensures $N^* \leq |\mathcal{I}_2|$.) To exclude the critical case, assume that $(N^* - 1)r + (|\mathcal{I}_2| - N^* + 1)\rho < 1 - \rho_{\mathcal{I}_1}$, so that the drift remains negative (and cannot be zero) when only $N^* - 1$ sources are On.

**Corollary 7.3.1** *If $A(\cdot) \in \mathcal{R}$, then*

$$\mathbb{P}\{V > x\} \sim \binom{|\mathcal{I}_2|}{N^*} \mathbb{P}\{\bar{V} > x\},$$

*with*

$$\mathbb{P}\{\bar{V} > x\} \sim p^{N^*} \sum_{n=0}^{N^*} \left( \begin{array}{c} N^* \\ n \end{array} \right) P_{\{1,\ldots,n\}}(x),$$

*where* $P_{\{1,\ldots,n\}}(x)$ *is given by (3.3). In particular,* $\mathbb{P}\{V > x\}$ *and* $P_{\{1,\ldots,n\}}(x)$ *are regularly varying of index* $-N^*(\nu - 1)$.

### 7.3.4 $K$ heterogeneous classes

We finally consider the important special case where each On-Off source in $\mathcal{I}_2$ belongs to one of $K$ heterogeneous classes. We will show how an approximate solution to the knapsack problem may be obtained using a simple index rule. The approximation is in fact asymptotically exact in the many-sources regime.

Specifically, consider the superposition of $n$ On-Off sources, each belonging to one of $K$ heterogeneous classes. Let $a_k$ be the fraction of sources of class $k \in \{1, \ldots, K\}$, with peak rate $r_k$, mean rate $\rho_k$, and an On-period distribution which is regularly varying of index $-\nu_k$. Let the service rate be $n$ (instead of 1), and let $V^{(n)}$ be the stationary workload. The knapsack problem then takes the form

$$\begin{array}{ll} \min\limits_{n_k \in \{0,\ldots,na_k\}} & \sum\limits_{k=1}^{K} n_k(\nu_k - 1) \\ \text{sub} & \sum\limits_{k=1}^{K} n_k r_k + \sum\limits_{k=1}^{K} (na_k - n_k)\rho_k > n. \end{array}$$

Unfortunately, the above problem cannot be easily solved due to the integrality constraints. Intuitively however, one may expect that as $n$ grows large, the integrality constraints should have a negligible effect, so that a continuous relaxation with $n_k \in [0, na_k]$ should give a good approximate solution.

This relaxation may be solved using a simple index rule. Index the $K$ classes in non-decreasing order of the ratios

$$\gamma_k := (\nu_k - 1)/(r_k - \rho_k).$$

For any $k \in \{1, \ldots, K\}$, define $\sigma_k := \sum\limits_{m=1}^{k-1} a_m r_m + \sum\limits_{m=k}^{K} a_m \rho_m$. Determine the (unique) index $\ell$ such that $1 \in (\sigma_{\ell-1}, \sigma_\ell]$. Then take $n_k^* = na_k$ for all classes $k < \ell$, $n_k^* = 0$ for all classes $k > \ell$, and $n_\ell^* = n(1 - \sigma_{\ell-1})/(r_\ell - \rho_\ell)$.

This yields the (crude) approximation

$$\mathbb{P}\{V^{(n)} > x\} \approx x^{-n\mu}, \tag{3.5}$$

with $\mu := \sum\limits_{k=1}^{\ell-1} a_k(\nu_k - 1) + (1 - \sigma_{\ell-1})\gamma_\ell$. In Section 7.7 we prove that the above approximation is logarithmically exact in the many-sources regime. In particular, one may show that the limits for $x \to \infty$ and $n \to \infty$ commute if one considers logarithmic asymptotics.

**Theorem 7.3.3** *(Robustness of logarithmic asymptotics)*

$$\lim_{n\to\infty}\lim_{x\to\infty}\frac{1}{n}\frac{\log\mathbb{P}\{V^{(n)}>nx\}}{\log x}=\lim_{x\to\infty}\lim_{n\to\infty}\frac{1}{n}\frac{\log\mathbb{P}\{V^{(n)}>nx\}}{\log x}.$$

The proof of the above theorem may be found in Section 7.7. Although logarithmically exact, the approximation (3.5) may not be appropriate from a practical perspective. In particular, it is shown in Section 7.7 that an analogue of Theorem 7.3.3 cannot hold if one considers exact asymptotics. This 'negative' result is reminiscent of a phenomenon occurring in heavy-traffic theory where two limiting regimes lead either to stable Lévy motion or to fractional Brownian motion, see e.g. Mikosch *et al.* [206] and references therein.

## 7.4 Bounds

In this section we collect some preliminary results (mostly lower and upper bounds) which will be used in later sections.
We first derive some simple bounds for the workload distribution $\mathbb{P}\{V_S^c>x\}$ for subsets $S\subseteq\mathcal{I}_2$. For any subset $S\subseteq\mathcal{I}_2$, $c<r_S$, define

$$P_S^c(x):=\prod_{j\in S}p_j\mathbb{P}\{A_j^r>\frac{x}{r_S-c}\}.$$

The next lemma gives a lower bound for $\mathbb{P}\{V_S^c>x\}$ which may also be found in Choudhury & Whitt [83].

**Lemma 7.4.1** *Let $S\subseteq\mathcal{I}_2$. For $c<r_S$,*

$$\mathbb{P}\{V_S^c>x\}\geq P_S^c(x).$$

**Proof**
Consider the event that at some arbitrary time $t$ all sources $j\in S$ have been On since time $t-\frac{x}{r_S-c}$ or longer. This event occurs with probability $P_s^c(x)$, and implies that the workload at time $t$ is larger than $\frac{r_Sx}{r_S-c}-\frac{cx}{r_S-c}=x$.

□

For any subset $S\subseteq\mathcal{I}_2$, $c<r_S$, define

$$K_S^c:=\prod_{j\in S}\frac{r_j-\rho_j}{r_j-\rho_j+c-r_S}.$$

The next lemma establishes an asymptotic upper bound for $\mathbb{P}\{V_S^c>x\}$ for the case where $S$ is a minimally-critical set with respect to the capacity $c$.

**Lemma 7.4.2** *Let $S \subseteq \mathcal{I}_2$. If $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$, and $A_j^r(\cdot) \in \mathcal{S}$ for all $j \in S$, then*

$$\mathbb{P}\{V_S^c > x\} \lesssim K_S^c P_S^c(x).$$

**Proof**
For any $i \in S$, denote $d_i := c - r_S + r_i$. Observe that $d_i > \rho_i$ since $c > r_S - (r_i - \rho_i)$. We apply the usual technique to obtain an upper bound: split the capacity. Formally, we have the sample-path upper bound

$$V_S^c(t) \leq V_i^{d_i}(t) + V_{S\setminus\{i\}}^{r_{S\setminus\{i\}}}(t) = V_i^{d_i}(t) \tag{4.1}$$

for all $i \in S$.
In the stationary regime, using Theorem 2.2.3,

$$\begin{aligned}
\mathbb{P}\{V_S^c > x\} &\leq \mathbb{P}\{V_j^{d_j} > x \text{ for all } j \in S\} \\
&= \prod_{j\in S} \mathbb{P}\{V_j^{d_j} > x\} \\
&\sim \prod_{j\in S} (1-p_j)\frac{\rho_j}{d_j - \rho_j}\mathbb{P}\{A_j^r > \frac{x}{r_j - d_j}\} \\
&= \prod_{j\in S} p_j \frac{r_j - \rho_j}{r_j - \rho_j + c - r_S}\mathbb{P}\{A_j^r > \frac{x}{r_S - c}\} \\
&= K_S^c P_S^c(x).
\end{aligned}$$

□

**Corollary 7.4.1** *Let $S \subseteq \mathcal{I}_2$. If $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$, and $A_j^r(\cdot) \in \mathcal{S}$ for all $j \in S$, then*

$$P_S^c(x) \leq \mathbb{P}\{V_S^c > x\} \lesssim K_S^c P_S^c(x).$$

**Proof**
The proof follows directly by combining Lemmas 7.4.1 and 7.4.2.

□

**Corollary 7.4.2** *Let $S \subseteq \mathcal{I}_2$. If $A_j^r(\cdot) \in \mathcal{IRV}$ for all $j \in S$, then for any closed interval $T \subseteq (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$ there exist constants $K^{(1)}$, $K^{(2)}$ independent of $c$, such that for all $c \in T$,*

$$K^{(1)} P_S(x) \lesssim \mathbb{P}\{V_S^c > x\} \lesssim K^{(2)} P_S(x),$$

*with*

$$P_S(x) := \prod_{j\in S} \mathbb{P}\{A_j^r > x\}.$$

**Proof**
The statement follows directly from Corollary 7.4.1 and the fact that $A_j^r(\cdot) \in \mathcal{IRV} \subset \mathcal{S}$ for all $j \in S$ when observing that $A_j^r(\cdot) \in \mathcal{IRV}$, $j \in S$ implies that

$$\limsup_{x\to\infty} \frac{P_S^{c_1}(x)}{P_S^{c_2}(x)} < \infty,$$

if $c_1, c_2 \in T$.

□

We now derive some general bounds for the total workload distribution $\mathbb{P}\{V > x\}$ which will be crucial in establishing the reduced-load equivalence.
For any $c \geq 0$, $E \subseteq \mathcal{I}$, define $Z_E^c(t) := \sup_{0\leq s\leq t} \{c(t-s) - A_E(s,t)\}$. For $c < \rho_E$, let $Z_E^c$ be a random variable with the limiting distribution of $Z_E^c(t)$ for $t \to \infty$. Let $\Omega \subseteq 2^{\mathcal{I}_2}$ be the collection of all minimally-critical sets.

We first present a lower bound. The idea is as follows: $V_E^{c_E}$ being large for some minimally-critical set $E \in \Omega$ basically implies that $V$ must be large too, unless the other sources $j \notin E$ persist in below-average behavior. Excluding such below-average behavior (reflected in large values of $Z_{\mathcal{I}\setminus E}^c$) from the event $\{V > x\}$ yields the following lower bound for $\mathbb{P}\{V > x\}$.

**Lemma 7.4.3** *Let $\Lambda \subseteq \Omega$. Then for any $\delta > 0$ and $y \geq 0$,*

$$\begin{aligned}\mathbb{P}\{V > x\} &\geq \sum_{E\in\Lambda} \mathbb{P}\{V_E^{c_E+\delta} > x + y\}\mathbb{P}\{Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y\} \\ &- \sum_{E_1,E_2\in\Lambda, E_1\neq E_2} \prod_{j\in E_1\cup E_2} \mathbb{P}\{V_j^{\rho_j+\delta} > x\}.\end{aligned}$$

**Proof**
Sample-path wise,

$$\begin{aligned}V(t) &= \sup_{0\leq s\leq t} \{A(s,t) - (t-s)\} \\ &= \sup_{0\leq s\leq t} \{A_E(s,t) + A_{\mathcal{I}\setminus E}(s,t) - (c_E+\delta)(t-s) - (\rho_{\mathcal{I}\setminus E} - \delta)(t-s)\} \\ &\geq \sup_{0\leq s\leq t} \{A_E(s,t) - (c_E+\delta)(t-s)\} + \inf_{0\leq s\leq t} \{A_{\mathcal{I}\setminus E}(s,t) - (\rho_{\mathcal{I}\setminus E} - \delta)(t-s)\} \\ &= \sup_{0\leq s\leq t} \{A_E(s,t) - (c_E+\delta)(t-s)\} - \sup_{0\leq s\leq t} \{(\rho_{\mathcal{I}\setminus E} - \delta)(t-s) - A_{\mathcal{I}\setminus E}(s,t)\} \\ &= V_E^{c_E+\delta}(t) - Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta}(t)\end{aligned}$$

for all $E \in \Lambda$.

In the stationary regime, for any $\delta > 0$ and $y \geq 0$, using the independence of $V_E^{c_E+\delta}$ and $Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta}$,

$$\begin{aligned}
& \mathbb{P}\{V > x\} \\
\geq\ & \mathbb{P}\{V_E^{c_E+\delta} - Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} > x \text{ for some } E \in \Lambda\} \\
\geq\ & \mathbb{P}\{V_E^{c_E+\delta} > x+y, Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y \text{ for some } E \in \Lambda\} \\
\geq\ & \mathbb{P}\{V_E^{c_E+\delta} > x+y, Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y \text{ for exactly one } E \in \Lambda\} \\
=\ & \sum_{E\in\Lambda} \mathbb{P}\{V_E^{c_E+\delta} > x+y, Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y\} \\
-\ & \sum_{E_1,E_2\in\Lambda, E_1\neq E_2} \mathbb{P}\{V_{E_1}^{c_{E_1}+\delta} > x+y, Z_{I\setminus E_1}^{\rho_{I\setminus E_1}-\delta} \leq y, V_{E_2}^{c_{E_2}+\delta} > x+y, Z_{I\setminus E_2}^{\rho_{I\setminus E_2}-\delta} \leq y\} \\
\geq\ & \sum_{E\in\Lambda} \mathbb{P}\{V_E^{c_E+\delta} > x+y\}\mathbb{P}\{Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y\} \\
-\ & \sum_{E_1,E_2\in\Lambda, E_1\neq E_2} \mathbb{P}\{V_{E_1}^{c_{E_1}+\delta} > x, V_{E_2}^{c_{E_2}+\delta} > x\}. \qquad (4.2)
\end{aligned}$$

As in (4.1),

$$V_E^{c_E+\delta}(t) \leq V_i^{c_E-r_{E\setminus\{i\}}+\delta}(t) + V_{E\setminus\{i\}}^{r_{E\setminus\{i\}}}(t) = V_i^{c_E-r_{E\setminus\{i\}}+\delta}(t) \qquad (4.3)$$

for all $i \in E$.
Note that $c_E - r_{E\setminus\{i\}} > \rho_i$ for all $i \in E$, $E \in \Lambda$, since $E$ is minimally-critical.
Hence,

$$V_E^{c_E+\delta}(t) \leq V_i^{\rho_i+\delta}(t)$$

for all $i \in E$, $E \in \Lambda$.
Thus,

$$\begin{aligned}
& \mathbb{P}\{V_{E_1}^{c_{E_1}+\delta} > x, V_{E_2}^{c_{E_2}+\delta} > x\} \qquad (4.4) \\
\leq\ & \mathbb{P}\{V_j^{\rho_j+\delta} > x \text{ for all } j \in E_1, V_j^{\rho_j+\delta} > x \text{ for all } j \in E_2\} \\
=\ & \mathbb{P}\{V_j^{\rho_j+\delta} > x \text{ for all } j \in E_1 \cup E_2\} \\
=\ & \prod_{j\in E_1\cup E_2} \mathbb{P}\{V_j^{\rho_j+\delta} > x\}. \qquad (4.5)
\end{aligned}$$

Substituting (4.5) into (4.2) completes the proof.

□

We now provide a corresponding upper bound, which is somewhat more involved. The idea is as follows: $V$ being large essentially means that $V_E^{c_E}$ must be large for some minimally-critical set $E \in \Lambda$ too, unless the other sources $j \notin E$ exhibit above-average

behavior. Extending the event $\{V > x\}$ with possible above-average behavior of the sources $j \notin E$ (manifesting itself in large values of $V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta}$) leads to the following upper bound for $\mathbb{P}\{V > x\}$.

**Lemma 7.4.4** *Let $\Lambda \subseteq \Omega$. Then for any $\delta, \epsilon > 0$ sufficiently small and $y$,*

$$\begin{aligned}\mathbb{P}\{V > x\} &\leq \sum_{E\in\Lambda} \mathbb{P}\{V_E^{c_E-\delta} > x - y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{E\in\Lambda} \mathbb{P}\{V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y\} \prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{E\in\Omega\setminus\Lambda} \prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\},\end{aligned}$$

*with $N := |\mathcal{I}|$ denoting the total number of sources.*

**Proof**

As before, we divide the capacity to obtain the sample-path upper bound

$$V(t) \leq V_E^{c_E-\delta}(t) + V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta}(t)$$

for all $E \in \Lambda$.

In addition, for $\epsilon > 0$ sufficiently small, $V(t) > x$ implies $V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon}(t) > x/N$, or there exists a minimally-critical set $S \in \Omega$ such that $V_j^{\rho_j+\epsilon}(t) > x/N$ for all $j \in S$.

This may be seen as follows: Suppose that it were not the case, i.e., $V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon}(t) \leq x/N$, and for every minimally-critical set $S \in \Omega$ there exists a $j$ (depending on $S$) such that $V_j^{\rho_j+\epsilon}(t) \leq x/N$. Then the set $\mathcal{J}(t) := \{j \in \mathcal{I}_2 : V_j^{\rho_j+\epsilon}(t) > x/N\}$ does not contain any minimally-critical set, hence $r_{\mathcal{J}(t)} + \rho_{\mathcal{I}\setminus\mathcal{J}(t)} < 1$. This means that $\rho_{\mathcal{I}\setminus\mathcal{J}(t)} + N\epsilon \leq 1 - r_{\mathcal{J}(t)}$ for $\epsilon > 0$ sufficiently small. Thus, noting that $\rho_{\mathcal{I}\setminus\mathcal{J}(t)} = \rho_{\mathcal{I}_1} + \rho_{\mathcal{I}_2\setminus\mathcal{J}(t)}$,

$$\begin{aligned}V(t) &\leq V_{\mathcal{J}(t)}^{r_{\mathcal{J}(t)}}(t) + V_{\mathcal{I}\setminus\mathcal{J}(t)}^{1-r_{\mathcal{J}(t)}}(t) \\ &= V_{\mathcal{I}\setminus\mathcal{J}(t)}^{1-r_{\mathcal{J}(t)}}(t) \\ &\leq V_{\mathcal{I}\setminus\mathcal{J}(t)}^{\rho_{\mathcal{I}\setminus\mathcal{J}(t)}+N\epsilon}(t) \\ &\leq V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon}(t) + \sum_{j\in\mathcal{I}_2\setminus\mathcal{J}(t)} V_j^{\rho_j+\epsilon}(t) \\ &\leq |\mathcal{I}\setminus\mathcal{J}(t)|\, x/N \\ &\leq x,\end{aligned}$$

contradicting the initial supposition.

In the stationary regime, for any $\delta, \epsilon > 0$ sufficiently small and $y$, using independence,

$$\begin{aligned}\mathbb{P}\{V > x\} \leq\ & \mathbb{P}\{V_E^{c_E-\delta} + V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > x \text{ for all } E \in \Lambda, \\ & V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N} \text{ or } V_j^{\rho_j+\epsilon} > x/\mathcal{N} \text{ for all } j \in S \text{ for some } S \in \Omega\}\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\{V_E^{c_E-\delta} > x-y \text{ or } V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y \text{ for all } E\in\Lambda,\\
&\qquad V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N} \text{ or } V_j^{\rho_j+\epsilon} > x/\mathcal{N} \text{ for all } j\in S \text{ for some } S\in\Omega\}\\
&\leq \sum_{E\in\Lambda}\mathbb{P}\{V_E^{c_E-\delta} > x-y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} +\\
&\qquad \sum_{S\in\Omega}\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N} \text{ for all } j\in S, V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y \text{ for all } E\in\Lambda\}\\
&\leq \sum_{E\in\Lambda}\mathbb{P}\{V_E^{c_E-\delta} > x-y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} +\\
&\qquad \sum_{E\in\Lambda}\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N} \text{ for all } j\in E, V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y\} +\\
&\qquad \sum_{E\in\Omega\setminus\Lambda}\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N} \text{ for all } j\in E\}\\
&\leq \sum_{E\in\Lambda}\mathbb{P}\{V_E^{c_E-\delta} > x-y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} +\\
&\qquad \sum_{E\in\Lambda}\mathbb{P}\{V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y\}\prod_{j\in E}\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}\\
&+ \sum_{E\in\Omega\setminus\Lambda}\prod_{j\in E}\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}.
\end{aligned}
$$

□

We conclude this section with the following lemma.

**Lemma 7.4.5** *Let $S\subseteq\mathcal{I}_2$. If $A_j(\cdot)\in\mathcal{R}$ for all $j\in S$ and $c\in(r_S-\min_{j\in S}\{r_j-\rho_j\}, r_S)$, then*

$$\lim_{M\to\infty}\limsup_{x\to\infty}\frac{\mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t)-(c-\epsilon)t\}>x\}}{\mathbb{P}\{V_S^c>x\}}=0,$$

*for any $\epsilon\in[0, r_S-c)$.*

**Proof**
For $t\geq Mx$, write

$$A_S(0,t)-(c-\epsilon)t = A_S(0,Mx)-(c-\epsilon)Mx + A_S(Mx,t)-(c-\epsilon)(t-Mx),$$

and observe that $A_S(Mx,t)\stackrel{d}{=}A_S(0,t-Mx)$ since the process $A_S(0,t)$ is stationary. Thus, for $\delta>0$ sufficiently small,

$$
\begin{aligned}
&\mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t)-(c-\epsilon)t\}>x\}\\
=\ &\mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,Mx)-(c-\epsilon)Mx + A_S(Mx,t)-(c-\epsilon)(t-Mx)\}>x\}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{P}\{A_S(0,Mx) - (c-\epsilon)Mx + \sup_{t\geq Mx}\{A_S(Mx,t) - (c-\epsilon)(t-Mx)\} > x\} \\
&\leq \mathbb{P}\{A_S(0,Mx) - (c-\epsilon)Mx > -\delta(c-\epsilon)Mx\} + \\
&\quad \mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t-Mx) - (c-\epsilon)(t-Mx)\} > (1+\delta(c-\epsilon)M)x\} \\
&= \mathbb{P}\{A_S(0,Mx) > (1-\delta)(c-\epsilon)Mx\} + \\
&\quad \mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t-Mx) - (c-\epsilon)(t-Mx)\} > (1+\delta(c-\epsilon)M)x\} \\
&\leq \mathbb{P}\{\sup_{t\geq 0}\{A_S(0,t) - (1-2\delta)(c-\epsilon)t\} > \delta(c-\epsilon)Mx\} + \\
&\quad \mathbb{P}\{\sup_{t\geq 0}\{A_S(0,t) - (c-\epsilon)t\} > (1+\delta(c-\epsilon)M)x\} \\
&= \mathbb{P}\{V_S^{(1-2\delta)(c-\epsilon)} > \delta(c-\epsilon)Mx\} + \mathbb{P}\{V_S^{c-\epsilon} > (1+\delta(c-\epsilon)M)x\} \\
&\leq 2\mathbb{P}\{V_S^{(1-2\delta)(c-\epsilon)} > \delta(c-\epsilon)Mx\}.
\end{aligned}
$$

Using Corollary 7.4.2, we then obtain for $\delta > 0$ sufficiently small,

$$
\frac{\mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t) - (c-\epsilon)t\} > x\}}{\mathbb{P}\{V_S^c > x\}} \leq 2\frac{K^{(2)}P_S(\delta(c-\epsilon)Mx)}{K^{(1)}P_S(x)}.
$$

Now let $x \to \infty$ and then $M \to \infty$ (use the fact that $P_S(\cdot)$ is of regular variation).

□

## 7.5 Reduced-load equivalence

In this section we provide the proofs of the various reduced-load equivalence results stated in Section 7.3. The proofs of the complementing results for the reduced system are presented in Section 7.6. In Subsection 7.5.1, we consider the case of a single dominant set, resulting in a proof of Equation (3.1), which is repeated as Theorem 7.5.1. Subsection 7.5.2 treats the case of several weakly dominant sets, culminating in a proof of Equation (3.4), see Theorem 7.5.2. In Subsection 7.5.3 we extend the results to the case of additional instantaneous, heavy-tailed input.

### 7.5.1 Single dominant set

In this subsection we prove the reduced-load equivalence result (3.1) for cases with a single dominant set.

**Theorem 7.5.1** *(Reduced-load equivalence)*
*Suppose $S^* \in \Omega$ satisfies Assumptions 7.5.1-7.5.5 as listed below with $c = c_{S^*}$. Then*

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}.$$

**Assumption 7.5.1** *For any $y$ and $\delta > 0$,*

$$F_S^c(\delta) := \liminf_{x\to\infty} \frac{\mathbb{P}\{V_S^{c+\delta} > x + y\}}{\mathbb{P}\{V_S^c > x\}},$$

*is independent of $y$. In addition, $\lim_{\delta\downarrow 0} F_S^c(\delta) = 1$.*

**Assumption 7.5.2** *For any $y$ and $\delta > 0$,*

$$G_S^c(\delta) := \limsup_{x\to\infty} \frac{\mathbb{P}\{V_S^{c-\delta} > x - y\}}{\mathbb{P}\{V_S^c > x\}},$$

*is independent of $y$. In addition, $\lim_{\delta\downarrow 0} G_S^c(\delta) = 1$.*

**Assumption 7.5.3** *For any $\epsilon > 0$,*

$$\lim_{x\to\infty} \frac{\mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_S^c > x\}} = 0.$$

**Assumption 7.5.4** *For any $\epsilon > 0$,*

$$H_S^c(\epsilon) := \limsup_{x\to\infty} \frac{\prod_{j\in S} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_S^c > x\}} < \infty.$$

**Assumption 7.5.5** *For any $E \in \Omega$, $E \neq S$, for any $\epsilon > 0$,*

$$\lim_{x\to\infty} \frac{\prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_S^c > x\}} = 0.$$

**Proof of Theorem 7.5.1**
The proof consists of deriving a lower bound and an upper bound which asymptotically coincide.
*Lower bound*
From Lemma 7.4.3, taking $\Lambda = \{S^*\}$, for any $\delta > 0$ and $y$,

$$\mathbb{P}\{V > x\} \geq \mathbb{P}\{V_{S^*}^{c_{S^*}+\delta} > x + y\}\mathbb{P}\{Z_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}-\delta} \leq y\}.$$

Thus, using Assumption 7.5.1,

$$\begin{aligned}\liminf_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} &\geq \mathbb{P}\{Z_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}-\delta} \leq y\} \liminf_{x\to\infty} \frac{\mathbb{P}\{V_{S^*}^{c_{S^*}+\delta} > x + y\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \\ &= F_{S^*}^{c_{S^*}}(\delta)\mathbb{P}\{Z_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}-\delta} \leq y\}.\end{aligned}$$

Letting $y \to \infty$, then $\delta \downarrow 0$,

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \geq 1,$$

which completes the proof of the lower bound.

*Upper bound*
From Lemma 7.4.4, taking $\Lambda = \{S^*\}$, for any $\delta, \epsilon > 0$ sufficiently small and $y$,

$$\begin{aligned}\mathbb{P}\{V > x\} &\leq \mathbb{P}\{V_{S^*}^{c_{S^*}-\delta} > x - y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} \\ &+ \mathbb{P}\{V_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}+\delta} > y\} \prod_{j\in S^*} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{E\in\Omega, E\neq S^*} \prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}.\end{aligned}$$

Thus, using Assumptions 7.5.2-7.5.5,

$$\begin{aligned}\limsup_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} &\leq \limsup_{x\to\infty} \frac{\mathbb{P}\{V_{S^*}^{c_{S^*}-\delta} > x - y\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} + \limsup_{x\to\infty} \frac{\mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \\ &+ \mathbb{P}\{V_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}+\delta} > y\} \limsup_{x\to\infty} \frac{\prod_{j\in S^*} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \\ &+ \sum_{E\in\Omega, E\neq S^*} \limsup_{x\to\infty} \frac{\prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \\ &= G_{S^*}^{c_{S^*}}(\delta) + H_{S^*}^{c_{S^*}}(\epsilon)\mathbb{P}\{V_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}+\delta} > y\}.\end{aligned}$$

Letting $y \to \infty$, then $\delta \downarrow 0$,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V_{S^*}^{c_{S^*}} > x\}} \leq 1,$$

which completes the proof.

□

In order to complete the proof of the reduced-load equivalence result (3.1), it remains to be shown that a dominant set $S^* \subseteq \mathcal{I}_2$ with $A_j(\cdot) \in \mathcal{R}$ for all $j \in S^*$ satisfies Assumptions 7.5.1-7.5.5. That is done in the following two propositions for $S = S^*$.

**Proposition 7.5.1** *Let $S \subseteq \mathcal{I}_2$. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then Assumptions 7.5.1 and 7.5.2 are satisfied for any $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$.*

**Proof**
We first prove that Assumption 7.5.2 is satisfied. It follows from Theorem 7.6.4 (see also Corollary 7.6.1; it is important to note here that the results in Section 7.6 do not rely on the results of this section) that if $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then $\mathbb{P}\{V_S^c > x\} \in \mathcal{IRV}$.

Since $\mathcal{IRV} \subset \mathcal{L}$ (see Lemma 2.1.9), it suffices to prove that the assumption is satisfied for $y = 0$.
Let $\epsilon \in [0, r_S - c)$, and let $\delta \in (0, \epsilon]$. Then

$$\begin{aligned}
& \mathbb{P}\{V_S^{c-\delta} > x\} \\
= \; & \mathbb{P}\{\sup_{t\geq 0}\{A_S(0,t) - (c-\delta)t\} > x\} \\
\leq \; & \mathbb{P}\{\sup_{t\leq x\delta^{-1/2}}\{A_S(0,t) - (c-\delta)t\} > x\} + \mathbb{P}\{\sup_{t\geq x\delta^{-1/2}}\{A_S(0,t) - (c-\delta)t\} > x\} \\
\leq \; & \mathbb{P}\{\sup_{t\leq x\delta^{-1/2}}\{A_S(0,t) - ct\} > (1-\delta^{1/2})x\} + \mathbb{P}\{\sup_{t\geq x\delta^{-1/2}}\{A_S(0,t) - (c-\epsilon)t\} > x\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\limsup_{x\to\infty} \frac{\mathbb{P}\{V_S^{c-\delta} > x\}}{\mathbb{P}\{V_S^c > x\}} \; \leq \; & \limsup_{x\to\infty} \frac{\mathbb{P}\{V_S^c > (1-\delta^{1/2})x\}}{\mathbb{P}\{V_S^c > x\}} \\
+ \; & \limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq x\delta^{-1/2}}\{A_S(0,t) - (c-\epsilon)t\} > x\}}{\mathbb{P}\{V_S^c > x\}}.
\end{aligned}$$

The fact that $\mathbb{P}\{V_S^c > x\} \in \mathcal{IRV}$ implies that the first term tends to 1 as $\delta \downarrow 0$, while Lemma 7.4.5 (with $M = \delta^{-1/2}$) shows that the second term then goes to 0.
The proof that Assumption 7.5.1 holds is similar, and therefore omitted.

□

**Proposition 7.5.2** *Let $S \subseteq \mathcal{I}_2$. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then Assumptions 7.5.3 and 7.5.4 are satisfied for any $c > \rho_S$. If in addition $S$ is a dominant set, then Assumption 7.5.5 is satisfied as well.*

**Proof**
Using Lemma 7.4.1,

$$\mathbb{P}\{V_S^c > x\} \geq \prod_{j\in S} p_j \mathbb{P}\{A_j^r > \frac{x}{r_S - c}\}.$$

Assumption 7.5.3 then follows from combining Assumption 7.2.1 and the assumption that $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$.
Theorem 2.2.3 gives

$$\mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\} \sim (1-p_j)\frac{\rho_j}{\epsilon}\mathbb{P}\{A_j^r > \frac{x/N}{r_j - \rho_j - \epsilon}\}$$

for all $j \in \mathcal{I}_2$.
Assumption 7.5.4 then follows from the assumption that $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, and so does Assumption 7.5.5 in case $S$ is a dominant set.

□

### 7.5.2 Several weakly dominant sets

In the previous subsection we considered a scenario with a single dominant set $S^* \subseteq \mathcal{I}_2$. In this subsection we prove the reduced-load equivalence result (3.4) for cases where no unique dominant set may exist. Recall that $\Upsilon$ denotes the collection of all weakly dominant sets, and that $\Omega$ represents the collection of all minimally-critical sets.
We first define a slightly modified version of Assumption 7.5.5.

**Assumption 7.5.6** *For any pair of sets $S \in \Upsilon$, $E \in \Omega \setminus \Upsilon$, for any $\epsilon > 0$,*

$$\lim_{x\to\infty} \frac{\prod\limits_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_S^c > x\}} = 0.$$

**Theorem 7.5.2** *(Generalized reduced-load equivalence; weakly dominant sets) Suppose the sets $S \in \Lambda$ satisfy Assumptions 7.5.1-7.5.4 and Assumption 7.5.6. Then*

$$\mathbb{P}\{V > x\} \sim \sum_{S\in\Lambda} \mathbb{P}\{V_S^{c_S} > x\}.$$

**Proof**
As before, the proof consists of a lower bound and an upper bound which asymptotically coincide. For compactness, denote $Q(x) := \sum\limits_{S\in\Lambda} \mathbb{P}\{V_S^{c_S} > x\}$.
(Lower bound) From Lemma 7.4.3, for any $\delta > 0$ and $y \geq 0$,

$$\begin{aligned}\mathbb{P}\{V > x\} &\geq \sum_{S\in\Lambda} \mathbb{P}\{V_S^{c_S+\delta} > x+y\}\mathbb{P}\{Z_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}-\delta} \leq y\} \\ &\quad - \sum_{S_1,S_2\in\Lambda, S_1\neq S_2} \prod_{j\in S_1\cup S_2} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}.\end{aligned}$$

Note that if $S_1, S_2 \in \Lambda$, $S_1 \neq S_2$, then $S_1 \cup S_2$ cannot be a minimally-critical set, so that $S_1 \cup S_2 \notin \Lambda$.
Thus, using Assumptions 7.5.1, 7.5.4, and the inequality

$$\frac{\sum_i a_i}{\sum_i b_i} \geq \min_i \frac{a_i}{b_i}$$

for $a_i, b_i > 0$, we obtain

$$\begin{aligned}\liminf_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{Q(x)} &\geq \liminf_{x\to\infty} \sum_{S\in\Lambda} \mathbb{P}\{Z_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}-\delta} \leq y\} \frac{\mathbb{P}\{V_S^{c_S+\delta} > x+y\}}{Q(x)} \\ &\quad - \sum_{S_1,S_2\in\Lambda, S_1\neq S_2} \limsup_{x\to\infty} \frac{\prod\limits_{j\in S_1\cup S_2} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{Q(x)} \\ &\geq \liminf_{x\to\infty} \min_{S\in\Lambda} \mathbb{P}\{Z_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}-\delta} \leq y\} \frac{\mathbb{P}\{V_S^{c_S+\delta} > x+y\}}{\mathbb{P}\{V_S^{c_S} > x\}}\end{aligned}$$

$$\begin{aligned} &\geq \min_{S\in\Lambda} \mathbb{P}\{Z_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}-\delta} \leq y\} \liminf_{x\to\infty} \frac{\mathbb{P}\{V_S^{c_S+\delta} > x+y\}}{\mathbb{P}\{V_S^{c_S} > x\}} \\ &= \min_{S\in\Lambda} F_S^{c_S}(\delta) \mathbb{P}\{Z_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}-\delta} \leq y\}. \end{aligned}$$

Letting $y \to \infty$, then $\delta \downarrow 0$, we obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{Q(x)} \geq 1,$$

which completes the proof of the lower bound.
(Upper bound) From Lemma 7.4.4, for any $\delta > 0$ and $y$,

$$\begin{aligned} \mathbb{P}\{V > x\} &\leq \sum_{S\in\Lambda} \mathbb{P}\{V_S^{c_S-\delta} > x-y\} + \mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{S\in\Lambda} \mathbb{P}\{V_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}+\delta} > y\} \prod_{j\in S} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{E\in\Omega\setminus\Lambda} \prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}. \end{aligned}$$

Thus, using Assumptions 7.5.2-7.5.4, 7.5.6, and the inequality

$$\frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}$$

for $a_i, b_i > 0$,

$$\begin{aligned} \mathbb{P}\{V > x\} &\leq \limsup_{x\to\infty} \sum_{S\in\Lambda} \frac{\mathbb{P}\{V_S^{c_S-\delta} > x-y\}}{Q(x)} + \limsup_{x\to\infty} \frac{\mathbb{P}\{V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\}}{Q(x)} \\ &+ \sum_{S\in\Lambda} \mathbb{P}\{V_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}+\delta} > y\} \limsup_{x\to\infty} \frac{\prod_{j\in S} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{Q(x)} \\ &+ \sum_{E\in\Omega\setminus\Lambda} \limsup_{x\to\infty} \frac{\prod_{j\in E} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{Q(x)} \\ &\leq \limsup_{x\to\infty} \max_{S\in\Lambda} \frac{\mathbb{P}\{V_S^{c_S-\delta} > x-y\}}{\mathbb{P}\{V_S^{c_S} > x\}} \\ &+ \sum_{S\in\Lambda} \mathbb{P}\{V_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}+\delta} > y\} \limsup_{x\to\infty} \frac{\prod_{j\in S} \mathbb{P}\{V_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{V_S^{c_S} > x\}} \\ &\leq \max_{S\in\Lambda} \limsup_{x\to\infty} \frac{\mathbb{P}\{V_S^{c_S-\delta} > x-y\}}{\mathbb{P}\{V_S^{c_S} > x\}} + \sum_{S\in\Lambda} H_S(\epsilon) \mathbb{P}\{V_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}+\delta} > y\} \\ &= \max_{S\in\Lambda} G_S^{c_S}(\delta) + \sum_{S\in\Lambda} H_S(\epsilon) \mathbb{P}\{V_{\mathcal{I}\setminus S}^{\rho_{\mathcal{I}\setminus S}+\delta} > y\}. \end{aligned}$$

Letting $y \to \infty$, then $\delta \downarrow 0$, we obtain

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{Q(x)} \leq 1,$$

which completes the proof.

