INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
Statistical Resource Sharing in the Presence of Heavy Tails

Petar Momčilović

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2003
ABSTRACT

Statistical Resource Sharing in the Presence of Heavy-Tails

Petar Momčilović

Self-similar/heavy-tailed phenomena are observed in a large number of natural and man-made systems ranging from earthquakes and genetics to telephone networks and UNIX operating systems. Statistical sharing of bottleneck resources, e.g., bandwidth, processing power and storage, is a common way of increasing the operating efficiency of network and service systems. While the benefits of sharing, i.e. multiplexing, are well understood for exponential traffic models, the implications of long-range dependence are just now being discovered. This thesis focuses on characterizing these implications.

First, we study the classical model of a network switching element, a finite buffer single server queue fed by On-Off traffic sources. The primary performance measures of this model are the loss rate and buffer overflow probability. These quantities indicate the quality of service provided by the system. For the case of heavy-tailed On-periods, explicit asymptotic formulas for the loss rate and buffer overflow probability are derived. The results provide important insight into qualitative tradeoffs between the performance measures and system parameters, the key element for proper dimensioning of the system. Furthermore, we quantify the benefits of buffer sharing and scheduling in the same model.

Second, we examine resource sharing with feedback control, such as TCP, the
 predominant transport protocol in the Internet. The processor sharing queue represents a baseline model of ideal flow control, i.e., when a number of users share bandwidth, each user receives an equal share. It is shown that job (file) transmission times admit an easy asymptotic characterization depending on whether the job size has a heavier or lighter tail than the Weibull distribution $e^{-\sqrt{x}}$. In other words, this fundamental model exhibits a phase transition at $e^{-\sqrt{x}}$. Furthermore, the newly developed large deviations approach also provides a mathematical framework for proving related results.
Contents

1 Introduction ............................................. 1
   1.1 Network Resource Sharing ............................. 1
   1.2 Notation .............................................. 5
   1.3 Heavy-Tailed Basics .................................... 6
   1.4 Thesis Outline ........................................ 9

2 Finite Buffer Fluid Queue ................................. 13
   2.1 Introduction ........................................... 13
   2.2 Definition and Sample Path Bounds ..................... 15
   2.3 Fluid Queue with an On-Off Arrival Process .......... 18
   2.4 Finite Buffer Fluid Queue ............................. 20
   2.5 Service Differentiation ................................. 36
   2.6 Numerical Examples ................................... 47
   2.7 Proofs ................................................. 50

3 Moderately Heavy Tails .................................. 57
3.1 Introduction ............................................. 57
3.2 Large Deviation Results ............................... 60
3.3 Gaussian Insensitivity ................................. 69
3.4 Engineering Applications ............................. 73
    3.4.1 Reduced Load Equivalence ....................... 73
    3.4.2 Independent Sampling ............................. 80
    3.4.3 Busy Period ................................... 83
    3.4.4 Processor Sharing ............................... 86
3.5 Proofs ................................................. 98
    3.5.1 Large deviation bounds ......................... 99
    3.5.2 Queueing Results ............................... 110
    3.5.3 Proof of Proposition 11 ......................... 114

4 Conclusions ........................................... 119

Bibliography ............................................. 121
List of Figures

1.1 Log/log plot of the empirical distribution of file sizes on five file servers in COMET laboratory at Columbia University. The tail of the empirical distribution (solid line) is well matched by a Pareto distribution $cx^{-\alpha}$ with $\alpha = 1.44$ (dashed line). ............................................. 4

2.1 Illustration for Remark 3 (ii). The long On period is shown with a dashed line. ................................................................. 26

2.2 An illustration of the buffer management policy. Two On-Off processes with peak rate $r$ share a server of capacity $\phi = r/2$ and buffer of size $B$. The GPS weights of the two processes are equal, $\varphi_1 = \varphi_2 = 0.5$. ........... 38

2.3 Illustration for Example 1 .................................................. 48

2.4 Illustration for Example 2 .................................................. 49

2.5 Illustration for Example 3 .................................................. 50

2.6 Illustration for Example 4 .................................................. 51
3.1 A typical sample path realization of function \( H \); \( T_n \)-s are the Poisson arrival points.
Acknowledgements

I would like to primarily thank my advisor, Professor Predrag Jelenković. I am truly indebted to him for his invaluable guidance and thought-provoking comments throughout my years at Columbia. Predrag’s perspective on research enabled my development beyond the pages of this thesis and I am deeply grateful for the opportunity to work with him.

Many thanks to Professor Edward Coffman for immensely enriching my experience as a graduate student. His comments and suggestions have always been of great value to me.

Finally, I would also like to thank the members of my defense committee, Professor Ioannis Karatzas, Professor Aurel Lazar and Dr. Mark Squillante for their valuable time.
To Maki and Mulja
Chapter 1

Introduction

1.1 Network Resource Sharing

Statistical resource sharing represents a principle for increasing the operating efficiency of communication networks. In addition, many other industries, including transportation and food service, operate under this rule. In some sense it is a natural solution when balancing between customer's desire to obtain the best possible service and provider's aim to maximize the profit. Typically, the higher degree of sharing leads to more efficient system operation at the expense of quality of service (QoS) observed by individual users.

Increased utilization in communication networks is achieved through sharing of network resources, e.g. link capacity and buffer space, among different user sessions. The benefits of sharing common resources are counterbalanced with potential in-
CHAPTER 1. INTRODUCTION

crease in congestion and degradation in quality of service perceived by individual sessions. Therefore, understanding the tradeoffs between the offered traffic load, perceived QoS measures, link capacity and buffer space is essential for the efficient design and provision of network switching elements.

Communication networks widely exploit resource sharing at different levels. The highest degree of sharing is implemented in datagram networks, e.g. the Internet, where no guarantees are provided to users on the quality of service. Hence, such networks are often referred to as best-effort networks. Modern networks carry a diverse spectrum of multimedia services, ranging from real-time traffic, such as voice and video, to various data and Web related applications. These services have different QoS requirements, e.g., real-time services have stringent delay requirements, but can tolerate relatively high losses. On the other hand, data related services typically could tolerate larger delays, but need minimal or no losses. Hence, it is essential for network operators to have the ability to provide QoS differentiation among traffic flows. This is usually achieved through priority scheduling mechanisms. The most popular scheduling schemes, e.g. the Weighted Fair Queueing, are based on the Generalized Processor Sharing algorithm. These algorithms offer the flexibility for providing high degree of service differentiation, extracting statistical multiplexing gains as well as protecting individual flows from the ones with high service demands.

Most of the early work on the resource sharing problem focuses on traffic models with exponential characteristics. However, repeated empirical measurements in
CHAPTER 1. INTRODUCTION

modern networks demonstrate the presence of heavy-tailed characteristics in network traffic streams. Early discoveries of the self-similar nature of Ethernet traffic were reported in [60]. Long-range dependence and subexponential properties of variable-bit-rate video streams (e.g., MPEG) were explored in [43, 50, 59]. Evidence and possible causes of heavy-tailed characteristics in World Wide Web traffic were presented in [29]. Here, we provide an additional confirmation of the existence of heavy tails in network traffic. We have measured the distribution of file sizes on five file servers in the COMET laboratory at Columbia University. The empirical distribution of 350,000 surveyed files is presented on a log/log scale in Figure 1.1. We find that the tail of the measured distribution is well matched by a Pareto distribution with parameter $\alpha = 1.44$; see the dashed line in Figure 1.1. This suggests that the corresponding ftp (file transfer protocol) traffic is heavy-tailed.

Informally, heavy-tailed phenomena are observed in systems that exhibit high degree of statistical variability. Loosely speaking, a random variable $X$ is heavy-tailed if $\mathbb{P}[X > x]$ decays in $x$ slower than any exponential function. Heavy-tailed variables are observed in multiple contexts including: the Web graph, UNIX holding times, peer-to-peer networks, Internet routing, gene networks, earthquakes, insurance claims, personal income, star brightness, sun activity, city population, stock market, river floods and even horse race betting in Korea [76]!

What leads to such behavior? Given the wide range of systems with heavy-tailed phenomena, it is unlikely that there will be a different system-specific explanation
for each case. Thus, most likely such behavior is due to a law of large numbers. Consider multiplicative noise, i.e.,

\[ \prod_{i=1}^{n} \left( 1 + \frac{X_i}{\sqrt{n}} \right), \]

where \( X_i \)'s are zero-mean random variables. Then as \( n \to \infty \) the product converges in distribution to a heavy-tailed random variable with lognormal distribution. Hence, the heavy-tailed nature is induced not by a single cause, but by many small causes in the same way Gaussian noise is due to additive noise. By slightly modifying the preceding product to include a barrier one obtains that the distribution is Pareto rather then lognormal [36].

![Figure 1.1: Log/log plot of the empirical distribution of file sizes on five file servers in COMET laboratory at Columbia University. The tail of the empirical distribution (solid line) is well matched by a Pareto distribution \( cx^{-\alpha} \) with \( \alpha = 1.44 \) (dashed line).](image)

At the end of this section, we note that currently the vast majority of Internet traffic is transferred by means of a closed-loop control protocol, i.e., feedback control.
CHAPTER 1. INTRODUCTION

Modeling analysis of systems with feedback control is particularly challenging task due to the fact that the offered load depends on the state of the system. Resorting to processor sharing algorithms is an elegant way of overcoming such challenges. These types of algorithms allow for efficient and fair distribution of resources. Early work on processor sharing [26] was motivated by the study of multi-user mainframe computer systems. Renewed interest in the processor sharing queues stems from their application in modeling of computer communication networks and Web servers that are designed with the notion of fairness in mind.

1.2 Notation

This section contains the most frequently used notation in this thesis.

Excess (residual, equilibrium) random variables (r.v.) play an important role in the analysis of renewal processes. For a nonnegative random variable $X$ with a finite mean, the excess distribution $F^{(e)}$ for $x \geq 0$ is defined by

$$F^{(e)}(x) = \frac{1}{E X} \int_{0}^{x} \mathbb{P}[X > u] \, du.$$ 

A random variable $X^{(e)}$ with distribution $F^{(e)}$ is called the excess (or residual) variable of $X$.

Throughout the thesis, for any two real functions $f(x)$ and $g(x)$, we use the standard notation $f(x) \sim g(x)$ as $x \to \infty$ to denote $\lim_{x \to \infty} f(x)/g(x) = 1$ or equivalently $f(x) = g(x)(1 + o(1))$ as $x \to \infty$. 

CHAPTER 1. INTRODUCTION

Variable $C$ denotes a sufficiently large positive constant, while $c$ denotes a sufficiently small positive constant. The values of $C$ and $c$ are generally different in different places. For example, $C/2 = C$, $C^2 = C$, $C + 1 = C$, etc.

1.3 Heavy-Tailed Basics

This section contains the definitions and some of the basic properties of heavy-tailed distributions.

First, we introduce a family of long-tailed distribution functions. This is the largest operational class of heavy-tailed distributions.

Definition 1. A nonnegative r.v. $X$ is called long-tailed, $X \in \mathcal{L}$, if

$$\lim_{x \to \infty} \frac{\mathbb{P}[X > x - y]}{\mathbb{P}[X > x]} = 1, \quad \forall y \in \mathbb{R}.$$ 

The following class of heavy-tailed distributions was introduced by Chistyakov [23].

Definition 2. A nonnegative r.v. $X$ is called subexponential, $X \in \mathcal{S}$, if

$$\lim_{x \to \infty} \frac{\mathbb{P}[X_1 + X_2 > x]}{\mathbb{P}[X > x]} = 2,$$

where $X_1$ and $X_2$ are independent copies of $X$.

It is well known that $\mathcal{S} \subset \mathcal{L}$ [9]. A survey on subexponential distributions can be found in [38]. The class of intermediately regularly varying distributions $\mathcal{IR}$ is a subclass of $\mathcal{S}$. 
CHAPTER 1. INTRODUCTION

Definition 3. A nonnegative r.v. $X$ is called intermediately regularly varying, $X \in \mathcal{IR}$, if

$$\lim_{\eta \uparrow 1} \lim_{x \to \infty} \frac{\mathbb{P}[X > \eta x]}{\mathbb{P}[X > x]} = 1.$$ 

Regularly varying distributions $\mathcal{R}_\alpha$, which contain Pareto distribution, are the best known examples from $\mathcal{IR}$ ($\mathcal{R}_\alpha \subset \mathcal{IR}$).

Definition 4. A nonnegative r.v. $X$ is called regularly varying with index $\alpha$, $X \in \mathcal{R}_\alpha$, if

$$\mathbb{P}[X > x] = l(x) x^{-\alpha}, \quad \alpha \geq 0,$$

where $l(x)$ is a function of slow variation, i.e., $l(\eta x) \sim l(x)$ as $x \to \infty$ for $\eta > 1$.

In addition to the above introduced classes of random variables we will refer to the following two as well.

Definition 5. A nonnegative r.v. $X$ belongs to the class $\mathcal{S}^*$, $X \in \mathcal{S}^*$, if $X$

$$\lim_{x \to \infty} \int_0^x \frac{\mathbb{P}[X > x - y]}{\mathbb{P}[X > x]} \mathbb{P}[X > y] \, dy = 2 \mathbb{E}X < \infty.$$ 

Definition 6. A nonnegative r.v. $X$ belongs to the class $\mathcal{D}$, $X \in \mathcal{D}$, of dominated-variation distributions if

$$\lim_{x \to \infty} \frac{\mathbb{P}[X > x]}{\mathbb{P}[X > 2x]} < \infty.$$ 

Next we state three basic lemmas on $\mathcal{IR}$ distributions.
CHAPTER 1. INTRODUCTION

Lemma 1. If $X \in \mathcal{I}\mathcal{R}$ then $X \in \mathcal{D}$; in particular for any $\eta \in (0, 1)$

$$\sup_{x \in [0, \infty)} \frac{\mathbb{P}[X > \eta x]}{\mathbb{P}[X > x]} < \infty.$$ 

Proof. Follows immediately from the definition. \qed

Lemma 2. If $X^{(e)} \in \mathcal{I}\mathcal{R}$, then

$$\lim_{z \to \infty} \frac{x \mathbb{P}[X^{(e)} \geq x]}{\mathbb{P}[X^{(e)} \geq x]} < \infty.$$ 

Proof. For any $\delta \in (0, 1)$ by definition of $F^{(e)}$

$$\frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X^{(e)} \geq x]} \leq \frac{\mathbb{P}[X^{(e)} \geq \delta x]}{\mathbb{P}[X^{(e)} \geq x]} x \mathbb{P}[X \geq x] \mathbb{E}X \int_0^\delta \frac{\mathbb{P}[X \geq u]}{u} du \leq \frac{\mathbb{P}[X^{(e)} \geq \delta x]}{\mathbb{P}[X^{(e)} \geq x]} \frac{\mathbb{E}X}{1 - \delta}.$$ 

Hence, the result follows by Lemma 1

$$\lim_{z \to \infty} \frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X^{(e)} \geq x]} \leq \frac{\mathbb{E}X}{1 - \delta} \lim_{z \to \infty} \frac{\mathbb{P}[X^{(e)} \geq \delta x]}{\mathbb{P}[X^{(e)} \geq x]} < \infty. \quad \Box$$

For any bounded nondecreasing function $F$ we say that $F \in \mathcal{I}\mathcal{R}$ if it satisfies

$$\lim_{\eta \downarrow 1} \lim_{z \to \infty} \frac{F(\eta x)}{F(x)} = 1.$$ 

Then, the following lemma follows directly from Definition 3.

Lemma 3. If $F_1, F_2 \in \mathcal{I}\mathcal{R}$, then

(i) $F_1 F_2 \in \mathcal{I}\mathcal{R},$

(ii) $w_1 F_1 + w_2 F_2 \in \mathcal{I}\mathcal{R}$, for $w_1, w_2 > 0.$
CHAPTER 1. INTRODUCTION

The section is concluded with two technical results on the properties of the class $S^*$

**Theorem 1 ([56]).** (a) If $X \in D \cap L$ has finite expectation, then $X \in S^*$. (b) If $X \in S^*$, then $X \in S$ and $X^{(e)} \in S$.

**Lemma 4 ([57]).** Let $\{X, X_i, i \geq 1\}$ be independent and identically distributed r.v.s. If $X \in S^*$, then

(i) for each $\epsilon > 0$ there exists a constant $K(\epsilon) > 0$ such that

$$
\frac{d\mathbb{P}\left[\sum_{i=1}^{n} X_i^{(e)} \leq x\right]}{dx} \leq K(\epsilon)(1 + \epsilon)^n \mathbb{P}[X > x], \quad x \geq 0, \ n \geq 1.
$$

(ii) for any fixed $n$, as $x \to \infty$

$$
\frac{d\mathbb{P}[\sum_{i=1}^{n} X_i^{(e)} \leq x]}{dx} \sim n \mathbb{P}[X > x].
$$

1.4 Thesis Outline

Analysis of resource sharing models undertaken in this thesis is of asymptotic nature, i.e., the obtained results are correct only in the limit as probabilities tend to zero. While obviously asymptotic results do not fully describe system behavior, we argue that they are of value since (i) in a well-designed system "undesirable" events occur with small probabilities, (ii) all obtained results are explicit and offer insights on the dependency of system performance and system parameters.
CHAPTER 1. INTRODUCTION

In Chapter 2 we consider a fluid queue with a finite buffer fed by a superposition of a number of independent On-Off processes. An On-Off process consists of a sequence of alternating independent activity and silence periods. For this queueing system, under the assumption that the excess activity periods belong to a class of heavy-tailed distributions, we derive explicit and asymptotically exact formulas for approximating the stationary overflow probability and loss rate. The analyzed queueing system represents a standard model of resource sharing in telecommunication networks. The results offer insight into qualitative tradeoffs between the overflow probability, offered traffic load, capacity and buffer space.