□

In order to complete the proof of the reduced-load equivalence result (3.4), it remains to be shown that the collection of all weakly dominant sets $S \in \Upsilon$ satisfies Assumptions 7.5.1-7.5.4 and Assumption 7.5.6. As shown in Proposition 7.5.1, any strictly critical set $S$ with $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$ satisfies Assumptions 7.5.1 and 7.5.2. Proposition 7.5.2 shows that any set $S$ with $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$ also satisfies Assumptions 7.5.3 and 7.5.4. Thus it suffices to prove that Assumption 7.5.6 is satisfied, which may be done in a similar fashion as for Assumption 7.5.5 (see Proposition 7.5.2).

### 7.5.3 Additional instantaneous input

So far we have considered a scenario with only *fluid* heavy-tailed input. We now extend the reduced-load equivalence to the case with additional *instantaneous*, heavy-tailed input. We thus allow for an additional subset of sources $\mathcal{I}_3 \subseteq \mathcal{I}$ which generate instantaneous traffic bursts according to independent renewal processes. The interarrival times between bursts of source $i$ are generally distributed with mean $1/\lambda_i$. The burst sizes $B_i$ have a heavy-tailed distribution $B_i(\cdot)$ with mean $\beta_i < \infty$. Thus the traffic intensity of source $i$ is $\rho_i := \lambda_i\beta_i$.

For each source $i \in \mathcal{I}_3$, we assume that the burst size distribution is regularly varying of index $-\nu_i$, i.e., $B_i(\cdot) \in \mathcal{R}_{-\nu_i}$ for some $\nu_i > 1$.

In order to formulate the results, we need to extend the concept of dominance introduced in Section 7.3. A source $i \in \mathcal{I}_3$ is said to (weakly) dominate a source $j \in \mathcal{I}_3$ if $\nu_i < (\leq)\nu_j$. A source $i \in \mathcal{I}_3$ is said to (weakly) dominate a critical set $S \subseteq \mathcal{I}_2$ if $\nu_i - 1 < (\leq) \sum_{j\in S}(\nu_j - 1)$. A critical set $S \subseteq \mathcal{I}_2$ is said to (weakly) dominate a source $i \in \mathcal{I}_3$ if $\nu_i - 1 > (\geq) \sum_{j\in S}(\nu_j - 1)$. A source $i \in \mathcal{I}_3$ is called (weakly) dominant if it (weakly) dominates all other sources $j \in \mathcal{I}_3$ as well as all critical sets $S \subseteq \mathcal{I}_2$. A critical set $S \subseteq \mathcal{I}_2$ is called (weakly) dominant if it (weakly) dominates any other critical set $U \subseteq \mathcal{I}_2$ as well as all sources $j \in \mathcal{I}_3$.

**Theorem 7.5.3** *Let $\mathcal{K} \subseteq \mathcal{I}_3$ and $\Upsilon \subseteq 2^{\mathcal{I}_2}$ be the collection of all weakly dominant sources and all weakly dominant sets, respectively. If $B_i(\cdot) \in \mathcal{R}$ for all $i \in \mathcal{K}$, and $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, $S \in \Upsilon$, then*

$$\mathbb{P}\{V > x\} \sim \sum_{i\in\mathcal{K}} \mathbb{P}\{V_i^{c_i} > x\} + \sum_{S\in\Upsilon} \mathbb{P}\{V_S^{c_S} > x\}, \tag{5.1}$$

*with $\mathbb{P}\{V_i^{c_i} > x\}$ and $\mathbb{P}\{V_S^{c_S} > x\}$ as in Theorem 2.2.1 and (3.2), (3.3), respectively.*

The proof of the above theorem is similar to that of Theorem 7.5.2 after a few modifications to Lemmas 7.4.3 and 7.4.4.

It may be worth mentioning that Theorem 7.5.3 continues to hold under the condition $B_i^r(\cdot) \in \mathcal{S}$ for all $i \in \mathcal{K}$, provided there are no weakly dominant sets of On-Off sources (the concept of dominance may be extended to subexponential distributions in a straightforward way). In particular, when there are simply no On-Off sources at all, one obtains the extension of Theorem 2.2.2 to the single-server queue fed by a superposition of renewal processes (which is not a renewal process). This result was obtained as Theorem 4.1 in Asmussen *et al.* [28], using a different approach.

Theorem 7.5.3 also provides an extension of a recent result in Boxma & Kurkova [73], who study an $M/G/1$ queue with two different speeds of service. They derive an expression for the transform of the workload distribution, which is then exploited to obtain the tail behavior of the workload using a Tauberian theorem.
A queue with two service speeds fits into our framework as follows. Consider a queue of unit capacity fed by two input sources:

**(i)** Instantaneous input with generic burst size $B$ and mean rate $\rho_1$;

**(ii)** Fluid input with generic On-period $A$, peak rate $r_2$, and mean rate $\rho_2$.

The above model is equivalent to a $GI/GI/1$ queue with service times $B$, two service speeds ($s_h := 1$ and $s_l := 1 - r_2$), the high-speed periods being generally distributed, and low-speed periods $A$. Assume that the distributions of $A$ and $B$ are both regularly varying (with respective indices $-\nu_1$ and $-\nu_2$) and that $\rho_1 \neq s_l$ (to exclude the critical case).
Theorem 7.5.3 then implies that the tail behavior of the workload distribution is determined by three different scenarios:

**(i)** $\nu_1 < \nu_2$ or $\nu_1 \geq \nu_2$ and $\rho_1 < s_l$: In this case the instantaneous input (source 1) is dominant, yielding

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_1^{1-\rho_2} > x\};$$

**(ii)** $\rho_1 > s_l$ and $\nu_1 > \nu_2$: In this case, the fluid input (source 2) is dominant, implying

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_2^{1-\rho_1} > x\};$$

**(iii)** $\rho_1 > s_l$ and $\nu_1 = \nu_2$: Now both input sources are weakly dominant, which gives

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{V_1^{1-\rho_2} > x\} + \mathbb{P}\{V_2^{1-\rho_1} > x\}.$$

The tail behavior of $\mathbb{P}\{V_1^{1-\rho_2} > x\}$ and $\mathbb{P}\{V_2^{1-\rho_1} > x\}$ is given by Theorems 2.2.2 and 2.2.3, respectively.

## 7.6 Tail asymptotics for the reduced system

In this section we derive the tail asymptotics for the reduced system. In particular, we give a proof of Equations (3.2) and (3.3).
For notational convenience, let $c$ be the capacity of the reduced system, let the set of sources be indexed as $\mathcal{J} = \{1, \ldots, N\}$, and denote $r := r_{\mathcal{J}}$ and $A(0,t) := A_{\mathcal{J}}(0,t)$. By definition, the reduced system satisfies the following two properties:

**(i)** The On-period distribution of source $i$ is regularly varying of index $-\nu_i < -1$, i.e., $A_i(\cdot) \in \mathcal{R}_{-\nu_i}$;

**(ii)** All sources must be On for the drift of the workload process to be positive, i.e., $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$.

We now state our main theorem.

**Theorem 7.6.1** *Consider a queue of capacity $c$ fed by $N$ On-Off sources. If $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$ with $r = \sum_{i=1}^{N} r_i$, and $A_j(\cdot) \in \mathcal{R}_{-\nu_j}$, $\nu_j > 1$, for all $j = 1, \ldots, N$, then*

$$\mathbb{P}\{V^c > x\} \sim \left(\prod_{j=1}^{N} p_j\right) \sum_{\mathcal{J}_0 \subseteq \{1,\ldots,N\}} P_{\mathcal{J}_0}(x),$$

*where $P_{\mathcal{J}_0}(x)$ is given by (with $\mathcal{J}_1 = \{1, \ldots, N\} \setminus \mathcal{J}_0$)*

$$\begin{aligned} & P_{\mathcal{J}_0}(x) \qquad (6.1)\\ = \; & \frac{1}{\prod_{i \in \mathcal{J}_1} \mathbb{E}\{A_i\}} \int_{y_i \in (0,\infty), i \in \mathcal{J}_1} \prod_{i \in \mathcal{J}_1} \mathbb{P}\{(r-c)A_i > \sum_{j \in \mathcal{J}_1} y_j(r_j - \rho_j) - (r-c)y_i + x\} \\ & \prod_{i \in \mathcal{J}_0} \mathbb{P}\{(r-c)A_i^r > \sum_{j \in \mathcal{J}_1} y_j(r_j - \rho_j) + x\} \prod_{i \in \mathcal{J}_1} \mathrm{d}y_i. \end{aligned}$$

An asymptotic characterization of $P_{\mathcal{J}_0}(x)$ which may be useful for further analysis is provided in Subsection 7.6.4. This characterization also shows that $\mathbb{P}\{V^c > x\}$ and $P_{\mathcal{J}_0}(x)$ are regularly varying, and gives an expression for the pre-factor in the asymptotic expansion of $\mathbb{P}\{V^c > x\}$.
With the framework provided in Section 2.4 in mind, we organize this section as follows: Detailed heuristic arguments are given in Subsection 7.6.1. In Subsection 7.6.2, we prove some preliminary results on the most probable behavior of the process $\{A(0,t) - ct\}$. The proof of Theorem 7.6.1 is then completed in Subsection 7.6.3. Subsection 7.6.4 deals with the asymptotic behavior of $P_{\mathcal{J}_0}(x)$.

### 7.6.1 Heuristic arguments

The proof of Theorem 7.6.1 is quite lengthy. Nevertheless, it is based on a simple intuitive argument: the most likely way for $V^c \equiv \sup_{t\geq 0}\{A(0,t)-ct\}$ to reach a large value is that all sources have been simultaneously On for a long time. Specifically, each source is likely to contribute through *exactly one* 'long' On-period; apart from these long On-periods, the sources show typical behavior.

The above heuristic argument may be used for computing $\sup_{t\geq 0}\{A(0,t)-ct\}$. Let us say that the long On-period of source $i$ begins at time $s_i$ and ends at time $s_i+t_i$. Define

$$t^* := \min_{i=1,\ldots,N}\{s_i+t_i\},$$

as the time epoch at which the first of the long On-periods finishes. One may also interpret $t^*$ as the time epoch at which the process $\{A(0,t)-ct\}$ reaches its largest value. Note that $A_i(0,s_i)\approx \rho_i s_i$, $A_i(s_i,s_i+t_i)=r_i t_i$, and $A_i(s_i+t_i,s_i+t_i+t)\approx \rho_i t$, $t\geq 0$. One thus obtains, using the fact that $c\in(r-\min_{i=1,\ldots,N}\{r_i-\rho_i\},r)$,

$$\begin{aligned}\sup_{t\geq 0}\{A(0,t)-ct\} &\approx A(0,t^*)-ct^*\\ &\approx \sum_{i=1}^{N}[\rho_i s_i + r_i(t^*-s_i)]-ct^*\\ &= \sum_{i=1}^{N}(\rho_i-r_i)s_i+(r-c)t^*. \qquad (6.2)\end{aligned}$$

The problem is thus reduced to calculating

$$\mathbb{P}\{\sum_{i=1}^{N}(\rho_i-r_i)s_i+(r-c)\min_{i=1,\ldots,N}\{s_i+t_i\}>x\}. \qquad (6.3)$$

Although the proof is based on the representation $V^c \equiv \sup_{t\geq 0}\{A(0,t)-ct\}$, it is useful to keep the original workload process $\sup_{0\leq s\leq t}\{A(s,t)-c(t-s)\}$ in mind as well. Figure 7.1 shows a typical scenario leading to a large workload level (so small fluctuations are ignored) in the case of two On-Off sources.

At a certain time $\omega_0$, the first long On-period begins. Before that time, both sources show average behavior. The queue starts to build (at rate $r_1+r_2-c$) at time $\omega_1$ when the second long On-period begins, and reaches its largest level at time $\omega_3$. Level $x$ is crossed at time $\omega_2$.

Between times $\omega_3$ and $\omega_4$, the queue drains at rate $c-r_1-\rho_2$: source 1 is still in its long On-period, and source 2 shows average behavior (remember small fluctuations are neglected). The process is still above level $x$ between times $\omega_4$ and $\omega_5$. However, here

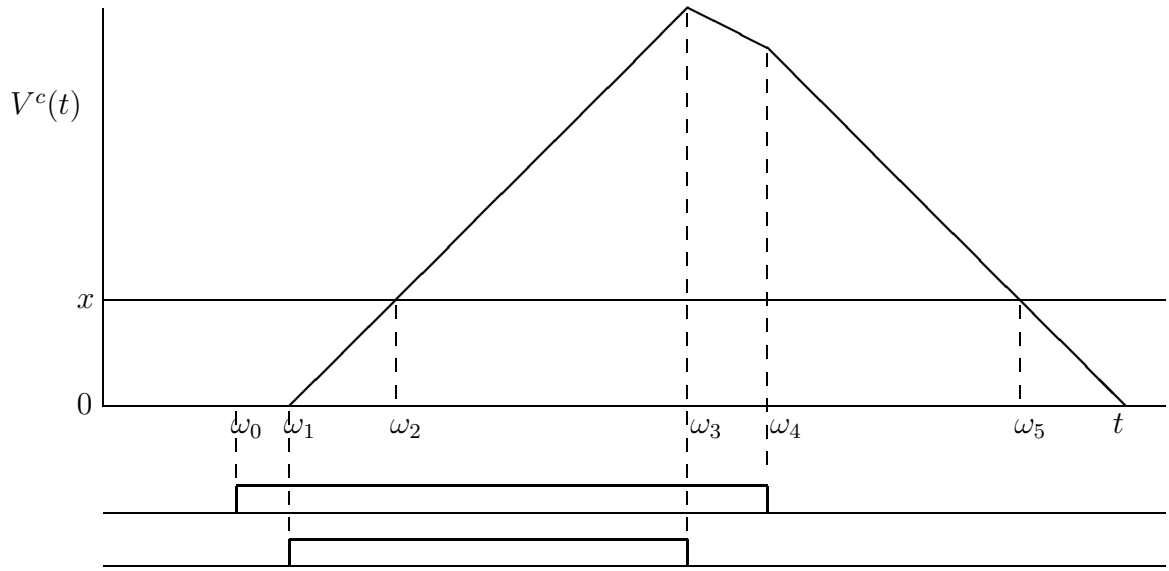


Figure 7.1: Typical overflow scenario for two On-Off sources

both sources show average behavior again, causing a negative drift $c - \rho_1 - \rho_2$.

The figure illustrates why the analysis of the reduced system is still quite complicated:

- Although the long On-periods must significantly overlap, the difference between the finishing times of these On-periods can be quite large (of order $x$, hence not negligible);

- Given that the observed workload is larger than $x$, it is not necessarily the case that all sources are in their long On-periods. In Figure 7.1, this is only the case in the time interval $(\omega_2, \omega_3)$. In fact, for any given source, its long On-period may have finished a long time ago. Consequently, there are $2^N$ different possibilities (corresponding to which sources are still in their long On-periods). Sample-path wise, there are $N + 1$ different time intervals in which the workload may be larger than $x$ (depending on how many of the sources are still in their long On-periods);

- Specifically, given that the observed workload is larger than $x$, it may still have been even larger at an earlier time epoch. In Figure 7.1, this is the case in the time intervals $(\omega_3, \omega_4)$ and $(\omega_4, \omega_5)$.

These complications do not arise if one considers a related problem, which concerns the overflow probability in a fluid queue with a *finite buffer* of size $x$. As is shown in a recent paper of Jelenković & Momčilović [165], the analysis of the reduced system is then

considerably simpler. It suffices to use bounds which are similar to Lemma 7.4.1 and Lemma 7.4.2, and to combine these with the asymptotic results for a single On-Off source in Jelenković [162] and Zwart [286].

### 7.6.2 Characterization of most probable behavior

In this subsection we prove some preliminary results characterizing the most probable behavior of the process $\{A(0,t) - ct\}$ given that it reaches a large value. In particular, we formalize the following two heuristic statements, resulting in a formal version of Equation (6.2).

**(i)** Each source contributes to $\sup_{t\geq 0}\{A(0,t) - ct\}$ through exactly one 'long' On-period;

**(ii)** Apart from these long On-periods, the sources show typical behavior.

An On period is referred to as 'long' when larger than $\epsilon x$, with $\epsilon$ some small, but positive constant. In order to formalize the above statements, we need to keep track of how many long On-periods occur.
With that in mind, we define $\mathcal{N}_i(A,B)$, for intervals $A, B \subseteq [0,\infty)$, as the number of On-periods of source $i$ of which the length is contained in $A$ and of which the beginning is contained in $B$. If $B$ contains 0, this number includes the possible activity period at time 0 (if its length is contained in $A$).
For compactness, denote

$$\mathcal{N}_i(u,t) \equiv \mathcal{N}_i((u,\infty),[0,t]).$$

We now proceed with a few preparatory lemmas.
First we show how to obtain an upper bound for the workload process in terms of a simple random walk. As in (4.1), we have $V^c(t) \leq V_i^{d_i}(t)$ for all $i = 1,\ldots,N$, with $d_i := c - r_{\mathcal{I}\setminus\{i\}} = c - r + r_i$. Recall that $V_i^{d_i}(t) \stackrel{d}{=} \sup_{0\leq s\leq t}\{A_i(0,s) - d_i s\}$. Now let, for fixed $i$, $S_{in} := X_{i1} + \ldots + X_{in}$ be a random walk with step sizes $X_{im} := (r_i - d_i)A_{im} - d_i U_{im}$, with $A_{im}$ and $U_{im}$ i.i.d. random variables distributed as the On- and Off-periods of source $i$, respectively.

Since $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$, we have $\rho_i < d_i$ for all $i = 1,\ldots,N$, so that $\mathbb{E}\{X_{i1}\} < 0$, i.e., the random walk has negative drift. Because of the saw-tooth nature of the process $A_i(0,s) - d_i s$, we have

$$\sup_{0\leq s\leq t}\{A_i(0,s) - d_i s\} \leq (r_i - d_i)(I_i A_{i0}^r + (1 - I_i)A_{i0}) + \sup_{n\leq N_i^A(t)} S_{in},$$

with $N_i^A(t)$ denoting the number of Off-periods of source $i$ elapsed during $[0,t]$ which started after time 0 (for a formal definition see Equation (2.1)).
The above observations are summarized in the following auxiliary lemma.

**Lemma 7.6.1** *For all $\epsilon > 0$, $t$ and $x$,*

$$\mathbb{P}\{V^c(t) > x, \mathcal{N}_i(\epsilon x, t) = 0\} \leq \mathbb{P}\{\sup_{n \leq N_i^A(t)} S_{in} > x(1 - \epsilon(r_i - d_i)), \mathcal{N}_i(\epsilon x, t) = 0\}.$$

**Proof**
We have

$$\begin{aligned}
& \mathbb{P}\{V^c(t) > x, \mathcal{N}_i(\epsilon x, t) = 0\} \\
\leq\ & \mathbb{P}\{V_i^{d_i}(t) > x, \mathcal{N}_i(\epsilon x, t) = 0\} \\
\leq\ & \mathbb{P}\{(r_i - d_i)(I_i A_{i0}^r + (1 - I_i)A_{i0}) + \sup_{n \leq N_i^A(t)} S_{in} > x, \mathcal{N}_i(\epsilon x, t) = 0\} \\
\leq\ & \mathbb{P}\{\sup_{n \leq N_i^A(t)} S_{in} > x(1 - \epsilon(r_i - d_i)), \mathcal{N}_i(\epsilon x, t) = 0\}.
\end{aligned}$$

The last inequality follows from the fact that $A_{i0}^r$ and $A_{i0}$ must be smaller than $\epsilon x$ if $\mathcal{N}_i(\epsilon x, t) = 0$.

□

To obtain upper bounds for probabilities as in Lemma 7.6.1, we will frequently apply the truncation Lemma 2.4.1, given in Section 2.4.
The final preparatory lemma is a simple consequence of Corollary 7.4.2, which will be used several times in combination with Lemma 2.4.1 to show that probabilities of certain events are of $o(\mathbb{P}\{V^c > x\})$. Define $P(x) := \prod_{j=1}^{N} \mathbb{P}\{A_j^r > x\} \in \mathcal{R}_{-\mu}$, $\mu := \sum_{j=1}^{N} (\nu_j - 1)$.

**Lemma 7.6.2** $$\limsup_{x \to \infty} \frac{P(x)}{\mathbb{P}\{V^c > x\}} < \infty,$$

We now show that, with overwhelming probability (as $x \to \infty$), the rare event $\{V^c > x\}$ occurs as follows.

**(i)** The process $\{A(0,t) - ct\}$ reaches level $x$ before time $Mx$ for some large $M$;

**(ii)** Up to time $Mx$, each source generates *exactly one* long On-period, i.e., $\mathcal{N}_i(\epsilon x, Mx) = 1$ for $i = 1, \ldots, N$.

**Proposition 7.6.1** $$\lim_{M\to\infty} \liminf_{x\to\infty} \frac{\mathbb{P}\{V^c(Mx) > x\}}{\mathbb{P}\{V^c > x\}} = 1.$$

**Proof**
By definition,

$$\begin{aligned}
\mathbb{P}\{V^c > x\} &= \mathbb{P}\{\sup_{t\geq 0}\{A(0,t) - ct\} > x\} \\
&\leq \mathbb{P}\{\sup_{0\leq t\leq Mx}\{A(0,t) - ct\} > x\} + \mathbb{P}\{\sup_{t\geq Mx}\{A(0,t) - ct\} > x\} \\
&= \mathbb{P}\{V^c(Mx) > x\} + \mathbb{P}\{\sup_{t\geq Mx}\{A(0,t) - ct\} > x\}.
\end{aligned}$$

Thus, it suffices to show

$$\lim_{M\to\infty} \limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq Mx}\{A(0,t) - ct\} > x\}}{\mathbb{P}\{V^c > x\}} = 0,$$

which however follows directly from Lemma 7.4.5.

□

Now suppose that the workload reaches level $x$. By the previous proposition, we may assume that this occurs before time $Mx$ (for $M$ sufficiently large). The next two propositions show that we may restrict the attention to a scenario where each source initiates *exactly one* long On-period before time $Mx$.

The first proposition indicates that each source has *at least* one long On-period.

**Proposition 7.6.2** *For all $i$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*]$ and all $M$,*

$$\mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 0\} = \mathrm{o}(\mathbb{P}\{V^c > x\}),$$

*as $x \to \infty$.*

**Proof**
Define $N_i^U(t) := \max\{n : \sum_{j=1}^{n} U_{ij} \leq t\} + 1$. Note that $N_i^A(t) \leq N_i^U(t)$.
Using Lemma 7.6.1, taking $t = Mx$,

$$\begin{aligned}
& \mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 0\} \\
\leq\ & \mathbb{P}\{\sup_{n\leq N_i^A(Mx)} S_{in} > x(1 - \epsilon(r_i - d_i)), \mathcal{N}_i(\epsilon x, Mx) = 0\} \\
\leq\ & \mathbb{P}\{\sup_{n\leq N_i^A(Mx)} S_{in} > x(1 - \epsilon(r_i - d_i)) | \mathcal{N}_i(\epsilon x, Mx) = 0\} \\
=\ & \mathbb{P}\{\sup_{n\leq N_i^A(Mx)} S_{in} > x(1 - \epsilon(r_i - d_i)) | A_{ij} < \epsilon x, j = 1, \ldots, N_i^A(Mx)\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\{\sup_{n\le N_i^A(Mx)} S_{in} > x(1-\epsilon(r_i-d_i))|A_{ij} < \epsilon x, j \ge 1\} \\
&\le \mathbb{P}\{\sup_{n\le N_i^U(Mx)} S_{in} > x(1-\epsilon(r_i-d_i))|A_{ij} < \epsilon x, j \ge 1\} \\
&= \mathbb{P}\{\sup_{n\le N_i^U(Mx)} S_{in} > x(1-\epsilon(r_i-d_i))|A_{ij} < \epsilon x, j = 1,\ldots,N_i^U(Mx)\} \\
&\le \mathbb{P}\{\sup_{n\le M_2 x} S_{in} > x(1-\epsilon(r_i-d_i))|A_{ij} < \epsilon x, j \ge 1\} + \mathbb{P}\{N_i^U(Mx) > M_2 x\}.
\end{aligned}
$$

The second term decays exponentially fast in $x$ if $M_2 > \lambda_i M$. The first term can be bounded by

$$
\sum_{m=1}^{M_2 x} \mathbb{P}\{S_{im} > x(1-\epsilon(r_i-d_i))|A_{ij} \le \epsilon x, j = 1,\ldots,m\}.
$$

According to Lemma 2.4.1, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\beta}$ with $\beta > \mu + 1$, such that for $\epsilon \in (0, \epsilon^*]$ the last quantity is upper bounded by $M_2 x \phi(x)$. The latter function is regularly varying of index $1 - \beta < -\mu$. Invoking Lemma 7.6.2 then completes the proof. □

The next proposition shows that each source has *at most* one long On-period.

**Proposition 7.6.3** *For all $i$, all $M$ and all $\epsilon > 0$,*

$$
\mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) \ge 2\} = \mathrm{o}(\mathbb{P}\{V^c > x\}),
$$

*as $x \to \infty$.*

**Proof**
Without loss of generality we may take $i = 1$. By Proposition 7.6.2 it suffices to show that

$$
\mathbb{P}\{V^c(Mx) > x, \mathcal{N}_1(\epsilon x, Mx) \ge 2, \mathcal{N}_i(\epsilon x, Mx) \ge 1, i \ge 2\} = \mathrm{o}(\mathbb{P}\{V^c > x\}).
$$

Note that the left hand side is bounded by

$$
\mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \ge 2\} \prod_{i=2}^{N} \mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \ge 1\}.
$$

Thus, invoking Lemma 7.6.2 it suffices to show that:
(i) $\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \ge 1\}$ is bounded by a function which is regularly varying of index $1-\nu_i$;
(ii) $\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \ge 2\} = \mathrm{o}(\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \ge 1\})$.
We will prove both assertions for $i = 1$. For assertion (i), note that

$$
\mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \ge 1\} \le p_1 \mathbb{P}\{A_1^r \ge \epsilon x\} + \mathbb{P}\{\#\{j \in \{1,\ldots,N_1^U(Mx)\} : A_{1j} \ge \epsilon x\} \ge 1\}.
$$

The first term is in $\mathcal{R}_{1-\nu_1}$. By conditioning upon $N_1^U(Mx)$, the second term can be bounded by $\mathbb{E}\{N_1^U(Mx)\}\mathbb{P}\{A_1 \geq \epsilon x\}$, which is also regularly varying of index $1-\nu_1$. To prove assertion (ii), note that

$$\begin{aligned}\mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \geq 2\} &\leq p_1 \mathbb{P}\{A_1^r \geq \epsilon x\}\mathbb{P}\{\mathcal{N}_1((\epsilon x, \infty), (0, Mx]) \geq 1\} \\ &+ \mathbb{P}\{\mathcal{N}_1((\epsilon x, \infty), (0, Mx]) \geq 2\}.\end{aligned}$$

Using $\mathbb{P}\{\mathcal{N}_1((\epsilon x, \infty), (0, Mx]) \geq 1\} \leq \mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \geq 1\}$ and assertion (i), it follows that the first term is of $o(\mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \geq 1\})$. To bound the second term, condition (again) on $N_1^U(Mx)$. This yields

$$\mathbb{P}\{\mathcal{N}_1((\epsilon x, \infty), (0, Mx)) \geq 2\} \leq \mathbb{E}\{N_1^U(Mx)^2\}\mathbb{P}\{A_1 \geq \epsilon x\}^2.$$

Finally, note that $\mathbb{E}\{N_1^U(Mx)^2\}$ is quadratic in $x$ for $x \to \infty$.

□

We have now shown that, with overwhelming probability, each source contributes to a large value of $\sup_{t\geq 0}\{A(0,t) - ct\}$ through exactly one long On-period. We thus proceed with the second statement (as indicated at the beginning of this subsection), implying that apart from these long On-periods, the sources show typical behavior. In order to formalize that statement, we need to introduce some notation. Define

$$\tau(y) := \inf\{t \geq 0 : A(0,t) - ct = y\}$$

as the first time at which the process $\{A(0,t) - ct\}$ reaches level $y$.
For fixed $\epsilon > 0$ and $x$, let $\tau_{s,i}(\epsilon x)$ and $\tau_{f,i}(\epsilon x)$ be the respective starting and finishing times of the first On-period of source $i$ exceeding length $\epsilon x$. Denote

$$\tau_s(\epsilon x) := \max_{i=1,\ldots,N} \tau_{s,i}(\epsilon x)$$

and

$$\tau_f(\epsilon x) := \min_{i=1,\ldots,N} \tau_{f,i}(\epsilon x).$$

Note that all sources are in the middle of their long On-periods between times $\tau_s(\epsilon x)$ and $\tau_f(\epsilon x)$. We will show that the fluctuations of the process $\{A(0,t) - ct\}$ away from the mean before time $\tau_s(\epsilon x)$ and after time $\tau_f(\epsilon x)$ can be neglected.
More formally, the next two propositions show that, given that the workload reaches level $x$ before time $Mx$, there exists for any small $\delta > 0$ an $\epsilon_\delta$ such that for all $\epsilon \in (0, \epsilon_\delta)$,

$$\tau_s(\epsilon x) \leq \tau(\delta x) < \tau((1-\delta)x) \leq \tau_f(\epsilon x).$$

Thus, the workload remains small up to time $\tau_s(\epsilon x)$, and reaches a level close to $x$ before time $\tau_f(\epsilon x)$, as depicted in Figure 7.2.
The first proposition indicates that it is most unlikely that the process $\{A(0,t) - ct\}$ reaches level $\delta x$ before time $\tau_s(\epsilon x)$.

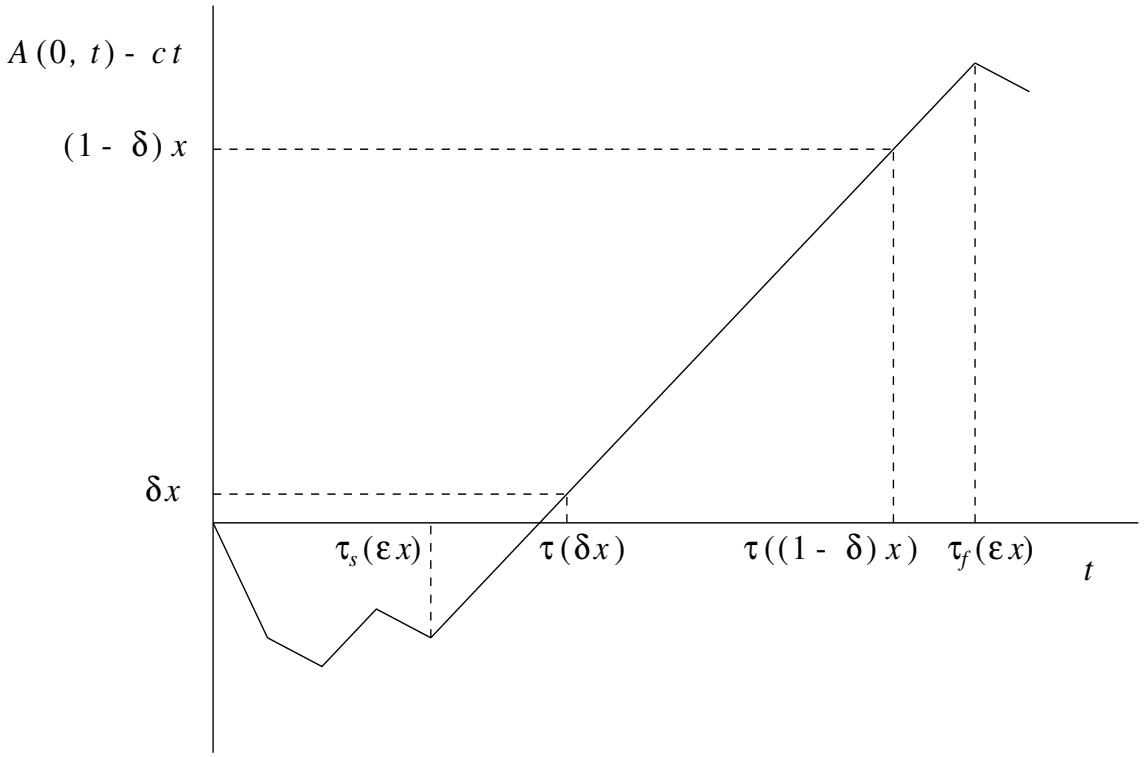


Figure 7.2: Typical path to overflow

**Proposition 7.6.4** *For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*]$,*

$$\mathbb{P}\{\tau(\delta x) < \tau_s(\epsilon x)\} = \mathrm{o}(\mathbb{P}\{V^c > x\}).$$

**Proof**
For compactness, denote $\tau_s \equiv \tau_s(\epsilon x)$, $\tau_{s,i} \equiv \tau_{s,i}(\epsilon x)$. Then

$$\mathbb{P}\{\tau(\delta x) < \tau_s\} = \mathbb{P}\{V^c(\tau_s) > \delta x\} \leq \sum_{i=1}^{N} \mathbb{P}\{V^c(\tau_{s,i}) > \delta x\}.$$

We bound each term in the last summation.
Define $N_i(\epsilon x) := N_i^A(\tau_{s,i}^-)$ as the number of On-periods initiated by source $i$ before the first On-period exceeding length $\epsilon x$. Note that $N_i(\epsilon x) + 1$ is geometrically distributed with parameter $\mathbb{P}\{A_i > \epsilon x\}$.
Using Lemma 7.6.1, taking $t = \tau_{s,i}$,

$$\begin{aligned}
& \mathbb{P}\{V^c(\tau_{s,i}) > \delta x\} \\
= \; & \mathbb{P}\{V^c(\tau_{s,i}) > \delta x, \mathcal{N}_i((\epsilon x, \infty), [0, \tau_{s,i})) = 0\} \\
\leq \; & \mathbb{P}\{\sup_{n \leq N_i(\epsilon x)} S_{in} > x(\delta - \epsilon(r_i - d_i)), A_{ij} \leq \epsilon x, j = 1, \ldots, N_i(\epsilon x)\} \\
\leq \; & \sum_{m=1}^{\infty} \mathbb{P}\{N_i(\epsilon x) = m\} \mathbb{P}\{\sup_{n \leq m} S_{in} > x(\delta - \epsilon(r_i - d_i)), A_{ij} \leq \epsilon x, j = 1, \ldots, m\}
\end{aligned}$$

$$\leq \sum_{m=1}^{\infty} \mathbb{P}\{N_i(\epsilon x) = m\}\mathbb{P}\{\sup_{n\leq m} S_{in} > x(\delta - \epsilon(r_i - d_i))|A_{ij} \leq \epsilon x, j = 1, \ldots, m\}.$$

According to Lemma 2.4.1, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\beta}$ with $\beta > 2\nu + 1$, such that for $\epsilon \in (0, \epsilon^*)$ the last quantity is upper bounded by

$$\mathbb{E}\{N_i(\epsilon x)\}\phi(x) = \frac{\phi(x)\mathbb{P}\{A_i \leq \epsilon x\}}{\mathbb{P}\{A_i > \epsilon x\}},$$

which is regularly varying of index $\nu_i - \beta < \mu + 1 - (2\mu + 1) = -\mu$.
Invoking Lemma 7.6.2 then completes the proof.

□

The next proposition shows that, given that the process $\{A(0,t) - ct\}$ reaches level $x$ before time $Mx$, most probably level $(1-\delta)x$ is crossed before time $\tau_f(\epsilon x)$.

**Proposition 7.6.5** *For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$ and $M < \infty$,*

$$\mathbb{P}\{\tau((1-\delta)x) > \tau_f(\epsilon x), V^c(Mx) > x\} = \mathrm{o}(\mathbb{P}\{V^c > x\}).$$

**Proof**
For conciseness, denote $\tau_f \equiv \tau_f(\epsilon x)$, $\tau_{f,i} \equiv \tau_{f,i}(\epsilon x)$. By Propositions 7.6.2 and 7.6.3, it suffices to show that

$$\begin{aligned} & \mathbb{P}\{\tau((1-\delta)x) > \tau_f, V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 1 \text{ for all } i = 1, \ldots, N\} \\ = \ & \mathrm{o}(\mathbb{P}\{V^c > x\}). \end{aligned}$$

Note that

$$\begin{aligned} & \mathbb{P}\{\tau((1-\delta)x) > \tau_f, V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 1 \text{ for all } i = 1, \ldots, N\} \\ = \ & \mathbb{P}\{V^c(\tau_f) > (1-\delta)x, V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 1 \text{ for all } i = 1, \ldots, N\} \\ \leq \ & \sum_{i=1}^{N} \mathbb{P}\{V^c(\tau_{f,i}) > (1-\delta)x, V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 1\}. \end{aligned}$$

As before, we bound each term in the last summation.

$$\begin{aligned} & \mathbb{P}\{V^c(\tau_{f,i}) > (1-\delta)x, V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 1\} \\ \leq \ & \mathbb{P}\{\sup_{0\leq t\leq \tau_{f,i}} \{A(0,t) - ct\} < (1-\delta)x, \\ & \qquad \sup_{0\leq t\leq Mx} \{A(0,t) - ct\} > x, \mathcal{N}_i((\epsilon x, \infty), (\tau_{f,i}, Mx]) = 0\} \\ \leq \ & \mathbb{P}\{\sup_{\tau_{f,i}\leq t\leq Mx} \{A(\tau_{f,i}, t) - c(t - \tau_{f,i})\} > \delta x, \mathcal{N}_i((\epsilon x, \infty), (\tau_{f,i}, Mx]) = 0\} \\ \leq \ & \mathbb{P}\{\sup_{\tau_{f,i}\leq t\leq Mx} \{A_i(\tau_{f,i}, t) - d_i(t - \tau_{f,i})\} > \delta x, \mathcal{N}_i((\epsilon x, \infty), (\tau_{f,i}, Mx]) = 0\}. \end{aligned}$$

The first inequality follows from the definitions. The second inequality follows from properties of the sup operator, and the last inequality is obtained by assuming that all sources but $i$ are On between times $\tau_{f,i}$ and $Mx$.
Note that the last probability is upper bounded by

$$\mathbb{P}\{\sup_{N_i(\epsilon x)+2\leq n\leq N_i^A(Mx)} S_{in} - S_{i,N_i(\epsilon x)+1} > \delta x, A_j \leq \epsilon x, N_i(\epsilon x)+2 \leq j \leq N_i^A(Mx)\}.$$

The latter probability can be upper bounded by a function which is regularly varying of index $-\beta < -\mu$ in a similar fashion as in the proof of Propositions 7.6.2 and 7.6.4.
The proof is completed by invoking Lemma 7.6.2.

□

Propositions 7.6.4, 7.6.5 may be used to obtain the following result.

**Corollary 7.6.1** *If $A_j(\cdot) \in \mathcal{R}$ for all $j = 1, \ldots, N$, then $\mathbb{P}\{V^c > x\} \in \mathcal{IRV}$.*

The above result suffices to prove the reduced-load equivalence (see Section 7.5, in particular Proposition 7.5.1, for the details). However, determining the exact asymptotic behavior of $\mathbb{P}\{V^c > x\}$ requires further analysis, to be found in Subsections 7.6.3 and 7.6.4. In particular, the analysis in Subsection 7.6.4 will lead to a sharper version of Corollary 7.6.1, showing that $\mathbb{P}\{V^c > x\} \in \mathcal{R}$ (which is a strict subset of $\mathcal{IRV}$).
Nevertheless, we sketch a direct proof of Corollary 7.6.1 which we believe is of independent interest. For the formal proof details we refer to Appendix 7.A.

**Sketch of proof**
The idea of the proof is as follows. If $V^c > x$, then Propositions 7.6.4 and 7.6.5 show that the process $\{A(0,t) - ct\}$ reaches the level $(1-\delta)x$ after all sources have been On for at least $\frac{(1-2\delta)x}{r-c}$ time units. Since $A_j(\cdot) \in \mathcal{R} \subseteq \mathcal{IRV}$ for all $j = 1, \ldots, N$, with high probability, all sources remain On for at least $\frac{2\delta x}{r-c}$ more time units. This yields

$$\lim_{\delta \downarrow 0} \liminf_{x\to\infty} \mathbb{P}\{V^c > (1+\delta)x | V^c > x\} = 1,$$

implying the desired statement (by definition).

□

## 7.6.3 Proof of Theorem 7.6.1

In this subsection we give a proof of Theorem 7.6.1. First we consolidate the key results from the previous subsection in the following theorem.

**Theorem 7.6.2** *For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$,*

$$\mathbb{P}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \leq \mathbb{P}\{V^c > x\} \lesssim \mathbb{P}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\}.$$

**Proof**
The lower bound is trivial. The upper bound follows from Propositions 7.6.1, 7.6.4, and 7.6.5.

□

In order to obtain tight bounds for the probabilities in Theorem 7.6.2, we condition upon $\tau_{s,i}$ for all $i$. Hence, for any $\mathcal{J}_0 \subseteq \mathcal{J}$, define the event $D_{\mathcal{J}_0}(\epsilon x)$ by

$$D_{\mathcal{J}_0}(\epsilon x) := \{\tau_{s,i}(\epsilon x) = 0 \text{ for all } i \in \mathcal{J}_0; \tau_{s,i}(\epsilon x) > 0 \text{ for all } i \notin \mathcal{J}_0\}.$$

The event $D_{\mathcal{J}_0}(\epsilon x)$ implies that the sources $i \in \mathcal{J}_0$ started their long On-period before time 0 (remember that we consider the system in stationarity). The sources $i \in \mathcal{J}_1$ start their long On-period at a later time epoch.
Denote $\mathbb{P}_{\mathcal{J}_0}\{\cdot\} = \mathbb{P}\{\cdot | D_{\mathcal{J}_0}(\epsilon x)\}$. The following two lemmas will be useful for providing tight upper and lower bounds for the probabilities in Theorem 7.6.2.

**Lemma 7.6.3** *(Upper bound) For any $\delta > 0$, there exists an $\epsilon_\delta > 0$ such that for all $\epsilon \in (0, \epsilon_\delta)$*

$$\mathbb{P}_{\mathcal{J}_0}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\} \prod_{i \in \mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\} \lesssim P_{\mathcal{J}_0}((1-\delta)x) \prod_{i \in \mathcal{J}_1} p_i,$$

*with $P_{\mathcal{J}_0}((1-\delta)x)$ as in (6.1).*

**Lemma 7.6.4** *(Lower bound) There exists an $\epsilon > 0$ such that*

$$\mathbb{P}_{\mathcal{J}_0}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \prod_{i \in \mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\} \gtrsim P_{\mathcal{J}_0}(x) \prod_{i \in \mathcal{J}_1} p_i,$$

*with $P_{\mathcal{J}_0}(x)$ as in (6.1).*

The proofs of these lemmas are quite technical, and are deferred to Appendices 7.B and 7.C. A brief sketch of the proofs is given at the end of this subsection.

We now have gathered all the ingredients for the proof of Theorem 7.6.1.

**Proof of Theorem 7.6.1**
The lower bound in Theorem 7.6.2 may be written as

$$\begin{aligned} & \mathbb{P}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \\ = & \sum_{\mathcal{J}_0 \subseteq \{1,\ldots,N\}} \mathbb{P}_{\mathcal{J}_0}\{A(0, \tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \mathbb{P}\{D_{\mathcal{J}_0}(\epsilon x)\}. \end{aligned}$$

Note that

$$\mathbb{P}\{D_{\mathcal{J}_0}(\epsilon x)\} \sim \prod_{i\in\mathcal{J}_0} p_i \mathbb{P}\{A_i^r > \epsilon x\}.$$

Using Lemma 7.6.4, we then obtain

$$\mathbb{P}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \gtrsim \left(\prod_{j=1}^N p_j\right) \sum_{\mathcal{J}_0\subseteq\{1,\ldots,N\}} P_{\mathcal{J}_0}(x).$$

Similarly, using Lemma 7.6.3,

$$\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\} \lesssim \left(\prod_{j=1}^N p_j\right) \sum_{\mathcal{J}_0\subseteq\{1,\ldots,N\}} P_{\mathcal{J}_0}((1-\delta)x).$$

Theorem 7.6.2 then gives

$$\left(\prod_{j=1}^N p_j\right) \sum_{\mathcal{J}_0\subseteq\{1,\ldots,N\}} P_{\mathcal{J}_0}(x) \lesssim \mathbb{P}\{V^c > x\} \lesssim \left(\prod_{j=1}^N p_j\right) \sum_{\mathcal{J}_0\subseteq\{1,\ldots,N\}} P_{\mathcal{J}_0}((1-\delta)x),$$

which implies Theorem 7.6.1, since $P_{\mathcal{J}_0}(x) \in \mathcal{R}$ as will be shown in Theorem 7.6.3.

□

In preparation for the proofs of Lemmas 7.6.3 and 7.6.4, we give a convenient representation for $A(0,\tau_f) - c\tau_f$ under the event $D_{\mathcal{J}_0}(\epsilon x)$.

**Lemma 7.6.5** *Under the event $D_{\mathcal{J}_0}(\epsilon x)$, $A(0,\tau_f) - c\tau_f$ can be represented as*

$$A(0,\tau_f) - c\tau_f = \min\{\min_{i\in\mathcal{J}_0} F_i, \min_{i\in\mathcal{J}_1} G_i\},$$

*where $\mathcal{J}_1 = \mathcal{J} \setminus \mathcal{J}_0$. The random variables $F_i$ and $G_i$ are given by*

$$\begin{aligned}
F_i &= (r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k \left[ I_k A_k^r(\epsilon x) + (1-I_k)[A_k(\epsilon x) + U_k^r] + \sum_{j=1}^{N_k(\epsilon x)} U_{kj} \right], \\
G_i &= (r-c)\bar{A}_i(\epsilon x) + (r-c)\left[ I_i A_i^r(\epsilon x) + (1-I_i)A_i(\epsilon x) + \sum_{j=1}^{N_i(\epsilon x)} A_{ij}(\epsilon x) \right] - \\
&\qquad d_i \left[ (1-I_i)U_i^r + \sum_{j=1}^{N_i(\epsilon x)} U_{ij} \right] - \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k \left[ (1-I_k)U_k^r + \sum_{j=1}^{N_k(\epsilon x)} U_{kj} \right].
\end{aligned}$$

*Here $\bar{A}_i(\epsilon x) = A_i | A_i > \epsilon x$, $\bar{A}_i^r(\epsilon x) = A_i^r | A_i^r > \epsilon x$, $A_{ij}(\epsilon x) \stackrel{d}{=} A_{ij} | A_{ij} \le \epsilon x$, and $A_i^r(\epsilon x) \stackrel{d}{=} A_i^r | A_i^r \le \epsilon x$.*

**Proof**
Under the event $D_{\mathcal{J}_0}(\epsilon x)$, the random variables $\tau_{s,i}$, $i \in \mathcal{J}_1$, can be represented as

$$\tau_{s,i} = I_i A_i^r(\epsilon x) + (1 - I_i)[U_i^r + A_i(\epsilon x)] + \sum_{j=1}^{N_i(\epsilon x)} [U_{ij} + A_{ij}(\epsilon x)], \qquad i \in \mathcal{J}_1.$$

Combined with the identities

$$\begin{aligned} A_i(0, \tau_{s,i}) &= r_i[I_i A_i^r(\epsilon x) + (1 - I_i)A_i(\epsilon x) + \sum_{j=1}^{N_i(\epsilon x)} A_{ij}(\epsilon x)], \\ \tau_f &= \min\{\min_{i \in \mathcal{J}_0} \bar{A}_i^r(\epsilon x), \min_{i \in \mathcal{J}_1}\{\bar{A}_i(\epsilon x) + \tau_{s,i}\}\}, \\ A_i(\tau_{s,i}, \tau_f) &= r_i(\tau_f - \tau_{s,i}), \end{aligned}$$

the representation for $A(0, \tau_f) - c\tau_f$ then easily follows.

□

We now give a brief sketch of the proofs of Lemmas 7.6.3 and 7.6.4. Both rely on the above representation for $A(0, \tau_f) - c\tau_f$ in terms of the variables $F_i$ and $G_i$. The proofs of the lemmas have a similar structure.

- The expressions for $F_i$ and $G_i$ are quite complicated, so an attempt to obtain the exact joint distribution does not seem promising. Therefore, the first step is to show that all random variables $A_{ij}(\epsilon x)$ and $U_{ij}$ can be replaced by their means;

- The above point indicates that $F_i$ and $G_i$ may be approximated as follows.

$$\begin{aligned} F_i &\approx (r - c)\bar{A}_i^r(\epsilon x) + \sum_{k \in \mathcal{J}_1} r_k \mathbb{E}\{U_k\} N_k(\epsilon x), \\ G_i &\approx (r - c)\bar{A}_i(\epsilon x) + [(r - c)\mathbb{E}\{A_i\} - d_i \mathbb{E}\{U_i\}]N_i(\epsilon x) - \\ &\quad \sum_{k \in \mathcal{J}_1 \setminus \{i\}} r_k \mathbb{E}\{U_k\} N_k(\epsilon x). \end{aligned}$$

  It will be useful to keep these approximations in mind. The formulas in Appendices 7.B and 7.C look much more cumbersome by the appearance of many additional, but small constants;

- The only random variables appearing in the above expressions are $\bar{A}_i(\epsilon x)$, $B_i^r(\epsilon x)$, and $N_i(\epsilon x)$, of which the distributions are known. What thus remains is a straightforward computation.

The first point causes the most technical difficulties. It requires a separate treatment in the proofs of Lemmas 7.6.3 and 7.6.4. Details may be found in the appendices.

### 7.6.4 Computation of the pre-factor

In this subsection we give an asymptotic characterization of $P_{\mathcal{J}_0}(x)$, which may be useful for further analysis. In particular, we establish that $P_{\mathcal{J}_0}(x)$ and $\mathbb{P}\{V^c > x\}$ are both regularly varying, and provide expressions for the pre-factors in their asymptotic expansions. Assume that $\mathcal{J}_0$ is a proper subset of $\mathcal{J}$, observing

$$P_{\mathcal{J}}(x) = \prod_{i\in\mathcal{J}} \mathbb{P}\{A_i^r > \frac{x}{r-c}\}.$$

For every set $\mathcal{J}_0$, define the $|\mathcal{J}_1|$-vector $g$ by

$$g := \left(\frac{r_j - \rho_j}{r-c}\right)_{j\in\mathcal{J}_1}.$$

Let $G$ be a (square) matrix with identical rows $g$, and let $\bar{G} := G - I$, with $I$ the identity matrix of dimension $|\mathcal{J}_1|$.
It can easily be shown that $\bar{G}$ is invertible; denote its inverse by $H$. A straightforward computation yields $H = \frac{1}{ge-1}G - I$, with $e = (1,\ldots,1)$, which implies that $gH = \frac{1}{ge-1}g$. A further straightforward computation shows $|\bar{G}| = eg - 1$.
Define $y = (y_i)_{\mathcal{J}_1}$ and $\mathrm{d}y = \prod\limits_{i\in\mathcal{J}_1} \mathrm{d}y_i$. Then we may write

$$P_{\mathcal{J}_0}(x) = \frac{1}{\prod\limits_{i\in\mathcal{J}_1}\mathbb{E}\{A_i\}} \int_{y\geq 0} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{A_i > (\bar{G}y)_i + \frac{x}{r-c}\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > gy + \frac{x}{r-c}\}\mathrm{d}y.$$

If we integrate w.r.t. $z := \bar{G}y$ (note that $\bar{G}$ is a positive matrix), then we obtain (defining $A_{\mathcal{J}_1} = (A_i^r)_{i\in\mathcal{J}_1}$)

$$\begin{aligned}
& P_{\mathcal{J}_0}(x) \\
= \;& \frac{1}{|\bar{G}|\prod\limits_{i\in\mathcal{J}_1}\mathbb{E}\{A_i\}} \int_{z\geq 0} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{A_i > z_i + \frac{x}{r-c}\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > gHz + \frac{x}{r-c}\}\mathrm{d}z \\
= \;& \frac{1}{eg-1} \int_{z\geq 0} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > \frac{1}{eg-1}gz + \frac{x}{r-c}\} \prod_{i\in\mathcal{J}_1} \mathrm{d}\mathbb{P}\{A_i^r \leq z_i + \frac{x}{r-c}\} \\
= \;& \frac{1}{eg-1}\mathbb{P}\{A_i^r \geq \frac{x}{r-c}, i\in\mathcal{J}; A_k^r - \frac{x}{r-c} \geq \frac{1}{eg-1}g\left(A_{\mathcal{J}_1}^r - e\frac{x}{r-c}\right), k\in\mathcal{J}_1\}.
\end{aligned}$$

We conclude that $P_{\mathcal{J}_0}(x)$ can be written (up to a constant) as the probability that $(A_i^r)_{i\in\mathcal{J}}$ belongs to a certain set. We now show that $P_{\mathcal{J}_0}(x)$ is regularly varying of index $-\mu$ (recall that $\mu = \sum\limits_{i=1}^{N}(\nu_i - 1)$). If $A_i$ is regularly varying of index $-\nu_i < -1$, then it is well-known and elementary to show that

$$\mathbb{P}\{\frac{A_i^r - \gamma x}{x} > y | A_i^r > \gamma x\} \to \left(1 + \frac{y}{\gamma}\right)^{1-\nu_i},$$

as $x \to \infty$. Let $Z_i$ be a random variable with the above limiting distribution, with $\gamma = \frac{1}{r-c}$ such that the $Z_i$, $i \in \mathcal{J}_1$ are independent. The above computations are summarized in the following theorem.

**Theorem 7.6.3** $$P_{\mathcal{J}_0}(x) \sim \kappa_{\mathcal{J}_0} \prod_{i=1}^{N} \mathbb{P}\{A_i^r > \frac{x}{r-c}\},$$

*with* $\kappa_{\mathcal{J}} = 1$ *and*

$$\kappa_{\mathcal{J}_0} = \frac{1}{eg-1}\mathbb{P}\{Z_i \geq \frac{1}{eg-1} g Z_{\mathcal{J}_1}, i \in \mathcal{J}_0\}$$

*if* $\mathcal{J}_0$ *is a proper subset of* $\mathcal{J}$. *In particular,* $P_{\mathcal{J}_0}(x)$ *is regularly varying of index* $-\mu$.

Combining Theorems 7.6.1 and 7.6.3, we obtain

**Theorem 7.6.4** $$\mathbb{P}\{V^c > x\} \sim \kappa \prod_{i=1}^{N} p_i \mathbb{P}\{A_i^r > \frac{x}{r-c}\},$$

*with*

$$\kappa = \sum_{\mathcal{J}_0 \subseteq \{1,\ldots,N\}} \kappa_{\mathcal{J}_0}.$$

*In particular,* $\mathbb{P}\{V^c > x\}$ *is regularly varying of index* $-\mu$.

The above theorem is used in proving the reduced-load equivalence (see Section 7.5), and may be potentially useful for computational purposes.

In particular, in the case of two On-Off sources, the computation of $\kappa$ is as difficult as the computation of $\kappa_1$ and $\kappa_2$. Using the probabilistic interpretation of these constants readily leads to an integral expression, which can be evaluated explicitly when both $\nu_1$ and $\nu_2$ are integer-valued. We omit the details.

## 7.7 $K$ heterogeneous classes: proofs

In this section we provide the proofs of the results in Section 7.3.4 for the case with $K$ heterogeneous classes of On-Off sources. In particular, we present a proof of Theorem 7.3.3. We start with the regime where we first let $x \to \infty$ and then $n \to \infty$. For every $n$ we have, using Theorem 7.3.2,

$$\lim_{x\to\infty} \frac{\mathbb{P}\{V^{(n)} > nx\}}{\log x} = -\mu^{(n)},$$

with $\mu^{(n)}$ denoting the optimal value of the criterion function of the associated knapsack problem. It thus remains to be shown that

$$\lim_{n\to\infty} \frac{\mu^{(n)}}{n\mu} = 1. \tag{7.1}$$

First observe that the optimal value of the continuous relaxation of the knapsack problem is $n\mu$, yielding a lower bound for $\mu^{(n)}$. On the other hand, the continuous relaxation may be used to construct a feasible solution of the knapsack problem. Take (use the notation of Section 7.3.4) $q_k = n_k = na_k$ for $k < \ell$, $q_k = n_k = 0$ for $k > \ell$, and $q_\ell = |n_\ell| + 1$. This is a feasible solution with a value at most $n\mu + \gamma_\ell$, giving an upper bound for $\mu^{(n)}$. In conclusion, we have

$$n\mu \le \mu^{(n)} \le n\mu + \gamma_\ell,$$

from which (7.1) directly follows.

We now turn to the regime where we first let $n \to \infty$ and then $x \to \infty$ (i.e., the many-sources regime). Define the 'decay rate'

$$I(x) := -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{V^{(n)} > nx\}.$$

It needs to be shown that $I(x) \sim \mu \log x$ as $x \to \infty$.
The above decay rate equals [62, page 300]

$$I(x) = \inf_{t\ge 0} \sup_{\theta} \left( \theta(x+t) - \sum_{k=1}^{K} a_k \log \mathbb{E}\{\mathrm{e}^{\theta A_k(t)}\} \right),$$

with $A_k(t) := A_k(0,t)$ representing the amount of traffic generated by a single class-$k$ source in a time interval of length $t$ in steady state. Replacing $\theta$ by $\theta(\log t)/t$, we obtain an alternative variational problem:

$$\inf_{t\ge 0} \log t \cdot J_t\left(\frac{x}{t} + 1\right), \text{ where } J_t(x) := \sup_{\theta} \left( \theta x - \sum_{k=1}^{K} a_k \frac{\log \mathbb{E}\{\mathrm{e}^{\theta(\log t) A_k(t)/t}\}}{\log t} \right), \quad (7.2)$$

for $x \in (0, \sum_{k=1}^{K} a_k r_k)$. The latter variational problem allows direct asymptotic analysis ($x \to \infty$) as in [195], which yields Theorem 7.7.1 below.
First, however, we state an auxiliary lemma. Recall that $\sigma_k = \sum_{m=1}^{k-1} a_k r_k + \sum_{m=k}^{K} a_k \rho_k$, and that the $K$ classes are indexed in non-decreasing order of the ratios $\gamma_k = (\nu_k - 1)/(r_k - \rho_k)$.

**Lemma 7.7.1** *For $\theta \ge 0$,*

$$\lim_{t\to\infty} \frac{\log \mathbb{E}\{\mathrm{e}^{\theta(\log t) A_k(t)/t}\}}{\log t} = \max\{\theta\rho_k, \theta r_k - \nu_k + 1\},$$

*so that the cumulant function of the superposition is piecewise linear:*

$$\sum_{k=1}^{K} a_k \lim_{t\to\infty} \frac{\log \mathbb{E}\{\mathrm{e}^{\theta(\log t) A_k(t)/t}\}}{\log t} = \sum_{k=1}^{K} a_k \max\{\theta\rho_k, \theta r_k - \nu_k + 1\}.$$

*Furthermore,*

$$\lim_{t\to\infty} J_t(x) = \gamma_{\ell(x)}x - \sum_{k=1}^{\ell(x)-1} a_k(\gamma_{\ell(x)}r_k - \nu_k + 1) - \sum_{k=\ell(x)}^{K} a_k\gamma_{\ell(x)}\rho_k, \tag{7.3}$$

*for* $x \in (0, \sum_{k=1}^{K} a_k r_k)$*, where* $\ell(x)$ *is such that* $x \in (\sigma_{\ell(x)-1}, \sigma_{\ell(x)})$*.*
*The function* $\lim_{t\to\infty} J_t(\cdot)$ *is increasing.*

The proof of the above lemma is analogous to that of Theorem 3.6 and Lemma 3.7 of [195].

**Theorem 7.7.1** *(Large-buffer asymptotics)*

$$\lim_{x\to\infty} \frac{I(x)}{\log x} = \mu,$$

*with* $\mu = \sum_{k=1}^{\ell-1} a_k(\nu_k - 1) + (1 - \sigma_{\ell-1})\gamma_\ell$ *and* $\ell := \ell(1)$*.*

**Proof**
The proof consists of deriving an upper bound and a lower bound which asymptotically coincide.

*Upper bound*
Using the representation (7.2),

$$\limsup_{x\to\infty} \frac{I(x)}{\log x} = \limsup_{x\to\infty} \inf_{t>0} \frac{\log t}{\log x} J_t\left(\frac{x}{t} + 1\right).$$

Substituting $t = x/s$, $s \in (0, \sum_{k=1}^{K} a_k r_k - 1)$, to obtain an upper bound, and using (7.3),

$$\begin{aligned}
& \limsup_{x\to\infty} \inf_{t>0} \frac{\log t}{\log x} J_t\left(\frac{x}{t} + 1\right) \\
\leq\ & \limsup_{x\to\infty} \frac{\log(x/s)}{\log x} J_{x/s}(s+1) \\
\leq\ & \limsup_{x\to\infty} \frac{\log(x/s)}{\log x} \limsup_{x\to\infty} J_{x/s}(s+1) \\
\leq\ & \limsup_{x\to\infty} J_{x/s}(s+1) \\
=\ & \gamma_{\ell(s+1)}(s+1) - \sum_{k=1}^{\ell(s+1)-1} (a_k\gamma_{\ell(s+1)}r_k - \nu_k + 1) - \sum_{k=\ell(s+1)}^{K} a_k\gamma_{\ell(s+1)}\rho_k.
\end{aligned}$$

The above inequality holds for any $s \in (0, \sum_{k=1}^{K} a_k r_k - 1)$. According to Lemma 7.7.1, the last term is increasing in $s+1$. Letting $s \downarrow 0$ to obtain the sharpest possible upper bound, we obtain

$$\limsup_{x\to\infty} \frac{I(x)}{\log x} \le \gamma_\ell - \sum_{k=1}^{\ell-1} a_k(\gamma_\ell r_k - \nu_k + 1) - \sum_{k=\ell}^{K} a_k \gamma_\ell \rho_k = \mu.$$

*Lower bound*
Using the representation (7.2), and taking $\theta = \gamma_\ell$, we obtain the lower bound

$$\begin{aligned} I(x) &= \inf_{t\ge 0} \log t \cdot \sup_{\theta} \left( \theta \left( \frac{x}{t} + 1 \right) - \sum_{k=1}^{K} a_k \frac{\log \mathbb{E}\{ \mathrm{e}^{\theta (\log t) A_k(t)/t} \}}{\log t} \right) \\ &\ge \inf_{t\ge 0} \log t \cdot \left( \gamma_\ell \left( \frac{x}{t} + 1 \right) - \sum_{k=1}^{K} a_k \frac{\log \mathbb{E}\{ \mathrm{e}^{\gamma_\ell (\log t) A_k(t)/t} \}}{\log t} \right). \end{aligned}$$

The optimizing value of $t$ in the above variational problem is *at least* linear in $x$, for large $x$. Formally, there exists a $d$ such that the above infimum needs to be taken only over $t > dx$, for large $x$. This may be proven analogously to case (iii) of [119, page 258]. Thus,

$$I(x) \ge \inf_{t>dx} \log t \cdot \left( \gamma_\ell \left( \frac{x}{t} + 1 \right) - \sum_{k=1}^{K} a_k \frac{\log \mathbb{E}\{ \mathrm{e}^{\gamma_\ell (\log t) A_k(t)/t} \}}{\log t} \right).$$

Using (7.3), we find that for any $\epsilon > 0$, and $x$ large enough, we have for all $t > dx$,

$$\sum_{k=1}^{K} a_k \frac{\log \mathbb{E}\{ \mathrm{e}^{\gamma_\ell \log t A_k(t)/t} \}}{\log t} \le (1+\epsilon) \sum_{k=1}^{K} a_k \max\{\gamma_\ell \rho_k, \gamma_\ell r_k - \nu_k + 1\}.$$

Thus,

$$\begin{aligned} & \liminf_{x\to\infty} \frac{I(x)}{\log x} \\ \ge\ & \liminf_{x\to\infty} \inf_{t>dx} \frac{\log t}{\log x} \left( \gamma_\ell \left( \frac{x}{t} + 1 \right) - (1+\epsilon) \sum_{k=1}^{K} a_k \max\left\{ \gamma_\ell \rho_k, \gamma_\ell r_k - \nu_k + 1 \right\} \right) \\ \ge\ & \liminf_{x\to\infty} \inf_{t>dx} \frac{\log t}{\log x} \inf_{t>dx} \left( \gamma_\ell \left( \frac{x}{t} + 1 \right) - (1+\epsilon) \sum_{k=1}^{K} a_k \max\{ \gamma_\ell \rho_k, \gamma_\ell r_k - \nu_k + 1 \} \right) \\ \ge\ & \gamma_\ell - (1+\epsilon) \sum_{k=1}^{K} a_k \max\{\gamma_\ell \rho_k, \gamma_\ell r_k - \nu_k + 1\}. \end{aligned}$$

Letting $\epsilon \downarrow 0$, we obtain

$$\begin{aligned}
\liminf_{x\to\infty} \frac{I(x)}{\log x} &\geq \gamma_\ell - \sum_{k=1}^{K} a_k \max\{\gamma_\ell\rho_k, \gamma_\ell r_k - \nu_k + 1\} \\
&= \gamma_\ell - \sum_{k=1}^{K} a_k(\gamma_\ell\rho_k + \max\{0, \gamma_\ell(r_k-\rho_k) - \nu_k + 1\}) \\
&= \gamma_\ell - \sum_{k=1}^{K} a_k(\gamma_\ell\rho_k + \max\{0, (\gamma_\ell-\gamma_k)(r_k-\rho_k)\}) \\
&= \gamma_\ell - \sum_{k=1}^{\ell-1} a_k(\gamma_\ell r_k - \nu_k + 1) - \sum_{k=\ell}^{K} a_k\gamma_\ell\rho_k \\
&= \mu.
\end{aligned}$$

□

As shown above, Theorem 7.3.3 implies that the limits $x \to \infty$ and $n \to \infty$ commute, as long as one considers 'rough' (i.e., logarithmic) asymptotics. However, in case of 'more refined' asymptotics, the limits do not necessarily commute. This may be seen as follows. Consider the case of $n$ homogeneous On-Off sources with Pareto($\nu$) distributed On-periods. In Mandjes [198], it is proven that

$$\lim_{x\to\infty}\left[\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\{V^{(n)} > nx\} + (\nu-1)\left(\frac{c-\rho}{r-\rho}\right)\log(x\log x) = H\right],$$

for some constant $H \in (0,\infty)$. Now reverse the limits. Denote by $k_n$ the number of sources sending at peak rate in the reduced-load approximation (in the notation of Section 7.3.3, we have $k_n = N^*$):

$$k_n := \left\lceil \frac{nc - n\rho}{r-\rho} \right\rceil.$$

Now with Theorem 7.3.1, we have for any finite $n$ and $x \to \infty$,

$$\mathbb{P}\{V^{(n)} > nx\} \sim f(n)x^{-(\nu-1)k_n},$$

for some function $f(\cdot)$. Hence,

$$\begin{aligned}
&\lim_{x\to\infty}\left[\frac{1}{n}\log\mathbb{P}\{V^{(n)} > nx\} + (\nu-1)\left(\frac{c-\rho}{r-\rho}\right)\log(x\log x)\right] \\
= \;& \log f(n) - \lim_{x\to\infty}(\nu-1)\left(\frac{k_n}{n} - \frac{c-\rho}{r-\rho}\right)\log x + (\nu-1)\frac{c-\rho}{r-\rho}\log\log x.
\end{aligned}$$

Since this limit does not exist in $\mathbb{R}$, we conclude that the limits do not necessarily commute.

# Appendix

## 7.A Proof of Corollary 7.6.1

In this appendix we give a formal proof of Corollary 7.6.1.

**Corollary 7.6.1**
*If $A_j(\cdot) \in \mathcal{R}$ for all $j = 1, \ldots, N$, then $\mathbb{P}\{V^c > x\} \in \mathcal{IRV}$.*

**Proof**
As described earlier, the idea behind the proof is as follows. If $V^c > x$, then Propositions 7.6.4 and 7.6.5 show that the process $\{A(0,t) - ct\}$ reaches the level $(1-\delta)x$ after all sources have been On for at least $\frac{(1-2\delta)x}{r-c}$ time units. Since $A_j(\cdot) \in \mathcal{R} \subseteq \mathcal{IRV}$ for all $j = 1, \ldots, N$, with high probability, all sources remain On for at least $\frac{2\delta x}{r-c}$ more time units. This yields

$$\lim_{\delta \downarrow 0} \liminf_{x \to \infty} \mathbb{P}\{V^c > (1+\delta)x | V^c > x\} = 1, \tag{A.1}$$

implying the desired statement (by definition).