After evaluating the overall system performance we turn our attention to service differentiation, i.e., performance observed by individual sources in the presence of buffer management and scheduling. The buffer sharing is unrestricted as long as there is available space. If the buffer is full, the necessary amount of fluid from the most demanding flows is discarded. The flows share a common server according to the Generalized Processor Sharing scheduling discipline. When each flow receives the minimum service rate guarantee that exceeds its long-term average demand, our result shows that the loss rate of a particular flow is asymptotically equal to the loss rate in a reduced system with a smaller capacity and the same buffer size, where this flow is served in isolation. In other words, each flow perceives to have the entire buffer to itself. This finding quantifies the benefits of having a shared buffer. This qualitative insight may prove useful in deciding on whether to engineer buffers at
the periphery or in the core of the network.

Chapter 3 deals with several problems that involve moderately heavy-tails, i.e., lognormal- and Weibull-like distribution. We establish a number of large deviation results for such distributions with a uniform bound being the main one. This study is motivated by recent findings that server access patterns and file sizes may have moderately heavy tails, e.g. lognormal [63,64,86]. The main focus from the technical point of view is on uniform bounds that allow us to bridge the gap between the Central Limit Theorem and large deviations regimes. The usefulness of such bounds is demonstrated on a number of problems in which the Weibull distribution $e^{-\sqrt{x}}$ plays an important role. The criticality of $e^{-\sqrt{x}}$ has appeared in a variety of settings, starting with early large deviation results of [69] and more recent analyses in [7, 35]. This phenomenon arises from a requirement that a distribution has to tolerate Gaussian deviations of order $\sqrt{x}$ that we refer to as square-root insensitivity. The Weibull tail $e^{-\sqrt{x}}$ represents a natural condition, since easy arguments show that our large deviation results do not hold for distributions lighter than $e^{-\sqrt{x}}$.

In the remaining part of Chapter 3 we investigate for problems. First, we examine the reduced load equivalence problem. Two processes are fed to an infinite buffer queue and probability of workload exceeding some large level is considered. A number of conditions is derived under which one of the processes can be replaced by its mean rate so that the asymptotic behavior of the total workload remains the same. Second, the problem of independent sampling is considered in which we
CHAPTER 1. INTRODUCTION

study the tail of the probability distribution of a function sampled at a random
time. Third, we look into the distribution of the busy period in a stable queue.
The busy period is a fundamental quantity the model since it is understanding is
essential is addressing a long list of queueing systems.

Finally, we investigate the distribution of the waiting time in a stable processor
sharing queue. We would like to mention prospects of modeling congested links
with transmission control protocol (TCP) traffic as a processor sharing queue.
More precisely, consider a number of independent TCP sessions that are running for
an extended period of time. Then, by fairness of TCP (e.g., see [55, 67]), it follows
that on a long run each session receives an equal share of bandwidth, which is exactly
captured by processor sharing discipline. In general, the processor sharing model
represents a prototypical model of all fair-share algorithms. When the distribution
of a customer service request belongs to a large class of subexponential distributions
with tails heavier than \(e^{-\sqrt{x}}\), it is shown that the waiting time is asymptotically the
same as if the customer were served in isolation at an equivalent rate.
Chapter 2

Finite Buffer Fluid Queue

2.1 Introduction

The fundamental switching components used for sharing bandwidth and buffer space are network multiplexers. An established baseline model of a network multiplexer is a single server queue with a constant capacity and finite buffer fed by a superposition of user sessions. Individual sessions can be modeled as On-Off processes, since a session can be either active, in which case it transmits data at a specified rate, or silent. The primary performance measures of this queueing system are the stationary overflow probability and loss rate. The analysis of a related infinite buffer queueing system dates back to [3, 27, 82].

The analysis of queueing models with multiplexed heavy-tailed renewal arrival sequences, e.g. On-Off processes, is difficult primarily due to the complex depen-
CHAPTER 2. FINITE BUFFER FLUID QUEUE

dency structure in the aggregate arrival process [40]. This stems from the fact that a superposition of renewal processes, in general, is not a renewal process. An intermediate case of multiplexing a single long-tailed arrival sequence with exponential processes was investigated in [2, 21, 49, 81]. An infinite limit of On-Off processes, the so-called \( M/G/\infty \) process, represents an instance of an analytically tractable model since it has both a renewal and Poisson structure. Recent results and additional references on both fluid and discrete time queues with \( M/G/\infty \) arrival processes can be found in [21, 31, 41, 46, 49, 62, 77, 80, 87].

On the other hand, the understanding of multiplexing a finite number of heavy-tailed On-Off arrival processes is quite limited, for general bounds see [24, 32]. In this chapter we derive explicit and asymptotically exact results for the stationary overflow probability and loss rate in a finite buffer queue with heterogeneous heavy-tailed On-Off arrival processes. The starting point of our analysis are the results from [45]. Recently the complementary results for the infinite buffer model were derived in [98].

In the second part of the chapter we examine the same system in the presence of scheduling and buffer management policy. In particular, we study individual user performance when the server capacity is divided among users by the means of the Generalized Processor Sharing (GPS) policy. Rigorous investigation of stochastic systems with GPS dates back to [34]; see also [34] for some earlier references. This work was centered around the problem of time-shared computer systems. Results for
traffic models with exponential characteristics can be found in [11,66,94,95]. Recent investigations of the behavior of GPS in the presence of heavy-tailed arrival streams can be found in [14–18]. The reader may consult the same papers for additional references on GPS. These papers consider a system with finite number of heavy-tailed sessions each of which is queued into an infinite buffer queue. The content of the queues is served by a single server that is scheduled using GPS. When the GPS weights are appropriately chosen, these studies show that each session experiences the same queueing behavior as if it were served in isolation with an appropriate constant capacity. On the other hand, if the weights are not properly engineered, the flows may experience induced burstiness [15,18]. These results stress the importance of properly selecting the GPS weights.

The chapter is based on [51,52].

2.2 Definition and Sample Path Bounds

Consider a fluid queue with a constant capacity \( \phi \), finite buffer \( B \) and arrival process \( A(t) \). Informally, at time \( t \), fluid is arriving at rate \( A(t) \) and is leaving the system at rate \( \phi \). When the queue level reaches the buffer limit \( B \), fluid arriving in excess of the draining rate \( \phi \) is lost. We use \( Q^B(t) \in [0,B] \) to denote the queue content at time \( t \).

In this chapter we only consider right continuous piece-wise constant processes with almost surely (a.s.) increasing jump times \( T_0 = 0 < T_1 < T_2 < \cdots \). Then, for
any initial value $Q^B(0)$ the evolution of $Q^B(t)$ is given by

$$Q^B(t) = (Q^B(T_n) + (t - T_n)(A(T_n) - \phi))^+ \land B, \quad t \in (T_n, T_{n+1}], \; n \geq 0,$$  \hspace{1cm} (2.1)

where $(x)^+ = \max(0, x)$ and $x \land y = \min(x, y)$. When necessary, we use the notation $Q^B_{A \phi} \equiv Q^B$ to mark the explicit dependence of $Q^B(t)$ on $A(t)$ and $\phi$.

In the case of $A(t)$, i.e. $\{(T_{n+1} - T_n), A(T_n)\}$, being stationary and ergodic, and $\mathbb{E}A(t) < \phi$, by using Loynes’ construction [65], one can show that recursion (2.1) has a unique stationary and ergodic solution. Furthermore, for all initial conditions $Q^B(0)$, the distribution of $Q^B(t)$ converges to that stationary solution as $t \to \infty$. Unless otherwise indicated, we assume throughout the thesis that all arrival processes are stationary, ergodic and that the corresponding queues are in their stationary regimes. Let $Q^B$ and $A$ be random variables that are equal in distribution to $Q^B(t)$ and $A(t)$, respectively.

The main focus of this chapter is the asymptotic evaluation, as $B \to \infty$, of the overflow probability $\mathbb{P}[Q^B \geq B - K]$, for finite $K$, and long time average loss rate $\Lambda^B$ given by

$$\Lambda^B \triangleq \lim_{t \to \infty} \frac{1}{t} \int_0^t \lambda^B(u) \, du,$$

where $\lambda^B(t) \triangleq (A(t) - \phi) 1\{Q^B(t) = B\}$ indicates the rate at which the buffer is overflowing at time $t$. We define the loss probability $\Lambda^B/\mathbb{E}A$ as the long time average fraction of fluid that is lost. Since there is a one to one correspondence between the loss rate and loss probability, we use the two terms interchangeably. An equivalent representation of $\Lambda^B$, which will be used for deriving our results, is $\Lambda^B = \mathbb{E} \lambda^B(t)$. 
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Similarly, the notation $\Lambda_A^{B,\phi} = \Lambda^B$ will be used to mark the explicit dependence of $\Lambda^B$ on $A(t)$ and $\phi$.

Next, we prove two useful sample path bounds. The first bound formalizes an intuitively expected notion that multiplexing reduces the aggregate queueing workload. Let $A_n(t)$ and $\phi_n$, $1 \leq n \leq N$, be arrival processes and service rates, respectively, with $A(t) = \sum_{n=1}^{N} A_n(t)$ and $\phi = \sum_{n=1}^{N} \phi_n$.

**Proposition 1.** If $Q_A^{B,\phi}(t) \leq \sum_{n=1}^{N} Q_{A_n}^{B,\phi_n}(t)$ for $t = 0$, then the inequality holds for all $t \geq 0$.

*Proof.* Let $0 = T_0 < T_1 < T_2 \cdots$ a.s. be the jump points in $A(t)$. Then, by the assumption and (2.1), the statement holds for any $t \in [0, T_1]$

$$Q_A^{B,\phi}(t) \leq \left( \sum_{n=1}^{N} \left( Q_{A_n}^{B,\phi_n}(0) + t(A_n(0) - \phi_n) \right) \right)^+ \wedge B$$

$$\leq \sum_{n=1}^{N} \left( Q_{A_n}^{B,\phi_n}(0) + t(A_n(0) - \phi_n) \right)^+ \wedge B = \sum_{n=1}^{N} Q_{A_n}^{B,\phi_n}(t),$$

where the last inequality follows from

$$\left( \sum_{n=1}^{N} x_n \right)^+ \wedge B \leq \left( \sum_{n=1}^{N} x_n^+ \right) \wedge B \leq \sum_{n=1}^{N} x_n^+ \wedge B. \quad \text{(2.2)}$$

Now, assume that the proposition holds for any $t \in [0, T_k], k \geq 1$. Hence, by the inductive assumption, (2.1) and (2.2), for any $t \in [T_k, T_{k+1}]$

$$Q_A^{B,\phi}(t) \leq \left( \sum_{n=1}^{N} \left( Q_{A_n}^{B,\phi_n}(T_k) + (t - T_k)(A_n(T_k) - \phi_n) \right) \right)^+ \wedge B$$

$$\leq \sum_{n=1}^{N} Q_{A_n}^{B,\phi_n}(t)$$

and, therefore, the result holds for all $t \geq 0$. \qed
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Next, we consider a stochastic process $Q_{\phi}^{\infty,A}(t)$ defined by the initial condition $Q_{\phi}^{\infty,A}(0)$ and

$$Q_{\phi}^{\infty,A}(t) = (Q_{\phi}^{\infty,A}(T_n) + (t - T_n)(\phi - A(T_n)))^+, \quad t \in (T_n, T_{n+1}], \quad n \geq 0. \quad (2.3)$$

Note that $Q_{\phi}^{\infty,A}(t)$ corresponds to an infinite buffer queueing process with constant arrival rate $\phi$ and service rate $A(t)$. We use $Q_{\phi}^{\infty,A}(t)$ to upper bound the amount of free buffer space, $B - Q_{A}^{\infty,B}(t)$, in the original system defined by (2.1).

**Lemma 5.** If $B - Q_{A}^{\infty,B}(t) \leq Q_{\phi}^{\infty,A}(t)$ for $t = 0$, then the inequality holds for all $t \geq 0$.

**Proof.** The proof is by induction and very similar to the proof of Proposition 1. From (2.1) for all $t \in (T_n, T_{n+1}]$, $n \geq 0$

$$Q_{A}^{\infty,B}(t) \geq \left(Q_{A}^{\infty,B}(T_n) + (t - T_n)(A(T_n) - \phi)\right) \land B,$$

and, therefore,

$$B - Q_{A}^{\infty,B}(t) \leq \left( B - Q_{A}^{\infty,B}(T_n) + (t - T_n)(\phi - A(T_n)) \right)^+.$$

The preceding inequality and the same arguments used in the proof of Proposition 1 imply the statement of the lemma. \qed

2.3 Fluid Queue with an On-Off Arrival Process

The results of this section characterize the asymptotic behavior of the finite buffer fluid queue fed by a single On-Off process. These results will be used for deriving
our main theorems in the subsequent section.

A stationary On-Off process $A(t)$ consists of a sequence of alternating independent On and Off periods. During the corresponding On and Off periods the process is equal to $A(t) = \tau$ and $A(t) = 0$. Successive On as well as Off periods are identically distributed and equal in distribution to $\tau$ and $\nu$, respectively. Random variables $\tau$ and $\nu$ have finite first moments and the process is in On state with probability $p = \mathbb{P}[A(t) = \tau] = \mathbb{E}\tau/(\mathbb{E}\tau + \mathbb{E}\nu)$. The average rate of the process is $\rho = \tau p$. For a detailed construction of such a process see e.g. [32].

The following proposition provides the asymptotic characterization of the overflow probability when the excess On period is in the subexponential class $S$.

**Proposition 2.** If $r > \phi > \rho$ and $\tau^{(e)} \in S$, then as $B \to \infty$

$$
\mathbb{P}[Q^B = B] \sim p \mathbb{P}\left[\tau^{(e)} > \frac{B}{r - \phi}\right].
$$

**Proof.** In [45] it was shown that $\Lambda^B \sim p(r - \phi)\mathbb{P}[\tau^{(e)} > B/(r - \phi)]$ as $B \to \infty$. Since $\Lambda^B = \mathbb{E}[(r - \phi)1\{Q^B = B\}] = (r - \phi)\mathbb{P}[Q^B = B]$ the statement holds. \hfill \Box

The next result characterizes the workload $Q^\infty \equiv Q_A^{\infty, \phi}$ in an infinite buffer system.

**Theorem 2 ([49]).** If $r > \phi > \rho$ and $\tau^{(e)} \in S$, then as $B \to \infty$

$$
\mathbb{P}[Q^\infty > B] \sim (1 - p)\frac{\rho}{\phi - \rho} \mathbb{P}\left[\tau^{(e)} > \frac{B}{r - \phi}\right].
$$

Note that quantities $\mathbb{P}[Q^B = B]$ and $\mathbb{P}[Q^\infty > B]$ are asymptotically proportional. We use this fact to obtain the following bound.
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Proposition 3. If \( r > \phi_i > \rho, \ i = 1,2 \) and \( \tau^{(e)} \in \mathcal{IR} \), then for \( 1 > \varepsilon > 0 \)

\[
\lim_{B \to -\infty} \frac{\mathbb{P}[Q_{A_{1}}^{B, \phi_1} \geq (1 - \varepsilon)B]}{\mathbb{P}[Q_{A_{2}}^{B, \phi_2} \geq \varepsilon B]} < \infty.
\]

Proof. Using sample path arguments it is easy to show that \( Q_{A_{1}}^{B, \phi} \) is stochastically dominated by \( Q_{A_{1}}^{\infty, \phi} \), and therefore

\[
\frac{\mathbb{P}[Q_{A_{1}}^{B, \phi} \geq (1 - \varepsilon)B]}{\mathbb{P}[Q_{A_{2}}^{B, \phi} \geq \varepsilon B]} \leq \frac{\mathbb{P}[Q_{A_{1}}^{\infty, \phi} \geq (1 - \varepsilon)B]}{\mathbb{P}[Q_{A_{2}}^{B, \phi} = B]}.
\]

Next, Proposition 2 and Theorem 2 yield

\[
\lim_{B \to -\infty} \frac{\mathbb{P}[Q_{A_{1}}^{\infty, \phi_1} \geq (1 - \varepsilon)B]}{\mathbb{P}[Q_{A_{2}}^{B, \phi_2} = B]} \leq \frac{(1 - p)\rho}{(\phi_1 - \rho)p} \lim_{B \to -\infty} \frac{\mathbb{P}[\tau^{(e)} > \frac{(1-\varepsilon)B}{\tau - \phi_1}]}{\mathbb{P}[\tau^{(e)} > \frac{B}{\tau - \phi_2}]} < \infty,
\]

where the last inequality is implied by Lemma 1. \( \square \)

The proposition below is the main technical result of this section. The proof is straightforward but tedious and, hence, it is deferred to Section 2.7.