In order to give a formal proof, define the event $C(\delta, \epsilon x)$ by

$$C(\delta, \epsilon x) := \{\tau_s(\epsilon x) \leq \tau(\delta x) < \tau((1-\delta)x) \leq \tau_f(\epsilon x)\}.$$

With $A_i^p(x,\delta)$ and $A_i^r(x,\delta)$ we denote the past and residual period that source $i$ is active at time $\tau((1-\delta)x) \in [\tau_s, \tau_f]$ (on $C(\delta, \epsilon x)$). Note that, given $C(\delta, \epsilon x)$,

$$A_i^p(x,\delta) \geq \tau((1-\delta)x) - \tau_s \geq \tau((1-\delta)x) - \tau(\delta x) = \frac{x(1-2\delta)}{r-c}, \tag{A.2}$$

and that

$$\begin{aligned}
\mathbb{P}\{V^c > (1+\delta)x\} &\geq \mathbb{P}\{V^c > (1+\delta)x, C(\delta,\epsilon x)\} \\
&\geq \mathbb{P}\{C(\delta,\epsilon x), A_i^r(x,\delta) \geq \frac{2\delta x}{r-c} \text{ for all } i = 1, \ldots, N\} \\
&\geq \mathbb{P}\{C(\delta,\epsilon x)\} - \sum_{i=1}^{N} \mathbb{P}\{C(\delta,\epsilon x), A_i^r(x,\delta) \leq \frac{2\delta x}{r-c}\}.
\end{aligned}$$

Hence,

$$\frac{\mathbb{P}\{V^c > (1+\delta)x\}}{\mathbb{P}\{V^c > x\}} \geq \left(1 - \sum_{i=1}^{N} \mathbb{P}\{A_i^r(x,\delta) \leq \frac{2\delta x}{r-c} | C(\delta,\epsilon x)\}\right) \frac{\mathbb{P}\{C(\delta,\epsilon x)\}}{\mathbb{P}\{V^c > x\}}. \tag{A.3}$$

In order to derive (A.1) we now develop (i) a lower bound for the ratio $\frac{\mathbb{P}\{C(\delta,\epsilon x)\}}{\mathbb{P}\{V^c>x\}}$ and (ii) an upper bound for the conditional probability $\mathbb{P}\{A_i^r(x,\delta) \leq \frac{2\delta x}{r-c}|C(\delta,\epsilon x)\}$. With $E^{\rm c}$ we denote the complement of a set $E$. (i) For all $M>0$,

$$\frac{\mathbb{P}\{C(\delta,\epsilon x)\}}{\mathbb{P}\{V^c>x\}} \geq \frac{\mathbb{P}\{V^c(Mx)>x\}}{\mathbb{P}\{V^c>x\}} - \frac{\mathbb{P}\{V^c(Mx)>x, C(\delta,\epsilon x)^{\rm c}\}}{\mathbb{P}\{V^c>x\}}.$$

Using Propositions 7.6.1, 7.6.4, and 7.6.5, we then obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{C(\delta,\epsilon x)\}}{\mathbb{P}\{V^c>x\}} \geq 1. \tag{A.4}$$

(ii) Conditioning upon $A_i^p(x,\delta)$, we obtain (using the Markov property in $(a)$ and a well-known identity from renewal theory in $(b)$ concerning the joint distribution of the past lifetime $A^p$ and residual lifetime $A^r$),

$$\begin{aligned}
& \mathbb{P}\{A_i^r(x,\delta) \leq \frac{2\delta x}{r-c}|C(\delta,\epsilon x)\} \\
= & \int_{x\frac{1-2\delta}{r-c}}^{\infty} \mathbb{P}\{A_i^r(x,\delta) \leq \frac{2\delta x}{r-c}|A_i^p(x,\delta)=y, C(\delta,\epsilon x)\}{\rm d}\mathbb{P}\{A_i^p(x,\delta)\leq y|C(\delta,\epsilon x)\} \\
\stackrel{(a)}{=} & \int_{x\frac{1-2\delta}{r-c}}^{\infty} \mathbb{P}\{A_i^r \leq \frac{2\delta x}{r-c}|A_i^p=y\}{\rm d}\mathbb{P}\{A_i^p(x,\delta)\leq y|C(\delta,\epsilon x)\} \\
\stackrel{(b)}{=} & \int_{x\frac{1-2\delta}{r-c}}^{\infty} \left(1-\frac{\mathbb{P}\{A_i>\frac{2\delta x}{r-c}+y\}}{\mathbb{P}\{A_i>y\}}\right){\rm d}\mathbb{P}\{A_i^p(x,\delta)\leq y|C(\delta,\epsilon x)\}.
\end{aligned}$$

Since $\mathbb{P}\{A_i>x\}$ is regularly varying, one can apply the Potter bound (see Lemma 2.1.6) to find positive constants $\eta$ and $K$, with $K$ arbitrarily close to 1, independent of $\delta$ such that for $x$ large enough and for all $y \geq \frac{x(1-2\delta)}{r-c}$,

$$\frac{\mathbb{P}\{A_i>\frac{2\delta x}{r-c}+y\}}{\mathbb{P}\{A_i>y\}} \geq K\left(\frac{\frac{2\delta x}{r-c}+y}{y}\right)^{-\eta} \geq K(1-2\delta)^{\eta}.$$

In view of (A.2), we conclude that for all $i$, $\epsilon>0$, and for any $\delta>0$ and $K<1$,

$$\mathbb{P}\{A_i^r(x,\delta)\leq \frac{2\delta x}{r-c}|C(\delta,\epsilon x)\} \leq 1-K(1-2\delta)^{\eta}$$

for $x$ large enough, so that

$$\lim_{\delta\downarrow 0}\liminf_{x\to\infty}\left(1-\sum_{i=1}^{N}\mathbb{P}\{A_i^r(x,\delta)\leq \frac{2\delta x}{r-c}|C(\delta,\epsilon x)\}\right)=1. \tag{A.5}$$

Combining (A.3), (A.4), and (A.5) now yields (A.1).

□

# 7.B Proof of Lemma 7.6.3

**Lemma 7.6.3** *(Upper bound)*
*For any $\delta > 0$, there exists an $\epsilon_\delta > 0$ such that for all $\epsilon \in (0, \epsilon_\delta)$,*

$$\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\} \lesssim P_{\mathcal{J}_0}((1-\delta)x) \prod_{i\in\mathcal{J}_1} p_i,$$

*with $P_{\mathcal{J}_0}((1-\delta)x)$ as in (6.1).*

**Proof**
As mentioned earlier, the first step is to replace all random variables $A_{ij}$ and $U_{ij}$ by their means. Let $\bar{\delta}$ and $\tilde{\delta}$ be two $\mathcal{J}_1$-vectors, of which the elements are positive, but arbitrarily small. Note that, for fixed $\mathcal{J}_0$,

$$\begin{aligned} F_i \;\leq\; & (r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} - \bar{\delta}_k] + \\ & \sum_{k\in\mathcal{J}_1} r_k \sum_{j=1}^{N_k(\epsilon x)} [\mathbb{E}\{U_k\} - \bar{\delta}_k - U_{kj}], \end{aligned}$$

$$\begin{aligned} G_i \;\leq\; & (r-c)\bar{A}_i(\epsilon x) + (r-c)\epsilon x + \\ & (r-c)N_i(\epsilon x)[\mathbb{E}\{A_i\} + \tilde{\delta}_i] + (r-c)\sum_{j=1}^{N_i(\epsilon x)} [A_{ij}(\epsilon x) - \mathbb{E}\{A_i\} - \tilde{\delta}_i] - \\ & d_i N_i(\epsilon x)[\mathbb{E}\{U_i\} - \tilde{\delta}_i] + d_i \sum_{j=1}^{N_i(\epsilon x)} [\mathbb{E}\{U_i\} - \tilde{\delta}_i - U_{ij}] - \\ & \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} - \bar{\delta}_k] + \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k \sum_{j=1}^{N_k(\epsilon x)} [\mathbb{E}\{U_k\} - \bar{\delta}_k - U_{kj}]. \end{aligned}$$

Define the event $E_1(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x)$ by

$$\left\{ \sum_{j=1}^{N_i(\epsilon x)} [\mathbb{E}\{U_i\} - \min\{\bar{\delta}_i, \tilde{\delta}_i\} - U_{ij}] \leq \gamma x/(2r), i \in \mathcal{J}_1 \right\} \bigcup$$

$$\left\{ \sum_{j=1}^{N_i(\epsilon x)} [A_{ij}(\epsilon x) - \mathbb{E}\{A_i\} - \min\{\bar{\delta}_i, \tilde{\delta}_i\}] \leq \gamma x/(2r) - (r-c)\epsilon x, i \in \mathcal{J}_1 \right\}.$$

A straightforward application of Lemma 2.4.1 (analogously to the proofs of Propositions 7.6.2, 7.6.4 and 7.6.5) shows that for any $\gamma, \bar{\delta}, \tilde{\delta} > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*]$,

$$\mathbb{P}_{\mathcal{J}_0}\{E_1(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x)^c\} = o(P(x)), \tag{B.1}$$

as $x \to \infty$ with $P(x) = \prod_{j=1}^{N} \mathbb{P}\{A_j^r > x\}$, as defined earlier.

From Equation (B.1) and Lemma 7.6.5, we conclude that, using the upper bounds for $F_i$ and $G_i$,

$$\begin{aligned}
& \mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f) - c\tau_f > (1-\delta)x\} \\
= \; & \mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f) - c\tau_f > (1-\delta)x; E_1(\gamma,\bar{\delta},\tilde{\delta},\epsilon,x)^c\} + \\
& \mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f) - c\tau_f > (1-\delta)x; E_1(\gamma,\bar{\delta},\tilde{\delta},\epsilon,x)\} \\
\leq \; & \mathbb{P}\{(r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} - \bar{\delta}_k] > (1-\gamma-\delta)x, i \in \mathcal{J}_0; \\
& (r-c)\bar{A}_i(\epsilon x) + (r-c)N_i(\epsilon x)[\mathbb{E}\{A_i\} + \tilde{\delta}_i] - d_i N_i(\epsilon x)[\mathbb{E}\{U_i\} - \tilde{\delta}_i] - \\
& \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} - \bar{\delta}_k] > (1-\gamma-\delta)x, i \in \mathcal{J}_1\} + \mathrm{o}(P(x)).
\end{aligned}$$

The last probability equals (condition on $N_i(\epsilon x)$, $i \in \mathcal{J}_1$),

$$\begin{aligned}
& \sum_{n_i\geq 1, i\in\mathcal{J}_1} \left( \prod_{i\in\mathcal{J}_1} \mathbb{P}\{N_i(\epsilon x) = n_i\} \right) \times \\
& \qquad \mathbb{P}\{(r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]n_k > (1-\gamma-\delta)x, i \in \mathcal{J}_0; \\
& \qquad\quad (r-c)\bar{A}_i(\epsilon x) + (r-c)[\mathbb{E}\{A_i\} + \tilde{\delta}_i]n_i - d_i[\mathbb{E}\{U_i\} - \tilde{\delta}_i]n_i - \\
& \qquad\quad \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]n_k > (1-\gamma-\delta)x, i \in \mathcal{J}_1\}.
\end{aligned}$$

Deconditioning upon $\bar{A}_i$ and $\bar{A}_i^r$ (i.e., dividing by $\prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{A_i > \epsilon x\}$), and noting that $\mathbb{P}\{N_i(\epsilon x) = n_i\} \leq \mathbb{P}\{A_i > \epsilon x\}$, we obtain that

$$\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f) - c\tau_f > (1-\delta)x\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\}$$

is upper bounded by (up to $\mathrm{o}(P(x))$)

$$\begin{aligned}
& \sum_{n_i\geq 0, i\in\mathcal{J}_1} \left( \prod_{i\in\mathcal{J}_0} \mathbb{P}\{(r-c)A_i^r > (1-\gamma-\delta)x + \sum_{k\in\mathcal{J}_1} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]n_k\} \right) \times \\
& \prod_{i\in\mathcal{J}_1} \mathbb{P}\{(r-c)A_i > (1-\gamma-\delta)x + [d_i\mathbb{E}\{U_i\} - (r-c)\mathbb{E}\{A_i\} - r_i\tilde{\delta}_i]n_i + \\
& \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]n_k\}.
\end{aligned}$$

It is important to note that this expression is independent of $\epsilon$.

Since all probabilities appearing in the right hand side are decreasing functions of $n_i$ (for $\bar{\delta}$ and $\tilde{\delta}$ small enough), the latter term is bounded by (with $y := (y_i)_{i\in\mathcal{J}_1}$ and $\mathrm{d}y := \prod_{i\in\mathcal{J}_1} \mathrm{d}y_i$)

$$\int_{y\geq 0} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{(r-c)A_i^r > (1-\gamma-\delta)x + \sum_{k\in\mathcal{J}_1} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]y_k\} \tag{B.2}$$
$$\prod_{i\in\mathcal{J}_1} \mathbb{P}\{(r-c)A_i > (1-\gamma-\delta)x + [d_i\mathbb{E}\{U_i\} - (r-c)\mathbb{E}\{A_i\} - r_i\tilde{\delta}_i]y_i$$
$$+ \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k[\mathbb{E}\{U_k\} - \bar{\delta}_k]y_k\}\mathrm{d}y.$$

We will rewrite this expression in terms of $P_{\mathcal{J}_0}(x)$. Apply the change of variables $z_i := y_i(\mathbb{E}\{A_i\} + \mathbb{E}\{U_i\})$. Redefine $\bar{\delta}_i := \bar{\delta}_i(\mathbb{E}\{A_i\} + \mathbb{E}\{U_i\})$ and similarly $\tilde{\delta}_i := \tilde{\delta}_i(\mathbb{E}\{A_i\} + \mathbb{E}\{U_i\})$. Note that $\frac{1}{\mathbb{E}\{A_i\}+\mathbb{E}\{U_i\}} = \frac{p_i}{\mathbb{E}\{A_i\}}$ and $r_i\frac{\mathbb{E}\{U_i\}}{\mathbb{E}\{A_i\}+\mathbb{E}\{U_i\}} = r_i(1-p_i) = r_i - \rho_i$. Then we obtain that (B.2) equals

$$\left(\prod_{i\in\mathcal{J}_1} \frac{p_i}{\mathbb{E}\{A_i\}}\right) \int_{z\geq 0} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{(r-c)A_i^r > (1-\gamma-\delta)x + \sum_{k\in\mathcal{J}_1} (r_k - \rho_k - \bar{\delta}_k)z_k\}$$

$$\prod_{i\in\mathcal{J}_1} \mathbb{P}\{(r-c)A_i > (1-\gamma-\delta)x + (d_i - \rho_i - \tilde{\delta}_i)z_i + \sum_{k\in\mathcal{J}_1\setminus\{i\}} (r_k - \rho_k - \bar{\delta}_k)z_k\}\mathrm{d}z.$$

If we take $\tilde{\delta}_i = \frac{d_i-\rho_i}{r_i-\rho_i}\bar{\delta}_i$ and integrate w.r.t. $z_i\frac{r_i-\rho_i-\bar{\delta}_i}{r_i-\rho_i}$, then we obtain

$$\left(\prod_{i\in\mathcal{J}_1} \frac{r_i-\rho_i}{r_i-\rho_i-\bar{\delta}_i}\frac{p_i}{\mathbb{E}\{A_i\}}\right) \int_{z\geq 0} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{(r-c)A_i^r > (1-\gamma-\delta)x + \sum_{k\in\mathcal{J}_1} (r_k - \rho_k)z_k\}$$

$$\prod_{i\in\mathcal{J}_1} \mathbb{P}\{(r-c)A_i > (1-\gamma-\delta)x + (d_i - \rho_i)z_i + \sum_{k\in\mathcal{J}_1\setminus\{i\}} (r_k - \rho_k)z_k\}\mathrm{d}z =$$

$$\prod_{i\in\mathcal{J}_1} p_i\frac{r_i-\rho_i}{r_i-\rho_i-\bar{\delta}_i}P_{\mathcal{J}_0}((1-\gamma-\delta)x).$$

Together with the fact that $P_{\mathcal{J}_0}(\cdot)$ is regularly varying, this completes the proof of the upper bound after dividing by $P_{\mathcal{J}_0}(x)$, letting $x\to\infty$, and noting that $\delta$, $\bar{\delta}$, and $\gamma$ may be chosen arbitrarily small.

□

# 7.C Proof of Lemma 7.6.4

**Lemma 7.6.4** *(Lower bound)*
*There exists an $\epsilon > 0$ such that*

$$\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{A_i^r > \epsilon x\} \gtrsim P_{\mathcal{J}_0}(x) \prod_{i\in\mathcal{J}_1} p_i,$$

*with $P_{\mathcal{J}_0}(x)$ as in (6.1).*

**Proof**
Like in Appendix 7.B, the first step is to replace the random variables $A_i(\epsilon x)$ and $U_i$ by their means. Adding and subtracting appropriate means, it is easy to see that, for fixed $\mathcal{J}_0$,

$$\begin{aligned}
F_i &= (r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] + \\
&\quad \sum_{k\in\mathcal{J}_1} r_k \sum_{j=1}^{N_k(\epsilon x)} [\mathbb{E}\{U_k\} - U_{kj} + \bar{\delta}_k] - \\
&\quad \sum_{k\in\mathcal{J}_1} r_k \left[I_k A_k^r(\epsilon x) + (1-I_k)(A_k(\epsilon x) + U_k^r)\right], \\
G_i &= (r-c)\bar{A}_i(\epsilon x) + (r-c)[I_i A_i^r(\epsilon x) + (1-I_i)A_i(\epsilon x)] - d_i(1-I_i)U_i^r - \\
&\quad \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k(1-I_k)U_k^r + (r-c)\sum_{j=1}^{N_i(\epsilon x)} [A_{ij}(\epsilon x) - \mathbb{E}\{A_i\} + \tilde{\delta}_i] + \\
&\quad (r-c)N_i(\epsilon x)[\mathbb{E}\{A_i\} - \tilde{\delta}_i] - d_i N_i(\epsilon x)[\mathbb{E}\{U_i\} + \tilde{\delta}_i] + \\
&\quad d_i \sum_{j=1}^{N_i(\epsilon x)} [\mathbb{E}\{U_i\} - U_{ij} + \tilde{\delta}_i] - \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] + \\
&\quad \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k \sum_{j=1}^{N_k(\epsilon x)} [\mathbb{E}\{U_k\} - U_{kj} + \bar{\delta}_k].
\end{aligned}$$

Define the event $E_2(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x)$ by

$$\left\{ \sum_{j=1}^{N_i(\epsilon x)} [\mathbb{E}\{U_i\} - U_{ij} + \min\{\bar{\delta}_i, \tilde{\delta}_i\}] \geq -\gamma x/(3r), i \in \mathcal{J}_1 \right\} \bigcup$$

$$\left\{ \sum_{j=1}^{N_i(\epsilon x)} [A_{ij}(\epsilon x) - \mathbb{E}\{A_i\} + \min\{\bar{\delta}_i, \tilde{\delta}_i\}] \geq -\gamma x/(3r), i \in \mathcal{J}_1 \right\} \bigcup$$

$$\left\{ \sum_{k\in\mathcal{J}_1} [I_k A_k^r(\epsilon x) + (1-_k)(A_k(\epsilon x) + U_k^r)] \leq \gamma x/(3r) \right\}.$$

We have the lower bound

$$\begin{aligned}
& \mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f) - c\tau_f > x\} \\
= \;& \mathbb{P}_{\mathcal{J}_0}\{F_i > x, i \in \mathcal{J}_0; G_i > x, i \in \mathcal{J}_1\} \\
\geq \;& \mathbb{P}_{\mathcal{J}_0}\{F_i > x, i \in \mathcal{J}_0; G_i > x, i \in \mathcal{J}_1; E_2(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x)\} \\
\geq \;& \mathbb{P}\{(r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] > (1+\gamma)x, i \in \mathcal{J}_0; \\
& \quad (r-c)\bar{A}_i(\epsilon x) + (r-c)N_i(\epsilon x)[\mathbb{E}\{A_i\} - \tilde{\delta}_i] - d_i N_i(\epsilon x)[\mathbb{E}\{U_i\} + \tilde{\delta}_i] - \\
& \quad \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] > (1+\gamma)x, i \in \mathcal{J}_1; E_2(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x)\}.
\end{aligned}$$

This probability is lower bounded by, for any $L$ (condition on $N_i(\epsilon x)$),

$$\begin{aligned}
& \sum_{0\leq n_i \leq Lx, i\in\mathcal{J}_1} \mathbb{P}\{E_2(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x) | N_i(\epsilon x) = n_i, i \in \mathcal{J}_1\} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{N_i(\epsilon x) = n_i\} \times \\
& \mathbb{P}\{(r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] > (1+\gamma)x, i \in \mathcal{J}_1; \\
& \quad (r-c)\bar{A}_i(\epsilon x) + (r-c)N_i(\epsilon x)[\mathbb{E}\{A_i\} - \tilde{\delta}_i] - d_i N_i(\epsilon x)[\mathbb{E}\{U_i\} + \tilde{\delta}_i] - \\
& \quad \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k N_k(\epsilon x)[\mathbb{E}\{U_k\} + \bar{\delta}_k] > (1+\gamma)x, i \in \mathcal{J}_1 | N_i(\epsilon x) = n_i, i \in \mathcal{J}_1\}. \quad \text{(C.1)}
\end{aligned}$$

Before proceeding, we first state a useful lemma (a proof is given at the end of this appendix).

**Lemma 7.C.1** *For all $\epsilon, \gamma, \bar{\delta}, \tilde{\delta} > 0$,*

$$\mathbb{P}\{E_2(\gamma, \bar{\delta}, \tilde{\delta}, \epsilon, x) | N_i(\epsilon x) = n_i, i \in \mathcal{J}_1\} \to 1, \quad \text{(C.2)}$$

*as $x \to \infty$ uniformly in $n_i \geq 0, i \in \mathcal{J}_1$, and*

$$\frac{\mathbb{P}\{N_i(\epsilon x) = n_i\}}{\mathbb{P}\{A_i > \epsilon x\}} \to 1 \quad \text{(C.3)}$$

*for all $i \in \mathcal{J}_1$ as $x \to \infty$ uniformly in $0 \leq n_i \leq Lx$.*

Equations (C.2) and (C.3) imply that for any $L < \infty$ and $\eta > 0$ one can lower bound Equation (C.1) for $x$ large enough by

$$\begin{aligned}
& (1-\eta) \sum_{n_i \leq Lx, i\in\mathcal{J}_1} \mathbb{P}_{\mathcal{J}_0}\{(r-c)\bar{A}_i^r(\epsilon x) - \sum_{k\in\mathcal{J}_1} r_k n_k[\mathbb{E}\{U_k\} + \delta_k]n > (1+\gamma)x, i \in \mathcal{J}_0; \\
& (r-c)\bar{A}_i(\epsilon x) + (r-c)n_i[\mathbb{E}\{A_i\} - \tilde{\delta}_i] - d_i n_i[\mathbb{E}\{U_i\} + \tilde{\delta}_i] - \\
& \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k n_k[\mathbb{E}\{U_k\} + \bar{\delta}_k] > (1+\gamma)x, i \in \mathcal{J}_1 | N_i(\epsilon x) = n_i, i \in \mathcal{J}_1\} \prod_{i\in\mathcal{J}_1} \mathbb{P}\{A_i > \epsilon x\}.
\end{aligned}$$

As before, deconditioning upon $\bar{A}_i$ and $\bar{A}_i^r$ and applying a similar change of variables as in Appendix 7.B, we obtain the lower bound

$$(1-\eta)\left(\prod_{i\in\mathcal{J}_1}\frac{p_i}{\mathbb{E}\{A_i\}}\right)\int_{1\le y_i\le Lx, i\in\mathcal{J}_1}\prod_{i\in\mathcal{J}_0}\mathbb{P}\{(r-c)A_i^r>(1+\gamma)x+\sum_{k\in\mathcal{J}_1}(r_k-\rho_k+\bar{\delta}_k)y_k\}$$

$$\prod_{i\in\mathcal{J}_1}\mathbb{P}\{(r-c)A_i>(1+\gamma)x+(d_i-\rho_i+\tilde{\delta}_i)y_i+\sum_{k\in\mathcal{J}_1\setminus\{i\}}(r_k-\rho_k+\bar{\delta}_k)y_k\}\mathrm{d}y.$$

Now write

$$(1-\eta)\int_{1\le y_i\le Lx, i\in\mathcal{J}_1}\ldots=(1-\eta)\int_{y_i\ge 0, i\in\mathcal{J}_1}\ldots-(1-\eta)\int_{\{1\le y_i\le Lx, i\in\mathcal{J}_1\}^c}\ldots$$

(the complement taken with respect to the non-negative orthant). The first term in the right hand side can be handled as in the proof of the upper bound (the only difference is the factor $1+\gamma$ instead of $1-\gamma-\delta$). The next lemma shows that the second term can be neglected.

□

**Lemma 7.C.2**

$$\lim_{L\to\infty}\limsup_{x\to\infty}$$
$$\frac{1}{P(x)}\int_{\{1\le y_i\le Lx, i\in\mathcal{J}_1\}^c}\prod_{i\in\mathcal{J}_0}\mathbb{P}\{(r-c)A_i^r>(1+\gamma)x+\sum_{k\in\mathcal{J}_1}(r_k-\rho_k+\bar{\delta}_k)y_k\}$$
$$\prod_{i\in\mathcal{J}_1}\mathbb{P}\{(r-c)A_i>(1+\gamma)x+(d_i-\rho_i+\tilde{\delta}_i)y_i+\sum_{k\in\mathcal{J}_1\setminus\{i\}}(r_k-\rho_k+\bar{\delta}_k)y_k\}\mathrm{d}y=0.$$

**Proof**

The integral over the regions in which at least one $y_i$ is smaller than 1 is easily shown to be of $o(P(x))$, so we concentrate on the set $\{0\le y_i\le Lx, i\in\mathcal{J}_1\}^c$. The integral

$$\int_{\{0\le y_i\le Lx, i\in\mathcal{J}_1\}^c}\prod_{i\in\mathcal{J}_0}\mathbb{P}\{(r-c)A_i^r>(1+\gamma)x+\sum_{k\in\mathcal{J}_1}(r_k-\rho_k+\bar{\delta}_k)y_k\}$$

$$\prod_{i\in\mathcal{J}_1}\mathbb{P}\{(r-c)A_i>(1+\gamma)x+(d_i-\rho_i+\tilde{\delta}_i)y_i+\sum_{k\in\mathcal{J}_1\setminus\{i\}}(r_k-\rho_k+\bar{\delta}_k)y_k\}\mathrm{d}y$$

is bounded from above by

$$\left(\prod_{i\in\mathcal{J}_0}\mathbb{P}\{(r-c)A_i^r>(1+\gamma)x\}\right)\sum_{j\in\mathcal{J}_1}\int_{y_j\ge Lx, y_i\ge 0, i\in\mathcal{J}_1, i\ne j}$$

$$\prod_{i\in\mathcal{J}_1}\mathbb{P}\{(r-c)A_i > (1+\gamma)x + (d_i-\rho_i+\tilde{\delta}_i)y_i + \sum_{k\in\mathcal{J}_1\setminus\{i\}}(r_k-\rho_k+\bar{\delta}_k)y_k, i\in\mathcal{J}_1\}\mathrm{d}y.$$

Observing that the integrals can be separated, we obtain the upper bound

$$\begin{aligned} & \mathrm{O}\left(\prod_{i\in\mathcal{J}_0}\mathbb{P}\{A_i^r > x\}\right)\sum_{j\in\mathcal{J}_1}\mathrm{O}\left(\mathbb{P}\{A_j^r > Lx\}\right)\prod_{i\in\mathcal{J}_1, i\neq j}\mathrm{O}\left(\prod_{i\in\mathcal{J}_0}\mathbb{P}\{A_i^r > x\}\right) \\ = \; & \mathrm{O}(P(x))\sum_{j\in\mathcal{J}_1}\frac{\mathbb{P}\{A_j^r > Lx\}}{\mathbb{P}\{A_j^r > x\}}. \end{aligned}$$

The result then follows immediately.

□

**Proof of Lemma 7.C.1**

Equation (C.2) follows immediately from the following result. Let $S_n := X_1 + \ldots + X_n$ be a random walk with i.i.d. step sizes with $\mathbb{E}\{X_1\} < 0$. Then

$$\limsup_{x\to\infty}\sup_{n\geq 1}\mathbb{P}\{S_n > x\} \leq \lim_{x\to\infty}\mathbb{P}\{\sup_{n\geq 1}S_n > x\} = 0,$$

since $\sup_{n\geq 1}S_n$ is a proper random variable. For every $i = 1,\ldots,N$, apply this result twice with $X_j := U_{ij} - \mathbb{E}\{U_i\} - \min\{\bar{\delta}_i, \tilde{\delta}_i\}$ and $X_j := \mathbb{E}\{A_i\} - A_{ij}(\epsilon x) - \min\{\bar{\delta}_i, \tilde{\delta}_i\}$. In order to prove Equation (C.3), note that for $n_i \leq Lx$,

$$\frac{\mathbb{P}\{N_i(\epsilon x) = n_i\}}{\mathbb{P}\{A_i > \epsilon x\}} = \mathbb{P}\{A_i \leq \epsilon x\}^{n_i} \leq \mathbb{P}\{A_i \leq \epsilon x\}^{Lx} = \left(1 - \frac{\mathrm{o}(1)}{x}\right)^{Lx} \to 1,$$

as $x\to\infty$. The last equality holds because $A_i$ has finite mean.

□

# Chapter 8

# Fluid queues with heavy-tailed $M/G/\infty$ input

## 8.1 Introduction

The previous two chapters have been devoted to the fluid queue fed by a finite number of On-Off sources. In the present chapter, we consider a closely related model: A fluid queue with $M/G/\infty$ input. The arrival dynamics in this system can be described as follows. Sessions arrive as a Poisson process, and remain in the system for a randomly distributed period of time. While in the system, each session generates traffic at some constant rate. Note that the number of active sessions behaves as the number of customers in an $M/G/\infty$ system, hence the term $M/G/\infty$ input. An $M/G/\infty$ input process may also be viewed as the limit of the superposition of On-Off sources when the number of sources grows large, and the fraction of On-time gets correspondingly small, as shown in Jelenković & Lazar [161].

While incorporating session-level dynamics, the $M/G/\infty$ model avoids the intricate temporal dependence structure of ordinary On-Off sources. At the same time, the $M/G/\infty$ model retains the usual versatility of fluid models in covering a wide spectrum of possible traffic characteristics through the distribution of the activity periods.

Fluid queues with heavy-tailed $M/G/\infty$ input have been extensively studied before. Likhanov [184] and Liu *et al.* [187] obtain asymptotic lower and upper bounds for the workload distribution. Under a certain peak rate condition, the bounds are shown to be tight (up to a constant factor) for Pareto-distributed session lengths, thus yielding the exact decay rate. The peak rate condition essentially implies that just a single long session is enough to cause overflow. Under roughly similar assumptions, Boxma [65], Jelenković & Lazar [161], and Resnick & Samorodnitsky [238] also determine the corresponding pre-factor, resulting in the exact workload asymptotics. Duffield [119] obtains logarithmic 'many-sources' asymptotics (as opposed to 'large-buffer' asymptotics) for a regime where the arrival rate, service rate, and buffer size are scaled up in proportion, see

also Mandjes [198].

Recently, several authors have considered heterogeneous heavy-tailed $M/G/\infty$ input, where sessions belong to one of several classes with distinct characteristics (arrival rates, session lengths, peak rates). Likhanov & Mazumdar [185] obtain asymptotic lower and upper bounds for the workload distribution, which are shown to be tight up to a constant factor. Under a similar peak rate condition as described above, the bounds coincide, yielding the exact asymptotics. An elegant treatment of this special case is also given in Jelenković [164]. Remarkably enough, the bounds in [185] are asymptotically exact for finite buffers as well.

As mentioned above, the $M/G/\infty$ model is closely related to the classical model with a fixed set of On-Off sources. Despite some subtle differences, the similarity manifests itself in the qualitative way that overflow occurs for heavy-tailed input, and is also reflected in the tail asymptotics of the workload. For example, the results in [120] for a fixed set of On-Off sources are reminiscent of the results in [185] for $M/G/\infty$ input. Also, the $M/G/\infty$ asymptotics in [65], [161], and [238] for the special case where a single long session can cause overflow are accompanied (in [65] and [161]) by conceptual counterparts for a scenario where a single regularly varying On-Off source is multiplexed with several light-tailed sources.

It is interesting to observe that the exact workload asymptotics for the $M/G/\infty$ model with infinite buffers have only been obtained under the condition that a single long session is sufficient to cause positive drift. Although technically convenient, this condition is rather restrictive from a practical perspective. The degree of multiplexing is typically so high, that the peak rate of an individual session is relatively small compared to the link rate. Thus, under moderate loading, several long sessions must coincide in order for the drift to turn positive. In the present chapter, we derive the exact asymptotic workload behavior under such general circumstances where a combination of several long sessions is involved in causing overflow. Besides the practical relevance, these scenarios are also theoretically challenging, since the combinatorial structure of the overlap of the various sessions significantly adds to the complexity. The analysis unifies and generalizes the results in [164], [185], and [238], and complements the exact tail asymptotics for a fixed set of On-Off sources which have been derived in Chapter 7 of this thesis. As in Chapter 7, we use the framework of Section 2.4.

The remainder of the chapter is organized as follows. In Section 8.2, we present a detailed model description. In Section 8.3, we provide some intuitive arguments, and summarize the main results of this chapter. Like in the previous chapters, the arguments are grounded on the large-deviations idea that overflow is typically due to some minimal combination of extremely long concurrent sessions causing positive drift. The typical configuration of long sessions is identified through a simple integer linear program, which corresponds to the set optimization problem defined in [185].

The subsequent sections are devoted to the detailed proofs. In particular, in Section 8.4,

we extend the probabilistic arguments developed in [238], enabling the exact calculation of the asymptotic workload behavior. In addition, the computations provide fundamental insight in the typical overflow scenario.

The analysis in fact focuses on the transient behavior, from which the steady-state asymptotics easily follow after showing in Section 8.5 that overflow occurs in linear time. As a by-product, we obtain asymptotically tight bounds for the transient workload distribution. The transient asymptotics in their full generality remain a challenging open problem, see Subsection 8.4.6. In Section 8.6, we combine our transient and steady-state asymptotics to obtain the limiting distribution of the most probable time to overflow.

## 8.2 Model description and preliminaries

In this section, we present a detailed model description, and introduce some notation.

### 8.2.1 Basic input and workload processes

We consider a fluid queue of unit capacity fed by $K$ heterogeneous $M/G/\infty$ input processes. Class-$k$ sessions arrive as a Poisson process of rate $\lambda_k$, and remain in the system for a random period $B_k$ having distribution $B_k(\cdot)$ with mean $\beta_k$, $k = 1, \ldots, K$. We assume that $B_k(\cdot)$ is regularly varying of index $-\nu_k < -1$ (this assumption can be relaxed somewhat, see Remark 8.3.1), so that $\beta_k < \infty$. While in the system, each class-$k$ session generates traffic at constant rate $r_k$.