Proposition 4. If \( r > \phi > \rho \) and \( \tau^{(e)} \in \mathcal{IR} \), then

\[
\lim_{\varepsilon \downarrow 1} \lim_{B \to -\infty} \frac{\mathbb{P}[Q_{A}^{B} \geq \varepsilon B]}{\mathbb{P}[Q_{A}^{B} = B]} = 1.
\]

2.4 Finite Buffer Fluid Queue

This section contains the main results of this chapter stated in Theorems 3 and 4. The theorems describe the asymptotic behavior of a finite buffer fluid queue fed by
CHAPTER 2. FINITE BUFFER FLUID QUEUE

$N$ independent On-Off processes $A_i(t), 1 \leq i \leq N$. Process $A_i(t)$ has On periods equal in distribution to $\tau_i$. Its peak rate, average rate and probability of being On are equal to $r_i$, $\rho_i$ and $p_i$, respectively. For convenience we define $n \triangleq \{1, 2, \ldots, N\}$ and the complement of $i \subseteq n$ is denoted by $-i \equiv n \setminus i$. Let for any collection of variables $X_i, 1 \leq i \leq N$ and $i \subseteq n$, $X_i$ stands for $\sum_{i \in I} X_i$.

Our proofs require the following minor technical assumption. Similar assumptions can be found in [41, 61, 98].

Assumption 1. The capacity of the system satisfies $r_n > \phi > \rho_n$ and $\phi \neq r_i + \rho_{-i}$ for all $i \subseteq n$.

Remark 1. (i) The first part of the assumption states that the queue is stable and that overflows are possible. (ii) If the second part of the assumption is not satisfied, by choosing an arbitrarily larger or lower capacity one can obtain a lower or upper bound on the queueing performance, respectively. The assumption ensures that the queue is not critically stable during periods of time when some of the processes have long On periods.

Before stating and proving our main results we introduce two preparatory lemmas. The first lemma derives an asymptotic expression for the overflow probability in the case when all processes need to be in the active state for a long period of time in order to have a buffer overflow. The proof makes use of Proposition 1. Although the bound provided by the proposition is weak, by combining $N$ weak bounds one obtains the desired result.
Lemma 6. If \( r_n - r_i + \rho_i < \phi < r_n \) for all \( 1 \leq i \leq N \), then for all \( B \geq 0 \) and \( 0 \leq \varepsilon \leq 1 \)

\[
\prod_{i=1}^{N} p_i \mathbb{P}\left[ \tau_i^{(c)} > \frac{\varepsilon B}{r_n - \phi} \right] \leq \mathbb{P}[Q_A^{B,\phi} \geq \varepsilon B] \leq \prod_{i=1}^{N} \mathbb{P}[Q_{A_i}^{B,\phi-r_n+r_i} \geq \varepsilon B].
\]

If in addition \( \tau_i^{(c)} \in S \) for \( 1 \leq i \leq N \), then as \( B \to \infty \)

\[
\mathbb{P}[Q_A^{B,\phi} = B] \sim \prod_{i=1}^{N} p_i \mathbb{P}\left[ \tau_i^{(c)} > \frac{B}{r_n - \phi} \right].
\]

Proof. Assume that at time \( t = 0 \) all the considered queues are empty. For all \( 1 \leq i \leq N \) Proposition 1 yields

\[
Q_A^{B,\phi}(t) = Q_{A_i}^{B,\phi-r_n+r_i}(t) + Q_{A-A_i}^{B,\phi-r_n-r_i}(t)
\]

where the equality follows from the fact that \( A(t) - A_i(t) \leq r_n - r_i \) for all \( t \) and, therefore, \( Q_{A-A_i}^{B,\phi-r_n-r_i}(t) \equiv 0, t \geq 0 \). Since (2.4) holds for all \( i \), then

\[
Q_A^{B,\phi}(t) \leq \bigwedge_{i=1}^{N} Q_{A_i}^{B,\phi-r_n+r_i}(t),
\]

which, by applying the operator \( \mathbb{P}[\cdot \geq \varepsilon B] \), using the independence of \( A_i \) and passing \( t \to \infty \), yields in stationarity

\[
\mathbb{P}[Q_A^{B,\phi} \geq \varepsilon B] \leq \prod_{i=1}^{N} \mathbb{P}[Q_{A_i}^{B,\phi-r_n+r_i} \geq \varepsilon B].
\]

Obtaining the lower bound is straightforward from evaluating the system in station-
arity at (say) $t = 0$; for simplicity the time index is omitted

\[
P[Q_{A}^{B,\phi} \geq \varepsilon B] \geq P \left[ \bigcap_{i=1}^{N} \left\{ A_i = \tau_i, \; \tau_i^{(e)} > \frac{\varepsilon B}{\tau_n - \phi} \right\} \right]
\]

\[
= P \left[ \bigcap_{i=1}^{N} \left\{ A_i = \tau_i, \; \tau_i^{(e)} > \frac{\varepsilon B}{\tau_n - \phi} \right\} \right]
\]

\[
= \prod_{i=1}^{N} p_i P \left[ \tau_i^{(e)} > \frac{\varepsilon B}{\tau_n - \phi} \right].
\]

By setting $\varepsilon = 1$ in the preceding upper and lower bounds and combining it with Proposition 2, we obtain the second statement of the proposition. \qed

The following set plays a crucial rule in establishing the main performance metrics of the fluid queue with a finite number of On-Off processes.

**Definition 7.** The minimum overflow set is defined as

\[
O \triangleq \{i \subseteq n : r_1 + \rho_{-1} - r_i + \rho_i < \phi < r_1 + \rho_{-1}, \forall i : i \in i\}.
\]

**Remark 2.** (i) Informally, the motivation behind this definition comes from the fact that only a few On-Off processes with long On periods are causing the most likely buffer overflows, while the remaining processes exhibit their average behavior. Hence, an element of $O$ indicates which processes need to have long On periods in order for a buffer overflow to occur. (ii) The definition of $O$ is analogous to the definition of the minimal set in [32].

Similarly, we define an underflow set $U$ of the combinations that do not cause an overflow

\[
U \triangleq \{i \subseteq n : r_1 + \rho_{-1} < \phi\}.
\]
CHAPTER 2. FINITE BUFFER FLUID QUEUE

This set does not appear in the statements of the main results but plays a role in their proof.

At this point, we are ready to state our last preparatory lemma.

**Lemma 7.** If $X_i \in IR$, $1 \leq i \leq N$, are independent, then as $x \to \infty$

$$
P \left[ \bigwedge_{i \in O} \sum_{i \in I} X_i > x \right] = o \left( \sum_{i \in I} \prod_{i \in I} P[X_i > x] \right).
$$

**Proof.** Observe that for all $j \in O \cup U$

$$
\left\{ \bigcap_{i \in I} \{ X_i > x/N \} \right\} \bigcap \left\{ \bigcap_{i \in I} \{ X_i \leq x/N \} \right\} \bigcap \left\{ \bigwedge_{i \in O \cup U} \sum_{i \in I} X_i > x \right\} = \emptyset, \quad (2.5)
$$
since on the first two events $\sum_{i \in I} X_i \leq x$. Next, $(2.5)$ implies

$$
P \left[ \bigwedge_{i \in O \cup U} \sum_{i \in I} X_i > x \right] \leq P \left[ \bigcup_{i \in O \cup U} \bigcap_{i \in I} \{ X_i > x/N \} \right],
$$

which yields by independence of $\{X_i\}$

$$
P \left[ \bigwedge_{i \in O \cup U} \sum_{i \in I} X_i > x \right] \leq \sum_{i \in O \cup U} \prod_{i \in I} P[X_i > x/N].
$$

The lemma follows from the preceding inequality, Lemma 1 and the definitions of $O$ and $U$, which imply that for every $i \notin O \cup U$ there exists $j \in O$ such that $j \subset i$. $\square$

At last, we arrive at our first main result.

**Theorem 3.** Let $\tau_i^{(e)} \in IR$ for $1 \leq i \leq N$ and

$$
\hat{P}(B) \triangleq \sum_{i \in O} \prod_{i \in I} p_i^{(e)} \left[ \tau_i^{(e)} > \frac{B}{\tau_i + \rho - \phi} \right].
$$
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Then, under Assumption 1,

$$\lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q_{A, \phi}^B \geq B - K]}{\hat{P}(B)} = \lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q_{A, \phi}^B \geq B - K]}{\hat{P}(B)} = 1.$$  

If, in addition, we assume $r_1 > \phi$, for all $i \in \mathcal{O}$, then for any $K > 0$

$$\mathbb{P}[Q_{A, \phi}^B \geq B - K] \sim \mathbb{P}[Q_{A, \phi}^B = B] \sim \hat{P}(B) \text{ as } B \to \infty.$$  

Remark 3. 
(i) Informally, the double limit implies that $\mathbb{P}[Q_{A, \phi}^B \geq B - K] \approx \hat{P}(B)$ for large $K$ and $B$ much larger than $K$. Hence, the result states that the fraction of time during which the buffer is effectively 100% full is asymptotically equal to $\hat{P}(B)$. (ii) The heuristic for this result can be explained by the following example.

Consider two i.i.d. On-Off processes with excess On periods in $\mathcal{T}_R$ and $r < \phi < r + \rho$. These assumptions result in the overflow set $\mathcal{O} = \{1\}, \{2\}$. In this case, the most probable way the buffer overflows is when one of the processes (say the first one) has a very long On period and the other behaves on average, i.e., $\int_0^1 A_2(u) \, du \approx \rho t$. During that long On period, the average amount of arriving fluid will be higher than the service rate, $r + \rho > \phi$, and the buffer will tend to fill. After the buffer fills, its content will stay close to the buffer boundary. When $r < \phi$, the queueing content will make small excursions away from the boundary during the Off periods in the second On-Off process, see Figure 2.1. In the proof we show that these excursions are almost surely finite and uniformly bounded for all $B$. (iii) In the last statement of the theorem, the values of $\rho_i$-s do not affect the computation of the minimal overflow set. Hence, during the most likely overflow event the arrival rate
Figure 2.1: Illustration for Remark 3 (ii). The long On period is shown with a dashed line.

is always higher than the capacity and, therefore, the buffer content $Q^B$ remains on the boundary $B$. This fact makes the asymptotic approximation of the probability that the buffer is full $P[Q^B = B]$ feasible. Also, due to the fluid nature of the model, $P[Q^B = B]$ represents the fraction of time that fluid is being lost. (iv) With additional assumptions on the ratios of tails of $\tau_i^{(e)}$ the minimum overflow set $\mathcal{O}$ in the statement of the theorem can be replaced by a smaller overflow set $\mathcal{O}_0$, which asymptotically yields the same value for $\hat{P}(B)$. For example, if $\tau_i^{(e)} \in R_{\alpha_i}$, then $\mathcal{O}_0 = \{i \in \mathcal{O} : \alpha_i = \bigwedge_{j \in \mathcal{O}} \alpha_j\}$.

Proof. Upper bound. For $\delta > 0$ consider queues $Q^{B,\phi-EA_{i-1}+\delta}_{A_i}$, $Q^{B,\rho_i+\delta/N}_{A_i}$ assuming that they are empty at time $t = 0$. For any $i \in \mathcal{O} \cup \mathcal{U}$ Proposition 1 yields

$$Q^{B,\phi}_{A_i}(t) \leq Q^{B,\phi-EA_{i-1}+\delta}_{A_i}(t) + \sum_{i \in \mathcal{I}} Q^{B,\rho_i+\delta/N}_{A_i}(t),$$

and, thus

$$Q^{B,\phi}_{A_i}(t) \leq \bigwedge_{i \in \mathcal{O} \cup \mathcal{U}} \left( Q^{B,\phi-EA_{i-1}+\delta}_{A_i}(t) + \sum_{i \in \mathcal{I}} Q^{B,\rho_i+\delta/N}_{A_i}(t) \right).$$

(2.6)

Next, by selecting sufficiently small $\delta$, such that all the queues in the preceding inequality have their capacities greater than the average arrival rates, applying the
operator \( \mathbb{P} [ \cdot \geq B - K ] \) in (2.6) and then passing \( t \to \infty \), we derive in stationarity

\[
\mathbb{P} [ Q_A^{B,\phi} \geq B - K ] \leq \mathbb{P} \left[ \bigwedge_{i \in \mathcal{O} \cup \mathcal{U}} \left( Q_{A_i}^{B,\phi - \mathcal{E}A_i - \delta} + \sum_{i \in \mathcal{I}} Q_{A_i}^{B,\rho_i + \delta} \right) \geq B - K \right].
\]

Now, let us select \( \delta < \bigwedge_{i \in \mathcal{U}} (\phi - \tau_i - \rho_i) \), such that \( Q_{A_i}^{B,\phi - \mathcal{E}A_i - \delta} \equiv 0 \) for all \( i \in \mathcal{U} \).

Then, the preceding inequality and union bound yield for \( 0 < \varepsilon < 1 \)

\[
\mathbb{P} [ Q_A^{B,\phi} \geq B - K ] \leq \mathbb{P} \left[ \bigcup_{i \in \mathcal{O}} \left\{ Q_{A_i}^{B,\phi - \mathcal{E}A_i - \delta} \geq \varepsilon (B - K) \right\} \right]
\]

\[
+ \mathbb{P} \left[ \bigwedge_{i \in \mathcal{O} \cup \mathcal{U}} \sum_{i \in \mathcal{I}} Q_{A_i}^{B,\rho_i + \delta} \geq (1 - \varepsilon) (B - K) \right]
\]

\[
\leq \sum_{i \in \mathcal{O}} \mathbb{P} \left[ Q_{A_i}^{B,\phi - \mathcal{E}A_i - \delta} \geq \varepsilon (B - K) \right]
\]

\[
+ \mathbb{P} \left[ \bigwedge_{i \in \mathcal{O} \cup \mathcal{U}} \sum_{i \in \mathcal{I}} Q_{A_i}^{B,\rho_i + \delta} \geq (1 - \varepsilon) (B - K) \right]
\]

\[
\leq (1 + o(1)) \sum_{i \in \mathcal{O}} \prod_{i \in \mathcal{I}} \mathbb{P} \left[ Q_{A_i}^{B,\phi_i + \tau_i - \delta} \geq \varepsilon (B - K) \right], \quad (2.7)
\]

as \( B \to \infty \), where \( \phi_i \triangleq \phi - \tau_i - \rho_i \) and the last inequality is due to Lemmas 6, 7 and Proposition 3. Here, by recalling that \( \tau_i^{(e)} \in \mathcal{I} \mathcal{R} \), one obtains from Propositions 2 and 4 for all \( i \in \mathcal{O} \) and \( i \in \mathcal{I} \)

\[
\lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 1} \lim_{B \to \infty} \frac{\mathbb{P} \left[ Q_{A_i}^{B,\phi_i + \tau_i - \delta} \geq \varepsilon (B - K) \right]}{\mathbb{P} \left[ Q_{A_i}^{B,\phi_i + \tau_i} = B \right]} = 1,
\]

which, by (2.7), Proposition 2 and Lemma 3, yields for any \( K \geq 0 \)

\[
\lim_{B \to \infty} \frac{\mathbb{P} [ Q_A^{B,\phi} \geq B - K ]}{\hat{P}(B)} \leq 1.
\]

**Lower bound for the first statement.** The lower bound is obtained by observing the queueing system in stationarity at (say) time \( t = 0 \). For any \( \varepsilon > 0 \) and all \( i \in \mathcal{O} \),
define an event indicating that all the processes \( A_i(t) \) with \( i \in \mathcal{i} \) are in the active state at time \( t = 0 \) and their On periods have lasted for an amount of time larger then \( t_1 \triangleq (1 + \varepsilon)B/(\tau_1 + \rho_{-1} - \phi) \), i.e.,

\[
\Psi_i \triangleq \bigcap_{i \in \mathcal{i}} \{ A_i(0) = r_i, \ inf\{ t > 0 : A_i(-t) = 0 \} > t_1 \} .
\]

(2.8)

We point out that \( \inf\{ t > 0 : A_i(-t) = 0 \} \) is equal in distribution to \( \tau_i^{(e)} \) on event \( \{ A_i(0) = r_i \} \). In a similar way, we define a corresponding event \( \Upsilon_i \) indicating that processes \( A_i(t) \) with \( i \not\in \mathcal{i} \) do not have long On periods at time \( t = 0 \), i.e.,

\[
\Upsilon_i \triangleq \bigcap_{i \not\in \mathcal{i}} \{ A_i(0) = r_i, \ inf\{ t > 0 : A_i(-t) = 0 \} > t_1 \} .
\]

Now, \( Q_A^{B,\phi}(-t_1) \) being nonnegative and Proposition 1 imply \( Q_A^{B,\phi}(0) \geq Q_{A_{-1}}^{B,\phi}(0) \) on event \( \Psi_i \), where \( Q_{A_{-1}}^{B,\phi}(-t_1) = 0 \). Then, since for all different \( i \in \mathcal{O} \) events \( \{ \Psi_i \cap \Upsilon_i \} \) are disjoint, by Lemma 5 one obtains

\[
P[Q_A^{B,\phi}(0) \geq B - K] \geq \sum_{i \in \mathcal{O}} P[Q_A^{B,\phi}(0) \geq B - K, \ Upsilon_i] \\
\geq \sum_{i \in \mathcal{O}} P[Q_{\phi-r_i}^{\infty,A_{-1}}(0) \leq K, \ Upsilon_i] \\
\geq \left( \bigwedge_{i \in \mathcal{O}} P[Q_{\phi-r_i}^{\infty,A_{-1}}(0) \leq K, \ Upsilon_i] \right) \sum_{i \in \mathcal{O}} P[\Psi_i],
\]