Let $\rho_k := \lambda_k \beta_k r_k$ be the traffic intensity associated with class-$k$ sessions. Define $\bar{\rho}_k := \lambda_k \beta_k = \rho_k / r_k$. Let $\rho := \sum_{k=1}^{K} \rho_k$ be the total traffic intensity. We assume $\rho < 1$ for stability.

Denote by $B_k^r(\cdot)$ the distribution of the residual life-time of $B_k$, and by $B_k^r$ a stochastic variable with that distribution.

Define $A_k(s,t)$ as the amount of class-$k$ traffic generated in the time interval $(s,t]$. Note that

$$A_k(s,t) \stackrel{d}{=} r_k \int_s^t N_k(u) \mathrm{d}u,$$

with $N_k(u), u \geq 0$, the number of customers at time $u$ in a stationary $M/G/\infty$ queue with arrival rate $\lambda_k$ and service time distribution $B_k(\cdot)$.

Denote by $A(s,t) := \sum_{k=1}^{K} A_k(s,t)$ the total amount of traffic generated in the time interval $(s,t]$. The workload in the system at time $t \geq 0$ is $V(t) := \sup_{0 \leq s \leq t} \{A(s,t) - (t-s)\}$, assuming the system is empty at time $t = 0$. Let $V$ be the weak limit of $V(t)$ for $t \to \infty$.

### 8.2.2 Auxiliary processes: separating short and long sessions

One of the first steps of the analysis will be to split the arriving sessions into two groups, short and long ones. In this subsection we introduce some notation for the corresponding processes.

We denote by $A_{k,\leq z}(s,t)$ the amount of traffic generated in $(s,t]$ by class-$k$ sessions of length at most $z$ (upon arrival). The corresponding traffic intensity is denoted by

$$\rho_{k,\leq z} := \lambda_k \mathbb{P}\{B_k \leq z\} r_k \mathbb{E}\{B_k \mid B_k \leq z\} = \rho_k B_k^r(z) - \lambda_k r_k z \mathbb{P}\{B^k > z\}.$$

Define $A_{\leq z}(s,t) := \sum_{k=1}^{K} A_{k,\leq z}(s,t)$, and $\rho_{\leq z} := \sum_{k=1}^{K} \rho_{k,\leq z}$. Similarly, we denote by $A_{k,>z}(s,t)$ the amount of traffic generated in $(s,t]$ by class-$k$ sessions of length exceeding $z$. The corresponding traffic intensity $\rho_{k,>z}$ is given by $\rho_k \mathbb{P}\{B_k^r > z\} + \lambda_k r_k z \mathbb{P}\{B_k > z\}$. Define $A_{>z}(s,t) := \sum_{k=1}^{K} A_{k,>z}(s,t)$, and $\rho_{>z} := \sum_{k=1}^{K} \rho_{k,>z}$. Denote $\bar{\rho}_{k,>z} = \rho_{k,>z}/r_k$.

Denote by $N_{k,>z}(t), t \geq 0$ the number of class-$k$ sessions exceeding length $z$ which are still active at time $t$. Note that the remaining lengths of these sessions at time $t$ may be smaller than $z$ (except for $t = 0$). The process $N_{k,>z}(t), t \geq 0$, is constructed from $N_k(t), t \geq 0$. In particular, it follows from basic $M/G/\infty$ theory that the random vector $N_{>z}(0) := (N_{1,>z}(0), \ldots, N_{K,>z}(t))$ has a multi-dimensional Poisson distribution with parameters $(\bar{\rho}_1 \mathbb{P}\{B_1^r > z\}, \ldots, \bar{\rho}_K \mathbb{P}\{B_K^r > z\})$, i.e.,

$$\mathbb{P}\{N_{>z}(0) = (n_1, \ldots, n_K)\} = \prod_{k=1}^{K} \mathrm{e}^{-\bar{\rho}_k \mathbb{P}\{B_k^r > z\}} \frac{\bar{\rho}_k^{n_k}}{n_k!} \mathbb{P}\{B_k^r > z\}^{n_k}. \tag{2.1}$$

Note that the steady-state distribution of $\{N_{k,>z}(t)\}$ is Poisson with rate $\bar{\rho}_{k,>z}$, and that $A_{k,>z}(0,t) \stackrel{d}{=} r_k \int_0^t N_{k,>z}(u)\mathrm{d}u$. For future purposes, we define the processes

$$\begin{aligned} V_{>z}^c(t) &:= \sup_{0\leq s\leq t} \{A_{>z}(0,s) - cs\}, \\ V_{>z}^c &:= \sup_{t\geq 0} \{A_{>z}(0,t) - ct\}. \end{aligned}$$

### 8.2.3 Representation for the workload

In this subsection we give a convenient representation for the transient and stationary workload. First, we consider the aggregate workload process $V^c(t)$. Using the expression $V^c(t) = \sup_{0\leq s\leq t} \{A(s,t) - c(t-s)\}$ and noting that the process $A(\cdot,\cdot)$ has stationary and reversible increments, the transient workload may be represented as

$$V^c(t) = \sup_{0\leq s\leq t} \{A(s,t) - c(t-s)\} \stackrel{d}{=} \sup_{0\leq s\leq t} \{A(0,s) - cs\}.$$

In the sequel, we proceed similarly as in [238] and Chapter 7 of this thesis, and use the latter expression as the *definition* of $V^c_{>z}(t)$. Accordingly, for $c > \rho$, the stationary workload as $t \to \infty$ may be expressed as

$$V^c := \sup_{t \geq 0}\{A(0,t) - ct\}.$$

## 8.3 Overview

In this section we present the main results of the chapter, which characterize the exact asymptotic behavior of $\mathbb{P}\{V > x\}$ as $x \to \infty$. This behavior, as well the corresponding intuition, is strongly reminiscent of the overflow scenarios (reduced-load) considered in Chapter 7. The reduced-peak scenario considered in Chapter 6 cannot occur.

### 8.3.1 Intuitive arguments

Before formally stating the results, we first provide some intuitive arguments. Large-deviations results for heavy-tailed distributions suggest that a large workload level is typically due to some 'minimal combination' of extremely long overlapping sessions causing positive drift. In a homogeneous context, the typical combination simply consists of the minimal *number* of long sessions needed for the drift to turn positive. However, in a heterogeneous setting, not only the number of long sessions counts, but also the class characteristics. Note that the number of long sessions required for a positive drift varies with the peak rates $r_k$ of the various classes. In addition, the relative frequency of long sessions differs across the various classes as governed by the tail exponents $\nu_k$.
Informally speaking, the typical combination may be interpreted as the one most likely to occur among those producing positive drift. Specifically, let the typical configuration of long sessions be $n = (n_1, \ldots, n_K)$. For the workload to reach a large level $x$, the associated drift must be strictly positive, i.e.,

$$\sum_{k=1}^{K} n_k r_k + \rho - 1 > 0. \tag{3.1}$$

In addition, the sessions must last for a period of the order $x$, which happens with probability of the order

$$x^{-\sum_{k=1}^{K} n_k(\nu_k - 1)}. \tag{3.2}$$

The supposition that $n = (n_1, \ldots, n_K)$ is the *most likely* combination, means that it should maximize (3.2) for large values of $x$, i.e., minimize the exponent $\sum_{k=1}^{K} n_k(\nu_k - 1)$,

subject to the drift condition (3.1). Thus, the most likely configuration of long sessions may be identified as follows.

$$\begin{aligned} \min \quad & \mu = \sum_{k=1}^{K} n_k(\nu_k - 1) \\ \text{sub} \quad & \sum_{k=1}^{K} n_k r_k \geq 1 - \rho \\ & n_k \in \mathbb{N}, \qquad k = 1, \ldots, K. \end{aligned}$$

The above integer linear program corresponds to the set optimization problem defined in [185]. In general, the optimal solution cannot be obtained in closed form due to the integrality constraints. However, if the integrality constraints are relaxed, then the optimization program may be easily solved. The optimal solution is then given by $n^* = (1-\rho)e_{k^*}/r_{k^*}$, with $k^* := \arg\max_{k=1,\ldots,K} r_k/(\nu_k - 1)$, and $e_k$ denoting the unit vector. This suggests that sessions of class $k^*$ are likely to be involved in the typical configuration of long sessions that causes overflow. This is especially the case when the peak rates $r_k$ are relatively small compared to the slack capacity $1-\rho$, so that the typical combination consists of a relatively large number of sessions. However, in general the optimal combination may include sessions of other classes as well due to the integrality constraints, and in extreme cases may not contain a single session of class $k^*$ at all. Let $S^* \subseteq \mathbb{N}^K$ be the set of optimal solutions (there may be several in general). Denote by $\mu^*$ the corresponding optimal value. Also, define $r^{\min} := \min_{n \in S^*} \sum_{k=1}^{K} n_k r_k$. Throughout the chapter, we assume that $r^{\min} > 1 - \rho$. This assumption ensures that the drift in all plausible overflow scenarios is strictly positive. (In general, some overflow scenarios may involve only *zero* drift.)

### 8.3.2 Steady-state workload asymptotics

We now state the central result of this chapter, which characterizes the exact asymptotic behavior of the stationary workload distribution. For given $n \in \mathbb{N}^K$, denote $d_n := \sum_{k=1}^{K} n_k r_k + \rho - 1$.

**Theorem 8.3.1** *Assume that $r^{\min} > 1 - \rho$. Then,*

$$\mathbb{P}\{V > x\} \sim \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n}(x), \tag{3.3}$$

*where $j = (j_1, \ldots, j_K)$, and $P_{j,n}(x)$ satisfies*

$$P_{j,n}(x) \sim \kappa_{j,n} \prod_{k=1}^{K} \mathbb{P}\{B_k^r > \frac{x}{d_n}\}^{n_k},$$

*for some constant* $\kappa_{j,n}$.
*In particular,* $\mathbb{P}\{V > x\}$ *is regularly varying of index* $-\mu^*$.

Explicit expressions for $P_{j,n}(x)$ and $\kappa_{j,n}$ are given in Subsection 8.4.5.

**Remark 8.3.1** Recall that we assumed $B_k(\cdot)$ to be regularly varying of index $-\nu_k < -1$ for all $k = 1, \ldots, K$. In fact, Theorem 8.3.1 continues to hold if, for some $k$, $1 - B_k(x) = o(x^{-\alpha})$ as $x \to \infty$ for any $\alpha$. Theorem 8.3.1 and all the results stated below which follow from it, formally go through if we simply define $\nu_k := \infty$ in this light-tailed case.

### 8.3.3 Single-session overflow scenario

The expressions for the coefficients $\kappa_{j,n}$ may in principle be computable, but are in general not very explicit. However, as described in the introduction, rather tractable results are available for scenarios where just a single long session can cause overflow. We now specialize the general result stated in Theorem 8.3.1 to these scenarios in order to obtain more explicit expressions, and recover these results. Let $e_k$ denote the $k$-th unit vector. Define $T^* = \{k : e_k \in S^*\}$.

**Theorem 8.3.2** *Assume that* $S^* \subseteq \{e_1, \ldots, e_K\}$. *If* $r^{\min} = \min_{k \in T^*} r_k > 1 - \rho$, *then*

$$\mathbb{P}\{V > x\} \sim \sum_{k \in T^*} \frac{\rho_k}{1-\rho} \mathbb{P}\{B_k^r > \frac{x}{r_k + \rho - 1}\}. \tag{3.4}$$

This result is obtained in [164] under the condition that $r_k > 1 - \rho$ and $B_k(\cdot)$ is of intermediate regular variation for all $k = 1, \ldots, K$. The discrete-time analogue for Pareto-distributed session lengths may be found in [185].

### 8.3.4 Single-class input

We now consider the important special case of a single input class, i.e., homogeneous input. For conciseness, we suppress the class index 1. We have $S^* = \{n^*\}$ and $\mu^* = n^*(\nu - 1)$, with $n^* := \lceil (1-\rho)/r \rceil$.

**Theorem 8.3.3** *Assume that* $r^{\min} = n^* r > 1 - \rho$. *Then,*

$$\mathbb{P}\{V > x\} \sim \sum_{j=0}^{n^*} \frac{\bar{\rho}^{n^*}}{j!} P_{j,n^*}(x), \tag{3.5}$$

*where* $P_{j,n^*}(x)$ *satisfies*

$$P_{j,n^*}(x) \sim \kappa_{j,n^*} \mathbb{P}\{B^r > \frac{x}{d_{n^*}}\}^{n^*},$$

*for some constant* $\kappa_{j,n^*}$.
*In particular,* $\mathbb{P}\{V > x\}$ *is regularly varying of index* $-n^*(\nu - 1)$.

An explicit expression for $\kappa_{j,n^*}$ is given in Subsection 8.4.5.

### 8.3.5 Single-class input with single-session overflow scenario

Finally, we consider the intersection of single-class input with a single-session overflow scenario. Taking $T^* = \{1\}$ in Theorem 8.3.2, or $n^* = 1$ in Theorem 8.3.3, we find

$$\mathbb{P}\{V > x\} \sim \frac{\rho}{1-\rho}\mathbb{P}\{B^r > \frac{x}{r+\rho-1}\}. \tag{3.6}$$

This result is also obtained in [164] and [238].

**Remark 8.3.2** It is worth observing that the qualitative resemblance of (3.6) with (3.4) is markedly stronger than with (3.5). Thus, the extension to a multiple-session overflow scenario has greater ramifications than the issue of heterogeneous input. This confirms that the fundamental problem lies in the plurality of the set $S^*$ rather than the heterogeneity of the input or non-uniqueness of the set $S^*$.

**Remark 8.3.3** It is also interesting to compare (3.6) with the corresponding result for a single On-Off source. Specifically, consider a fluid queue of capacity $c$ fed by a single On-Off source with the same On-periods $B$, Off-periods with mean $1/\lambda'$, peak rate $r'$, fraction Off-time $p = (1+\lambda'\beta)^{-1}$, and traffic intensity $\rho' = (1-p)r'$, with $\rho' < c < r'$. Then the asymptotic behavior of the workload is given by Theorem 2.2.3,

$$\mathbb{P}\{V' > x\} \sim p\frac{\rho'}{c-\rho'}\mathbb{P}\{B^r > \frac{x}{r'-c}\}. \tag{3.7}$$

Now suppose that we choose $r = r' - \rho' = pr'$, $\lambda = (1/\lambda' + \beta)^{-1}$, so that $\rho = \lambda\beta r = (1-p)r = (1-p)pr' = p\rho'$, and $c = 1 + \rho' - \rho$. Then (3.7) agrees with (3.6). In other words, if $r + \rho > 1$, then the workload in a queue of unit capacity fed by $M/G/\infty$ input with $\lambda = (1/\lambda' + \beta)^{-1}$ and $r = r' - \rho'$ is asymptotically equivalent to that in a queue of capacity $c = 1 + \rho' - \rho$ fed by a single On-Off source with the same On-periods $B$, peak rate $r'$, and Off-periods with mean $1/\lambda'$.

This may be understood as follows. In both situations, a large workload level is most likely due to a single extreme event causing a persistent positive drift, either a long session in the $M/G/\infty$ case, or a long On-period in the On-Off case. By assumption, the sessions in the $M/G/\infty$ case have the same distribution as the On-periods in the On-Off case. The chosen parameter values imply that the frequency of sessions and On-periods is also equal. The mean number of On-periods per unit of time is $(1/\lambda' + \beta)^{-1} = \lambda$, the rate at which sessions arrive. As a result, the occurrence of long sessions and long On-periods matches. The workload dynamics during long sessions and long On-periods coincide as well. With $M/G/\infty$ input, the workload has positive drift $r + \rho - 1$ when a long session is active, and negative drift $\rho - 1$ otherwise. With On-Off input, the workload increases at rate $r' - c = r + \rho - 1$ during a long On-period, and decreases approximately at rate $\rho' - c = \rho - 1$ otherwise. Unfortunately, this equivalence does not seem to extend to more general scenarios.

## 8.4 Proof of Theorem 8.3.1

In this section we analyze the asymptotic behavior of $\mathbb{P}\{V(ax) > x\}$ for fixed $a$ and $x \to \infty$. As the next theorem shows, this directly yields the steady-state asymptotics after letting $a \to \infty$.

**Theorem 8.4.1** *If* $r^{\min} > 1 - \rho$, *then*

$$\lim_{a\to\infty} \lim_{x\to\infty} \frac{\mathbb{P}\{V(ax) > x\}}{\mathbb{P}\{V > x\}} = 1.$$

The proof of the above theorem is deferred to Section 8.5.

In order to analyze $\mathbb{P}\{V(ax) > x\}$, it will be convenient to use the representation

$$V(ax) = \sup_{0\le s\le ax} \{A(0,s) - s\},$$

see Subsection 8.2.3. For the tail behavior of $\mathbb{P}\{V(ax) > x\}$, similar heuristic arguments apply as those sketched in Subsection 8.3.1. The only difference is that in general a positive drift alone is not enough for the process $\{A(0,s) - s\}$ to reach level $x$ before time $ax$. Instead, the drift should be at least $\frac{1}{a}$. Therefore, the integer linear program as formulated in Subsection 8.3.1 needs to be modified as follows.

$$\begin{aligned} \min \quad & \mu = \sum_{k=1}^{K} n_k(\nu_k - 1) \\ \text{sub} \quad & \sum_{k=1}^{K} n_k r_k \ge 1 - \rho + \frac{1}{a} \\ & n_k \in \mathbb{N}, \qquad k = 1, \ldots, K. \end{aligned}$$

Let $S_a^* \subseteq \mathbb{N}^K$ be the set of optimal solutions of the above linear program. Denote by $\mu_a^*$ the corresponding optimal value. Also, define $r_a^{\min} := \min_{n\in S_a^*} \sum_{k=1}^{K} n_k r_k$.

The analysis of the tail behavior of $\mathbb{P}\{V(ax) > x\}$ involves several steps.

- We first separate 'short' and 'long' sessions. A session is called 'long' if it exceeds length $\epsilon x$, with $\epsilon$ some small positive constant, independent of $x$. Otherwise, it is called 'short'. We show that the 'short' sessions can be asymptotically ignored if the capacity is reduced by $\rho$, in the sense that for $\epsilon$ sufficiently small,

$$\mathbb{P}\{V(ax) > x\} \sim \mathbb{P}\{V_{>\epsilon x}^{1-\rho}(ax) > x\}.$$

- Next, we determine the typical combination of long sessions involved in causing overflow. Specifically, we prove that, for overflow of level $x$ to occur within time $ax$, the configuration of long sessions in the interval $[0, ax]$ must be $n = (n_1, \ldots, n_K)$, for some $n \in S_a^*$.

- Subsequently, we identify a stopping time $\bar{\tau}_f^n(\epsilon x)$ (conditional upon the event that the configuration of long sessions is $n \in S_a^*$) such that for $a$ sufficiently large and $c$ sufficiently close to $1 - \rho$,

$$\mathbb{P}\{\sup_{0 \le s \le ax} \{A_{>\epsilon x}(0,s) - cs\} > x\} \sim \mathbb{P}\{A_{>\epsilon x}(0, \bar{\tau}_f^n(\epsilon x)) - c\bar{\tau}_f^n(\epsilon x) > x\}.$$

- Last, we compute the asymptotic behavior of $\mathbb{P}\{A_{>\epsilon x}(0, \bar{\tau}_f^n(\epsilon x)) - c\bar{\tau}_f^n(\epsilon x) > x\}$ as $x \to \infty$, which involves a rather tedious but straightforward calculation.

Subsections 8.4.1-8.4.4 elaborate upon the above four steps, which prepare the way for the proof of Theorem 8.3.1 in Subsection 8.4.5. As a by-product of the analysis, we obtain asymptotically tight lower and upper bounds for the transient workload distribution in Subsection 8.4.6. The various steps involve similar probabilistic arguments as developed in [239] for the special case where a single long session is enough to cause overflow. The first two steps are also used in [185] to derive asymptotic lower and upper bounds for $\mathbb{P}\{V > x\}$ which coincide up to a constant factor. The exact asymptotics for infinite buffers however entail a detailed calculation as in the last two steps listed above.

### 8.4.1 Discarding short sessions

As a first step, we separate short and long sessions. We show that – as far as asymptotic behavior is concerned – the short sessions can be deleted if the capacity is reduced by $\rho$. Formally, we derive asymptotic lower and upper bounds for $\mathbb{P}\{V(ax) > x\}$ of the form $\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1 \pm \theta)x\}$ for arbitrarily small $\delta$, $\theta$.

We first establish a simple sample-path lower bound. For any $c > 0$, define $Z_{\le z}^c(t) := \sup_{0 \le s \le t} \{cs - A_{\le z}(0,s)\}$.

**Proposition 8.4.1** *For any $c \in (0, \rho_{\le z})$,*

$$\mathbb{P}\{V(t) > x\} \ge \mathbb{P}\{V_{>z}^{1-c}(t) > x + y\}\mathbb{P}\{Z_{\le z}^c(t) \le y\}.$$

**Proof**
Sample-path wise,

$$\begin{aligned} V(t) &= \sup_{0\le s\le t}\{A(0,s)-s\} \\ &= \sup_{0\le s\le t}\{A_{>z}(0,s)-(1-c)s+A_{\le z}(0,s)-cs\} \\ &\ge \sup_{0\le s\le t}\{A_{>z}(0,s)-(1-c)s\}-\sup_{0\le s\le t}\{cs-A_{\le z}(0,s)\} \\ &= V_{>z}^{1-c}(t)-Z_{\le z}^{c}(t). \end{aligned}$$

□

We now use the above sample-path bound to obtain an asymptotic lower bound for $\mathbb{P}\{V(ax)>x\}$ as $x\to\infty$.

**Proposition 8.4.2** *For any $\delta>0$, $\epsilon>0$, $\theta>0$,*

$$\mathbb{P}\{V(ax)>x\} \stackrel{>}{\sim} \mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta}(ax)>(1+\theta)x\}.$$

**Proof**
Since $\rho_{\le z}\uparrow\rho$ for $z\to\infty$, there exists an $x_0$ such that $\rho_{\le\epsilon x}>\rho-\delta$ for all $x\ge x_0$.
From Proposition 8.4.1, taking $c=\rho-\delta$, $y=\theta x$, $z=\epsilon x$, for all $x\ge x_0$,

$$\frac{\mathbb{P}\{V(ax)>x\}}{\mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta}(ax)>(1+\theta)x\}} \ge \mathbb{P}\{Z_{\le\epsilon x}^{\rho-\delta}(ax)\le\theta x\} \ge \mathbb{P}\{Z_{\le\epsilon x_0}^{\rho-\delta}(ax)\le\theta x\}.$$

The statement then easily follows.

□

We now proceed with a simple sample-path upper bound.

**Proposition 8.4.3** *For any $c\in(\rho_{\le z},1-\rho_{>z})$,*

$$\mathbb{P}\{V(t)>x\} \le \mathbb{P}\{V_{>z}^{1-c}(t)>x-y\}+\mathbb{P}\{V_{\le z}^{c}(t)>y\}.$$

**Proof**
Sample-path wise,

$$\begin{aligned} V(t) &= \sup_{0\le s\le t}\{A(0,s)-s\} \\ &= \sup_{0\le s\le t}\{A_{>z}(0,s)-(1-c)s+A_{\le z}(0,s)-cs\} \\ &\le \sup_{0\le s\le t}\{A_{>z}(0,s)-(1-c)s\}+\sup_{0\le s\le t}\{A_{\le z}(0,s)-cs\} \\ &= V_{>z}^{1-c}(t)+V_{\le z}^{c}(t). \end{aligned}$$

□

The next proposition provides an upper bound which indicates that the workload from the short sessions can be asymptotically neglected.

**Proposition 8.4.4** *For any $c > \rho$, $\theta > 0$, $\mu > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,*

$$\mathbb{P}\{V^c_{\leq \epsilon x}(ax) > \theta x\} = \mathrm{o}(x^{-\mu})$$

*as $x \to \infty$.*

**Proof**
Define $\delta := (c - \rho)/K$. Then

$$\begin{aligned} V^c_{\leq \epsilon x}(ax) &= \sup_{0 \leq s \leq ax} \{A_{\leq \epsilon x}(0,s) - cs\} \\ &= \sup_{0 \leq s \leq ax} \{\sum_{k=1}^{K} A_{k,\leq \epsilon x}(0,s) - \sum_{k=1}^{K} (\rho_k + \delta)s\} \\ &\leq \sum_{k=1}^{K} \sup_{0 \leq s \leq ax} \{A_{k,\leq \epsilon x}(0,s) - (\rho_k + \delta)s\} \\ &= \sum_{k=1}^{K} V^{\rho_k+\delta}_{k,\leq \epsilon x}(ax). \end{aligned}$$

This implies

$$\mathbb{P}\{V^c_{\leq \epsilon x}(ax) > x\} \leq \sum_{k=1}^{K} \mathbb{P}\{V^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) > x/K\}.$$

Thus, it suffices to show that

$$\mathbb{P}\{V^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) > x/K\} = \mathrm{o}(x^{-\mu})$$

as $x \to \infty$ for all $k = 1, \ldots, K$.
Now observe that

$$V^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) \leq A^{(0)}_{k,\leq \epsilon x} + \sup_{0 \leq s \leq ax} \{A^{(>0)}_{k,\leq \epsilon x}(0,s) - (\rho_k + \delta)s\},$$

where the two terms correspond to the traffic generated by the sessions already active at and starting after time 0, respectively. Hence,

$$\begin{aligned} \mathbb{P}\{V^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) > x/K\} &\leq \mathbb{P}\{A^{(0)}_{k,\leq \epsilon x} > x/(2K)\} \\ &+ \mathbb{P}\{\sup_{0 \leq s \leq ax} \{A^{(>0)}_{k,\leq \epsilon x}(0,s) - (\rho_k + \delta)s\} > x/(2K)\} \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

In the remainder of the proof, we bound the terms I and II.

We first consider Term I. Let $\alpha \in (0,1)$ such that $\mathbb{E}\{(B_k^r)^\alpha\} < \infty$. Let $\beta \in (0,\alpha)$. Note that $A^{(0)}_{k,\leq \epsilon x}$ is stochastically smaller than $r_k \sum_{i=1}^{N_k(0)} B^r_{k,i}(\epsilon x)$, where $B^r_{k,i}(\epsilon x) \stackrel{d}{=} B^r_{k,i} \mid B^r_{k,i} \leq \epsilon x$. Thus,

$$\mathrm{I} \leq \mathbb{P}\{r_k \sum_{i=1}^{x^\beta} B^r_{k,i}(\epsilon x) > x/(2K)\} + \mathbb{P}\{N_k(0) > x^\beta\}.$$

Since $N_k(0)$ is Poisson distributed, the second term decays exponentially fast in $x$. Using Lemma 2.4.1 (applied to $B^r_{k,i}(\epsilon x)^\alpha$), the first term can be bounded as follows, for $\epsilon$ sufficiently small:

$$\begin{aligned}
& \mathbb{P}\{r_k \sum_{i=1}^{x^\beta} B^r_{k,i}(\epsilon x) > x/(2K)\} \\
= \; & \mathbb{P}\{(r_k \sum_{i=1}^{x^\beta} B^r_{k,i}(\epsilon x))^\alpha > (x/2K)^\alpha\} \\
\leq \; & \mathbb{P}\{r_k^\alpha \sum_{i=1}^{x^\beta} B^r_{k,i}(\epsilon x)^\alpha > (x/2K)^\alpha\} \\
= \; & \mathbb{P}\{r_k^\alpha \sum_{i=1}^{x^\beta} [B^r_{k,i}(\epsilon x)^\alpha - 2\mathbb{E}\{(B^r_{k,i})^\alpha\}] > (x/(2K))^\alpha - 2\mathbb{E}\{(B^r_{k,i})^\alpha\} x^\beta\} \\
\leq \; & \phi(x^\alpha/(2Kr_K)^\alpha - 2\mathbb{E}\{(B_k^r)^\alpha\} x^\beta / r_k^\alpha),
\end{aligned}$$

with $\phi(\cdot) \in \mathcal{R}_{-\eta}$, $\eta > \mu/\alpha$.

We now turn to Term II. Note that $\sup_{0\leq s\leq ax} \{A^{(>0)}_{k,\leq \epsilon x}(0,s) - (\rho_k+\delta)s\}$ is stochastically smaller than $W^{\rho_k+\delta}_{k,\leq \epsilon x}(ax)$, where the latter quantity represents the workload if the entire amount of traffic generated over the duration of a session were released instantaneously upon the arrival of the session. Thus,

$$\mathrm{II} \leq \mathbb{P}\{W^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) > x/(2K)\}.$$

Now observe that $W^{\rho_k+\delta}_{k,\leq \epsilon x}(ax)$ is the workload at time $ax$ in an $M/G/1$ queue of capacity $\rho_k + \delta$ with arrival rate $\lambda_k B_k(\epsilon x)$ and service time distribution $B_k(y/r_k)/B_k(\epsilon x)$, $0 \leq y \leq \epsilon x r_k$. Let $B'_{k,n}(\epsilon x)$, $n \geq 1$, be an i.i.d. sequence of random variables with this distribution, and let $U_{k,n}$, $n \geq 1$, be an i.i.d. sequence of interarrival times. Denote by $N_k(ax) := \sup\{n : U_{k,1} + \ldots + U_{k,n} \leq ax\}$ the number of arrivals in this $M/G/1$ queue up to time $ax$. Define $S_{k,n}(\epsilon x) := \sum_{i=1}^{n} X_{k,i}$, with $X_{k,i} := B'_{k,i}(\epsilon x) - (\rho_k + \delta)U_{k,i}$. Then, for any $\Lambda$,

$$\begin{aligned}
\mathbb{P}\{W^{\rho_k+\delta}_{k,\leq \epsilon x}(ax) > x/(2K)\} &= \mathbb{P}\{\sup_{n\leq N_k(ax)} S_{k,n}(\epsilon x) > x/(2K)\} \\
&\leq \mathbb{P}\{\sup_{n\leq \Lambda ax} S_{k,n}(\epsilon x) > x/(2K)\} + \mathbb{P}\{N_k(ax) > \Lambda ax\}.
\end{aligned}$$

The second term decays exponentially fast in $x$ for $\Lambda > \lambda_k$. Using the truncation Lemma 2.4.1, noting that $\mathbb{E}\{X_1\} < 0$, the first term can be bounded by, for $\epsilon^* > 0$ sufficiently small,

$$\sum_{n=1}^{\Lambda a x} \mathbb{P}\{S_{k,n}(\epsilon x) > x/(2K)\} \leq \Lambda a x \phi(x/(2K)),$$

with $\phi(\cdot) \in \mathcal{R}_{-\alpha}$, $\alpha > \mu + 1$. This completes the proof.

□

We now combine the above two bounds to obtain an asymptotic upper bound for $\mathbb{P}\{V(ax) > x\}$ as $x \to \infty$.

**Proposition 8.4.5** *For any $\delta > 0$, $\theta > 0$, $\mu > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,*

$$\mathbb{P}\{V(ax) > x\} \leq \mathbb{P}\{V_{>\epsilon x}^{1-\rho-\delta}(ax) > (1-\theta)x\} + \mathrm{o}(x^{-\mu})$$

*as $x \to \infty$.*

**Proof**
The proof follows directly from Propositions 8.4.3 and 8.4.4 taking $c = \rho + \delta$.

□

Combined, Propositions 8.4.2 and 8.4.5 allow us to restrict the attention to long sessions only, and focus on probabilities of the form $\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1 \pm \theta)x\}$.

### 8.4.2 Configuration of long sessions

In this subsection, we determine the typical combination of long sessions involved in causing overflow. Specifically, we show that, for overflow of level $x$ to occur within time $ax$, the configuration of long sessions in the interval $[0, ax]$ must be in the set $S_a^*$. As we argued before, these configurations of long sessions may be interpreted as the most likely ones to occur among those producing sufficiently high drift. All other combinations are unlikely to cause overflow, either because the resulting drift is simply too low, or because the corresponding probability is too small (or both).
In order to formalize these statements, we need to keep track of the number of long sessions in the time interval $[0, ax]$. With minor abuse of notation, define $N_{k,>\epsilon x}(T)$ as the number of class-$k$ sessions exceeding length $\epsilon x$ in the time interval $T$. Denote

$N_{>\epsilon x}(T) := (N_{1,>\epsilon x}(T), \ldots, N_{K,>\epsilon x}(T))$. Formally, we will show that for $\delta$, $\theta$ sufficiently small,

$$\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1\pm\theta)x\} \sim \sum_{n\in S_a^*} \mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1\pm\theta)x; N_{>\epsilon x}([0,ax]) = n\}.$$

We first exclude the possibility that overflow is caused by some configuration which fails to generate at least a drift $1/a$.
Let $S_a^*(c)$ be the set of optimal solutions of the integer linear program formulated at the beginning of this section with the constraint value $1-\rho+1/a$ replaced by $c+1/a$. Denote by $\mu_a^*(c)$ the corresponding optimal value. Define $S_a^-(c) := \{n \in \mathbb{N}^K : \sum_{k=1}^K n_k r_k < c+1/a\}$, $S_a^+(c) := \{n \in \mathbb{N}^K : \sum_{k=1}^K n_k r_k > c+1/a\}$, and $r_a^{\max}(c) := \max_{n\in S_a^-(c)} \sum_{k=1}^K n_k r_k$.

**Proposition 8.4.6** *For $\theta$ sufficiently small, and all $\epsilon > 0$, $x > 0$,*

$$\mathbb{P}\{V_{>\epsilon x}^{c}(ax) > (1\pm\theta)x; N_{>\epsilon x}([0,ax]) \in S_a^-(c)\} = 0.$$

**Proof**
The idea of the proof is as follows. If $N_{>\epsilon x}([0,ax]) \in S_a^-(c)$, then during the time interval $[0,ax]$ the drift of the workload is always less than $1/a$. Hence, the workload cannot reach level $(1\pm\theta)x$ before time $ax$ for $\theta$ sufficiently small.
Formally, denote $u_a(c) := c + 1/a - r_a^{\max}(c) > 0$. If $N_{>\epsilon x}([0,ax]) \in S_a^-(c)$, then the left derivative $\frac{\mathrm{d}}{\mathrm{d}s} A_{>\epsilon x}(0,s) \le r_a^{\max}(c)$ for all $s \in [0,ax]$, so that $A_{>\epsilon x}(0,s) \le r_a^{\max}(c)s$ for all $s \in [0,ax]$. Therefore,

$$\begin{aligned} V_{>\epsilon x}^{c}(ax) &= \sup_{0\le s\le ax} \{A_{>\epsilon x}(0,s) - cs\} \le \sup_{0\le s\le ax} \{(r_a^{\max}(c) - c)s\} \\ &= \sup_{0\le s\le ax} \{(1/a - u_a(c))s\} = (1/a - u_a(c))ax. \end{aligned}$$

The latter quantity is less than $(1\pm\theta)x$ for $\theta < au_a(c)$.

□

We now eliminate all configurations of long sessions that do generate at least a drift $1/a$, but that are relatively unlikely compared to other combinations that do so.

**Proposition 8.4.7** *There exists a $\mu > \mu_a^*(c)$ such that, for all $\epsilon > 0$, $n \in S_a^+(c) \setminus S_a^*(c)$,*

$$\mathbb{P}\{N_{>\epsilon x}([0,ax]) \ge n\} = \mathrm{o}(x^{-\mu}),$$

*as $x \to \infty$.*

**Proof**
Note that $N_{k,>\epsilon x}([0,ax])$ has a Poisson distribution with parameter $\bar{\rho}_k \mathbb{P}\{B_k^r > \epsilon x\} + \lambda_k a x \mathbb{P}\{B_k > \epsilon x\}$. A straightforward computation then shows that $\mathbb{P}\{N_{k,>\epsilon x}([0,ax]) \geq n_k\}$ is upper bounded by a function which is regularly varying of index $-n_k(\nu_k - 1)$. Since

$$\mathbb{P}\{N_{>\epsilon x}([0,ax]) \geq n\} = \prod_{k=1}^{K} \mathbb{P}\{N_{k,>\epsilon x}([0,ax]) \geq n_k\},$$

the left hand side is upper bounded by a function which is regularly varying of index $-\sum_{k=1}^{K} n_k(\nu_k - 1)$. The fact that $n \in S_a^+(c) \setminus S_a^*(c)$ implies $\sum_{k=1}^{K} n_k(\nu_k - 1) > \mu_a^*(c)$, because otherwise $n \in S_a^*(c)$.

□

Combined, the above two propositions allow us to limit the attention to scenarios with $N_{>\epsilon x}([0,ax]) \in S_a^*(c)$, as formalized in the following lemma.

**Lemma 8.4.1** *Assume that $r_a^{\min} > 1 - \rho$. Then there exists a $\mu > \mu_a^*$ such that for $\delta$, $\theta$ sufficiently small, and all $\epsilon > 0$,*

$$\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1 \pm \theta)x\} = \sum_{n \in S_a^*} \mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n\} + \mathrm{o}(x^{-\mu}).$$

**Proof**
The proof follows directly from Propositions 8.4.6, 8.4.7, noting that $S_a(1 - \rho \pm \delta) = S_a^*$ for $\delta$ sufficiently small as $r_a^{\min} > 1 - \rho$.

□

Combined with the earlier results, we have now obtained asymptotic lower and upper bounds for $\mathbb{P}\{V > x\}$ in terms of the probabilities $\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1\pm\theta)x; N_{>\epsilon x}([0,ax]) = n\}$. What thus remains is to determine the asymptotic behavior of these probabilities as $x \to \infty$, which is the subject of the next subsection.

### 8.4.3 Identifying a stopping time

In this subsection we identify a stopping time $\bar{\tau}_f^n(\epsilon x)$ (conditional upon the event $N_{>\epsilon x}([0,ax]) = n$) such that for $a$ sufficiently large and $c$ sufficiently close to $1 - \rho$,

$$\mathbb{P}\{\sup_{0 \leq s \leq ax} \{A_{>\epsilon x}(0,s) - cs\} > x\} \sim \mathbb{P}\{A_{>\epsilon x}(0, \bar{\tau}_f^n(\epsilon x)) - c\bar{\tau}_f^n(\epsilon x) > x\}.$$

We first introduce some additional notation. Assume that $N_{>\epsilon x}(0) \leq n$. In this case, we define $A^n_{>\epsilon x}(0,t)$ as the amount of traffic generated up to time $t$ by the first $n_k$ class-$k$ sessions only, $k = 1, \ldots, K$. Define $V^{c,n}_{>\epsilon x}(t) := \sup_{0 \leq s \leq t} \{A^n_{>\epsilon x}(0,s) - cs\}$.
Let $\tau^n_{s,k}(\epsilon x)$ and $\tau^n_{f,k}(\epsilon x)$ be the respective starting and finishing times of the $n$-th class-$k$ session exceeding length $\epsilon x$. For any $n \in \mathbb{N}^K$, let

$$\tau^n_s(\epsilon x) := \max_{k=1,\ldots,K} \tau^{n_k}_{s,k}(\epsilon x),$$

and

$$\tau^n_f(\epsilon x) := \min_{k=1,\ldots,K} \tau^{n_k}_{f,k}(\epsilon x).$$

Thus, for a configuration $n \in \mathbb{N}^K$ of long sessions, $\tau^n_s(\epsilon x)$ is the time at which the last long session begins, and $\tau^n_f(\epsilon x)$ is the time at which the first long session ends. To account for the case $\tau^n_f(\epsilon x) > ax$, we define $\bar{\tau}^n_f(\epsilon x) := \min\{ax, \tau^n_f(\epsilon x)\}$. This turns out to be the relevant stopping time, as is demonstrated by the following lemma.

**Lemma 8.4.2** *There exists a $\mu > \mu^*_a(c)$ such that for $\theta$ sufficiently small and all $n \in S^*_a(c)$,*

$$\mathbb{P}\{V^c_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n\} \stackrel{>}{\sim}$$

$$\mathbb{P}\{N_{>\epsilon x}(0) \leq n; A^n_{>\epsilon x}(0, \bar{\tau}^n_f(\epsilon x)) - c\bar{\tau}^n_f(\epsilon x) > (1-\theta)x\} + \mathrm{o}(x^{-\mu}).$$

*In case $r^{\max}_a(c) < c$, there also exists a $\mu > \mu^*_a(c)$ such that for $\theta$ sufficiently small and all $n \in S^*_a(c)$,*

$$\mathbb{P}\{V^c_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n\} \stackrel{<}{\sim}$$

$$\mathbb{P}\{N_{>\epsilon x}(0) \leq n; A^n_{>\epsilon x}(0, \bar{\tau}^n_f(\epsilon x)) - c\bar{\tau}^n_f(\epsilon x) > (1 \pm \theta)x\} + \mathrm{o}(x^{-\mu}).$$

**Proof**
We first prove the second statement. Since $V^{c,n}_{>\epsilon x}(ax) \leq V^c_{>\epsilon x}(ax)$, with strict equality under the event $N_{>\epsilon x}([0,ax]) = n$, and the latter event also implies that $N_{>\epsilon x}(0) \leq n$, we have

$$\mathbb{P}\{V^c_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n\} =$$

$$\mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n; N_{>\epsilon x}(0) \leq n\}.$$

First observe that

$$\begin{aligned} & \mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0,ax]) = n; N_{>\epsilon x}(0) \leq n\} \\ \leq\ & \mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1 \pm \theta)x; N_{>\epsilon x}(0) \leq n\} \\ =\ & \mathbb{P}\{\sup_{0 \leq s \leq ax} \{A^n_{>\epsilon x}(0,s) - cs\} > (1 \pm \theta)x; N_{>\epsilon x}(0) \leq n\}. \end{aligned}$$

Note that before time $\tau_s^n(\epsilon x)$ and after time $\tau_f^n(\epsilon x)$ the drift of the process $A^n_{>\epsilon x}(0,s)$ is at most $r_a^{max}(c) < c$. Thus, the drift of the process $\{A^n_{>\epsilon x}(0,s) - cs\}$ is only positive between times $\tau_s^n(\epsilon x)$ and $\tau_f^n(\epsilon x)$. Hence, $\sup_{0\le s\le ax}\{A^n_{>\epsilon x}(0,s) - cs\} > (1\pm\theta)x$ implies that $A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x$. Thus, the last probability in the above display is smaller than $\mathbb{P}\{N_{>\epsilon x}(0)\le n; A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\}$.
We now turn to the first statement. Observe that $V^{c,n}_{>\epsilon x}(ax) = 0$, unless $N_{>\epsilon x}([0,ax]) \ge n$, so for $\theta$ sufficiently small, using Proposition 8.4.7,

$$\begin{aligned}
& \mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1\pm\theta)x; N_{>\epsilon x}([0,ax]) = n; N_{>\epsilon x}(0)\le n\}\\
\ge\ & \mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1\pm\theta)x; N_{>\epsilon x}(0)\le n\} - \mathbb{P}\{N_{>\epsilon x}([0,ax]) > n\}\\
\ge\ & \mathbb{P}\{V^{c,n}_{>\epsilon x}(ax) > (1\pm\theta)x; N_{>\epsilon x}(0)\le n\} + \mathrm{o}(x^{-\mu})\\
=\ & \mathbb{P}\{\sup_{0\le s\le ax}\{A^n_{>\epsilon x}(0,s) - cs\} > (1\pm\theta)x; N_{>\epsilon x}(0)\le n\} + \mathrm{o}(x^{-\mu})\\
\ge\ & \mathbb{P}\{N_{>\epsilon x}(0)\le n; A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\} + \mathrm{o}(x^{-\mu}).
\end{aligned}$$

□

Combined with the earlier results, we have now obtained asymptotic lower and upper bounds for $\mathbb{P}\{V > x\}$ in terms of the probabilities $\mathbb{P}\{N_{>\epsilon x}(0)\le n; A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\}$ with $c = 1-\rho\pm\delta$. What thus remains is to determine the asymptotic behavior of these probabilities as $x\to\infty$, which is the subject of the next subsection.

### 8.4.4 Computation of the pre-factor

As a final step, we compute the asymptotic behavior of $\mathbb{P}\{N_{>\epsilon x}(0)\le n; A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\}$ for fixed $n\in S_a^*(c)$ and $x\to\infty$. Throughout this subsection, we assume that $a$ is large enough for the condition $r_a^{\max}(c) < c$ to hold.
We start by conditioning upon the configuration of long sessions active at time 0. For $j = (j_1,\ldots,j_K)$, define the event $D_j(\epsilon x)$ by $D_j(\epsilon x) := \{N_{>\epsilon x}(0) = j\}$. In words, $D_j(\epsilon x)$ is the event that the number of long class-$k$ sessions active at time 0 is $j_k$, $k = 1,\ldots,K$. Denote $\mathbb{P}_j\{\cdot\} = \mathbb{P}\{\cdot|D_j(\epsilon x)\}$. Then

$$\begin{aligned}
& \mathbb{P}\{N_{>\epsilon x}(0)\le n; A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\}\\
=\ & \sum_{j\le n}\mathbb{P}\{D_j(\epsilon x)\}\mathbb{P}_j\{A^n_{>\epsilon x}(0,\bar\tau_f^n(\epsilon x)) - c\bar\tau_f^n(\epsilon x) > (1\pm\theta)x\}.
\end{aligned}$$

Note that

$$\mathbb{P}\{D_j(\epsilon x)\} = \prod_{k=1}^K \frac{(\bar\rho_k\mathbb{P}\{B_k^r > \epsilon x\})^{j_k}}{j_k!}\mathrm{e}^{-\bar\rho_k\mathbb{P}\{B_k^r>\epsilon x\}} \sim \prod_{k=1}^K \frac{(\bar\rho_k\mathbb{P}\{B_k^r > \epsilon x\})^{j_k}}{j_k!}.$$

It remains to compute the asymptotic behavior of $\mathbb{P}_j\{A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x))-c\bar\tau^n_f(\epsilon x) > (1\pm\theta)x\}$ as $x\to\infty$. In order to do so, we need to condition upon the arrival times of the remaining sessions as well. Denote the interarrival times of the class-$k$ sessions by $E_{ki}(\epsilon x)$, $k = 1,\ldots,K$, $i = 1,2,\ldots$. Note that $E_{ki}(\epsilon x)$ is an exponentially distributed random variable with parameter $\lambda_k\mathbb{P}\{B_k > \epsilon x\}$.
To obtain an expression for $A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x)) - c\bar\tau^n_f(\epsilon x)$ under the event $D_j(\epsilon x)$, note that, if all long sessions had been active already at time 0, the expression would equal $c_n\bar\tau^n_f(\epsilon x)$, with $c_n := \sum_{k=1}^{K} n_k r_k - c$. However, some sessions may have started later. To account for this, it is not hard to see that we need to subtract $H(\epsilon x)$, which is defined by

$$H(\epsilon x) := \sum_{k=1}^{K} r_k \sum_{i=1}^{n_k-j_k}\sum_{l=1}^{i} E_{kl}(\epsilon x).$$

This is summarized in the following lemma.

**Lemma 8.4.3** *Under the event $D_j(\epsilon x)$, $A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x)) - c\bar\tau^n_f(\epsilon x)$ can be represented as*

$$A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x)) - c\bar\tau^n_f(\epsilon x) = c_n\bar\tau^n_f(\epsilon x) - H(\epsilon x),$$

*with $\bar\tau^n_f(\epsilon x) = \min\{ax, \tau^n_f(\epsilon x)\}$ the stopping time defined earlier and*

$$\tau^n_f(\epsilon x) = \min_{k=1,\ldots,K}\min\{\min_{i=1,\ldots,j_k}\bar B^r_{ki}(\epsilon x), \min_{i=1,\ldots,n_k-j_k}\big[E_{k1}(\epsilon x)+\ldots+E_{ki}(\epsilon x)+\bar B_{ki}(\epsilon x)\big]\}.$$

*Here $\bar B^r_{ki}(\epsilon x) \stackrel{d}{=} B^r_{ki} \mid B^r_{ki} > \epsilon x$, and $\bar B_{ki}(\epsilon x) \stackrel{d}{=} B_{ki} \mid B_{ki} > \epsilon x$.*

We proceed to compute the asymptotic behavior of $\mathbb{P}_j\{A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x)) - c\bar\tau^n_f(\epsilon x) > (1\pm\theta)x\}$, using the above representation.
Define the sets $E_0 := \{(k,i) : k = 1,\ldots,K, i = 1,\ldots,j_k\}$ and $E_1 := \{(k,i) : k = 1,\ldots,K, i = 1,\ldots,n_k - j_k\}$. Write $y = y_{(k,i)\in E_0}$ (we interpret $y$ as a vector), and let $h(y)$ be a realization of $H(\epsilon x)$, i.e., if $E_{ki}(\epsilon x) = y_{ki}$ for $(k,i)\in E_1$, then

$$h(y) = \sum_{k=1}^{K} r_k \sum_{i=1}^{n_k-j_k}\sum_{l=1}^{i} y_{kl}.$$

Let $t(y)$ be distributed as $\tau^n_f(\epsilon x)$ conditional upon $E_{ki}(\epsilon x) = y_{ki}$ for $(k,i)\in E_1$. Note that $t(y)$ is still a random variable. Hence, using Lemma 8.4.3,

$$\begin{aligned}
&\mathbb{P}_j\{A^n_{>\epsilon x}(0,\bar\tau^n_f(\epsilon x)) - c\bar\tau^n_f(\epsilon x) > (1\pm\theta)x\}\\
&= \int_{y\ge 0}\left[\prod_{(k,i)\in E_1}\big(\lambda_k\mathbb{P}\{B_k>\epsilon x\}\mathrm{e}^{-y_{ki}\lambda_k\mathbb{P}\{B_k>\epsilon x\}}\big)\right]\times\\
&\qquad\qquad \mathbb{P}\{c_n\min\{ax,t(y)\} > (1\pm\theta)x + h(y)\}\mathrm{d}y
\end{aligned}$$

$$= \frac{1}{\prod_{k=1}^{K} \mathbb{P}\{B_k^r > \epsilon x\}^{j_k}} \int_{y\geq 0, h(y)\leq (c_n a-1)x} \left[\prod_{(k,i)\in E_1} \left(\lambda_k \mathrm{e}^{-y_{ki}\lambda_k \mathbb{P}\{B_k>\epsilon x\}}\right)\right]$$
$$\left[\prod_{(k,i)\in E_0} \mathbb{P}\{c_n B_k^r > (1\pm\theta)x + h(y)\}\right]$$
$$\left[\prod_{(k,i)\in E_1} \mathbb{P}\{c_n(y_{k1}+\ldots+y_{ki}+B_k) > (1\pm\theta)x + h(y)\}\right] \mathrm{d}y.$$

This implies (using bounded convergence)

$$\mathbb{P}\{D_j(\epsilon x)\}\mathbb{P}_j\{A^n_{>\epsilon x}(0,\tau_f^n(\epsilon x)) - c\tau_f^n(\epsilon x) > (1\pm\theta)x\}$$
$$\sim \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} \prod_{k=1}^{K} \frac{1}{\beta_k^{n_k-j_k}} \int_{y\geq 0, h(y)\leq (c_n a-1)x} \left[\prod_{(k,i)\in E_0} \mathbb{P}\{c_n B_k^r > (1\pm\theta)x + h(y)\}\right]$$
$$\left[\prod_{(k,i)\in E_1} \mathbb{P}\{c_n(y_{k1}+\ldots+y_{ki}+B_k) > (1\pm\theta)x + h(y)\}\right] \mathrm{d}y.$$

For given $n$ and $j$, define the $|E_1|$-dimensional row vector $g = g^{c,j,n}$ by $g = (g_1, \ldots, g_K)$. Here $g_k$ is a row vector of dimension $n_k - j_k$ with all elements equal to $r_k/c_n$. In the sequel, we write $g := (g_{(k,i)})_{(k,i)\in E_1}$. Let $G$ be a square matrix with all rows equal to $g$. Define $\bar{G} := G - I$. Note that $|\bar{G}| = eg - 1$ and that the inverse $H$ of $\bar{G}$ is given by $H = \frac{1}{eg-1}G - I$. Here $e := (1, \ldots, 1)$ is the unit vector with all elements equal to 1. Note that $gH = \frac{1}{eg-1}g$. Set $z := (z_{ki})_{(k,i)\in E_1}$, where $z_{ki} = y_{k1} + \ldots + y_{ki}$. Define $w := \bar{G}z$. Note that $h(y) = c_n gz$.

Straightforward computations yield

$$\prod_{k=1}^{K} \frac{1}{\beta_k^{n_k-j_k}} \int_{y\geq 0, h(y)\leq (c_n a-1)x} \left[\prod_{(k,i)\in E_0} \mathbb{P}\{c_n B_k^r > (1\pm\theta)x + h(y)\}\right]$$
$$\left[\prod_{(k,i)\in E_1} \mathbb{P}\{c_n(z_{ki}+B_k) > (1\pm\theta)x + h(y)\}\right] \mathrm{d}y$$
$$= \prod_{k=1}^{K} \frac{1}{\beta_k^{n_k-j_k}} \int_{z\geq 0, gz\leq (a-1/c_n)x} \left[\prod_{(k,i)\in E_0} \mathbb{P}\{B_{ki}^r > (1\pm\theta)\frac{x}{c_n} + gz\}\right]$$
$$\left[\prod_{(k,i)\in E_1} \mathbb{P}\{B_{ki} > (1\pm\theta)\frac{x}{c_n} + (\bar{G}z)_{(k,i)}\}\right] \mathrm{d}z$$

$$
\begin{aligned}
&= \prod_{k=1}^{K} \frac{1}{\beta_k^{n_k-j_k}} \frac{1}{eg-1} \int\limits_{w\geq 0, gw\leq (eg-1)(a-1/c_n)x} \left[ \prod_{(k,i)\in E_0} \mathbb{P}\{B_{ki}^r > (1\pm\theta)\frac{x}{c_n} + \frac{gw}{eg-1}\} \right] \\
&\qquad \left[ \prod_{(k,i)\in E_1} \mathbb{P}\{B_{ki} > (1\pm\theta)\frac{x}{c_n} + w_{(k,i)}\} \right] \mathrm{d}w \\
&= \frac{1}{eg-1} \int\limits_{w\geq 0, gw\leq (eg-1)(a-1/c_n)x} \left[ \prod_{(k,i)\in E_0} \mathbb{P}\{B_{ki}^r > (1\pm\theta)\frac{x}{c_n} + \frac{1}{eg-1}gw\} \right] \\
&\qquad \mathrm{d} \prod_{(k,i)\in E_1} \mathbb{P}\{B_{ki}^r > (1\pm\theta)\frac{x}{c_n} + w_{(k,i)}\} \\
&= \mathbb{P}\{B_{ki}^r > (1\pm\theta)\frac{x}{c_n}, k=1,\ldots,K, i=1,\ldots,n_k; \\
&\qquad B_{ki}^r - (1\pm\theta)\frac{x}{c_n} \geq \frac{1}{eg-1} g\left( B_{E_1}^r - (1\pm\theta)\frac{x}{c_n} e \right), (k,i)\in E_0; \\
&\qquad \frac{1}{eg-1} g\left( B_{E_1}^r - (1\pm\theta)\frac{x}{c_n} e \right) \leq (1\pm\theta)x(a-\frac{1}{c_n})\} \\
&=: P_{j,n,a}^c((1\pm\theta)x).
\end{aligned}
$$

In the last expression, $B_{E_1}^r := (B_{ki}^r)_{(k,i)\in E_1}$.

Using the fact that $B_{ki}^r$ is regularly varying of index $1-\nu_k$, it is easy to show that

$$\mathbb{P}\{B_{ki}^r - \frac{x}{c_n} > yx | B_{ki}^r > \frac{x}{c_n}\} \sim \mathbb{P}\{Z_{ki} > y\} := (1+c_n y)^{1-\nu_k}.$$

Take the $Z_{ki}$ independent. Then, with obvious notation,

$$
\begin{aligned}
&(eg-1)P_{j,n,a}^c(x) \\
&= \mathbb{P}\{B_{ki}^r > \frac{x}{c_n}, k=1,\ldots,K, i=1,\ldots,n_k; B_{ki}^r - \frac{x}{c_n} \geq \frac{1}{eg-1} g\left( B_{E_1}^r - \frac{x}{c_n} e \right), \\
&\qquad (k,i)\in E_0; \frac{1}{eg-1} g\left( B_{E_1}^r - \frac{x}{c_n} e \right) \leq x(a-\frac{1}{c_n})\} \\
&\sim \mathbb{P}\{Z_{ki} \geq \frac{1}{eg-1} g Z_{E_1}, (k,i)\in E_0; (a-\frac{1}{c_n}) \geq \frac{1}{eg-1} g Z_{E_1}\} \prod_{k=1}^{K} \mathbb{P}\{B_k^r > \frac{x}{c_n}\}^{j_k}.
\end{aligned}
$$

The above calculations are summarized in the following lemma.

**Lemma 8.4.4** *For $n \in S_a^*(c)$ there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,*

$$\mathbb{P}\{N_{>\epsilon x}(0) \leq n; A_{>\epsilon x}^n(0, \bar{\tau}_f^n(\epsilon x)) - c\bar{\tau}_f^n(\epsilon x) > (1\pm\theta)x\} \sim \sum_{j\leq n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n,a}^c((1\pm\theta)x),$$

*where*

$$\begin{aligned} P^c_{j,n,a}(x) &= \frac{1}{eg-1}\mathbb{P}\{B^r_{ki} > \frac{x}{c_n}, k=1,\ldots,K, i=1,\ldots,n_k; \\ &\qquad B^r_{ki} - \frac{x}{c_n} \geq \frac{1}{eg-1} g\left(B^r_{E_1} - \frac{x}{c_n}e\right), (k,i) \in E_0; \\ &\qquad \frac{1}{eg-1} g\left(B^r_{E_1} - \frac{x}{c_n}e\right) \leq x(a - \frac{1}{c_n})\}, \end{aligned}$$

*with $g = g^{c,j,n}$ as defined earlier.*
*In particular, we have*

$$P^c_{j,n,a}(x) \sim \kappa^c_{j,n,a} \prod_{k=1}^{K} \mathbb{P}\{B^r_k > \frac{x}{c_n}\}^{n_k},$$

*with $\kappa^c_{n,n,a} = 1$, and for $j \leq n$, $j \neq n$,*

$$\kappa^c_{j,n,a} = \frac{1}{eg-1}\mathbb{P}\{Z_{ki} \geq \frac{1}{eg-1} g Z_{E_1}, (k,i) \in E_0; (a - \frac{1}{c_n}) \geq \frac{1}{eg-1} g Z_{E_1}\}.$$

*The coefficient $\kappa^c_{j,n,a}$ is a continuous function of $c$ in a neighborhood of $c = 1-\rho$.*

The continuity property of the coefficient $\kappa^c_{j,n,a}$ follows immediately from its definition.

### 8.4.5 Proof of Theorem 8.3.1

We have now gathered all the ingredients for the proof of Theorem 8.3.1, which is restated below in extended form. Recall that $d_n = \sum_{k=1}^{K} n_k r_k + \rho - 1$.

**Theorem 8.3.1**
*Assume that $r^{\min} > 1-\rho$. Then,*

$$\mathbb{P}\{V > x\} \sim \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n}(x),$$

*where $j = (j_1, \ldots, j_K)$ and $P_{j,n}(x) := \lim_{a\to\infty} P^{1-\rho}_{j,n,a}(x)$ satisfies*

$$P_{j,n}(x) \sim \kappa_{j,n} \prod_{k=1}^{K} \mathbb{P}\{B^r_k > \frac{x}{d_n}\}^{n_k},$$

*for some constant $\kappa_{j,n} := \lim_{a\to\infty} \kappa_{j,n,a}$, with $\kappa_{j,n,a} := \kappa^{1-\rho}_{j,n,a}$, which is given by*

$$\kappa_{j,n} = \frac{1}{eg-1}\mathbb{P}\{Z_{ki} \geq \frac{1}{eg-1} g Z_{E_1}, (k,i) \in E_0\},$$

*with $g = g^{1-\rho,j,n}$ as defined earlier.*
*In particular, $\mathbb{P}\{V > x\}$ is regularly varying of index $-\mu^*$.*

**Proof**
For compactness, denote

$$P_a^c(x) := \sum_{n\in S_a^*} \sum_{j\le n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n,a}^c(x),$$

and

$$P(x) := \sum_{n\in S^*} \sum_{j\le n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n}(x).$$

We need to show that

$$\lim_{x\to\infty} \frac{\mathbb{P}\{V > x\}}{P(x)} = 1.$$

We may write, for any $a > 0$,

$$\frac{\mathbb{P}\{V > x\}}{P(x)} = \frac{\mathbb{P}\{V > x\}}{\mathbb{P}\{V(ax) > x\}} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)}.$$

Because of Theorem 8.4.1, it thus suffices to show that

$$\lim_{a\to\infty} \lim_{x\to\infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} = 1. \tag{4.8}$$

First observe that if $r^{\min} > 1-\rho$, then there exists an $a_0$ such that $S_a^* = S^*$ for all $a \ge a_0$. Also, combining Lemmas 8.4.1, 8.4.2, 8.4.4, we have that for $\delta$, $\theta$ sufficiently small,

$$\mathbb{P}\{V_{>\epsilon x}^{1-\rho\pm\delta}(ax) > (1\pm\theta)x\} \sim P_a^{1-\rho\pm\delta}((1\pm\theta)x). \tag{4.9}$$

The proof of (4.8) consists of a lower and an upper bound.

*Lower bound*
Using Proposition 8.4.2 and Equation (4.9), we obtain that for $\delta > 0$, $\theta > 0$ sufficiently small,

$$\mathbb{P}\{V(ax) > x\} \stackrel{>}{\sim} P_a^{1-\rho+\delta}((1+\theta)x).$$

Thus, for all $a \ge a_0$,

$$\frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \stackrel{>}{\sim} \frac{\sum_{n\in S^*} \sum_{j\le n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n,a}^{1-\rho+\delta}((1+\theta)x)}{\sum_{n\in S^*} \sum_{j\le n} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} P_{j,n}(x)} \ge \min_{n\in S^*, j\le n} \frac{P_{j,n,a}^{1-\rho+\delta}((1+\theta)x)}{P_{j,n}(x)}.$$

Letting $\theta \downarrow 0$, using the fact that $P^c_{j,n,a}(x)$ is regularly varying, we find

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \geq \min_{n\in S^*, j\leq n} \frac{\kappa^{1-\rho+\delta}_{j,n,a}}{\kappa_{j,n}}.$$

Letting $\delta \downarrow 0$, recalling that $\kappa^c_{j,n,a}$ is continuous in $c$ in a neighborhood of $1-\rho$, and then $a \to \infty$, the desired lower bound follows.

*Upper bound*
Using Proposition 8.4.4 and Equation (4.9), we obtain that for $\delta > 0$, $\theta > 0$ sufficiently small,

$$\mathbb{P}\{V(ax) > x\} \stackrel{<}{\sim} P^{1-\rho-\delta}_a((1-\theta)x).$$

Thus, for all $a \geq a_0$,

$$\frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \leq \frac{\sum_{n\in S^*}\sum_{j\leq n}\prod_{k=1}^{K}\frac{\bar{\rho}_k^{n_k}}{j_k!}P^{1-\rho-\delta}_{j,n,a}((1-\theta)x)}{\sum_{n\in S^*}\sum_{j\leq n}\prod_{k=1}^{K}\frac{\bar{\rho}_k^{n_k}}{j_k!}P_{j,n}(x)} \leq \max_{n\in S^*, j\leq n}\frac{P^{1-\rho-\delta}_{j,n,a}((1-\theta)x)}{P_{j,n}(x)}.$$

Letting $\theta \downarrow 0$, using the fact that $P^c_{j,n,a}(x)$ is regularly varying, we conclude

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \leq \max_{n\in S^*, j\leq n} \frac{\kappa^{1-\rho-\delta}_{j,n,a}}{\kappa_{j,n}}.$$

Letting $\delta \downarrow 0$, recalling that $\kappa^c_{j,n,a}$ is continuous in $c$ in a neighborhood of $1-\rho$, and then $a \to \infty$, the desired upper bound follows.

□

### 8.4.6 Transient workload asymptotics

Recall that the steady-state workload asymptotics were obtained from an analysis of the asymptotic behavior of $\mathbb{P}\{V(ax) > x\}$ for $x \to \infty$ after letting $a \to \infty$. This raises the question whether it is possible to obtain the exact asymptotics of $\mathbb{P}\{V(ax) > x\}$ for $x \to \infty$ for any value of $a$.
To answer this question, we first consider the case where $a$ is large enough for the condition $r^{\max}_a(1-\rho) < 1-\rho$ to hold, which implies that the overflow scenarios in the transient and steady-state case coincide.

**Theorem 8.4.2** *If $r^{\max}_a(1-\rho) < 1-\rho$, then*

$$\mathbb{P}\{V(ax) > x\} \sim P^{1-\rho}_a(x).$$

**Proof**
The proof is largely similar to that of Theorem 8.3.1 in the previous subsection, except that the use of Theorem 8.4.1 is not needed now.

□

Unfortunately, it seems difficult to remove the condition $r_a^{\max}(1-\rho) < 1-\rho$ in the above theorem. This condition is induced by the use of Lemma 8.4.2, where it is needed to ensure that the process $\{A^n_{>\epsilon x}(0,s) - cs\}$ reaches its supremum over the interval $[0, ax]$ at time $\bar{\tau}^n_f(\epsilon x)$.
This is no longer guaranteed to be the case when $r_a^{\max}(1-\rho) > 1-\rho$. In that case, the event $A^n_{>\epsilon x}(0, \bar{\tau}^n_f(\epsilon x)) - c\bar{\tau}^n_f(\epsilon x) > x$ is by far not necessary for the event $V^{c,n}_{>\epsilon x}(ax) > x$ to occur because the drift of the process $\{A^n_{>\epsilon x}(0,s) - cs\}$ may remain positive after time $\bar{\tau}^n_f(\epsilon x)$. This necessitates a detailed analysis of the process $\{A^n_{>\epsilon x}(0,s) - cs\}$ after time $\bar{\tau}^n_f(\epsilon x)$, which seems rather difficult, even in the single-class case $K = 1$.
Nevertheless, it is possible to apply the earlier results to obtain asymptotically tight lower and upper bounds for $\mathbb{P}\{V(ax) > x\}$ as $x \to \infty$, which hold for any value of $a$, under the considerably milder condition $r_a^{\min} > 1-\rho$.

**Theorem 8.4.3** *Assume that* $r_a^{\min} > 1-\rho$. *Then,*

$$P_a^{1-\rho}(x) \lesssim \mathbb{P}\{V(ax) > x\} \lesssim P_a^{1-\rho}(x(1 - a(r_a^{\max} + \rho - 1))).$$

**Proof**
The lower bound follows directly from Lemmas 8.4.2 and 8.4.4 and the results of Subsection 8.4.1. For the upper bound, note that the drift of the process $\{A^n_{>\epsilon x}(0,s) - cs\}$ is at most $r_a^{\max} - c$ after time $\bar{\tau}^n_f(\epsilon x)$. Hence, this process can increase by at most $a(r_a^{\max} - c)x$ until time $ax$. This implies that one must have $A^n_{>\epsilon x}(0, \bar{\tau}^n_f(\epsilon x)) - c\bar{\tau}^n_f(\epsilon x) > x(1 - a(r_a^{\max} - c))$ in order for the event $V^{c,n}_{>\epsilon x}(ax) > x$ to occur. The proof of the upper bound is then completed by using Lemma 8.4.4.

□

Note that the upper bound in the above theorem is non-trivial because $r_a^{\max} + \rho - 1 < \frac{1}{a}$. Moreover, the bounds asymptotically coincide up to a constant factor, since the function $P_a^{1-\rho}(\cdot)$ is regularly varying of index $-\mu_a^*$.

## 8.5 Proof of Theorem 8.4.1

In this section we provide the proof of Theorem 8.4.1. We first collect some preparatory results. For conciseness, we drop $a = \infty$ from the previously introduced notation to

denote steady-state quantities. For example $S^*(c) := S^*_\infty(c)$ is the set of optimal solutions of the linear program formulated at the beginning of Section 8.4, $r^{\min}(c) := r^{\min}_\infty(c) = \min_{n \in S^*(c)} \sum_{k=1}^K n_k r_k$, $S^-(c) := S^-_\infty(c) = \{n \in \mathbb{N}^K : \sum_{k=1}^K n_k r_k < c\}$, $S^+(c) := S^+_\infty(c) = \{n \in \mathbb{N}^K : \sum_{k=1}^K n_k r_k \geq c\}$, and $r^{\max}(c) := r^{\max}_\infty(c) = \max_{n \in S^-(c)} \sum_{k=1}^K n_k r_k < c$.

**Proposition 8.5.1** *Assume that* $r^{\min}(c) > c$.
*Then for all* $\epsilon < \frac{1}{r^{\min}(c)-c}$,

$$\mathbb{P}\{V^c_{>\epsilon x} > x\} \geq \sum_{n \in S^*(c)} \prod_{k=1}^K \mathrm{e}^{-\bar\rho_k} \frac{\bar\rho_k^{n_k}}{n_k!} (\mathbb{P}\{B^r_k > \frac{x}{r^{\min}(c) - c}\})^{n_k}.$$

**Proof**
Consider the event that at some arbitrary time $t$ there are exactly $n_k$ active class-$k$ sessions, $k = 1, \ldots, K$, $n \in S^*(c)$, which all started before time $t - \frac{x}{r^{\min}(c)-c}$.
Since $\epsilon < \frac{1}{r^{\min}(c)-c}$, this event implies that $V^c_{>\epsilon x}(t)$ is larger than

$$(\sum_{k=1}^K n_k r_k - c) \frac{x}{r^{\min}(c) - c} \geq x,$$

while it occurs with probability

$$\prod_{k=1}^K \mathrm{e}^{-\bar\rho_k} \frac{\bar\rho_k^{n_k}}{n_k!} (\mathbb{P}\{B^r_k > \frac{x}{r^{\min}(c) - c}\})^{n_k}.$$

□

**Proposition 8.5.2** *Consider a queue of capacity c fed by a process which generates traffic at rate* $r_n$ *for a fraction of the time* $p_n$, $n = 1, \ldots, N$ *(possibly* $N = \infty$*). Assume* $r_1 \leq r_2 \leq \ldots r_{K-1} < c \leq r_K \leq \ldots \leq r_N$, *and* $\sum_{n=1}^N p_n r_n < c$ *for stability. Let* $V^c$ *be the stationary workload. Then for any* $x > 0$

$$\mathbb{P}\{V^c > 0\} \leq \frac{1}{c - r_{K-1}} \sum_{n=K}^N p_n (r_n - r_{K-1}).$$

**Proof**
First observe that $\mathbb{P}\{V^c > 0\} \leq \pi_{>0}$, where the latter quantity represents the stationary probability that the workload is non-zero if the rate $r_n$ were increased to $r_{K-1}$ for all $n = 1, \ldots, K-1$.