(2.9)

where \( Q_{\phi-r_i}^{\infty,A_{-1}} \) is defined by recursion (2.3) and the initial condition \( Q_{\phi-r_i}^{\infty,A_{-1}}(-t_1) = B \); the last inequality follows from the independence of \( A_i \) and \( A_{-1} \). The preceding inequality and \( P[\Upsilon_i] \rightarrow 1 \) as \( B \rightarrow \infty \) lead to

\[
\lim_{B \rightarrow \infty} \frac{P[Q_A^{B,\phi} \geq B - K]}{P(B)} \geq \lim_{B \rightarrow \infty} \bigwedge_{i \in \mathcal{O}} P[Q_{\phi-r_i}^{\infty,A_{-1}}(0) \leq K] \lim_{B \rightarrow \infty} \frac{\sum_{i \in \mathcal{O}} P[\Psi_i]}{P(B)}.
\]
At this point, by recalling the definition of $\Psi_i$, using the fact that $\tau_i(\epsilon) \in \mathcal{I}_K$, Lemma 3 and passing $\epsilon \downarrow 0$ in the preceding inequality, we obtain

\[
\lim_{B \to -\infty} \frac{\mathbb{P}[Q^B_A \geq B - K]}{\hat{P}(B)} \geq \lim_{\epsilon \to 0} \lim_{B \to -\infty} \bigwedge_{i \in O} \mathbb{P}[Q^{\infty,A_{-1}}_{\phi_{-r_i}}(0) \leq K]. \quad (2.10)
\]

Finally, using the standard queueing reflection mapping argument, quantity $Q^{\infty,A_{-1}}_{\phi_{-r_i}}(0)$ can be represented as

\[
Q^{\infty,A_{-1}}_{\phi_{-r_i}}(0) = \sup_{-t_i \leq s \leq 0} \left\{ (\phi - r_i)|s| - \int_s^0 A_{-1}(u) \, du \right\} \lor \left( B + (\phi - r_i)t_i - \int_{-t_i}^0 A_{-1}(u) \, du \right),
\]

where $\lor$ denotes the maximum. Then, the stationarity of $A_{-1}$ leads to

\[
\mathbb{P}[Q^{\infty,A_{-1}}_{\phi_{-r_i}}(0) \leq K] \geq \mathbb{P} \left[ \sup_{s \leq 0} \left\{ (\phi - r_i)|s| - \int_s^0 A_{-1}(u) \, du \right\} \leq K \right]
- \mathbb{P} \left[ t_i^{-1} \int_0^{t_i} A_{-1}(u) \, du - E A_{-1} \leq -\frac{\epsilon}{1+\epsilon} (\phi - r_i + \rho_{-1}) \right] \quad (2.11)
\]

and by the facts that the supremum in the first term is equal to the workload in a stable queue and that process $A_{-1}$ is stationary and ergodic one obtains

\[
\lim_{K \to -\infty} \lim_{\epsilon \to 0} \lim_{B \to -\infty} \mathbb{P}[Q^{\infty,A_{-1}}_{\phi_{-r_i}}(0) \leq K] = 1. \quad (2.12)
\]

The preceding limit and (2.10) yield the lower bound for the first statement.

*Lower bound for the second statement.* Note that in this case $\Psi_i \subseteq \{Q^B_A(0) = B\}$ and, therefore, (2.9) simplifies to

\[
\mathbb{P}[Q^B_A(0) = B] \geq \sum_{i \in O} \mathbb{P}[\Psi_i].
\]
By dividing the preceding inequality with \( \hat{P}(B) \), taking \( \lim \) as \( B \to \infty \), recalling (2.8), using \( \tau_i^{(e)} \in \mathcal{IR} \) and letting \( \varepsilon \downarrow 0 \) we obtain the lower bound for the second statement and conclude the proof of the theorem.

Our second primary result characterizes the asymptotic behavior of the average loss rate.

**Theorem 4.** If \( \tau_i^{(e)} \in \mathcal{IR} \) for \( 1 \leq i \leq N \), then under Assumption 1 as \( B \to \infty \)

\[
\Lambda^B \sim \hat{\Lambda}(B) \triangleq \sum_{i \in \mathcal{O}} (r_i + \rho - \phi) \prod_{i \in \mathcal{I}} p_i \mathbb{P} \left[ \tau_i^{(e)} > \frac{B}{r_i + \rho - \phi} \right].
\]

**Remark 4.** (i) Recall that the loss probability is computable from \( \Lambda^B / E A \). (ii) A related result for a discrete time finite buffer queue with a Pareto-like M/G/\( \infty \) arrival process can be found in [61]. In their proofs the authors exploit the Poisson decomposition property of the arrival processes, which does not hold for the multiplexed On-Off processes. In addition, in [61] it is assumed that the buffer overflows in a unique way.

**Proof.** Since the proof is very similar to the proof of Theorem 3, we omit some details.

**Upper bound.** Let \( \delta > 0 \) be sufficiently small, such that the queues \( Q_{A_i}^{B,\phi-EA_{i-1}-\delta} \), \( i \in \mathcal{O} \cup \mathcal{U} \), \( Q_{A_i}^{B,\rho_i+\delta/N} \) have their service rates greater than the mean arrival rates, and \( Q_{A_i}^{B,\phi-EA_{i-1}-\delta} \equiv 0 \) for all \( i \in \mathcal{U} \). Then, recalling the definition of \( \phi_i = \phi - r_i - \rho_i \),
(2.6) yields for $0 < \varepsilon < 1$

$$
\Lambda^B = \mathbb{E}[(A - \phi)1\{Q^{B,\phi}_{A_1} = B\}] \\
\leq \sum_{i \in \mathcal{O}} \mathbb{E}\left[(A - \phi)1\{Q^{B,\phi-E_{A_{n-1}}-\delta}_{A_{n}} \geq \varepsilon B\}\right] \\
\quad + (r_n - \phi) \mathbb{P}\left[\bigwedge_{i \in \mathcal{O} \cup \mathcal{U}} Q^{B,E_{A_{n-1}}+\delta}_{A_{n}} \geq (1 - \varepsilon)B\right] \\
\leq (1 + o(1)) \sum_{i \in \mathcal{O}} (r_i + \rho_{n-1} - \phi) \prod_{i \in \pi} \mathbb{P}\left[Q^{B,\phi_1+\rho_{n-1}-\delta}_{A_i} \geq \varepsilon B\right],
$$

where the last inequality follows from the independence of arrival processes, Lemmas 6, 7 and Proposition 3. By dividing both sides of the preceding expression with $\hat{\Lambda}(B)$, taking $\lim$ as $B \to \infty$, and then passing $\varepsilon \uparrow 1$ and $\delta \downarrow 0$, we obtain the upper bound.

Lower bound. Assume that all processes are in their stationary regimes. Note that for all $T > 0$

$$
\Lambda^B = \mathbb{E}\lambda^B(0) = T^{-1} \int_0^T \lambda^B(u) \, du
$$

and recall the definitions of events $\Psi_1$, $\Upsilon_1$ and the initial condition for $Q^{\infty,A_{n-1}}_{\phi-r_{n-1}}$ from the proof of the lower bound in Theorem 3. Then, by using $B - Q^{B,\phi}_A(0) \leq Q^{\infty,A_{n-1}}_{\phi-r_{n-1}}(0)$

on event $\Psi_1$, we derive

$$
\Lambda^B \geq T^{-1} \sum_{i \in \mathcal{O}} \mathbb{E}\left[\int_0^T \lambda^B(u) \, du \, 1\{Q^{B,\phi}_A(0) \geq B - K, \, \Upsilon_1, \, \Psi_1\}\right] \\
\quad \geq T^{-1} \sum_{i \in \mathcal{O}} \mathbb{E}\left[\left(-K + \int_0^T A(u) \, du - \phi T\right) \, 1\{Q^{\infty,A_{n-1}}_{\phi-r_{n-1}}(0) \leq K, \, \Upsilon_1, \, \Psi_1\}\right] \\
\quad \triangleq \sum_{i \in \mathcal{O}} L_i. \tag{2.13}
$$
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Next, let \( t_i \triangleq \bigwedge_{i \in \mathcal{I}} \inf \{ t > 0 : A_i(t) = 0 \} \), i.e., \( t_i \) is the first time after \( t = 0 \) that one of the processes with a large On period is equal to zero. By the independence of arrival processes we lower bound \( L_i \) for every \( i \in \mathcal{O} \) as follows

\[
L_i \geq -(KT^{-1} + \phi)\mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K, \ \tau_i] \mathbb{P}[^{\Psi_i}]
+ T^{-1} \mathbb{E} \left[ \int_0^T A_i(u) \, du \, 1\{Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K, \ \tau_i \} \right] \mathbb{P}[^{\Psi_i}]
+ r_i T^{-1} \mathbb{E} \left[ (T \wedge t_i) 1\{\Psi_i\} \right] \mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K, \ \tau_i],
\]

(2.14)

where in the last term we used \( \int_0^T A_i(u) \, du \geq (T \wedge t_i) r_i \). Now, for all finite \( T \), due to \( \tau_i^{(c)} \in \mathcal{I} \mathcal{R} \subset \mathcal{L} \) for all \( i \) and the independence of arrival processes, it follows that

\[
\lim_{B \to -\infty} \frac{\mathbb{E} \left[ (T \wedge t_i) 1\{\Psi_i\} \right]}{\mathbb{P}[\Psi_i]} \geq T \lim_{B \to -\infty} \mathbb{P}[T < t_i | \Psi_i] = T.
\]

Inequality (2.14), together with the preceding limit and \( \mathbb{P}[\Psi_i] \to 1 \) as \( B \to \infty \) implies

\[
\lim_{B \to -\infty} \frac{L_i}{\mathbb{P}[\Psi_i]} \geq (-KT^{-1} - \phi + r_i + \rho_{-i}) \lim_{B \to -\infty} \mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K]
- r_i \lim_{B \to -\infty} \mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) > K].
\]

Hence, by setting \( T = K^2 \), in view of (2.11), for any \( \delta > 0 \) there exist large enough \( K \) and \( B_\delta \), such that uniformly for all \( B \geq B_\delta \) and \( i \in \mathcal{O} \)

\[
L_i \geq (1 - \delta) \mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K](r_i + \rho_{-i} - \phi)\mathbb{P}[\Psi_i],
\]

which, when replaced in (2.13), yields

\[
\Lambda^B \geq (1 - \delta) \left( \bigwedge_{i \in \mathcal{O}} \mathbb{P}[Q_{\phi - \tau_i}^{\infty, A_i}(0) \leq K] \right) \sum_{i \in \mathcal{O}} (r_i + \rho_{-i} - \phi)\mathbb{P}[\Psi_i].
\]
CHAPTER 2. FINITE BUFFER FLUID QUEUE

By dividing the preceding equation with $\tilde{\Lambda}(B)$ and taking $\lim$ as $B \to \infty$ we derive

$$\lim_{B \to \infty} \frac{\Lambda^B}{\tilde{\Lambda}(B)} \geq (1 - \delta) \lim_{B \to \infty} \left( \bigwedge_{i \in \mathcal{O}} \mathbb{P}[Q_{\phi^{-1}}(0) \leq K] \right) \lim_{B \to \infty} \frac{\sum_{i \in \mathcal{O}} (r_i + \rho_i - \phi) \mathbb{P}[\Psi_i]}{\tilde{\Lambda}(B)},$$

which after passing $\varepsilon \downarrow 0$ and using Lemma 3 results in

$$\lim_{B \to \infty} \frac{\Lambda^B}{\tilde{\Lambda}(B)} \geq (1 - \delta) \lim_{\varepsilon \downarrow 0} \lim_{B \to \infty} \left( \bigwedge_{i \in \mathcal{O}} \mathbb{P}[Q_{\phi^{-1}}(0) \leq K] \right).$$

Finally, by setting first $\delta \downarrow 0$, recalling (2.12) and then setting $K \to \infty$ the lower bound follows. This concludes the proof of the theorem. \(\square\)

For the case of homogeneous arrival processes, the expressions for the loss rate and overflow probability admit the following simple forms since every element in the overflow set is of cardinality one. In this case it is easy to see that both $\hat{P}(B)$ and $\Lambda^B$ are decreasing exponentially in the capacity $\phi$ and only polynomially in the buffer size $B$.

**Corollary 1.** Homogeneous sources. Let

$$\hat{P}(B) \triangleq \binom{N}{m} \left( p \mathbb{P} \left[ \tau^{(\phi)} > \frac{B}{mr + (N - m)\rho - \phi} \right] \right)^m.$$

If $\rho N < \phi < r N$, $\tau^{(\phi)} \in \mathcal{T}_r$ and there is an integer $m \geq 1$ such that

$$0 < mr + (N - m)\rho - \phi < r - \rho,$$

then

$$\lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q^{B,\phi}_A \geq B - K]}{\hat{P}(B)} = \lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q^{B,\phi}_A \geq B - K]}{\hat{P}(B)} = 1,$$
and as $B \to \infty$

$$
\Lambda^B \sim (mr + (N - m)\rho - \phi)\hat{P}(B).
$$

If in addition $mr > \phi$, then $P[Q^B_A \geq B - K] \sim P[Q^B_A = B] \sim \hat{P}(B)$ as $B \to \infty$ for all $K \geq 0$.

Next, we allow for some of the multiplexed arrival processes to have lighter than polynomial tails; we term these processes subpolynomial. A stationary, ergodic and right-continuous process $A(t)$ is subpolynomial ($A \in SP$) if for all $\phi > \mathbb{E}A(t)$ and $\beta > 0$ the stationary workload of the corresponding infinite buffer queue $Q^\infty_A$ satisfies

$$
\lim_{B \to \infty} B^\beta P[Q^\infty_A \geq B] = 0.
$$

This is satisfied for a general class of exponentially bounded arrival processes (see [37, 89]) as well as for some heavy-tailed processes, e.g., On-Off processes with Weibull On periods, $P[\tau > x] = e^{-x^b}$, $0 < b < 1$, $x \geq 0$ (see Theorem 2). Note that if $A_1, A_2 \in SP$ then $A_1 + A_2 \in SP$. This easily follows from the well known fact that $Q^\infty_{A_1 + A_2}$ is stochastically dominated by $Q^\infty_{A_1} + Q^\infty_{A_2}$, $\phi_i > \mathbb{E}A_i$ (an infinite buffer equivalent of Proposition 1). Thus, we will use $A_{SP}$ to denote the aggregate process of all subpolynomial arrival processes. The following corollary yields the reduce load equivalence results for multiplexing subpolynomial and intermediately regularly varying processes.

**Corollary 2.** Suppose that $A_{SP} \in SP$ and Assumption 1 is satisfied with $(\phi - \mathbb{E}A_{SP})$
in place of $\phi$. If $\tau_i^{(e)} \in \mathcal{IR}$ for $1 \leq i \leq M$, then

$$
\lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q_{A+\lambda}^{B,\phi} \geq B - K]}{\mathbb{P}[Q_{A}^{B,\lambda - \mathbb{E}_{A_{SP}}} \geq B - K]} = \lim_{K \to \infty} \lim_{B \to \infty} \frac{\mathbb{P}[Q_{A+\lambda}^{B,\phi} \geq B - K]}{\mathbb{P}[Q_{A}^{B,\lambda - \mathbb{E}_{A_{SP}}} \geq B - K]} = 1,
$$

and, if for some $\delta > 0$, $\lambda_{A_{SP}}^{1+\delta} < \infty$, then as $B \to \infty$

$$
\Lambda_{A+\lambda}^{B,\phi} \sim \Lambda_{A}^{B,\lambda - \mathbb{E}_{A_{SP}}}.
$$

Proof. First, by stochastic dominance, for any $\phi > \mathbb{E}_{A_{SP}}$, $\mathbb{P}[Q_{A_{SP}}^{\phi \leftarrow B} \geq B - K]$ $\geq$ $\mathbb{P}[Q_{A_{SP}}^{B,\phi} \geq B - K]$, and therefore, for any $\beta > 0$

$$
\lim_{B \to \infty} B^\beta \mathbb{P}[Q_{A_{SP}}^{B,\phi} \geq B - K] = 0. \quad (2.15)
$$

Then, by Proposition 1, for any $0 < \delta < \phi - \mathbb{E}_{A_{SP}}$ and $0 < \epsilon < 1$

$$
\mathbb{P}[Q_{A+\lambda}^{B,\phi} \geq B - K] \leq \mathbb{P}[Q_{A}^{B,\lambda - \mathbb{E}_{A_{SP}} - \delta} \geq \epsilon(B - K)] + \mathbb{P}[Q_{A_{SP}}^{B,\lambda + \delta} \geq (1 - \epsilon)(B - K)]. \quad (2.16)
$$

Next, recall the definition of $\hat{P}(B) \equiv \hat{P}(B, \phi')$, $\phi' = \phi - \mathbb{E}_{A_{SP}}$ from Theorem 3. Clearly, $\hat{P}(B)$ belongs to $\mathcal{IR}$, and thus, there exists a finite $\alpha$ such that for all sufficiently large $B$

$$
\hat{P}(B) \geq B^{-\alpha}; \quad (2.17)
$$

see equation (1.6) of [79]. Now, by dividing (2.16) with $\hat{P}(B)$, taking $\lim$ as $B \to \infty$, using Theorem 3, (2.15) and (2.17), and then passing $\epsilon \uparrow 1$, $\delta \downarrow 0$ we complete the proof of the upper bound.