From a simple balance argument, noting that the workload cannot be zero when traffic is generated at a rate $p_n > c$,

$$\sum_{n=K}^{N} p_n(r_n - c) = (\pi - \sum_{n=K}^{N} p_n)(c - r_{K-1}),$$

yielding

$$\pi_{>0} = \frac{1}{c - r_{K-1}} \sum_{n=K}^{N} p_n(r_n - r_{K-1}),$$

which completes the proof.

□

**Proposition 8.5.3** *For each $\epsilon > 0$ there exists a finite $M_\epsilon$ such that*

$$\mathbb{P}\{V^c_{>\epsilon x} > 0\} \lesssim \frac{\max_{k=1,\ldots,K} r_k}{c - r^{\max}(c)} \sum_{n \in S^*(c)} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{n_k}.$$

**Proof**

Since $B_k$ is regularly varying, it is possible to construct a finite constant $M_\epsilon$ such that $\bar{\rho}_{k,>\epsilon x} \leq \bar{\rho}_k M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\}$. Using Proposition 8.5.2 (noting that $p_n = \prod_{k=1}^{K} \mathrm{e}^{-\bar{\rho}_{k,>\epsilon x}} \frac{\bar{\rho}_{k,>\epsilon x}^{n_k}}{n_k!}$),

$$\begin{aligned}
& \mathbb{P}\{V^c_{>\epsilon x} > 0\} \\
\leq\ & \frac{1}{c - r^{\max}(c)} \sum_{n \in S^+(c)} (\sum_{k=1}^{K} n_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \mathrm{e}^{-\bar{\rho}_{k,>\epsilon x}} \frac{\bar{\rho}_{k,>\epsilon x}^{n_k}}{n_k!} \\
\leq\ & \frac{1}{c - r^{\max}(c)} \sum_{n \in S^+(c)} (\sum_{k=1}^{K} n_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{n_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{n_k} \\
=\ & \frac{1}{c - r^{\max}(c)} \sum_{n \in S^*(c)} (\sum_{k=1}^{K} n_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{n_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{n_k} \\
+\ & \frac{1}{c - r^{\max}(c)} \sum_{m \in S^+(c) \setminus S^*(c)} (\sum_{k=1}^{K} m_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \frac{\bar{\rho}_k^{m_k}}{j_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{m_k}.
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{\sum_{n \in S^*(c)} (\sum_{k=1}^{K} n_k r_k - r^{\max}(c)) \frac{\bar{\rho}_k^{n_k}}{n_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{n_k}}{\sum_{n \in S^*(c)} \frac{\bar{\rho}_k^{n_k}}{n_k!} (M_\epsilon \mathbb{P}\{B^r_k > \epsilon x\})^{n_k}} &\leq \max_{n \in S^*(c)} \sum_{k=1}^{K} n_k r_k - r^{\max}(c) \\
&\leq \max_{k=1,\ldots,K} r_k.
\end{aligned}$$

From the definition of $S^*(c)$ it follows that there exists an $x_0$ such that for all $x \geq x_0$,

$$\prod_{k=1}^{K} \mathbb{P}\{B_k^r > \epsilon x\})^{m_k} \leq H(x) \prod_{k=1}^{K} \frac{(M_\epsilon \bar{\rho}_k)^{n_k}}{n_k!} \mathbb{P}\{B_k^r > \epsilon x\})^{n_k},$$

for all $m \in S^+(c) \setminus S^*(c)$, $n \in S^*(c)$, with $H(x) = \mathrm{o}(1)$ as $x \to \infty$, so that

$$\begin{aligned}
& \frac{\sum_{m \in S^+(c) \setminus S^*(c)} (\sum_{k=1}^{K} m_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \frac{(M_\epsilon \bar{\rho}_k)^{m_k}}{m_k!} (\mathbb{P}\{B_k^r > \epsilon x\})^{m_k}}{\sum_{n \in S^*(c)} \frac{(M_\epsilon \bar{\rho}_k)^{n_k}}{n_k!} (\mathbb{P}\{B_k^r > \epsilon x\})^{n_k}} \\
\leq \; & H(x) \sum_{m \in S^+(c) \setminus S^*(c)} (\sum_{k=1}^{K} m_k r_k - r^{\max}(c)) \prod_{k=1}^{K} \frac{(M_\epsilon \bar{\rho}_k)^{m_k}}{m_k!} \\
\leq \; & H(x) \sum_{m \geq 0} (\sum_{k=1}^{K} m_k r_k) \prod_{k=1}^{K} \frac{(M_\epsilon \bar{\rho}_k)^{m_k}}{m_k!} \\
= \; & H(x) (\sum_{k=1}^{K} \bar{\rho}_k r_k) \mathrm{e}^{M_\epsilon \sum_{k=1}^{K} \bar{\rho}_k} = \rho H(x) \mathrm{e}^{M_\epsilon \sum_{k=1}^{K} \bar{\rho}_k}.
\end{aligned}$$

Hence,

$$\limsup_{x \to \infty} \frac{\mathbb{P}\{V_{>\epsilon x}^c > 0\}}{\sum_{n \in S^*(c)} \prod_{k=1}^{K} \frac{\bar{\rho}_k^{n_k}}{j_k!} (M_\epsilon \mathbb{P}\{B_k^r > \epsilon x\})^{n_k}} \leq \frac{\max_{k=1,\ldots,K} r_k}{c - r^{\max}(c)}.$$

□

We have now gathered all the ingredients for the proof of Theorem 8.4.1 which is repeated below.

**Theorem 8.4.1**
If $r^{\min} > 1 - \rho$, then

$$\lim_{a \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}\{V(ax) > x\}}{\mathbb{P}\{V > x\}} = 1.$$

**Proof**
By definition,

$$\begin{aligned}
\mathbb{P}\{V > x\} &= \mathbb{P}\{\sup_{t \geq 0}\{A(0,t) - t\} > x\} \\
&\leq \mathbb{P}\{\sup_{t \leq ax}\{A(0,t) - t\} > x\} + \mathbb{P}\{\sup_{t \geq ax}\{A(0,t) - t\} > x\} \\
&= \mathbb{P}\{V(ax) > x\} + \mathbb{P}\{\sup_{t \geq ax}\{A(0,t) - t\} > x\}.
\end{aligned}$$

Thus, it suffices to show that

$$\lim_{a\to\infty} \limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq ax}\{A(0,t)-t\} > x\}}{\mathbb{P}\{V>x\}} = 0.$$

For $t \geq ax$, write

$$A(0,t) - t = A(0,ax) - ax + A(ax,t) - (t-ax),$$

and observe that $A(ax,t) \stackrel{d}{=} A(0,t-ax)$ since the process $A(0,t)$ has stationary increments. Thus, for $\delta > 0$ sufficiently small,

$$\begin{aligned}
& \mathbb{P}\{\sup_{t\geq ax}\{A(0,t)-t\} > 0\} \\
= \; & \mathbb{P}\{\sup_{t\geq ax}\{A(0,ax) - ax + A(ax,t) - (t-ax)\} > 0\} \\
= \; & \mathbb{P}\{A(0,ax) - ax + \sup_{t\geq ax}\{A(ax,t) - (t-ax)\} > 0\} \\
\leq \; & \mathbb{P}\{A(0,ax) - ax > -\delta ax\} + \mathbb{P}\{\sup_{t\geq ax}\{A(0,t-ax) - (t-ax)\} > \delta ax\} \\
= \; & \mathbb{P}\{A(0,ax) - (1-2\delta)ax > \delta ax\} + \mathbb{P}\{\sup_{t\geq ax}\{A(0,t-ax) - (t-ax)\} > \delta ax\} \\
\leq \; & \mathbb{P}\{\sup_{t\geq 0}\{A(0,t) - (1-2\delta)t\} > \delta ax\} + \mathbb{P}\{\sup_{t\geq 0}\{A(0,t) - t\} > \delta ax\} \\
= \; & \mathbb{P}\{V^{1-2\delta} > \delta ax\} + \mathbb{P}\{V > \delta ax\} \\
\leq \; & 2\mathbb{P}\{V^{1-2\delta} > \delta ax\}.
\end{aligned}$$

Hence, using Propositions 8.4.2, 8.4.5, for $\theta > 0$ sufficiently small,

$$\begin{aligned}
\limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq ax}\{A(0,t)-t\} > x\}}{\mathbb{P}\{V>x\}} & \leq 2\limsup_{x\to\infty} \frac{\mathbb{P}\{V^{1-2\delta} > \delta ax\}}{\mathbb{P}\{V>x\}} \\
& \leq 2\limsup_{x\to\infty} \frac{\mathbb{P}\{V_{>\epsilon x}^{1-\rho-3\delta} > (1-\theta)\delta ax\}}{\mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta} > (1+\theta)x\}} \\
& \leq 2\limsup_{x\to\infty} \frac{\mathbb{P}\{V_{>\epsilon x}^{1-\rho-3\delta} > (1-\theta)\delta x/(1+\theta)\}}{\mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta} > x/a\}} \\
& \leq 2\limsup_{x\to\infty} \frac{\mathbb{P}\{V_{>\epsilon x}^{1-\rho-3\delta} > 0\}}{\mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta} > x/a\}}.
\end{aligned}$$

The assumption that $r^{\min} > 1-\rho$ ensures that there exists a $\delta^*$ such that $r_{\min} > 1-\rho+\delta^*$, $r_{\max} < 1-\rho-3\delta^*$, and $S^*(1-\rho-3\delta^*) = S^*(1-\rho+\delta^*) = S^*$.
Using Propositions 8.5.1, 8.5.3, we then find that there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq ax}\{A(0,t)-t\} > x\}}{\mathbb{P}\{V>x\}} \leq$$

$$2\limsup_{x\to\infty}\frac{\max\limits_{k=1,\ldots,K} r_k}{1-\rho-r^{\max}-3\delta^*}\frac{\sum\limits_{n\in S^*}\prod\limits_{k=1}^{K}\frac{\bar\rho_k^{n_k}}{n_k!}(M_\epsilon\mathbb{P}\{B_k^r>\epsilon x\})^{n_k}}{\sum\limits_{n\in S^*}\prod\limits_{k=1}^{K}\mathrm{e}^{-\bar\rho_k}\frac{\bar\rho_k^{n_k}}{n_k!}(\mathbb{P}\{B_k^r>\frac{x}{a(r^{\min}+\rho-1-\delta^*)}\})^{n_k}}\leq$$

$$2\frac{\max\limits_{k=1,\ldots,K} r_k}{1-\rho-r^{\max}-3\delta^*}\mathrm{e}^{\sum\limits_{k=1}^{K}\bar\rho_k}\max_{n\in S^*}\limsup_{x\to\infty}\prod_{k=1}^{K}\left(\frac{M_\epsilon\mathbb{P}\{B_k^r>\epsilon x\}}{\bar\rho^{n_k}\mathbb{P}\{B_k^r>\frac{x}{a(r_{\min}+\rho-1-\delta^*)}\}}\right)^{n_k}.$$

Now first let $x\to\infty$ and then $a\to\infty$ (use the fact that $\mathbb{P}\{B_k^r>x\}$ is of regular variation). □

## 8.6 Most probable time to overflow

As a direct application of the workload asymptotics which we derived in the previous sections, we now establish a conditional limit theorem for the most probable time to overflow, given that the process $\{A(0,t)-ct\}$ reaches a large level $x$. Define $\tau(x)=\inf\{t\geq 0: A(0,t)-ct>x\}$. Note that $V\geq x$ iff $\tau(x)<\infty$. We will give an expression for the asymptotic distribution of $\tau(x)$ conditional upon $\tau(x)<\infty$ for $x\to\infty$. Define the probability measure $\mathbb{P}_x\{\cdot\}:=\mathbb{P}\{\cdot\mid\tau(x)<\infty\}$. In this section we compute the limiting $\mathbb{P}_x$-distribution of $\frac{\tau(x)}{x}$ for $x\to\infty$.

A similar problem has been investigated by Asmussen & Klüppelberg [21] for random walks and Lévy processes with negative drift and heavy-tailed jumps. As has been shown in [21], this class of processes allows for a general subexponential jump size distribution. Here though, like in the rest of the chapter, we consider the case of regular variation. In fact, since slowly varying functions may be difficult to compare in the multi-class case, we assume that the session lengths are Pareto distributed, i.e.,

$$\mathbb{P}\{B_k^r>x\}\sim\gamma_k x^{1-\nu_k},\qquad k=1,\ldots,K.$$

This assumption may be weakened, as will be discussed below.

In order to state the result, we need to introduce some additional notation. For given $a$, define the set $S_a$ as $S_a:=\{n\in S^*:\sum\limits_{k=1}^{K}r_kn_k\geq 1-\rho+\frac{1}{a}\}$. We will also make extensive use of the coefficients $\kappa_{j,n}$ and $\kappa_{j,n,a}$ defined earlier. The definition of $\kappa_{j,n,a}$ as given in Subsection 8.4.4 only makes sense for $\sum\limits_{k=1}^{K}r_kn_k>1-\rho+\frac{1}{a}$. If $\sum\limits_{k=1}^{K}r_kn_k=1-\rho+\frac{1}{a}$, we define $\kappa_{j,n,a}=1_{\{j=n\}}$.

**Theorem 8.6.1** *The quantity $\frac{\tau(x)}{x}$ converges in $\mathbb{P}_x$-distribution for $x\to\infty$ to a random*

*variable $Y$, which has distribution function*

$$G(a) := \mathbb{P}\{Y \le a\} = \frac{\sum_{n\in S_a}\sum_{j\le n} d_n^{\mu^*}\kappa_{j,n,a}\prod_{k=1}^{K}\frac{(\bar\rho_k\gamma_k)^{n_k}}{j_k!}}{\sum_{n\in S^*}\sum_{j\le n} d_n^{\mu^*}\kappa_{j,n}\prod_{k=1}^{K}\frac{(\bar\rho_k\gamma_k)^{n_k}}{j_k!}},$$

*with* $d_n = \sum_{k=1}^{K} n_k r_k + \rho - 1$ *as before.*

**Proof**

First observe that the extended definition of $\kappa_{j,n,a}$ ensures that $\kappa_{j,n,a}$ is right-continuous in $a$ if $a$ is such that $\sum_{k=1}^{K} r_k n_k = 1 - \rho + \frac{1}{a}$. This then implies that the function $G(\cdot)$ is right-continuous. From the analysis in the previous sections, it follows that $G(\cdot)$ is non-decreasing and that $G(a) \to 1$ as $a \to \infty$. Hence, $G(\cdot)$ is a proper distribution function, so that $Y$ is a well-defined random variable.

We need to show that $\mathbb{P}_x\{\tau(x) < ax\} \to G(a)$ as $x \to \infty$ for each continuity point of $G(\cdot)$. Using the definition of $S_a$ and the (extended) definition of $\kappa_{j,n,a}$, it is easy to see that $G(\cdot)$ is continuous in $a$ iff $\sum_{k=1}^{K} r_k n_k > 1 - \rho + \frac{1}{a}$ for all $n \in S_a$ (look at the structure of $S_a$).

Hence, we may assume that $a$ is such that $\sum_{k=1}^{K} r_k n_k > 1 - \rho + \frac{1}{a}$ for all $n \in S_a$.

Now write

$$\mathbb{P}\{\tau(x) \le ax \mid \tau(x) < \infty\} = \frac{\mathbb{P}\{\tau(x) \le ax\}}{\mathbb{P}\{\tau(x) < \infty\}} = \frac{\mathbb{P}\{V(ax) \ge x\}}{\mathbb{P}\{V \ge x\}} \sim \frac{\mathbb{P}\{V(ax) > x\}}{\mathbb{P}\{V > x\}}.$$

Note that $\mathbb{P}\{V > x\}$ is regularly varying of index $-\mu^*$. If $\sum_{k=1}^{K} r_k n_k < 1 - \rho + \frac{1}{a}$ for all $n \in S^*$ (i.e. $S_a = \emptyset$), then it is obvious that $\mathbb{P}\{V(ax) > x\}$ is regularly varying of index $-\mu_a^* < -\mu^*$. This implies that $\mathbb{P}\{V(ax) > x\}/\mathbb{P}\{V > x\} \to 0$ if $a$ is small enough for $S_a$ to be empty.

Now suppose that $a$ is large enough such that $S_a$ is non-empty. It is then easy to see that $S_a = S_a^*$. If we combine this identity with Theorems 8.3.1 and 8.4.2, we find, noting that $\sum_{k=1}^{K}(\nu_k - 1) = \mu^*$ for all $n \in S^*$,

$$\begin{aligned}\mathbb{P}_x\{\tau(x) \le ax\} &= \frac{\mathbb{P}\{V(ax) \ge x\}}{\mathbb{P}\{V > x\}} \\ &\sim \frac{\sum_{n\in S_a}\sum_{j\le n}\kappa_{j,n,a}\prod_{k=1}^{K}\frac{\bar\rho_k^{n_k}}{j_k!}(\mathbb{P}\{B_k^r > \frac{x}{d_n}\})^{n_k}}{\sum_{n\in S^*}\sum_{j\le n}\kappa_{j,n}\prod_{k=1}^{K}\frac{\bar\rho_k^{n_k}}{j_k!}(\mathbb{P}\{B_k^r > \frac{x}{d_n}\})^{n_k}}\end{aligned}$$

$$\sim \frac{\sum_{n\in S_a}\sum_{j\le n}\kappa_{j,n,a}\prod_{k=1}^{K}\frac{\bar{\rho}_k^{n_k}}{j_k!}\gamma_k^{n_k}\left(\frac{x}{d_n}\right)^{-n_k(\nu_k-1)}}{\sum_{n\in S^*}\sum_{j\le n}\kappa_{j,n}\prod_{k=1}^{K}\frac{\bar{\rho}_k^{n_k}}{j_k!}\gamma_k^{n_k}\left(\frac{x}{d_n}\right)^{-n_k(\nu_k-1)}}$$

$$\sim \frac{\sum_{n\in S_a} d_n^{\mu^*}\sum_{j\le n}\kappa_{j,n,a}\prod_{k=1}^{K}\frac{(\bar{\rho}_k\gamma_k)^{n_k}}{j_k!}}{\sum_{n\in S^*} d_n^{\mu^*}\sum_{j\le n}\kappa_{j,n}\prod_{k=1}^{K}\frac{(\bar{\rho}_k\gamma_k)^{n_k}}{j_k!}}.$$

□

If the set $S^*$ is a singleton, then it is easy to see that regular variation suffices in the last two lines of the above proof. In particular, this is true in the single-class case $K = 1$.

We conclude the section with the most basic single-class scenario where overflow is caused by a single long session, which occurs when $r > 1 - \rho$. In this case, the distribution of $Y$ takes the explicit form

$$\mathbb{P}\{Y \le a\} = \frac{1-\rho}{r} + \left(1 - \frac{1-\rho}{r}\right)\mathbb{P}\{\frac{r}{1-\rho}Z \le a - \frac{1}{r-(1-\rho)}\},$$

where $\mathbb{P}\{Z > a\} = (1 + (r - (1-\rho))a)^{1-\nu}$. This expression reduces to the results for the case of compound Poisson input in [21] when we let $r \to \infty$. The results in [21] further include conditional limit theorems for the behavior of the process $\{A(0,t) - ct\}$ up to time $\tau(x)$. It should be possible to derive similar results for the case of $M/G/\infty$ input considered here as well.

# Bibliography


[1] Abate, J., Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5–88.

[2] Abate, J., Choudhury, G.L., Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems* **16**, 311–338.

[3] Abate, J., Choudhury, G.L., Whitt, W. (1995). Exponential approximations for tail probabilities in queues. I. Waiting times. *Operations Research* **43**, 885–901.

[4] Abate, J., Whitt, W. (1997). Asymptotics for $M/G/1$ low-priority waiting-time tail probabilities. *Queueing Systems* **25**, 173–233.

[5] Abate, J., Whitt, W. (1997). Limits and approximations for the $M/G/1$ LIFO waiting-time distribution. *Operations Research Letters* **20**, 199–206.

[6] Abate, J., Whitt, W. (1999). Explicit M/G/1 waiting-time distributions for a class of long-tail service-time distributions. *Operations Research Letters* **25**, 25–31.

[7] Abate, J., Whitt, W. (1999). Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS Journal on Computing* **11**, 394–405.

[8] Abate, J., Whitt, W. (1999). Infinite-series representations of Laplace transforms of probability density functions for numerical inversion. *Journal of the Operations Research Society of Japan* **42**, 268–285.

[9] Abramovitz, M., Stegun, I. (1965). *Handbook of Mathematical Functions.* Dover, New York.

[10] Abry, P., Flandrin, P., Taqqu, M.S., Veitch, D. (2000). Wavelets for the analysis, estimation, and synthesis of scaling data. In: Park, K., Willinger, W. (editors) *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 39–88.

[11] Adler, R., Feldman, R., Taqqu, M.S. (editors) (1998). *A Practical Guide to Heavy Tails.* Birkhäuser, Boston.

[12] Agrawal, R., Makowski, A.M., Nain, Ph. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33**, 5–41.

[13] Anantharam, V. (1988). How large delays build up in a GI/G/1 queue. *Queueing Systems* **5**, 345–368.

[14] Anantharam, V. (1996). Networks of queues with long-range dependent traffic streams. In: Glasserman, P., Sigman, K., Yao, D.D. (editors). *Stochastic Networks: Stability and Rare Events.* Springer, New York, 237–256.

[15] Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems* **33**, 73–89.

[16] Anick, D., Mitra, D., Sondhi, M.M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical Journal* **61**, 1871–1894.

[17] Arvidsson, A., Karlsson, P. (1999). On traffic models for TCP/IP. In: Key, P., Smith, D. (editors). *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, North-Holland, Amsterdam, 457–466.

[18] Asmussen, S. (1982). Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue. *Advances in Applied Probability* **14**, 143–170.

[19] Asmussen, S. (1987). *Applied Probability and Queues.* Wiley, Chichester.

[20] Asmussen, S. (1994). Busy period analysis, rare events and transient behavior in fluid flow models. *Journal of Applied Mathematics and Stochastic Analysis* **7**, 269–299.

[21] Asmussen, S., Klüppelberg, C. (1996). Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic Processes and their Applications* **64**, 103–125.

[22] Asmussen, S., Klüppelberg, C. (1996). Stationary $M/G/1$ excursions in the presence of heavy tails. *Journal of Applied Probability* **33**, 208–212.

[23] Asmussen, S., Teugels, J. (1996). Convergence rates for $M/G/1$ queues and ruin problems with heavy tails. *Journal of Applied Probability* **33**, 1181–1190.

[24] Asmussen, S. (1997). Subexponential asymptotics for stochastic processes: extremal behaviour, stationary distributions and first passage times. *Annals of Applied Probability* **8**, 354–374.

[25] Asmussen, S., Möller, J. (1999). Tail asymptotics for M/G/1 type queueing processes with subexponential increments. *Queueing Systems* **33**, 153–176.

[26] Asmussen, S., Collamore, J. (1999). Exact asymptotics for a large deviations problem in the GI/GI/1 queue. *Markov Processes and Related Fields* **5**, 451–476.

[27] Asmussen, S., Klüppelberg, C., Sigman, K. (1999). Sampling at subexponential times, with queueing applications. *Stochastic Processes and their Applications* **79**, 265–286.

[28] Asmussen, S., Schmidli, H., Schmidt, V. (1999). Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Advances in Applied Probability* **31**, 422–447.

[29] Asmussen, S. (2000) *Ruin Probabilities.* World Scientific, Singapore.

[30] Asmussen, S., Binswanger, K., Hojgaard, B. (2000). Rare events simulation for heavy-tailed distributions. *Bernoulli* **6**, 303–322.

[31] Athreya, K.B., Ney, P.E. (1972). *Branching Processes.* Springer-Verlag, Berlin.

[32] Awater, G.A. (1994). *Broadband Communication – Modeling, Analysis and Synthesis of an ATM Switching Element,* Ph.D thesis, Delft University.

[33] Baccelli, F., Schlegel, S., Schmidt, V. (1999). Asymptotics of stochastic networks with subexponential service times. *Queueing Systems* **33**, 205–232.

[34] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios-Gomez, F. (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260.

[35] Baltrunas, A. (2001). Some asymptotic results for transient random walks with applications to insurance risk. *Journal of Applied Probability* **38**, 108–121.

[36] Beran, J. (1994). *Statistics for Long-Memory Processes.* Chapman & Hall, New York.

[37] Beran, J., Sherman, R., Taqqu, M.S., Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions in Communications* **43**, 1566–1579.

[38] van den Berg, J.L. (1990). *Sojourn Times in Feedback and Processor Sharing Queues.* Ph.D Thesis, Utrecht University.

[39] Bertoin, J., Doney, R.A. (1994). On conditioning a random walk to stay nonnegative. *Annals of Probability* **22**, 2152–2167.

[40] Bertoin, J., Doney, R.A. (1996). Some asymptotic results for transient random walks. *Advances in Applied Probability* **28**, 207–226.

[41] Billingsley, P. (1996). *Probability and Measure.* Wiley, New York.

[42] Bingham, N.H., Doney, R.A. (1974). Asymptotic properties of supercritical branching processes I: The Galton-Watson process. *Advances in Applied Probability* **6**, 711–731.

[43] Bingham, N.H., Doney, R.A. (1975). Asymptotic properties of supercritical branching processes II: Crump-Mode and Jirina processes. *Advances in Applied Probability* **7**, 66–82.

[44] Bingham, N.H., Goldie, C., Teugels, J. (1987). *Regular Variation.* Cambridge University Press, Cambridge, UK.

[45] Boots, N.K., Tijms, H.C. (1998). A multi-server queueing system with impatient customers. *Management Science* **45**, 444–448.

[46] Boots, N.K., Shahabuddin, P. (2000). Simulating $GI/GI/1$ queues and insurance risk processes with subexponential distributions. Unpublished manuscript, Free University, Amsterdam. Shortened version in: *Proceedings of the 2000 Winter Simulation Conference.*

[47] Boots, N.K., Shahabuddin, P. (2001). A framework for efficiently simulating small ruin probabilities in insurance risk processes with subexponential claims. Unpublished manuscript, Free University, Amsterdam.

[48] Borovkov, A.A. (1964). Some limit theorems in the theory of mass service, I. *Theory of Probability and its Applications* **9**, 550–565.

[49] Borovkov, A.A. (1970). Factorization identities and properties of the distribution of the supremum of sequential sums. *Theory of Probability and its Applications* **15**, 359–402.

[50] Borovkov, A.A. (1976). *Stochastic Processes in Queueing Theory.* Springer, New York.

[51] Borovkov, A.A. (1984). *Asymptotic Methods in Queueing Theory.* Wiley, New York.

[52] Borovkov, A.A. (1998). *Ergodicity and Stability of Stochastic Processes.* Wiley, Chichester.

[53] Borovkov, A.A., Borovkov, K. (1999). On large deviation probabilities for random walks. Preprint no. 62, Sobolev Institute of Mathematics, Novosibirsk.

[54] Borst, S.C., Boxma, O.J., Jelenković, P.R. (1999). Generalized processor sharing with long-tailed traffic sources. In: Key, P., Smith, D. (editors). *Teletraffic Engineering in a Competitive World, Proc. ITC-16.* North-Holland, Amsterdam, 345–354.

[55] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Coupled processors with regularly varying service times. In: *Proceedings of Infocom 2000,* Tel-Aviv, Israel, 157–164.

[56] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Asymptotic behavior of generalized processor sharing with long-tailed traffic sources. In: *Proceedings of Infocom 2000,* Tel-Aviv, Israel, 912–921.

[57] Borst, S.C. Boxma, O.J., Jelenković, P.R. (2000). Induced burstiness in generalized processor sharing with long-tailed traffic flows. In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing.*

[58] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. Report PNA-R0016, CWI, Amsterdam.

[59] Borst, S.C., Zwart, A.P. (2000). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows. SPOR-Report 2000-04, Eindhoven University of Technology.

[60] Borst, S.C., Zwart, A.P. (2001). Fluid queues with heavy-tailed $M/G/\infty$ input. SPOR-Report 2001-02, Eindhoven University of Technology.

[61] Borst, S.C., Boxma, O.J., Núñez-Queija, R. (2001). Personal communication.

[62] Botvich, D.D., Duffield, N.G. (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* **20**, 293–320.

[63] Boxma, O.J. (1978). On the longest service time in a busy period of the $M/G/1$ queue. *Stochastic Processes and their Applications* **8**, 93–100.

[64] Boxma, O.J. (1980). The longest service time in a busy period. *Zeitschrift für Operations Research* **24**, 235–242.

[65] Boxma, O.J. (1996). Fluid queues and regular variation. *Performance Evaluation* **27 & 28**, 699–712.

[66] Boxma, O.J. (1997). Regular variation in a multi-source fluid queue. In: Ramaswami, V., Wirth, P. (editors). *Teletraffic Contributions for the Information Age, Proc. ITC-15.* North-Holland, Amsterdam, 391–402.

[67] Boxma, O.J., Dumas, V. (1998). Fluid queues with heavy-tailed activity period distributions. *Computer Communications* **21**, 1509–1529.

[68] Boxma, O.J., Dumas, V. (1998). The busy period in the fluid queue. *Performance Evaluation Review* **26**, 100–110.

[69] Boxma, O.J., Cohen, J.W. (1998). The M/G/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications* **16**, 349–363.

[70] Boxma, O.J., Deng, Q., Zwart, A.P. (1999). Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers. Technical Memorandum COSOR 99-20, Eindhoven University of Technology. *Queueing Systems,* to appear.

[71] Boxma, O.J., Cohen, J.W. (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems* **33**, 177–204.

[72] Boxma, O.J., Cohen, J.W., Deng, Q. (1999). Heavy-traffic analysis of the $M/G/1$ queue with priority classes. In: Key, P., Smith, D. (editors). *Teletraffic Engineering in a Competitive World, Proc. ITC-16.* North-Holland, Amsterdam, 1157–1167.

[73] Boxma, O.J., Kurkova, I. (1999). The $M/G/1$ queue with two speeds of service. Technical Report 99-057, EURANDOM, Eindhoven. *Advances in Applied Probability,* to appear.

[74] Boxma, O.J., Kurkova, I. (2000). The $M/M/1$ queue in a heavy-tailed random environment. *Statistica Neerlandica* **54**, 221–236.

[75] Boxma, O.J., Cohen, J.W. (2000). The single server queue: heavy tails and heavy traffic. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 143–170.

[76] Boxma, O.J., Deng, Q. (2000). Asymptotic behaviour of the tandem queueing system with identical service times at both queues. *Mathematical Methods of Operations Research* **52**, 307–323.

[77] Boxma, O.J, Deng, Q., Resing, J.A.C. (2000). Polling systems with regularly varying service and/or switchover times. *Advances in Performance Analysis* **3**, 71–107.

[78] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications* **10**, 323–331.

[79] Brichet, F., Roberts, J., Simonian, A., Veitch, D. (1996). Heavy traffic analysis of a storage model with long-range dependent On-Off sources. *Queueing Systems* **23**, 197–215.

[80] Brockwell, P.J., Davis, R.A. (1991). *Time Series: Theory and Methods*. Springer, New York.

[81] Choe, J., Shroff, N. (1999). On the supremum distribution of integrated stationary Gaussian processes with negative linear drift. *Advances in Applied Probability* **31**, 135–157.

[82] Chistyakov, V.P. (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.

[83] Choudhury, G. L., Whitt, W. (1997). Long-tail buffer-content distributions in broadband networks. *Performance Evaluation* **30**, 177–190.

[84] Chover, J., Ney, P., Wainger, S. (1973). Functions of probability measures. *Journal d'Analyse Mathématique* **26**, 255–302.

[85] Cline, D. (1986). Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields* **72**, 529–557.

[86] Cline, D. (1994). Intermediate regular and $\Pi$ variation. *Proceedings of the London Mathematical Society* **68**, 594–616.

[87] Cline, D., Samorodnitsky, G. (1994). Subexponentiality of the product of two random variables. *Stochastic Processes and their Applications* **49**, 75–98.

[88] Coffman, E.G., Muntz, R.R., Trotter, H. (1970). Waiting-time distributions for processor-sharing systems. *Journal of ACM* **17**, 123–130.

[89] Cohen, J.W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability* **10**, 343–353.

[90] Cohen, J.W. (1973). Asymptotic relations in queueing theory. *Stochastic Processes and their Applications* **1**, 107–124.

[91] Cohen, J.W. (1974). Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications* **2**, 31–58.

[92] Cohen, J.W. (1976). *Regenerative Processes in Queueing Theory.* Springer, Berlin.

[93] Cohen, J.W. (1977). On up- and downcrossings. *Journal of Applied Probability* **14**, 405–410.

[94] Cohen, J.W. (1978). On the maximal content of a dam and logarithmic concave renewal functions. *Stochastic Processes and their Applications* **6**, 291–304.

[95] Cohen, J.W. (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* **12**, 245–284.

[96] Cohen, J.W., Hooghiemstra, G. (1981). Brownian excursion, the $M/M/1$ queue and their occupation times. *Mathematics of Operations Research* **6**, 608–629.

[97] Cohen, J.W. (1982). *The Single Server Queue.* North-Holland, Amsterdam.

[98] Cohen, J.W., Boxma, O.J. (1985). A survey of the evolution of queueing theory. *Statistica Neerlandica* **39**, 143–158.

[99] Cohen, J.W. (1994). On the effective bandwidth in buffer design for the multi-server channels. Report BS-R9406, CWI, Amsterdam.

[100] Cohen, J.W. (1997). Heavy-traffic limit theorems for the heavy-tailed GI/G/1 queue. Report PNA-R9719, CWI, Amsterdam.

[101] Cohen, J.W. (1998). A heavy-traffic theorem for the $GI/G/1$ queue with a Pareto-type service time distribution, *Journal of Applied Mathematics and Stochastic Analysis* **11**, 339-355.

[102] Cohen, J.W. (1998). Heavy-traffic theory for the heavy-tailed M/G/1 queue and $\nu$-stable Lévy noise traffic. Report PNA-R9805, CWI, Amsterdam.

[103] Cohen, J.W. (1998). The $\nu$-stable Lévy motion in heavy-traffic analysis of queueing models with heavy-tailed distributions. Report PNA-R9808, CWI, Amsterdam.

[104] Cohen, J.W. (2000). Random walk with a heavy-tailed jump distribution. Report PNA-R0010. CWI, Amsterdam. *Queueing Systems,* to appear.

[105] Courcoubetis, C., Weber, R. (1996). Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability* **33**, 886–903.

[106] Cox, D.R. (1984). Long-range dependence: a review. In: David, H.A., David, H.T. (editors). *Statistics: An Appraisal.* Iowa State University Press, 55–74.

[107] Cramèr, H. (1930). *On the Mathematical Theory of Risk.* Skandia Jubilee Volume, Stockholm. Reprinted in: Martin-Löf, A. (editor). *Harald Cramèr Collected Works Volume I.* Springer, Berlin, 601–678.