The upper bound for the loss rate is obtained by using the same approach and,
instead of (2.16),

\[
\Lambda_{A+A_{SP}}^{B,\phi} = \mathbb{E}[(A + A_{SP} - \phi)1\{Q_{A+A_{SP}}^{B,\phi - E_{A_{SP}} - \delta} \geq \varepsilon B\}]
\]

\[
\leq \mathbb{E}[(A + A_{SP} - \phi)1\{Q_{A_{SP}}^{B,\phi - E_{A_{SP}} - \delta} \geq (1 - \varepsilon)B\}]
\]

+ \mathbb{E}[(A + A_{SP} - \phi)1\{Q_{A_{SP}}^{B,\phi - E_{A_{SP}} - \delta} \geq \varepsilon B\}]

\leq \mathbb{E}[(A + \mathbb{E}A_{SP} - \phi)1\{Q_{A_{SP}}^{B,\phi - E_{A_{SP}} - \delta} \geq (1 - \varepsilon)B\}]

+ (\mathbb{E}A - \phi) \mathbb{P}[Q_{A_{SP}}^{B,\phi - E_{A_{SP}} + \delta} \geq (1 - \varepsilon)B]

+ (\mathbb{E}A_{SP}^\delta + \mathbb{E}A_{SP}^\delta)^{\frac{1}{1+\delta}} \left( \mathbb{P}[Q_{A_{SP}}^{B,\phi - E_{A_{SP}} + \delta} \geq (1 - \varepsilon)B] \right)^{\frac{1}{1+\delta}} ,
\]

where the last inequality follows from the independence of \( A \) and \( A_{SP} \), and Hölder's inequality.

The proofs of the lower bounds can be obtained in the same spirit as in Theorems 3 and 4. These proofs only require that \( A_{SP} \) satisfies the Strong Law of Large Numbers, which follows from the stationarity and ergodicity. To avoid repetition we omit the details. \( \square \)

### 2.5 Service Differentiation

In the previous section we analyzed the main performance measures of a single server queue fed by multiple processes. If each of these processes represents a user, then a natural question to look into is the system performance observed by individual users. As it is often observed, a satisfactory overall performance does not imply a satisfactory performance observed by individual users.
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Recall for the previous section that $N$ On-Off processes share a common server of capacity $\phi$ and a buffer space of size $B$. Let $W_i(t)$ be the unfinished work of process $A_i$ at time $t$, $\sum W_i(t) \leq B$. The server capacity is distributed among processes according to the Generalized Processor Sharing (GPS) scheme. Each process $A_i$ is assigned a weight $\varphi_i > 0$ such that $\sum_{i=1}^{N} \varphi_i = 1$. Weight $\varphi_i$ represents the guaranteed share of the server capacity for flow $A_i$. Available excess capacity is redistributed among processes according to the GPS weights $\varphi_i$. Service rates $\phi_i(t)$ for each flow $i$ at time $t$ can be computed by using a recursive algorithm described in [33]. Let $E(t)$ be the set of processes with $W_i(t) = 0$, which are receiving service at rate $\phi_i(t) = A_i(t)$, i.e. $E(t) \triangleq \{i : W_i(t) = 0, \phi_i(t) = A_i(t)\}$. Then, it is not difficult to see that $E(t) = \{i : W_i(t) = 0\}$ almost everywhere (a.e.) Lebesgue. Therefore, using this property and the characteristics of GPS for any process $i \not\in E(t)$, rate $\phi_i(t)$ satisfies

$$\phi_i(t) = \frac{\varphi_i \left( \phi - \sum_{j \in E(t)} A_j(t) \right)}{\sum_{j \not\in E(t)} \varphi_j} \geq \varphi_i \phi \quad \text{a.e.} \quad (2.18)$$

Buffer sharing is unrestricted as long as there is available space, i.e., the workloads evolve as if they were in the infinite buffer system. When the buffer fills up, the processes with the maximum amount of fluid $W_i(t)$ in the buffer are subject to penalty. They will experience the minimum necessary loss of fluid such that the processes with smaller workloads can be accommodated. We exemplify this policy for two On-Off processes with equal GPS weights in Figure 2.2. Possible extensions to more general buffer sharing policies will be briefly discussed at the end of this sec-
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Figure 2.2: An illustration of the buffer management policy. Two On-Off processes with peak rate $r$ share a server of capacity $\phi = r/2$ and buffer of size $B$. The GPS weights of the two processes are equal, $\varphi_1 = \varphi_2 = 0.5$.

More formally, following the approach from [33], the evolution of $W_i$'s can be described with a set of differential equations. In order to account for the finiteness of the buffer space we define

$$D(t) = \left\{ i : W_i(t) \leq \sum_{j=1}^{N} W_j(t), \sum_{j=1}^{N} W_j(t) = B \right\}$$
CHAPTER 2. FINITE BUFFER FLUID QUEUE

with $|D(t)|$ denoting the cardinality of $D(t)$; note that $D(t)$ is nonempty only if the buffer is full. The elements of $D(t)$ are processes that could potentially experience losses at time $t$. Let $M(t)$ be the largest subset of $D(t)$ such that for all $i \in M(t)$ the following inequality holds

$$A_i(t) - \phi_i(t) > -|M(t)|^{-1} \sum_{j \notin M(t)} (A_j(t) - \phi_j(t)).$$

The workloads of the processes $A_i(t)$ with $i \in M(t)$ are regulated according to

$$\dot{W}_i(t) = -|M(t)|^{-1} \sum_{j \notin M(t)} (A_j(t) - \phi_j(t));$$

this adjustment is necessary to ensure $\sum W_i = B$. Note that $\dot{W}_i(t)$ can be both positive and negative. For example, consider a system of three process with equal GPS weights. Let $W_1(0) = W_2(0) = 2B/5$, $W_3(0) = B/5$, $A_1(0) = A_2(0) = \phi$ and $A_3(0) = 0$. Then, $M(0) = \{1, 2\}$ and, hence, $\dot{W}_1(0) = \dot{W}_2(0) = \phi/6$, i.e., the workloads of the first two processes are increasing at rate $\phi/6$. However, without any adjustment the workloads would increase at rate $2\phi/3$.

Next, similarly as for the set $E(t)$, we observe that $M(t) = D(t)$ a.e. (Lebesgue).

Thus, the evolution of the individual workloads is a.e. described by the following set of differential equations

$$\dot{W}_i(t) =
\begin{cases}
0 & \text{if } i \in E(t), \\
-|D(t)|^{-1} \sum_{j \notin D(t)} (A_j(t) - \phi_j(t)) & \text{if } i \in D(t), \\
A_i(t) - \phi_i(t) & \text{otherwise}.
\end{cases}
$$

(2.19)
CHAPTER 2. FINITE BUFFER FLUID QUEUE

The functions $W_i(t)$ are absolutely continuous and therefore sets of Lebesgue measure zero can be ignored for purpose of their characterization [33]. Hence, the system of equations (2.19) completely defines the behavior of the system. We assume that this system of equations has a unique stationary solution and, unless otherwise specified, $W_i(t)$ will be used to denote this solution. Let $A_i$, $\phi_i$, $W_i$ be random variables equal in distribution to $A_i(t)$, $\phi_i(t)$, $W_i(t)$ in stationarity, respectively.

The main objective is the asymptotic computation of the long-term average loss rate for a given process. We say that process $i$ is overflowing at time $t$ if $A_i(t) - \phi_i(t) > \hat{W_i}(t)$. The instantaneous loss rate process for process $A_i(t)$ is defined as

$$\gamma_i^B(t) \triangleq A_i(t) - \phi_i(t) - \hat{W_i}(t),$$

and its expected loss rate is equal to $\Gamma_i^B = \mathbb{E}\gamma_i^B(t)$. Note that the buffer management policy implies

$$\left\{A_i(t) - \phi_i(t) > \hat{W_i}(t) \right\} \subseteq \left\{W_i(t) \geq B/N, \sum_{i=1}^{N} W_i(t) = B \right\}.$$ 

This inclusion is a simple consequence of the fact that if the buffer is full and $W_i(t) < B/N$, then there exists a process $j$ with higher workload than that of process $i$, and, therefore, process $i$ can not experience losses. It is worth mentioning that the described queueing system is work- and buffer-conserving. Work-conservation follows from the properties of GPS; buffer-conservation stands for no process experiencing loss of fluid unless the buffer is full.
CHAPTER 2. FINITE BUFFER FLUID QUEUE

The following sample path bound relates the amount of unfinished work of process $A_i(t)$ in the GPS system and the workload of a queue in which process $i$ is served in isolation.

**Proposition 5.** If $W_i(t) \leq Q_{A_i}^{B,\phi}(t)$ for $t = 0$, then the inequality holds for all $t \geq 0$.

*Proof.* Follows from inequality (2.18) and the fact that the portion of the buffer available to $W_i(t)$ is not greater than $B$. □

Finally, we are ready to state our main result. It establishes the buffer equivalence between the GPS system and a finite buffer queue in which a process is served in isolation. For any set of variables $\{X_i\}$ let $X_{-i} \overset{\Delta}{=} X_n - X_i$.

**Theorem 5.** Let $\phi_j > \rho_j$ for all flows. If $r_i^{(c)} \in I\mathcal{R}$ and $r_i > \phi - \rho_{-i}$ for some flow $i$, then as $B \to \infty$

$$\Gamma_i^B \sim A_{A_i}^{B,\phi-\rho_{-i}}.$$

The strict stability $\phi_j > \rho_j$ for all processes ensures that the workload build-up of process $i$ is unlikely to be caused by other processes. This fact and the heavy-tailed nature of On periods of process $i$ result in the most likely overflow scenario being due to a single long On period in process $i$. Therefore, during an overflow, with very high probability all other processes exhibit average behavior, while process $i$ transmits at its peak rate $r_i$. This implies that the buffer fills up at rate $r_i + \rho_{-i} - \phi$. 
In addition, the average behavior of processes $j \neq i$ yields $W_j = O(1)$, and, thus, process $i$ can potentially occupy up to $B - O(1)$ buffer space. This heuristic is made rigorous in the following proof that consists of an upper and lower bound. The proof of the upper bound is based on the fact that during the described events the only process that experiences losses is $i$. Equivalently, the total losses in the system are equal to the losses of process $i$

$$\gamma^B_i(t) \approx (r_i + A_{-i} - \phi)_1 \{\text{overflow due to } i\}.$$ 

The lower bound is conceptually easier, yet more tedious. The event of having process $i$ experiencing losses is intersected with events that process $i$ has a long On period and all other processes exhibit its average behavior and have $O(1)$ workloads.

**Proof. Upper bound.** Assume that all queueing processes below are equal to zero at time $t = 0$. Based on (2.20) of the instantaneous loss rate and its non-negativity, the following holds

$$\gamma^B_i(t) = (A_i(t) - \phi_i(t) - \hat{W}_i(t))1\{A_i(t) - \phi_i(t) > \hat{W}_i(t)\}$$

$$\leq (r_i + A_{-i}(t) - \phi)_1 A_i(t) - \phi_i(t) > \hat{W}_i(t),$$  

(2.22)

where the last inequality follows from the fact that the buffer is full on event \{$A_i(t) - \phi_i(t) > \hat{W}_i(t)$\} and the instantaneous loss rate of a single flow is upper bounded by the total loss rate in the system $A(t) - \phi \leq r_i + A_{-i}(t) - \phi$. Inequality (2.22), inclusion (2.21), Propositions 5 and the work-conserving property of the
GPS scheduling scheme yield for all \( t \geq 0 \)

\[
\gamma_i^B(t) \leq (r_i + A_{-i}(t) - \phi)1\{W_i(t) \geq B/N, \sum_{i=1}^{N} W_i(t) \geq B\} \\
\leq (r_i + A_{-i}(t) - \phi)1\{Q_{A_i}^{B,\phi}\rho_i^*(t) \geq B/N, Q_{A_i}^{B,\phi}(t) \geq B\}.
\]

Next, the preceding inequality and Proposition 1 result in

\[
\gamma_i^B(t) \leq (r_i + A_{-i}(t) - \phi)1\{Q_{A_i}^{B,\phi-\rho_i-\delta}(t) \geq \varepsilon B\} \\
+ (r_n - \phi)1\{Q_{\infty}^{\infty,\rho_n}(t) \geq B/N\}1\{Q_{\infty,\rho_n}(t) \geq 1 - \varepsilon B\}
\]

where we set \( \varepsilon \in (0, 1) \) and \( \delta \in (0, \phi - \rho_n) \). Note that for any choice of \( \varepsilon \) and \( \delta \) in the given intervals, all queueing processes in the last inequality are stable and converge in distribution to proper random variables as \( t \to \infty \). Now, by independence of arrival processes, the last inequality renders

\[
\Gamma_i^B \leq (r_i + \rho_{-i} - \phi)\mathbb{P}[Q_{A_i}^{B,\phi-\rho_{-i}-\delta} \geq \varepsilon B] \\
+ (r_n - \phi)\mathbb{P}[Q_{\infty}^{\infty,\rho_n} \geq B/N]\mathbb{P}[Q_{\infty,\rho_n} \geq 1 - \varepsilon B].
\]

Then, Theorem 2, Proposition 2 and Lemma 1 imply

\[
\lim_{B \to \infty} \frac{\Gamma_i^B}{Q_{A_i}^{B,\phi-\rho_{-i}}} \leq \lim_{B \to \infty} \frac{\mathbb{P}[Q_{A_i}^{B,\phi-\rho_{-i}-\delta} \geq \varepsilon B]}{\mathbb{P}[Q_{A_i}^{B,\phi-\rho_{-i}} = B]}.\]

To complete the proof of the upper bound first pass \( \varepsilon \uparrow 1 \), then \( \delta \downarrow 0 \) and use Proposition 4.

**Lower bound.** This part of the proof is very similar to the proof of the lower bound in Theorem 4. Assume that all processes are in their stationary regimes
unless otherwise specified. Recall that the loss rate is defined as $\mathbb{E} \gamma^B_i(t)$, where $\gamma^B_i(t)$ is the instantaneous loss rate process of flow $i$. Clearly, for any $0 < T < \infty$

$$\Gamma^B_i = \mathbb{E} \gamma^B_i(0) = T^{-1} \mathbb{E} \int_0^T \gamma^B_i(t) \, dt.$$ 

Next, recall the definitions of events $\Psi_{\{i\}} \equiv \Psi_i$, $\Upsilon_{\{i\}} \equiv \Upsilon_i$ from the proof of Theorem 4. Then the loss rate for flow $i$ can be lower bounded by

$$\Gamma^B_i \geq \frac{1}{T} \mathbb{E} \left[ \int_0^T \gamma^B_i(t) \, dt \left\{ \Psi_i, \Upsilon_i, \bigcap_{j \neq i} \left\{ W_j(0) \leq \frac{B}{2N} \right\}, W_n(0) \geq B - K \right\} \right]. \quad (2.23)$$

Now, by using Lemma 5 and recalling that the considered system is work- and buffer-conserving we obtain

$$\left\{ \Psi_i, \Upsilon_i, \bigcap_{j \neq i} \left\{ W_j(0) \leq \frac{B}{2N} \right\}, K \geq B - W_n(0) \right\}$$

$$\geq \left\{ \Psi_i, \Upsilon_i, \bigcap_{j \neq i} \left\{ W_j(0) \leq \frac{B}{2N} \right\}, K \geq \mathbb{Q}_{\phi - r_i}^{\infty}(0) \right\}, \quad (2.24)$$

where $\mathbb{Q}_{\phi - r_i}^{\infty}(t)$ denotes the workload of an infinite buffer queue with constant arrival rate $\phi - r_i$, service rate $A_{-i}(t)$ and the initial condition $\mathbb{Q}_{\phi - r_i}^{\infty}(-t_i) = B$.

Next, for all $B > 2NT \tau_n$ on event $\Upsilon_{\{i\}}$ the workloads of all processes other than $i$ are smaller than $B/N$ for all $t \in (0, T)$ and, therefore, those processes do not experience loss of fluid. This leads to

$$\int_0^T \gamma_i(t) \, dt = \int_0^T \sum_{j=1}^N \gamma_j(t) \, dt \geq \int_0^T A_n(t) \, dt - \phi T - K,$$

where the last summand is due to the fact that at time $t = 0$ the buffer can accommodate an additional $K$ units of fluid (see (2.23)). Hence, the preceding inequality,
CHAPTER 2. FINITE BUFFER FLUID QUEUE

(2.23) and (2.24) yield for all \( B > 2NT\kappa_n \)

\[
\Gamma_i^B \geq \mathbb{E} \left[ \left( -KT^{-1} + T^{-1} \int_0^T A_\kappa(u) \, du - \phi \right) \times \mathbf{1} \left\{ \Psi_i, T_i, \bigcap_{j \neq i} \left\{ W_j(0) \leq \frac{B}{2N} \right\}, K \geq Q_{\phi-r_i}^{\infty,A-i}(0) \right\} \right].
\]

The rest of the proof closely follows steps of the lower bound proof of Theorem 4 after (2.13). To avoid repetition we omit the details. \( \square \)

At this point we would like to discuss possible generalizations of the preceding theorem. The strict stability condition \( \varphi_i \phi > \rho_i \) for all \( i \) represents a natural engineering condition. However, from a theoretical perspective the behavior of the system remains unclear if one or more processes have higher average demands than their minimum guaranteed rates. The following corollary represents an easy extension of this type. It states that a process with guaranteed service rate lower than its expected rate will not be asymptotically affected by other processes if the tail of the On period of that process is sufficiently heavy.

**Corollary 3.** Let \( \tau_j^{(e)} \in \mathcal{I}\mathcal{R} \) for \( 1 \leq j \leq N \) and \( \tau_i + \rho_{-i} > \phi > \rho_n \) for some flow \( i \). If \( \varphi_j \phi > \rho_j \) for all \( j \neq i \) and \( \prod_{j \neq i} \mathbb{P}[\tau_j^{(e)} > x] = o \left( \mathbb{P}[\tau_i^{(e)} > x] \right) \) for all \( i \leq n \) that satisfy \( r_i + \rho_{-i} \geq \phi \), then as \( B \to \infty \)

\[
\Gamma_i^B \sim \Lambda_{A_i}^{B_\phi-\rho_{-i}}.
\]

**Proof.** The upper bound is a direct consequence of inclusion (2.21) and Theorem 4. The proof of the lower bound is the same as in Theorem 5. \( \square \)
CHAPTER 2. FINITE BUFFER FLUID QUEUE

Careful examination of the proof of Theorem 5 shows that it holds for much more general buffer management policies. In particular, this includes the GPS-like rules for buffer management, which can be utilized to improve the behavior of the system for small workload sizes. More generally, Theorem 5 is expected to hold for all buffer-conserving policies under which the guaranteed buffer space for each process is an increasing function of the total buffer size.

It is worth mentioning that when condition \( r_i + \rho_{-i} > \phi \) fails, the most likely overflow for process \( i \) may be caused jointly by process \( i \) and a subset of other processes. Unfortunately, it appears that the expression for the loss rate in the general case depends on the buffer management policy as well as the parameters of other processes. Hence, the answers and analysis under these assumptions are expected to be more involved.

The main novelty of our result is that the analyzed system behaves as if it had \( N \) times larger buffer than it actually did. The result indicates that the buffer sharing can be very beneficial in the case of heavy-tailed traffic. This insight provides additional guideline for deciding on the buffer distribution between the access and core network switching elements. Furthermore, we observe that the result raises an interesting problem of socially fair buffer pricing. Everybody gets a full share of the resource, but it is not quite clear who and by how much one should pay for it.
2.6 Numerical Examples

In this section we illustrate through simulation experiments the accuracy of derived asymptotic results in approximating the overflow probabilities and loss rates for finite buffer sizes. Since the asymptotic results are insensitive to the distribution of Off periods we choose their distribution to be exponential $\mathbb{P}[\nu > x] = e^{-\lambda x}, x \geq 0$; On periods are selected from Pareto family $\mathbb{P}[r > x] = x^{-\alpha}, x \geq 1, \alpha > 1$. We select $\alpha$ in the range of measured file sizes ($\alpha = 1.44$, see Figure 1.1). The asymptotic approximation from Corollary 1 computes explicitly to

$$\hat{P}(B) = \binom{N}{m} \left[ \frac{p}{\alpha B^{\alpha-1}} (mr + (N - m)\rho - c)^{\alpha-1} \right]^m,$$

where $p = \lambda\alpha/ (\lambda\alpha + \alpha - 1)$ and $\rho = rp$. To ensure the accuracy of our simulation experiments we select the length of the simulated sample path to be $t = 10^{10}$.

Example 1. Here, we select $N = 10$ i.i.d. On-Off processes with parameters $\lambda = 0.012, \alpha = 1.3$ and $r = 2$, which yield $p = 0.05$ and $\rho = 0.1$. For the choice of capacity $c = 5$ we simulated the overflow probabilities for buffer sizes $B = 100i, i = 1, \ldots, 10$. The results of the simulation are presented in Figure 2.3 with "+" symbols. The selected parameters render $m = 3$ for the asymptotic approximation $\hat{P}(B)$, as defined in (2.25). Note that $mr > c$ and, therefore, we can use the last statement of Corollary 1 to approximate $\mathbb{P}[Q^B = B]$. The accuracy of the approximation, plotted on the same figure with dashed lines, is apparent.

Example 2. In this example we choose $N = 50, \lambda = 3.37 \times 10^{-3}, \alpha = 1.5, r = 3$, which
imply $p = 0.01$ and $\rho = 0.03$. Now, for the same capacity $c = 5$, we readily compute $m = 2$, the asymptotic formula $\hat{\Lambda}(B) = (mr + (N - m)\rho - c)\hat{P}(B)$ and repeat the same simulation procedure as in Example 1. The results of the simulation and approximation are plotted with "+" symbols and dashed lines, respectively. Again, the approximation matches well the simulated estimates even for relatively high probabilities.

Next, consider a fluid queue with capacity $c = 2.5$ and finite buffer $B$ shared by five On-Off flows. As earlier in this section, we choose $\mathbb{P}[\nu > x] = e^{-\nu x}$, $x \geq 0$ and $\mathbb{P}[r > x] = x^{-\alpha}$, $x \geq 1$, $\alpha > 1$. The peak rates of On-Off flows are defined by $r_1 = 4$, $r_2 = r_3 = 2$, $r_4 = 3$, $r_5 = 1$ and the On probabilities are chosen to be $p_i = 0.1$ for all five flows. The length of the simulated sample path in both examples is set to $10^9$.

Example 3. Let the distributions of On periods be defined by $\alpha_1 = \alpha_3 = \alpha_5 = 1.6$, $\alpha_2 = \alpha_4 = 1.5$. The work of the GPS mechanism is fully determined by the weights
\( \phi_1 = \phi_3 = 0.3, \phi_2 = \phi_5 = 0.1 \) and \( \phi_4 = 0.2 \). Clearly, the conditions of Theorem 5 are satisfied for the first flow. For this flow we simulated the loss rates for buffer sizes \( B = 100, 200, \ldots, 1000 \). The results of the simulation are presented in Figure 2.5 with "o" symbols. The approximation of the loss rate is plotted in the same figure with a solid line.

**Example 4.** Here, consider the system from the previous example with \( \alpha_1 = 1.6, \alpha_2 = \alpha_3 = 3.0, \alpha_4 = 3.4, \alpha_5 = 1.9 \) and \( \phi_1 = \phi_5 = 0.1, \phi_2 = 0.2, \phi_3 = \phi_4 = 0.3 \). In this example it is easy to verify that the conditions of Corollary 3 are satisfied for the first flow. The simulation results ("o" symbols) and the approximation (solid line) for the loss rate are plotted on Figure 2.6.

It is apparent from both figures that the derived approximations are in agreement with the simulated results. Furthermore, we would like to point out that the asymptotic formulas are quite precise even though the estimated probabilities are
relatively high, in range of $10^{-2} - 10^{-3}$.

2.7 Proofs

This section contains the proof of Proposition 4. The proof is based on the subsequent three lemmas which derive preliminary results on a discrete time finite buffer queue, defined as follows. Consider two i.i.d. sequences of positive random variables \( \{X, X_n\} \) and \( \{Y, Y_n\}, n \in \mathbb{N} \). Let \( W_0^B = 0 \) and

\[
W_{n+1}^B = \left((W_n^B + X_{n+1}) \wedge B - Y_{n+1}\right)^+.
\] (2.26)

Assuming that \( P[X_n = Y_n] < 1 \), in Chapter III.5 of [28] it was shown that \( W_n^B \) has a unique stationary distribution, and that for any initial condition \( W_0^B, W_n^B \) converges to that stationary distribution. Let \( W^B \) be a random variable that is equal in distribution to \( W_n^B \) in stationarity.
Lemma 8. If $X^{(e)} \in \mathcal{IR}$ and $\mathbb{E} X < \mathbb{E} Y$, then

$$\lim_{\delta \to 1} \lim_{B \to \infty} \frac{\mathbb{P}[W^B \geq \delta B]}{\mathbb{P}[X^{(e)} \geq B]} = 0.$$ 

Proof. Instead of proving the statement for $W^B$ directly, we will consider an easier to analyze queueing variable $V^B$ that stochastically upper bounds $W^B$. Let $\{V^B_n\}_{n=0}^\infty$ be defined by $V^B_0 = 0$ and

$$V^B_{n+1} = (V^B_n + X_{n+1} - Y_{n+1})^+ \land B.$$ 

The above recursion is similar to (2.26) and under the non-triviality condition $\mathbb{P}[X_n = Y_n] < 1$, $V^B_n$ converges to a stationary distribution (see Chapter III.4 of [28]) again, let $V^B$ be a random variable that is equal in distribution to $V^B_n$ in stationarity. Now, we show that $W^B_n \leq V^B_n$ for all $n \geq 0$. Clearly, this is implied by
$W_0^B = V_0^B = 0$ and the next inductive step

\[
W_{n+1}^B = \left( (W_n^B + X_{n+1}) \land B - Y_{n+1} \right)^+ \\
\leq (V_n^B + X_{n+1} - Y_{n+1})^+ \land B = V_{n+1}^B,
\]

where the inequality follows from

\[(x \land B - y)^+ \leq (x \land B - y \land B)^+ \leq (x - y)^+ \land B,
\]

for $x, y \geq 0$. Therefore, it will be enough to prove the statement of the lemma with $W^B$ being replaced by $V^B$.

First, we restrict our attention to the case of $\{X_n\}_{n=1}^\infty$, $\{Y_n\}_{n=1}^\infty$ being lattice valued and $Y$ bounded. Without loss of generality we assume that $X_n$ and $Y_n$ are integer valued. Let $q_i^B \triangleq \mathbb{P}[V^B = i]$ and $\eta_i^B \triangleq q_i^B / q_0^B$ for $0 \leq i \leq B \leq \infty$, $B \in \mathbb{N}$ ($B = \infty$ represents the infinite buffer case). Lemma 1 from [45] yields for all $B \geq 0$

\[
\mathbb{P}[V^B \geq \delta B] \leq \sum_{i=\lfloor \delta B \rfloor}^B q_0^B \eta_i^B \leq \sum_{i=\lfloor \delta B \rfloor}^B \eta_i^B \\
\leq \sum_{i=\lfloor \delta B \rfloor}^B \left( \eta_i^\infty + K_2 \mathbb{P}[V^B + X - Y > B] \mu^{B-i} \right),
\]

where $K_2$ is a positive constant, $0 < \mu < 1$ if $\mathbb{P}[Y_n = 0] + \mathbb{P}[Y_n = 1] < 1$, and $\mu = 0$, otherwise. Next, the preceding inequality and Lemma 2 from [45] imply

\[
\mathbb{P}[V^B \geq \delta B] \leq \frac{1}{q_0^\infty} \mathbb{P}[\delta B \leq V^\infty \leq \delta B] + 1\{\mu \neq 0\} \frac{K_2}{1-\mu} \left( \mathbb{P}[X^c > B] \right).
\]

Thus,

\[
\lim_{B \to \infty} \frac{\mathbb{P}[V^B \geq \delta B]}{\mathbb{P}[X^c > B]} \leq \frac{1}{q_0^\infty} \left( \lim_{B \to \infty} \frac{\mathbb{P}[V^\infty \geq \delta B]}{\mathbb{P}[X^c > B]} - \lim_{B \to \infty} \frac{\mathbb{P}[V^\infty \geq B]}{\mathbb{P}[X^c > B]} \right).
\]
which, by Pakes’ theorem [75] and \( X^{(e)} \in IR \) results in

\[
\lim_{\delta \downarrow 1} \lim_{B \to \infty} \frac{P[V^B \geq \delta B]}{P[X^{(e)} \geq B]} = 0.
\]

This proves the result for lattice valued \( X \) and \( Y \), with \( Y \) being bounded. Next, we use the technique from [45, pp.98-99] to extend the result to the general case of non-lattice valued \( X, Y \) with \( Y \) unbounded.

If \( Y \) is unbounded we can always choose a truncated service variable \( Y_n^d = Y_n \wedge d \), with \( d \) being sufficiently large to satisfy the stability condition \( \mathbb{E}X_n < \mathbb{E}Y_n^d \). Let \( W_n^{B,d} \) be a process satisfying recursion (2.26) with the process \( Y_n^d \) in place of \( Y_n \).

It is clear that the stationary workload for this queue \( W_n^{B,d} \) is stochastically larger than the original workload \( W_n^B \), i.e., \( P[W_n^B \geq \delta B] \leq P[W_n^{B,d} \geq \delta B] \). Therefore, since the lemma holds for process \( W_n^{B,d} \), it is true for \( W_n^B \) as well. When \( X \) and \( Y \) are non-lattice, we can approximate them with lattice valued random variables \( X' \) and \( Y' \). For any \( \Delta > 0 \) such that \( \mathbb{E}Y - \mathbb{E}X > 2\Delta < 0 \), define distribution functions for \( Y' \) and \( X' \) as

\[
P[Y' = i\Delta] = P[i\Delta \leq Y < (i + 1)\Delta], \quad i \geq 0,
\]

\[
P[X' = i\Delta] = P[(i - 1)\Delta \leq X < i\Delta], \quad i \geq 1.
\]

From these definitions it follows that for all \( x \geq 0 \)

\[
P[Y > x + \Delta] \leq P[Y' > x] \leq P[Y > x],
\]

\[
P[X > x] \leq P[X' > x] \leq P[X > x - \Delta],
\]
which implies that $X' - Y'$ is stochastically larger than $X - Y$, $\mathbb{E}X' \leq \mathbb{E}X + \Delta < \mathbb{E}Y - \Delta \leq \mathbb{E}Y'$, and

$$\int_B^\infty \mathbb{P}[X' > u] \, du \sim \int_B^\infty \mathbb{P}[X > u] \, du \quad \text{as} \quad B \to \infty. \quad (2.27)$$

Next, let $\{X'_n\}_{n=1}^\infty$ and $\{Y'_n\}_{n=1}^\infty$ be two independent i.i.d. sequences with probability distributions equal to distributions of $X'$ and $Y'$, respectively. Consider a queue $W'^B$ with buffer $B$ that corresponds to sequences $\{X'_n\}_{n=1}^\infty$ and $\{Y'_n\}_{n=1}^\infty$. Then, the newly constructed queueing process, $W'^B$, dominates the original queueing process in distribution

$$\mathbb{P}[W'^B \geq \delta B] \leq \mathbb{P}[W'^B \geq \delta B].$$

Finally, the preceding inequality and (2.27) imply

$$\lim_{B \to \infty} \frac{\mathbb{P}[W^B \geq \delta B]}{\mathbb{P}[X^{(e)} \geq B]} \leq \lim_{B \to \infty} \frac{\mathbb{P}[W'^B \geq \delta B]}{\mathbb{P}[X'^{e} \geq B]},$$

which, by using the already proved lattice case and letting $\delta \uparrow 1$ yields the desired result. This concludes the proof. \hfill \Box

**Lemma 9.** If $X^{(e)}$ is independent of $W^B$, $X^{(e)} \in \mathcal{IR}$ and $\mathbb{E}X < \mathbb{E}Y$, then

$$\lim_{B \to \infty} \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}[X^{(e)} + W^B \geq \varepsilon B]}{\mathbb{P}[X^{(e)} \geq B]} = 1.$$

**Proof.** Let $\varepsilon \in (0, 1)$. For all $\delta \in (0, \varepsilon/2)$ a simple argument leads to

$$\mathbb{P}[X^{(e)} + W^B \geq \varepsilon B]$$

$$\leq \mathbb{P}[X^{(e)} \geq (\varepsilon - \delta)B] + \mathbb{P}[X^{(e)} + W^B \geq \varepsilon B, X^{(e)} < (\varepsilon - \delta)B]$$

$$\leq \mathbb{P}[X^{(e)} \geq (\varepsilon - \delta)B] + \frac{1}{\mathbb{E}X} \int_0^{(\varepsilon - \delta)B} \mathbb{P}[W^B \geq \varepsilon B - x] \mathbb{P}[X \geq x] \, dx. \quad (2.28)$$
Next, we bound the second term in (2.28) as follows

\[
\int_0^{\delta B} \mathbb{P}[W^B \geq \varepsilon B - x] \mathbb{P}[X \geq x] \, dx \leq \mathbb{E}X \mathbb{P}[W^B \geq (\varepsilon - \delta)B],
\]

and

\[
\int_{\delta B}^{(\varepsilon - \delta)B} \mathbb{P}[W^B \geq \varepsilon B - x] \mathbb{P}[X \geq x] \, dx \leq \mathbb{P}[W^B \geq \delta B](\varepsilon - 2\delta)B \mathbb{P}[X \geq \delta B],
\]

which together with Lemma 2 results in

\[
\lim_{B \to \infty} \int_0^{(\varepsilon - \delta)B} \frac{\mathbb{P}[W^B \geq \varepsilon B - x] \mathbb{P}[X \geq x] \, dx}{\mathbb{P}[X^{(\varepsilon)} \geq B]} \leq \mathbb{E}X \lim_{B \to \infty} \frac{\mathbb{P}[W^B \geq (\varepsilon - \delta)B]}{\mathbb{P}[X^{(\varepsilon)} \geq B]} \tag{2.29}
\]