[108] Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In: *Proceedings of ACM Sigmetrics '96*, 160–169.

[109] Crovella, M., Taqqu, M.S., Bestavros, A. (1998). Heavy tails in the World Wide Web. In: Adler, R., Feldman, R., Taqqu, M.S. (editors). *A Practical Guide to Heavy Tails.* Birkhäuser, Boston, 3–25.

[110] Daley, D.J. (1964). General customer impatience in the queue $GI/G/1$. *Journal of Applied Probability* **2**, 186–205.

[111] Daley, D.J., Vesilo, R. (1997). Long range dependence of point processes, with queueing examples. *Stochastic Processes and their Applications* **70**, 265–282.

[112] Daniëls, T. (1999). *Asymptotic Behaviour of Queueing Systems.* Ph.D Thesis, Antwerpen University.

[113] Davis, R.A., Resnick, S.I. (1996). Limit theory for bilinear processes with heavy-tailed noise. *Annals of Applied Probability* **6**, 1191–1210.

[114] Debiçki, K., Rolski, T. (2000). Gaussian fluid models; a survey. Preprint, Mathematical Institute, University of Wroclaw. Available at

`http://www.math.uni.wroc.pl/~rolski/publications.html`

[115] Doetsch, G. (1974). *Introduction to the Theory and Applications of the Laplace Transformation.* Springer, New York.

[116] Doney, R.A. (1989). On the asymptotic behavior of first passage times for transient random walks. *Probability Theory and Related Fields* **81**, 239–246.

[117] Duffield, N.G., O'Connell, N. (1995). Large deviations and overflow probabilities for the general single server queue, with applications. *Proceedings of the Cambridge Philosophical Society* **118**, 363–374.

[118] Duffield, N.G. (1996). Economies of scale in queues with sources having power-law large deviation scalings. *Journal of Applied Probability* **33**, 840–857.

[119] Duffield, N.G. (1998). Queueing at large resources driven by long-tailed $M/G/\infty$ modulated processes. *Queueing Systems* **28**, 245–266.

[120] Dumas, V., Simonian, A. (2000). Asymptotic bounds for the fluid queue fed by subexponential on/off sources. *Advances in Applied Probability* **32**, 244–255.

[121] Elwalid, A., Mitra, D. (1993). Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks. *IEEE/ACM Transactions on Networking* **1**, 329–343.

[122] Embrechts, P., Goldie, C.M., Veraverbeke, N. (1979). Subexponentiality and infinite divisibility. *Probability Theory and Related Fields* **49**, 335–347.

[123] Embrechts, P., Goldie, C. (1980). On closure and factorization properties of subexponential and related distributions. *Journal of the Australian Mathematical Society Series A* **29**, 243–256.

[124] Embrechts, P., Goldie, C. (1982). On convolution tails. *Stochastic Processes and their Applications* **13**, 263–278.

[125] Embrechts, P., Veraverbeke, N. (1982). Estimates for the probability of ruin with special emphasis on the possibility of large claims. *Insurance: Mathematics and Economics* **1**, 55–72.

[126] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events.* Springer, Berlin.

[127] Erlang, A.K. (1917). Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges. *Elektroteknikeren* **13**.

[128] Erramilli, A., Singh, R.P., Pruthi, P. (1994). Chaotic maps as models of packet traffic. In: Labetoulle, J., Roberts, J.W. (editors). *Proceedings of ITC-14*. North-Holland, Amsterdam, 329–338.

[129] Erramilli, A., Narayan, O., Willinger, W. (1996). Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking* **4**, 209–223.

[130] Feldmann, A., Whitt, W. (1998). Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation* **31**, 245–279.

[131] Feller, W. (1971) *An Introduction to Probability Theory and its Applications, Volume II.* Wiley, New York.

[132] Figueiredo,R., Liu, B., Misra, V., Towsley, D. (2000). On the autocorrelation structure of TCP traffic. UMass CMPSCI Technical Report TR 00-55, University of Massachusetts at Amherst.

[133] Foss, S.G., Korshunov, D.A. (1999). On waiting time distribution in $GI/GI/2$ queue system with heavy tailed service times. Unpublished manuscript. Novosibirsk University.

[134] Foss, S.G., Korshunov, D.A. (2000). Sampling at a random time with a heavy-tailed distribution. *Markov Processes and Related Fields* **6**, 643–658.

[135] Furrer, H., Michna, Z., Weron, A. (1997). Stable Lévy motion approximation in collective risk theory. *Insurance: Mathematics and Economics* **20**, 97–114.

[136] Gautam, N., Kulkarni, V.G., Palmowski, Z., Rolski, T. (1998). Bounds for fluid models driven by semi-Markov inputs. *Probability in the Engineering and Informational Sciences* **13**, 429–475.

[137] Gaver, D., Jacobs, P. (1999). Waiting times when service times have stable laws: tamed and wild. In: Shanthikumar, J.G., Sumita, U. (editors). *Applied Probability and Stochastic Processes.* Kluwer, Boston, 219–229.

[138] Glynn, P.W., Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* **31A**, 131–156.

[139] Goldie, C.M., Klüppelberg, C. (1998). Subexponential distributions. In: Adler, R., Feldman, R., Taqqu, M.S. (editors). *A Practical Guide to Heavy Tails.* Birkhäuser, Boston, 435–460.

[140] Gouweleeuw, F. (1996). *A General Approach to Computing Loss Probabilities,* Ph.D Thesis, Free University, Amsterdam.

[141] Gouweleeuw, F. (1996). Calculating the loss probability in a $BMAP/G/1/N+1$ queue. *Stochastic Models* **12**, 473–492.

[142] Grandell, J. (1997). *Mixed Poisson Processes.* Chapman & Hall, London.

[143] Green, L. (1982). A limit theorem on subintervals of interrenewal times. *Operations Research* **30**, 210–216.

[144] Greenberg, A.G., Madras, N. (1992). How fair is fair queueing? *Journal of the ACM* **39**, 568–598.

[145] Greenwood, P., Monroe, I. (1977). Random stopping preserves regular variation of process distributions. *Annals of Probability* **5**, 42–51.

[146] Grishechkin, S. (1992). On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability* **24**, 653–698.

[147] Grishechkin, S. (1994). $GI/G/1$ processor sharing queue in heavy traffic. *Advances in Applied Probability* **26**, 539–555.

[148] Grossglauser, M., Bolot, J.-C. (1999). On the relevance of long-range dependence in network traffic. IEEE/ACM *Transactions on Networking* **7**, 629–640.

[149] De Haan, L. (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes.* CWI Tract **32**, Amsterdam.

[150] Haji, R., Newell, G.F. (1971). A relation between stationary queue length and waiting time distributions. *Journal of Applied Probability* **8**, 617–620.

[151] Heath, D., Resnick, S., Samorodnitsky, G. (1997). Patterns of buffer overflow in a class of queues with long memory in the input stream. *Annals of Applied Probability* **7**, 1021–1057.

[152] Heath, D., Resnick, S., Samorodnitsky, G. (1998). Heavy tails and long-range dependence in on-off processes and associated fluid models. *Mathematics of Operations Research* **23**, 145–165.

[153] Heath, D., Resnick, S., Samorodnitsky, G. (1999). How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Annals of Applied Probability* **9**, 352–375.

[154] Heyman, D., Lakshman, T.V. (1996). What are the implications of long-range dependence for traffic engineering? IEEE/ACM *Transactions on Networking* **4**, 301–317.

[155] Hooghiemstra, G. (1979). *Brownian Excursion and Limit Theorems for the $M/G/1$ Queue*. Ph.D Thesis, Utrecht University.

[156] Hooghiemstra, G. (1987). A path construction for the virtual waiting time of an $M/G/1$ queue. *Statistica Neerlandica* **41**, 175–181.

[157] Huang, T., Sigman, K. (1999). Delay asymptotics for tandem, split & match, and other feedforward queues with heavy tailed service. *Queueing Systems* **33**, 233–259.

[158] Hui, J. (1988). Resource allocation for broadband networks. *IEEE Journal on Selected Areas in Communications* **6**, 1598–1608.

[159] Iglehart, D. (1965). Limit theorems for traffic with intensity one. *Annals of Mathematical Statistics* **36**, 1437–1449.

[160] Jelenković, P.R., Lazar, A.A. (1998). Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *Journal of Applied Probability* **35**, 325–347.

[161] Jelenković, P.R., Lazar, A.A. (1999). Asymptotic results for multiplexing subexponential on-off processes. *Advances in Applied Probability* **31**, 394–421.

[162] Jelenković, P.R. (1999). Subexponential loss rates in a GI/GI/1 queue with applications. *Queueing Systems* **33**, 91–123.

[163] Jelenković, P.R. (1999). Network multiplexer with truncated heavy-tailed arrival streams. In: *Proceedings of Infocom '99*, New York, 625–632.

[164] Jelenković, P.R. (2000). On the asymptotic behavior of a fluid queue with a heavy-tailed $M/G/\infty$ arrival process. Preprint, Columbia University. Submitted for publication.

[165] Jelenković, P.R., Momčilović, P. (2000). Asymptotic loss probability in a finite buffer queue with heterogeneous heavy-tailed fluid on-off processes. Technical Report, Columbia University. Submitted for publication.

[166] Kalashnikov, V. (1997). *Geometric Sums: Bounds for Rare Events with Applications.* Kluwer, Dordrecht.

[167] Kalashnikov, V., Tsitsiashvili, G. (1999). Tails of waiting times and their bounds. *Queueing Systems* **32**, 257–283.

[168] Kella, O., Whitt, W. (1992). A storage model with a two-stage random environment. *Operations Research* **40**, S257–S262.

[169] Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, Chichester.

[170] Kendall, D.G. (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.

[171] Kennedy, D. (1973). Limit theorems for finite dams. *Stochastic Processes and their Applications* **1**, 269–278.

[172] Kiefer, J., Wolfowitz, J. (1956). On the characteristics of the general queueing process with applications to a random walk. *Annals of Mathematical Statistics* **27**, 147–161.

[173] Kingman, J.F.C. (1965). The heavy traffic approximation in the theory of queues. In: Smith, W.L., Wilkinson, W.E. (editors). *Proceedings of the Symposium on Congestion Theory.* The University of North Carolina Press, Chapel Hill, 137–159.

[174] Kleinrock, L. (1975), (1976). *Queueing Systems*, Volume I, II. Wiley, New York.

[175] Klüppelberg, C. (1998). Subexponential distributions and integrated tails. *Journal of Applied Probability* **25**, 132–141.

[176] Knessl, C., Matkowsky, B.J., Schuss, Z., Tier, C. (1990). An integral equation approach to the M/G/2 queue. *Operations Research* **38**, 506–518.

[177] Korshunov, D.A. (1997). On distribution tail of the maximum of a random walk. *Stochastic Processes and their Applications* **72**, 97–103.

[178] Kosten, L. (1974). Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F* **1**, 10–18.

[179] Kosten, L. (1984). Stochastic theory of data-handling systems with groups of multiple sources. In: Rudin, H., Bux, W. (editors). *Performance of Computer-Communication Systems.* Elsevier, Amsterdam, 321–331.

[180] Kulkarni, V.G. (1997). Fluid models for single buffer systems. In: Dshalalow, J. (editor). *Frontiers in Queueing*, CRC Press, Boca Raton, 321–338.

[181] Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V. (1994). On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking* **2**, 1–15.

[182] Leslie, J. (1989). On the non-closure under convolution of the class of subexponential distributions. *Journal of Applied Probability* **26**, 58–66.

[183] Likhanov, N., Mazumdar, R.R. (1999). Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability* **36**, 86–96.

[184] Likhanov, N. (2000). Bounds on the buffer occupancy probability with self-similar input traffic. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 193–213.

[185] Likhanov, N., Mazumdar, R.R. (2000). Loss asymptotics in large buffers fed by heterogeneous long-tailed sources. *Advances in Applied Probability* **32**, 1168–1189.

[186] Lindley, D.V. (1959). Discussion of a paper by C.B. Winsten. *Journal of the Royal Statistical Society Series B* **21**, 22–23.

[187] Liu, Z., Nain, Ph., Towsley, D., Zhang, Z.-L. (1999). Asymptotic behavior of a multiplexer fed by a long-range dependent process. *Journal of Applied Probability* **36**, 105–118.

[188] Loève, M. (1960). *Probability Theory.* Van Nostrand, New York.

[189] Loynes, R.M. (1962). The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philosophical Society* **58**, 497–520.

[190] Loynes, R.M. (1965). On a property of the random walks describing simple queues and dams. *Journal of the Royal Statistical Society Series B* **27**, 125–129.

[191] Lundberg, F. (1926). *Försäkringsteknisk Riskutjämning.* F. Englunds boktryckeri A.B., Stockholm.

[192] Mandelbrot, B.B., Van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review* **10**, 422–437.

[193] Mandelbrot, B.B. (1982). *The Fractal Geometry of Nature.* W.H. Freeman, New York.

[194] Mandjes, M. (1996). *Rare Event Analysis of Communication Networks.* Ph.D Thesis, Free University, Amsterdam.

[195] Mandjes, M., Borst, S.C. (2000). Overflow behavior in queues with many long-tailed inputs. *Advances in Applied Probability* **32**, 1150–1167.

[196] Mandjes, M., Kim, J.-H. (2001). Large deviations for small buffers: an insensitivity result. *Queueing Systems* **37**, 349–362.

[197] Mandjes, M., Ridder, A. (1999). Optimal trajectory to overflow in a queue fed by a large number of sources. *Queueing Systems* **31**, 137–170.

[198] Mandjes, M. (2000). A note on fluid queues with $M/G/\infty$ input. *Operations Research Letters*, to appear.

[199] Mannersalo, P., Norros, I., Riedi, R. (1999). Multifractal products of stochastic processes: a preview. Unpublished manuscript, VTT, Helsinki.

[200] Mansfield, P., Rachev, S., Samorodnitsky, G. (1999). Long strange segments of a stochastic process and long-range dependence. Technical Report TR1252, Cornell University.

[201] Maulik, K., Resnick, S., Rootzén, H. (2000). A network traffic model with random transmission rate. Technical Report TR1278, Cornell University.

[202] De Meyer, A., Teugels, J.L. (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.

[203] De Meyer, A. (1982). On a theorem of Bingham and Doney. *Journal of Applied Probability* **19**, 217–220.

[204] Mikosch, T., Nagaev, A.V. (1998). Large deviations of heavy-tailed sums with applications in insurance. *Extremes* **1**, 81–110.

[205] Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory.* Report 99-013. EURANDOM, Eindhoven.

[206] Mikosch, T., Resnick, S., Rootzén, H., Stegeman, A.W. (1999). Is network traffic approximated by stable Lévy motion or fractional Brownian motion? Technical Report TR1247, Cornell University. *Annals of Applied Probability*, to appear.

[207] Mikosch, T., Nagaev, A.V. (2000). Rates in approximations to ruin probabilities for heavy-tailed distributions. *Extremes,* to appear.

[208] Mikosch, T., Samorodnitsky, G. (2000). The supremum of a negative drift random walk with dependent heavy-tailed steps. *Annals of Applied Probability* **10**, 1025–1064.

[209] Minh, D.L. (1980). The $GI/G/1$ queue with uniformly limited virtual waiting times; the finite dam. *Advances in Applied Probability* **12**, 501–516.

[210] Mitra, D. (1988). Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability* **20**, 646–676.

[211] Morrison, J.A. (1985). Response-time distribution for a processor sharing system. *SIAM Journal of Applied Mathematics* **45**, 152–167.

[212] Norros, I. (1994). A storage model with self-similar input. *Queueing Systems* **16**, 387–396.

[213] Norros, I. (1995). On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications* **13**, 953–962.

[214] Norros, I. (2000). Queueing behavior under fractional Brownian traffic. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 101–114.

[215] Núñez Queija, R. (2000). *Processor Sharing Models for Integrated Services Networks.* Ph.D Thesis, Eindhoven University of Technology.

[216] Olver, F.W.J. (1974). *Asymptotics and Special Functions.* Academic Press, New York.

[217] Ott, T.J. (1984). The sojourn-time distribution in the $M/G/1$ queue with processor sharing. *Journal of Applied Probability* **21**, 360–378.

[218] Van Ommeren, J.C.W. (1989). *Asymptotic Analysis of Queueing Systems.* Ph.D Thesis, Free University, Amsterdam.

[219] Pakes, A.G. (1975). On the tails of waiting-time distributions. *Journal of Applied Probability* **12**, 555–564.

[220] Palmowksi, Z., Rolski, T. (1998). The superposition of alternating on-off flows and a fluid model. *Annals of Applied Probability* **8**, 524–541.

[221] Palmowski, Z. (1999). *Bounds for Steady-State Buffer Content in Fluid Models.* Ph.D Thesis, Mathematical Institute, University of Wroclaw, Poland.

[222] Park, K., Willinger, W. (2000). Self-similar network traffic: an overview. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 1–38.

[223] Park, K., Willinger, W. (eds.) (2000). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York.

[224] Parulekar, M., Makowski, A.M. (1996). Tail probabilities for a multiplexer with self-similar traffic. In: *Proceedings of Infocom 1996*, San Francisco CA, 1452–1459.

[225] Parulekar, M., Makowski, A.M. (1997). $M/G/\infty$ input processes: a versatile class of models for network traffic. In: *Proceedings of Infocom 1997*, Kobe, Japan, 419–426.

[226] Parulekar, M., Makowski, A.M. (1997). Tail probabilities for a multiplexer driven by $M/G/\infty$ input processes. (I): preliminary asymptotics. *Queueing Systems* **27**, 271–296.

[227] Parulekar, M., Makowski, A.M. (2000). Buffer asymptotics for $M/G/\infty$ input processes. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation*, Wiley, New York, 215–248.

[228] Paxson, V., Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. IEEE/ACM *Transactions on Networking* **3**, 226–244.

[229] Pitman, E.J.G. (1980). Subexponential distribution functions. *Journal of the Australian Mathematical Society Series A* **29**, 337–347.

[230] Prabhu, N.U. (1980). *Stochastic Storage Processes.* Springer, New York.

[231] Pruthi, P. (1995). *An Application of Chaotic Maps to Packet Traffic Modelling.* Ph.D Thesis, Royal Institute of Technology, Kista, Sweden.

[232] Resnick, S. (1986). Point processes, regular variation and weak convergence. *Advances in Applied Probability* **18**, 66–138.

[233] Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer, New York.

[234] Resnick, S. (1997). Heavy tail modeling and teletraffic data. With discussion and a rejoinder by the author. *Annals of Statistics* **25**, 1805–1869.

[235] Resnick, S., Samorodnitsky, G. (1997). Performance decay in a single server exponential queueing model with long range dependence. *Operations Research* **45**, 235–243.

[236] Resnick, S., Samorodnitsky G. (1999). Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems* **33**, 43–71.

[237] Resnick, S., Rootzén, H. (2000). Self-similar communication models and very heavy tails. *Annals of Applied Probability* **10**, 753–778.

[238] Resnick, S., Samorodnitsky, G. (2000). A heavy traffic approximation for workload processes with heavy tailed service requirements. *Management Science* **46**, 1236–1248.

[239] Resnick, S., Samorodnitsky, G. (2001). Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. *Bernoulli* **7**, 191–210.

[240] Riedi, R., Willinger, W. (2000). Towards an improved understanding of network traffic dynamics. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 507–530.

[241] Roberts, J.W., Mocci, U., Virtamo, J. (1996). *Broadband Network Traffic - Final Report of COST Action 242.* Springer, Berlin.

[242] Roberts, J.W. (2000). Engineering for Quality of Service. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 401–420.

[243] Rolski, T., Schlegel, S., Schmidt, V. (1999). Asymptotics of Palm-stationary buffer content distributions in fluid flow queues. *Advances in Applied Probability* **31**, 235–253.

[244] Ryu, B., Elwalid, A.I. (1996). The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Computer Communications Review* **13**, 1017–1027.

[245] Sakata, M., Noguchi, S., Oizumi, J. (1969). Analysis of a processor shared queueing model for time sharing systems. *Proceedings of the Second Hawaii International Conference on System Sciences*, 625–628.

[246] Samorodnitsky, G., Taqqu, M.S. (1994). *Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance.* Chapman & Hall, London.

[247] Schassberger, R. (1984). A new approach to the $M/G/1$ processor sharing queue. *Advances in Applied Probability* **16**, 802–813.

[248] Schwefel, H., Lipsky, L. (2001). Impact of aggregated, self-similar on-off traffic on delay in stationary queueing models (extended version). *Performance Evaluation*, to appear.

[249] Sengupta, B. (1992). An approximation for the sojourn-time distribution for the $GI/G/1$ processor-sharing queue. *Stochastic Models* **8**, 35–57.

[250] Scheinhardt, W.R.W. (1998). *Markov-modulated and Feedback Fluid Queues.* Ph.D thesis. University of Twente.

[251] Scheller-Wolf, A., Sigman, K. (1997). Delay moments for FIFO GI/GI/s queues, *Queueing Systems* **25**, 77–95.

[252] Scheller-Wolf, A. (2000). Further delay moment results for FIFO multiserver queues. *Queueing Systems* **34**, 387–400.

[253] Schmidli, H. (1999). Compound sums and subexponentiality. *Bernoulli* **5**, 999–1012.

[254] Shwartz, A., Weiss, A. (1995). *Large Deviations for Performance Analysis.* Chapman & Hall, London.

[255] Sigman, K. (editor) (1999). *Queueing Systems* **33**. Special Issue on Queues with Heavy-Tailed Distributions.

[256] Sigman, K. (1999). Appendix: A primer on heavy-tailed distributions. *Queueing Systems* **33**, 261–275.

[257] Stanford, R.E. (1979). Reneging phenomena in single channel queues. *Mathematics of Operations Research* **4**, 162–178.

[258] Starobinski, D., Sidi, M. (2000). Modeling and analysis of power-tail distributions via classical teletraffic methods. *Queueing Systems* **36**, 243–267.

[259] Stegeman, A. (2000). Heavy tails versus long-range dependence in self-similar network traffic. *Statistica Neerlandica* **54**, 293–314.

[260] Sutton, W.L.G. (1934). The asymptotic expansion of a function whose operational equivalent is known. *Journal of the London Mathematical Society* **9**, 131–137.

[261] Stern, T.E., Elwalid, A.I. (1991). Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability* **23**, 105–139.

[262] Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes.* Wiley, New York.

[263] Taqqu, M.S. (1986). A bibliographical guide to self-similar processes and long-range dependence. In: Eberlein, E., Taqqu, M.S. (editors). *Dependence in Probability and Statistics.* Birkhäuser, Boston, 137–162.

[264] Taqqu, M.S., Willinger, W., Sherman, R. (1997). Proof of a fundamental result in teletraffic modelling. *Computer Communications Review* **27**, 5–23.

[265] Teugels, J. (1976). The class of subexponential distributions. *Annals of Probability* **3**, 1000–1011.

[266] Tijms, H.C. (1994). *Stochastic Models – An Algorithmic Approach.* Wiley, Chichester.

[267] Titchmarsh, E.C. (1952). *The Theory of Functions.* Oxford University Press, London.

[268] Vamvakos, S., Anantharam, V. (1996). On the departure process of a leaky bucket system with long-range dependent input traffic. Preprint, EECS Department, University of California, Berkeley.

[269] Veraverbeke, N. (1977). Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes and their Applications* **5**, 27–37.

[270] Walrand, J., Varaiya, P. (2001). *High-Performance Communication Networks.* Morgan Kaufmann, San Francisco.

[271] Weiss, A. (1986). A new technique of analyzing large traffic systems. *Advances in Applied Probability* **18**, 506–532.

[272] Whitt, W. (1974). Heavy traffic limit theorems for queues: a survey. In: Clarke, A.B. (editor). *Mathematical Methods in Queueing Theory.* Springer, Berlin, 307–350.

[273] Whitt, W. (1991). A review of $L = \lambda W$ and extensions. *Queueing Systems* **9**, 235–268.

[274] Whitt, W. (2000). The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. *Queueing Systems* **36**, 71–87.

[275] Whitt, W. (2001). *Limits, Jumps, and Queues.* Draft of a book, available at `http://www.research.att.com/~wow`

[276] Willekens, E., Teugels, J. (1992). Asymptotic expansions for waiting time probabilities in an $M/G/1$ queue with long-tailed service time. *Queueing Systems* **10**, 295–312.

[277] Willinger, W., Taqqu, M.S., Leland, W.E., Wilson, D.V. (1995). Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements. *Statistical Science* **10**, 67–85.

[278] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM *Transactions on Networking* **5**, 71–86.

[279] Wischik, D. (2000). Sample-path large deviations for queues with many inputs. *Annals of Applied Probability,* to appear.

[280] Wischik, D. (2000). Moderate deviations in queueing theory. Unpublished manuscript, Statistical Laboratory, University of Cambridge.

[281] Yashkov, S.F. (1983). A derivation of response time distribution for a M/G/1 processor sharing queue. *Problems of Control and Information Theory* **12**, 133–148.

[282] Yashkov, S.F. (1987). Processor-sharing queues: some progress in analysis. *Queueing Systems* **2**, 1–17.

[283] Yashkov, S.F. (1992). Mathematical problems in the theory of shared-processor systems. *Journal of Soviet Mathematics* **58**, 101–147.

[284] Yashkov, S.F. (1993). On a heavy traffic limit theorem for the M/G/1 processor sharing queue. *Stochastic Models* **9**, 467–471.

[285] Zwart, A.P. (1999). Sojourn times in a multiclass processor sharing queue. In: Key, P., Smith, D. (editors). *Proceedings of ITC-16.* North-Holland, Amsterdam, 335–344.

[286] Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Advances in Applied Probability* **32**, 221–243.

[287] Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* **35,** 141–166.

[288] Zwart, A.P., Borst, S.C., Mandjes, M. (2000). Exact asymptotics for fluid queues fed by multiple heavy-tailed On-Off flows. SPOR-Report 2000-14, Eindhoven University of Technology. Shortened version in: *Proceedings of Infocom 2001,* Anchorage AK, 279–288.

[289] Zwart, A.P. (2001). Tail asymptotics for the busy period in the $GI/G/1$ queue. *Mathematics of Operations Research*, to appear.

[290] Zwart, A.P. (2001). Subexponential asymptotics for a fluid model. In preparation.

# Samenvatting (Summary)

Het onderwerp van dit proefschrift is de probabilistische analyse van wachtrijmodellen en vloeistofmodellen, waarbij bijvoorbeeld de bedieningsvraag van een klant of de aanperiode van een aan-uit bron zwaarstaartig is.

De belangstelling voor deze klasse van wachtrijmodellen is ingegeven door recente metingen aan moderne communicatienetwerken, zoals het Internet. Deze metingen hebben uitgewezen dat het verkeer in deze systemen zich extreem grillig gedraagt. Essentiële kenmerken van dit verkeer zijn onder meer het 'fractale' karakter (*self-similarity*) en significante correlaties op grote tijdschalen (*long-range dependence*). Een algemeen geaccepteerde verklaring voor deze verschijnselen is 'zwaarstaartigheid' van verdelingen van diverse grootheden, zoals lengtes van telefoongesprekken en filegroottes in het dataverkeer. Hoofdstuk 1 van dit proefschrift gaat dieper in op bovenstaande motivatie en plaatst de in dit proefschrift gevolgde aanpak in een breder kader.

Hoofdstuk 2 gaat dieper in op de wiskundige aspecten van wachtrijen met zware staarten. Er worden diverse klassen en eigenschappen van zwaarstaartige verdelingen geïntroduceerd. Daarnaast worden diverse in de literatuur bekende resultaten voor basismodellen gegeven. Dit hoofdstuk heeft als rode draad de aandacht voor de intuïtieve verklaring van deze resultaten en bevat ook een heuristische afleiding van de staartkans van de wachttijd in een wachtrij met twee heterogene bedienden. Het hoofdstuk besluit met een algemeen recept dat in latere hoofdstukken als leidraad dient om deze heuristische afleidingen te vertalen in een bewijs.

In Hoofdstuk 3 analyseren we de verblijftijd van een klant in de $M/G/1$ wachtrij met de Processor Sharing (PS) bedieningsdiscipline, voor het geval dat de bedieningsduurverdeling van een klant een regulier variërende staart heeft. Het belangrijkste resultaat van dit hoofdstuk is dat de staarten van de bedieningsduurverdeling en verblijftijdverdeling *even zwaar* zijn. Dit staat in schril contrast met de traditionele First-Come-First-Served (FCFS) bedieningsdiscipline, waarbij een zwaarstaartige bedieningsduurverdeling leidt tot een nog zwaardere staart van de verdeling van de verblijftijd. De resultaten in dit hoofdstuk geven duidelijk aan dat een lange bedieningstijd van een klant slechts een beperkte invloed heeft op de verblijftijd van andere klanten.

De in dit proefschrift bestudeerde modellen hebben vrijwel allemaal een oneindig grote buffer. Een uitzondering op deze regel wordt gemaakt in Hoofdstuk 4: Dit hoofdstuk

richt zich op het evalueren van de verliesfractie in een vloeistofmodel, gebruikmakend van relaties met het vloeistofmodel met oneindig grote buffer. Daarnaast wordt de verdeling van de bufferinhoud, in het bijzonder de gemiddelde bufferinhoud, bestudeerd. De resultaten worden toegepast om het asymptotische gedrag van de verliesfractie en gemiddelde bufferinhoud te bepalen, voor het geval dat de buffer groot is. De resultaten laten zien dat in sommige gevallen een extreem grote buffer nodig is om een kleine verliesfractie te garanderen, hetgeen een direct gevolg is van de zwaarstaartige input.

Het centrale onderwerp in Hoofdstuk 5 is de lengte van de 'bezige periode' in de $G/G/1$ wachtrij; dit is de periode dat de bediende onafgebroken aan het werk is. We concentreren ons in het bijzonder op de staartkans van de bijbehorende kansverdeling in het geval dat de bedieningsduurverdeling regulier variërend is; de tussenaankomsttijd heeft een willekeurige verdeling. Een belangrijke bijdrage van dit hoofdstuk is de manier waarop het staartgedrag van de bezige periode wordt afgeleid. Eerst wordt heuristisch beargumenteerd dat een lange bezige periode het gevolg is van een extreem grote hoeveelheid werk in het systeem aan het 'begin' van die bezige periode. Vervolgens wordt deze intuïtie gebruikt in het bewijs.

In Hoofdstukken 6 en 7 van het proefschrift analyseren we het vloeistofmodel met meerdere aan-uit bronnen met zwaarstaartige (regulier variërende) aan-tijden. Naast deze bronnen laten we ook verkeer met een lichtstaartig karakter toe. Voor deze superpositie analyseren we het staartgedrag van de stationaire verdeling van de bufferinhoud.

Er is een duidelijk criterium aan te geven dat het kwalitatieve gedrag van deze staart bepaalt. Als de capaciteit van het systeem groter is dan een bepaalde kritieke waarde, dan is de kansverdeling van de bufferinhoud lichtstaartig; in het andere geval heeft de verdeling van de bufferinhoud een zware staart. Deze regimes zijn het respectievelijke onderwerp van Hoofdstukken 6 en 7. Beide hoofdstukken leunen zwaar op intuïtieve verklaringen voor de totstandkoming van extreem grote vertragingen.

In Hoofdstuk 6 laten we zien dat een grote bufferinhoud het gevolg is van het feit dat alle zwaarstaartige bronnen tegelijkertijd een lange aan-periode beleven. In dit regime is de overgebleven capaciteit voor de lichtstaartige input nog steeds genoeg om het systeem stabiel te houden. Dit systeem wordt geanalyseerd met behulp van bestaande resultaten uit de theorie van grote afwijkingen.

In het regime van hoofdstuk 7 komt een extreem grote bufferinhoud op geheel andere wijze tot stand. We laten zien dat een bepaalde 'dominante' verzameling aan-uit bronnen verantwoordelijk is. Deze verzameling kan worden beschreven als de oplossing van een 'knapsack' probleem. De bronnen die niet tot deze verzameling behoren kunnen vervangen worden door hun gemiddelde input en oefenen zo geen invloed uit op de zeldzame gebeurtenis. Dit hoofdstuk laat zien dat de bronnen uit de dominante verzameling elk één lange aan-periode genereren. Deze aan-periodes treden vrijwel gelijktijdig op.

In voorgaande studies is alleen het geval opgelost waarbij de dominante verzameling uit één aan-uit bron bestaat. Deze aanname vereenvoudigt de analyse, maar is uit praktisch

oogpunt onbevredigend. In dit hoofdstuk wordt deze restrictie opgeheven, hetgeen een aanmerkelijk gecompliceerder bewijs met zich meebrengt. Het bewijs leunt zwaar op de gegeven intuïtie en op het recept in Hoofdstuk 2.
De analyse in Hoofdstuk 8 van dit proefschrift is nauw gerelateerd aan die van Hoofdstuk 7, maar de input van het vloeistofmodel wordt nu gereguleerd door het aantal klanten in een $M/G/\infty$ wachtrij, d.w.z., het inputproces kan gezien worden als de superpositie van oneindig veel aan-uit bronnen die elk één aan-periode genereren. De structuur van dit inputproces is aanmerkelijk eenvoudiger dan dat van Hoofdstuk 7 en is daarom buitengewoon populair. Evenals in Hoofdstuk 7 zijn in de literatuur slechts exacte resultaten bekend voor de staart van de bufferinhoud wanneer één lange aan-periode genoeg is om het systeem instabiel te maken. Dit hoofdstuk geeft exacte asymptotische resultaten voor het algemenere geval waarbij meerdere lange aan-periodes nodig zijn. De mooie structuur van het inputproces maakt het mogelijk om ook asymptotische resultaten af te leiden voor de transiënte verdeling van de bufferinhoud.

# Curriculum Vitae

Bert (Albertus Petrus) Zwart was born in Hilversum (the Netherlands) on April 16, 1974. He graduated from Grammar School (College Stad en Lande in Huizen) in June 1992. Thereafter, he started his study in Econometrics at the Free University in Amsterdam. From February 1996 until August 1997 he was a research assistant at the department of Spatial Economics of the same university. In August 1997 he received his masters' degree in econometrics (cum laude).
After this, he became a Ph.D student at CWI (Center for Mathematics and Computer Science, Amsterdam). In September 1998, he continued his Ph.D research at Eindhoven University of Technology. He hopes to defend his Ph.D thesis at the same university on September 11, 2001. In October 2001, Bert plans to work as a post-doc at INRIA (the French national institute for research in computer science and control) in Rocquencourt, France.

Bert (voluit: Albertus Petrus) Zwart werd geboren op 16 april 1974 in Hilversum. In juni 1992 behaalde hij het VWO diploma aan het College Stad en Lande in Huizen, waarna hij econometrie ging studeren aan de Vrije Universiteit in Amsterdam. Van februari 1996 tot augustus 1997 was hij student-assistent op dezelfde universiteit bij de vakgroep Ruimtelijke Economie. In augustus 1997 behaalde hij het doctoraal diploma Econometrie (cum laude). Hierna werd hij onderzoeker-in-opleiding aan het Centrum voor Wiskunde en Informatica (CWI) in Amsterdam. Vanaf september 1998 zette hij zijn werk voort aan de Technische Universiteit Eindhoven, alwaar hij zijn proefschrift hoopt te verdedigen op 11 september 2001. Vanaf oktober 2001 werkt Bert als post-doc bij INRIA (Het Frans nationaal onderzoeksinstituut voor informatica en regeltheorie) in Rocquencourt, Frankrijk.