By dividing (2.28) with \(\mathbb{P}[X^{(\varepsilon)} \geq B]\), taking \(\lim\) with respect to \(B\), and using (2.29) we obtain

\[
\lim_{B \to \infty} \frac{\mathbb{P}[X^{(\varepsilon)} + W^B \geq \varepsilon B]}{\mathbb{P}[X^{(\varepsilon)} \geq B]} \leq \lim_{B \to \infty} \frac{\mathbb{P}[X^{(\varepsilon)} \geq (\varepsilon - \delta)B] + \mathbb{P}[W^B \geq (\varepsilon - \delta)B]}{\mathbb{P}[X^{(\varepsilon)} \geq B]}.
\]

Hence, by letting \(\delta \downarrow 0, \varepsilon \uparrow 1\) and invoking Lemma 8 in the preceding inequality the desired statement follows.

\[\square\]

**Lemma 10.** Let \(X, Y^{(\varepsilon)}\) and \(W^B\) be mutually independent. If \(X^{(\varepsilon)} \in \mathcal{IR}\) and \(\mathbb{E}X < \mathbb{E}Y\), then

\[
\lim_{\varepsilon \downarrow 1} \lim_{B \to \infty} \frac{\mathbb{P}[(W^B + X) \wedge B - Y^{(\varepsilon)} \geq \varepsilon B]}{\mathbb{P}[X^{(\varepsilon)} \geq B]} = 0.
\]

**Proof.** Let \(\varepsilon \in (0, 1)\). For all \(\delta \in (0, \varepsilon)\) we write

\[
\mathbb{P}[(W^B + X) \wedge B - Y^{(\varepsilon)} \geq \varepsilon B] \leq \mathbb{P}[W^B + X \geq \varepsilon B] \leq \mathbb{P}[W^B \geq (\varepsilon - \delta)B] + \mathbb{P}[X \geq \delta B],
\]
which, jointly with Lemmas 2 and 8, leads to
\[
\lim_{\varepsilon \downarrow 1} \lim_{B \to \infty} \frac{\mathbb{P}[X^{(e)} + W^B \geq \varepsilon B]}{\mathbb{P}[X^{(e)} \geq B]} \leq \lim_{\varepsilon \downarrow 1} \lim_{B \to \infty} \frac{\mathbb{P}[X \geq \delta B] + \mathbb{P}[W^B \geq (\varepsilon - \delta)B]}{\mathbb{P}[X^{(e)} \geq B]} = 0.
\]

At this point we are able to provide a proof of Proposition 4.

Proof of Proposition 4. Let \(X_{n+1} = (r - \phi)\tau_{n+1}, Y_{n+1} = \phi \nu_{n+1}\) in (2.26) and assume that \(W^B_n\) is in its stationary regime. Then, \(W^B_n\) represents the amount of fluid in a queue with a single On-Off arrival process observed at the beginnings of On periods. Thus, by evaluating \(Q^B(t)\) in stationarity at time (say) \(t = 0\) we derive (for simplicity of notation we leave out the time index)
\[
\mathbb{P}[Q^B \geq \varepsilon B] = \mathbb{P}[A = 0, (W^B + (r - \phi)\tau) \wedge B - \phi \nu^{(e)} \geq \varepsilon B] \\
+ \mathbb{P}[A = \tau, W^B + (r - \phi)\tau^{(e)} \geq \varepsilon B] \\
= (1 - p)\mathbb{P}[W^B + (r - \phi)\tau \wedge B - \phi \nu^{(e)} \geq \varepsilon B] \\
+ p\mathbb{P}[W^B + (r - \phi)\tau^{(e)} \geq \varepsilon B].
\]

Then, by dividing the preceding equality with \(\mathbb{P}[(r - \phi)\tau^{(e)} \geq B]\), applying the operator \(\lim_{\varepsilon \downarrow 1} \lim_{B \to \infty}\), and using Lemmas 9, 10 and Proposition 2 we complete the proof. \(\square\)
Chapter 3

Moderately Heavy Tails

3.1 Introduction

Fair-share algorithms play an important role in designing modern systems, e.g., majority of kernel schedulers in operating systems are designed with a notion of fairness in mind. The processor sharing scheduling discipline is a prime example of a policy that has fairness build in it. Unfortunately, the exact meaning of "fair" in a specific context is not always easy to define. A possible example of desired properties in the context of file transfers over a single link could be, for example, (i) file transfers are as short as possible, and (ii) when a file transfer is long it is due to the fact that the file being transferred is long, i.e., not due to other transfers. Ideally, these properties would hold for every transfer but one should hope that they hold at least on "average". The primary question considered here
CHAPTER 3. MODERATELY HEAVY TAILS

is the conditions on the input probability distribution under which the desirable
properties hold under low-complexity sharing algorithms, e.g., processor sharing. In
the cases when low-complexity algorithms do not achieve desirable results one needs
to resort to increased complexity of the system, e.g., implement priority mechanisms
or reservation schemes. As it is shown in this chapter, the Weibull distribution $e^{-\sqrt{x}}$
defines the taxonomy in qualitative system behavior.

First, we study probabilities of large deviations for sums of subexponential ran-
dom variables for which the distribution function decays faster than any polynomial.
This question is central to understanding many important problems of probability
theory and has been extensively investigated over the years, originating with the
classical results of [42, 69–71]. Recently, in [58], the authors consider large devi-
tions of random renewal sums of variables with polynomially decaying distributions;
see also [58] for additional references on large deviations of heavy-tailed sums.

In the second part of the chapter we investigate problems that involve moderately
heavy-tailed random variables considered in the first part.

Integrated service networks are capable of carrying various types of traffic in-
cluding voice and video. Provisioning network links or implementing admission
control in order to assure required QoS guarantees is a challenging task. It is quite
clear that proper decisions whether to admit a session or not depend on stochastic
characteristics of the session. It is known that video streams exhibit heavy-tailed
behavior while voice streams are of exponential nature. Given that video streams
CHAPTER 3. MODERATELY HEAVY TAILS

are much more bursty than voice ones, we consider to what extent stochastic nature of voice streams can be neglected. More formally, given two input processes to a queue we examine conditions under which one of the processes can be replaced by its average rate so that the probability of buffer content exceeding some large level does not change.

Busy period is one of the primary quantities of the fundamental GI/G/1 queueing model. Its understanding is essential in addressing a long list of queueing systems, including the processor sharing (Section 3.4.4), generalized processor sharing [18], coupled processors [17], static priority [1] and fluid [20] queues, as well as in estimating ruin probabilities [8] in insurance risk theory. Hence, it is of general interest to provide understanding of the behavior of this quantity.

Finally we turn our attention to processor sharing. The literature on M/G/1 PS queue is extensive; a comprehensive survey with more than 200 references on mathematical problems of shared-processor systems can be found in [92]. Early investigations of processor sharing systems used Laplace transform technique, e.g. see [26,68,74,84,91]. In the case of the M/M/1 PS system the conditional Laplace transform of the waiting time was derived in [26]; further analysis of this system was carried out in [68]. Representative studies of the M/G/1 PS queue can be found in [74,84,91]. Recently, prediction methods for processor-sharing queues were developed in [88]. Waiting times in GI/G/1 PS queue were shown to be associated in [10]. The heavy-traffic and fluid approximations were studied in [39,
85, 93] and [22], respectively. In contrast to most of the preceding analyses, of both exponential (e.g., [26, 68]) and heavy-tailed [96] systems, that were based on the Laplace transform technique, in this paper we develop a novel sample path large deviation approach. Using this approach, we first provide a direct sample path proof of the result from [72, 96]. Then, we extend this result, in Theorem 13, to a large class of subexponential distributions with tails lighter than polynomial and heavier than $e^{-\sqrt{x}}$. Furthermore, we demonstrate that the result does not hold for service distributions with tails lighter than $e^{-\sqrt{x}}$.

The chapter is based on [47, 48, 53].

3.2 Large Deviation Results

In this chapter we focus on a class of distributions that belongs to a large set of subexponential distributions $S^*$. This class of distribution functions was first introduced by A.V. Nagaev in [70].

Definition 8. A nonnegative r.v. $X$ (or its hazard function) belongs to class $S^* \subset SC$ (subexponential concave) if its hazard function $Q(x) \triangleq -\log P[X > x]$ is eventually concave, such that, as $x \to \infty$

$$Q(x)/\log x \to \infty$$

(3.1)

and for $x \geq x_0$, $\beta x \leq u \leq x$,

$$\frac{Q(x) - Q(u)}{Q(x)} \leq \alpha \frac{x - u}{x},$$

(3.2)
where $0 < \alpha < 1$, $0 < \beta < 1$.

Examples of random variables in $\mathcal{S}$ include random variables with hazard functions of type (i) $c(\log x)^\gamma$, $\gamma > 1$ and (ii) $c(\log x)^\gamma x^\alpha$, $\gamma > 0$, $0 < \alpha < 1$, i.e. widely used lognormal and Weibull distributions belong to $\mathcal{S}$. Condition (3.1) implies that if $X \in \mathcal{S}$ then $\mathbb{E}X^\beta < \infty$ for all finite $\beta$.

In the case when $Q(x)$ is absolutely continuous with hazard rate $q(x) \triangleq dQ(x)/dx$, the eventual concavity of $Q(x)$ is implied by $q(x)$ being eventually decreasing and the condition (3.2) is equivalent to $xq(x)/Q(x) \leq \alpha$, $\forall x \geq x_0$.

The next lemma summarizes the basic properties of r.v.s in $\mathcal{S}$.

**Lemma 11.** Let $X \in \mathcal{S}$ and $Q$ be its hazard function, then:

(i) $Q(x) \leq Q(u)(x/u)^\alpha$ for all $x_0 \leq u \leq x$,

(ii) $\mathbb{P}[X > x - x^\delta] \sim \mathbb{P}[X > x]$ as $x \to \infty$ for any $0 < \delta < 1 - \alpha$,

(iii) $X \in \mathcal{S}^*$,

(iv) for any $0 < \xi < 1$ there is $\delta > 0$ such that for some $\varepsilon > 0$ and all sufficiently large $x$

$$Q((\xi - \delta)x) + Q((1 - \xi)x) \geq (1 + \varepsilon)Q(x).$$

**Remark 5.** Clearly, for $\alpha < 1/2$, the second part of the preceding lemma implies $\mathbb{P}[X > x - \sqrt{x}] \sim \mathbb{P}[X > x]$ as $x \to \infty$; this property will be examined in more detail in Section 3.3.
CHAPTER 3. MODERATELY HEAVY TAILS

Proof. (i) For any \( x_0 \leq u \leq x \), we can choose \( \beta < \gamma < 1 \) and \( n \) such that \( \gamma^{n+1}x \leq u \leq \gamma^nx \). Then (3.2) implies

\[
Q(x)(1 - \alpha + \alpha\gamma)^{n+1} \leq Q(\gamma^{n+1}x) \leq Q(u)
\]

and, therefore,

\[
Q(x)(1 - \alpha + \alpha\gamma)^{1 + \frac{\log x/u}{\log 1/\gamma}} \leq Q(u).
\]

The last inequality can be restated in the following equivalent form

\[
Q(x) \leq Q(u) (1 - \alpha(1 - \gamma))^{-1} \left( \frac{x}{u} \right)^{\frac{\log(1 - \alpha(1 - \gamma))^{-1}}{\log \gamma^{-1}}}
\]

The statement (i) follows from the preceding inequality and the following limit

\[
\lim_{\gamma \to 1} \frac{\log(1 - \alpha(1 - \gamma))^{-1}}{\log \gamma^{-1}} = \alpha.
\]

(ii) Set \( u = x - x^\delta \) in (3.2) to obtain

\[
Q(x - x^\delta) \geq Q(x)(1 - \alpha x^{\delta-1}).
\]

Part (i) implies \( Q(x) \leq Cx^\alpha \) for \( x \geq x_0 \). Then for \( x \geq x_0 \)

\[
e^{-Q(x - x^\delta) + Q(x)} \leq e^{\alpha x^{\delta-1}Q(x)} \leq e^{Cax^{\alpha + \delta-1}}
\]

and the claim (ii) follows.

(iii) For any \( \delta > 0 \) and sufficiently large \( x \) the symmetry and concavity of \( Q(x) \)
yields
\[
\int_0^z P[X > u] P[X > x - u] du \leq 2 \int_0^z P[X > u] P[X > x - u] du \\
+ \int_{x^\delta}^{z - x^\delta} e^{-Q(u)} e^{-Q(x - u)} du \\
\leq 2\mathbb{E}X e^{-Q(z - x^\delta)} + xe^{-Q(z - x^\delta)} e^{-Q(x^\delta)}.
\]

Next, set $\delta < 1 - \alpha$ and use (i) and (3.1) to obtain the upper bound. The lower bound follows from
\[
\int_0^z P[X > u] P[X > x - u] du \geq 2 \int_0^{z - x^\delta} P[X > u] P[X > x - u] du \\
\geq 2P[X > x] \int_0^{z - x^\delta} P[X > u] du.
\]

(iiv) Directly from (i)
\[
Q((\xi - \delta)x) \geq Q(x)(\xi - \delta)^\alpha,
\]
\[
Q((1 - \xi)x) \geq Q(x)(1 - \xi)^\alpha.
\]

Then, summing the last two inequalities results in
\[
Q((\xi - \delta)x) + Q((1 - \xi)x) \geq Q(x)((\xi - \delta)^\alpha + (1 - \xi)^\alpha)
\]
and the statement follows.

In the remaining part of this section, we assume that \(\{X, X_i, i \geq 1\}\) is a sequence of independent and identically distributed (i.i.d.) r.v.s.
Next, let \( \{J, J_i, i \geq 2\} \) be a sequence of non-negative i.i.d. r.v. independent of \( \{X_n\} \) with \( \mathbb{E}J = \lambda^{-1} < \infty, \mathbb{E}J^2 < \infty \) and define \( N_x \) to be a counting process

\[
N_x = \max \left\{ n : \sum_{i=1}^{n} J_i < x \right\}.
\] (3.3)

Note that the distribution of \( J_1 \geq 0 \) (\( J_1 \) is independent of all \( J_i, i \geq 2 \)) is in general different from the one of \( J \). When \( J_1 \equiv J(e) \) then process \( N_x \) is stationary; if in addition \( J \) is exponential of rate \( \lambda \) then \( N_x \) is Poisson. Deviations of \( N_x \) from its expected value are estimated in the following lemma.

**Lemma 12.** There exists \( \delta > 0 \) such that for all \( x \) and \( 0 \leq u \leq \delta x \)

\[
\mathbb{P}[N_u - \lambda u > x] \leq Ce^{-c\frac{x^2}{\lambda}}.
\]

**Proof.** See Section 3.5. \( \square \)

In what follows we derive various large deviation bounds for sums of random variables with distribution in \( SC \). Related large deviation bounds can be found in [71].

**Theorem 6.** Let \( \mathbb{E}e^{Q(X)} < \infty \) for some \( Q \in SC \). Then

(i) for all \( x \) and \( u \) and Poisson \( N \)

\[
\mathbb{P} \left[ \sum_{i=1}^{N_u} X_i - \mathbb{E}XN_u > x \right] \leq C \left( e^{-c\frac{x^2}{\lambda}} + ue^{-\frac{1}{2}Q(x)} \right).
\]
(ii) for all $x$ and $u$

\[
\mathbb{P}\left[ \bigvee_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n\mathbb{E}X \right\} > x \right] \leq C \left( e^{-c\frac{x^2}{u}} + ue^{-\frac{1}{2}Q(x)} \right).
\]

(iii) for any integer $k > 0$ there exists $1 > \gamma > 0$ such that for all $1 \leq n \leq Cx$

\[
\mathbb{P}\left[ \sum_{i=1}^{n} X_i \wedge \gamma x - n\mathbb{E}X > x \right] \leq Ce^{-kQ(x)}.
\]

Remark 6. The conditions on the hazard function $Q$ can be relaxed. In particular, condition (3.1) can be replaced with

\[
\lim_{x \to \infty} \frac{Q(x)}{\log x} > 2,
\]

which implies the existence of $(2 + \varepsilon)$ moment for some $\varepsilon > 0$. (ii) By Markov’s inequality condition $\mathbb{E}e^{Q(X)} < \infty$ easily implies $\mathbb{P}[X > x] \leq Ce^{-Q(x)}$.

Proof. See Section 3.5. \hfill \Box

Lemma 13. If $\mathbb{E}e^{Q(X)} < \infty$ for some $Q \in SC$, then for any $\phi > \mathbb{E}X$

\[
\lim_{x \to \infty} e^{\phi Q(x)} \mathbb{P}\left[ \sup_{n \geq 1} \left\{ \sum_{i=1}^{n} X_i - \phi \right\} > x \right] = 0.
\]

Proof. The lemma follows from stochastic dominance, Lemma 11 and Pakes’ theorem [75]. \hfill \Box
CHAPTER 3. MODERATELY HEAVY TAILS

Proposition 6. If \( X \in SC \) with \( \alpha < 1/2 \), then, for any \( 0 < \delta < 1/2 - \alpha \) and \( y > 0 \), as \( x \to \infty \)

\[
P \left[ \sum_{n=0}^{N_x} X_n - \lambda x \mathbb{E}X > yx, \bigvee_{n=0}^{N_x} X_n \leq y(x - x^{\frac{1}{2}+\delta}) \right] = o(P[X > yx]).
\]

Proof. See Section 3.5. \( \square \)

Using the preceding proposition, the next large deviation theorem easily follows.

Theorem 7. If \( X \in SC \) with \( \alpha < 1/2 \), then for \( y > 0 \) as \( x \to \infty \)

\[
P \left[ \sum_{n=0}^{N_x} X_n - \lambda x \mathbb{E}X > yx \right] \sim \lambda x P[X > yx].
\]

Remark 7. (i) Straightforward examination of the proof shows that the result holds assuming that the first renewal interval is almost surely finite \( A_1 < \infty \). (ii) \( N_x \) does not have to be renewal as long as its right tail is exponentially bounded, i.e., it is necessary that \( N_x \) satisfies the bound of Lemma 12. (iii) Using the same arguments as in the following proof of the lower bound one can show that this result fails to hold for distributions with tails lighter than \( e^{-\sqrt{x}} \).

Proof. The upper bound is a direct consequence of Proposition 6 and Lemma 11 (ii).
In proving the lower bound, for $\eta > 0$, let $x_\eta \triangleq yx + 2\eta \mathbb{E}X \sqrt{x}$ to derive
\[
\mathbb{P} \left[ \sum_{n=0}^{N_x} X_n - \lambda x \mathbb{E}X > yx \right] \geq \mathbb{P}[N_x > \lfloor \lambda x - \eta \sqrt{x} \rfloor] \mathbb{P} \left[ \sum_{n=0}^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} X_n - \lambda x \mathbb{E}X > yx \right] \\
\geq \mathbb{P}[N_x > \lfloor \lambda x - \eta \sqrt{x} \rfloor] \left( \lambda x - \eta \sqrt{x} \right) \mathbb{P}[X > x_\eta] \times \\
\times \mathbb{P} \left[ \sum_{n=1}^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} X_n - \lambda x \mathbb{E}X > yx - x_\eta, \bigvee_{n=1}^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} X_n \leq x_\eta \right].
\]
(3.4)

Since $\mathbb{E}X^2 < \infty$, by Markov's inequality one has
\[
\mathbb{P} \left[ \bigvee_{n=1}^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} X_n \leq x_\eta \right] = (1 - \mathbb{P}[X > x_\eta])^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} \\
\geq \left( 1 - \frac{\mathbb{E}X^2}{x_\eta^2} \right)^{\lfloor \lambda x - \eta \sqrt{x} \rfloor} \longrightarrow 1,
\]
as $x \to \infty$. By taking lim as $x \to \infty$ in (3.4), using Lemma 11 (ii), the Central Limit Theorem, the preceding inequality and passing $\beta \to \infty$ the lower bound follows. \hfill \Box

**Lemma 14.** If $Q \in SC$, $\mathbb{E}e^{Q(X)} < \infty$, then for any $1/4 > \varepsilon > 0$, all $x$ and $xy \leq u \leq (1/2 - \varepsilon)xy$

we have
\[
\mathbb{P} \left[ \sum_{i=1}^{\lfloor \lambda x + x_\eta \varepsilon \rfloor} X_i \wedge u - \lambda x \mathbb{E}X > yx - u \right] \leq Ce^{-Q(xy-u)}.
\]

**Proof.** Presented in Section 3.5. \hfill \Box
We assume that \( B \) is a regenerative process with \( B_0 = 0 \). The length of the \( n \)th regenerative cycle is denoted by \( \nu_n > 0 \). Random variables \( \{\nu_n\}_{n=1}^\infty \) are i.i.d. independent of a.s. finite \( \nu_0 \) and have finite second moment, \( \mathbb{E} \nu_1^2 < \infty \). With each \( \nu_n, n \geq 1 \), we associate random variables \( \gamma_n \) and \( \gamma_n^* \). If \( T_n = \sum_{i=0}^n \nu_i \) then

\[
\gamma_n = B_{T_n} - B_{T_{n-1}},
\]

\[
\gamma_n^* = \sup_{T_{n-1} \leq t \leq T_n} B_t - B_{T_{n-1}},
\]

with \( \gamma_0 = B_{\nu_0} < \infty \) and \( \gamma_0^* = \sup_{0 \leq t \leq \nu_0} B_t < \infty \) a.s. Random variables \( \{\gamma_n\}_{n=1}^\infty \) and \( \{\gamma_n^*\}_{n=1}^\infty \) are i.i.d., independent of \( \gamma_0 \) and \( \gamma_0^* \), with finite first moments. Note that, by the SLLN, the mean rate \( b = \mathbb{E} \gamma_1/\mathbb{E} \nu_1 = \lim_{t \to \infty} B_t/t \) a.s. The next Proposition extends the result of Theorem 6 to regenerative processes.

**Proposition 7.** If \( \mathbb{E} e^{Q(\gamma_i)} < \infty, i = 0, 1 \) for some \( Q \in SC \), then for all \( x \) and \( u \geq 0 \)

\[
\mathbb{P} \left[ \sup_{0 \leq t \leq u} \{B_t - bt\} > x \right] \leq C \left( e^{-\frac{c}{u} x^2} + e^{-cu} + xe^{-cQ(x)} \right).
\]

**Proof.** Let \( K_u = \max\{n : \sum_{i=1}^n \nu_i < u\} \). Since for all \( t \geq 0 \)

\[
B_t - bt \leq \gamma_0^* + \gamma_{K_t-\nu_0+1} + \sum_{i=1}^{K_t-\nu_0} \gamma_i - b \sum_{i=1}^{K_t-\nu_0} \nu_i,
\]

one concludes

\[
\mathbb{P} \left[ \sup_{0 \leq t \leq u} \{B_t - bt\} > x \right] \leq \mathbb{P} \left[ \gamma_0^* > \frac{x}{4} \right] + \mathbb{P} \left[ \gamma_1^* > \frac{x}{4} \right]
\]

\[
+ \mathbb{P} \left[ \bigvee_{1 \leq n \leq K_u} \sum_{i=1}^{n} (\gamma_i - b \nu_i) > \frac{x}{2} \right].
\]

(3.6)
Lemma 12 provides a bound on $K_x$ for all $\delta$ small enough
\[ \mathbb{P}[K_u - u/\mathbb{E}u_1 > \delta u] \leq C e^{-cu} \]
and, hence, the third term in (3.6) can be upper bound as follows
\[ \mathbb{P}\left[ \bigvee_{0 \leq n < K_u} \sum_{i=1}^{n} (\gamma_i - b \nu_i) > \frac{x}{2} \right] \leq C e^{-cu} + \mathbb{P}\left[ \bigvee_{0 \leq n < (\delta + 1/\mathbb{E}\nu_1)u} \sum_{i=1}^{n} (\gamma_i - b \nu_i) > \frac{x}{2} \right] \]
\[ \leq C e^{-cu} + C \left( e^{-\frac{2}{u^2}} + ue^{-\frac{1}{4}Q(x/2)} \right) , \]
where the last inequality is due to Theorem 6. Substituting the preceding bound in (3.6) leads to
\[ \mathbb{P}\left[ \sup_{0 \leq t \leq x} \{ B_t - bt \} > x \right] \leq C e^{-Q(x/4)} + C e^{-cu} + C \left( e^{-\frac{2}{u^2}} + ue^{-\frac{1}{4}Q(x/2)} \right) \]
and the statement follows by Lemma 11. \hfill \Box

3.3 Gaussian Insensitivity

With the Central Limit Theorem being one of the pillars of the probability theory, it is of certain interest to understand what kind of distributions are asymptotically immune to the CLT effects. In this regard we consider the following class of distributions.

Definition 9. Random variable $X$ is called square-root insensitive if
\[ \mathbb{P}[X > x - \sqrt{x}] \sim \mathbb{P}[X > x] \quad \text{as} \quad x \to \infty. \]
CHAPTER 3. MODERATELY HEAVY TAILS

Remark 8. (i) The square-root insensitivity appears in work of [7, 35]. (ii) By monotonicity it follows that if $X$ is square-root insensitive then for any $k$

$$\mathbb{P}[X > x - k\sqrt{x}] \sim \mathbb{P}[X > x] \quad \text{as } x \to \infty.$$ 

Lemma 15. If $X$ is square-root insensitive, then its hazard function $Q$ satisfies

$$Q(x) = o(\sqrt{x}) \quad \text{as } x \to \infty.$$

Proof. The assumption of the lemma implies that for any $\varepsilon > 0$ there exists $x_\varepsilon$, such that for all $x \geq x_\varepsilon \geq 1$

$$\frac{\mathbb{P}[X > x]}{\mathbb{P}[X > x + \sqrt{x}]} \leq 1 + \varepsilon. \quad (3.7)$$

Next, define recursively $f^{(n)}(u) = f^{(n-1)}(u) + \sqrt{f^{(n-1)}(u)}$ for integers $n \geq 1$ with $f^{(0)}(u) = u$. Then,

$$f^{(2)}(x_\varepsilon) = x_\varepsilon + \sqrt{x_\varepsilon} + \sqrt{x_\varepsilon + \sqrt{x_\varepsilon}}$$

$$\geq x_\varepsilon + \sqrt{x_\varepsilon} + 1/4$$

$$= (\sqrt{x_\varepsilon} + 1/2)^2.$$

From the last inequality, it is easy to show by induction that

$$f^{(2n)}(x_\varepsilon) \geq (\sqrt{x_\varepsilon} + n/2)^2. \quad (3.8)$$

Now, let $n_x$ be the smallest integer such that $(\sqrt{x_\varepsilon} + n_x/2)^2 \geq x$, implying

$$n_x < 2(\sqrt{x} - \sqrt{x_\varepsilon}) + 1. \quad (3.9)$$
Next, due to (3.8), the choice of \( n_x \) and the monotonicity of \( \mathbb{P}[X > \cdot] \)

\[
\frac{\mathbb{P}[X > x]}{\mathbb{P}[X > x]} \leq \frac{\mathbb{P}[X > x]}{\mathbb{P}[X > f^{(2)}(x)]} \frac{\mathbb{P}[X > f^{(3)}(x)]}{\mathbb{P}[X > f^{(4)}(x)]} \cdots \frac{\mathbb{P}[X > f^{(2n_x-2)}(x)]}{\mathbb{P}[X > f^{(2n_x)}(x)]}.
\]

Observe that by (3.7) each of the ratios in the preceding inequality is upper bounded by \((1 + \varepsilon)^2\) and, hence, recalling (3.9) yields

\[
\frac{\mathbb{P}[X > x]}{\mathbb{P}[X > x]} \leq (1 + \varepsilon)^{2n_x} \leq (1 + \varepsilon)^{4(\sqrt{x} - \sqrt{x}) + 2}
\]

implying as \( \varepsilon \to 0 \)

\[
\lim_{x \to \infty} \frac{Q(x)}{\sqrt{x}} \leq 4 \log(1 + \varepsilon) \to 0.
\]

\[\square\]

**Theorem 8.** Let \( Z \) be the absolute value of a standard normal random variable.

Then \( X \) is square-root insensitive if and only if

\[
\mathbb{P}[X > x - Z\sqrt{x}] \sim \mathbb{P}[X > x] \text{ as } x \to \infty.
\]

**Remark 9.** By monotonicity it follows that if \( X \) is square-root insensitive then for any \( k \)

\[
\mathbb{P}[X > x - kZ\sqrt{x}] \sim \mathbb{P}[X > x] \text{ as } x \to \infty.
\]

**Proof.** (only if part) For \( k > 0 \) conditioning on \( Z \) results in

\[
\mathbb{P}[X > x - Z\sqrt{x}] \leq \mathbb{P}[X > x - k\sqrt{x}] + C \int_{k}^{\sqrt{x}/2} e^{-\frac{z^2}{2}} \mathbb{P}[X > x - z\sqrt{x}]dz \\
+ \mathbb{P}[Z > \sqrt{x}/2].
\]

(3.10)
CHAPTER 3. MODERATELY HEAVY TAILS

In order to estimate the integral term of (3.10) we choose \( x_{\varepsilon} \geq 1 \) such that (3.7) holds for all \( x \geq x_{\varepsilon} \). Observe that \( x - z\sqrt{x} \geq x_{\varepsilon} \) for all \( x \geq 2x_{\varepsilon} \) and \( z \leq \sqrt{x}/2 \).

Now, let \( n_{z} \) be the smallest integer such that \((\sqrt{x} - z\sqrt{x} + n_{z}/2)^{2} \geq x\), i.e.

\[
n_{z} < 2\left(\sqrt{x} - \sqrt{x - z\sqrt{x}}\right) + 1 \\
\leq 2z + 1.
\]

(3.11)

Next, recall the definition of \( f^{(n)} \) from the proof of Lemma 15. Due to (3.8), the choice of \( n_{z} \) and the monotonicity of \( \mathbb{P}[X > \cdot] \) for all \( z \leq \sqrt{x}/2 \) and \( x \geq 2x_{\varepsilon} \)

\[
\frac{\mathbb{P}[X > x - z\sqrt{x}]}{\mathbb{P}[X > x]} \leq \frac{\mathbb{P}[X > f^{(0)}(x - z\sqrt{x})]}{\mathbb{P}[X > f^{(2)}(x - z\sqrt{x})]} \cdots \frac{\mathbb{P}[X > f^{(2n_{z}-2)}(x - z\sqrt{x})]}{\mathbb{P}[X > f^{(2n_{z})}(x - z\sqrt{x})]} \\
\leq (1 + \varepsilon)^{2n_{z}} \\
\leq (1 + \varepsilon)^{4z+2},
\]

where the last inequality follows from (3.11). Therefore, the upper bound for the second term in (3.10) is as follows

\[
\int_{k}^{\sqrt{x}/2} e^{-\frac{x^{2}}{2}} \mathbb{P}[X > x - z\sqrt{x}]dz \leq C\mathbb{P}[X > x] \int_{k}^{\sqrt{x}/2} e^{-\frac{x^{2}}{2}} e^{x^{2}}dz \\
\leq C\mathbb{P}[X > x] \int_{k}^{\infty} e^{-\frac{(x-\varepsilon)^{2}}{2}}dz.
\]

Combining the preceding bound and Lemma 15 with (3.10) easily yields the "only if" part of the theorem.

(if part) Note that

\[
\mathbb{P}[X > x - Z\sqrt{x}] \geq \mathbb{P}[Z > 1]\mathbb{P}[X > x - \sqrt{x}] + \mathbb{P}[Z \leq 1]\mathbb{P}[X > x],
\]
which, in conjunction with the assumption, implies $P[X > x - \sqrt{x}] \sim P[X > x]$ as $x \to \infty$. 

3.4 Engineering Applications

The present section demonstrates the applicability of results from the previous section to queueing problems with moderately heavy tails. These problems include the reduced load equivalence, independent sampling, busy period and processor sharing.

3.4.1 Reduced Load Equivalence

In this subsection we investigate the tail behavior of the stationary workload of a stable queue. The stationary workload $W_X^\phi$ of a queue with service rate $\phi$ fed by a process $X$ with stationary increments, is known to satisfy

$$W_X^\phi \overset{d}{=} \sup_{t \geq 0} \{X_t - \phi t\},$$

where $\overset{d}{=}$ denotes equality in distribution and $X_t$ represents the amount of work that arrives to the queue in $(-t, 0)$; throughout the paper $X$ will be considered equal to $A$, $B$ or $A + B$, where process $B$ is defined as in Section 3.2. In the queueing context, a natural assumption on $B$ is that sample paths are a.s. increasing, i.e., in this subsection $\gamma_i = \gamma_i^*$ for $i \geq 0$. For convenience, we put $W_X \equiv W_X^1$. Let $a$ denote the mean rate of process $A$.

**Theorem 9.** Let $\mathbb{E}^Q(\gamma_i) < \infty$, $i = 0, 1$ for some $Q \in SC$ and $\mathbb{E}\nu_i^2 < \infty$. If $W_A^{-b} \in$
 CHAPTER 3. MODERATELY HEAVY TAILS

$S$ is square-root insensitive, $\mathbb{P}[W_{A}^{1-b} > x] = e^{-o(Q(x))}$ and for some $a < \phi < 1 - b$

$$\lim_{k \to \infty} \lim_{z \to \infty} \frac{\mathbb{P}[W_{A}^{k} > kz]}{\mathbb{P}[W_{A}^{1-b} > x]} = 0,$$

then as $x \to \infty$

$$\mathbb{P}[W_{A+B} > x] \sim \mathbb{P}[W_{A}^{1-b} > x].$$

When the regenerative increments of $B$ do not have tails heavier than $e^{-\theta \sqrt{x}}$, $\theta > 0$, the conditions of the preceding theorem can be relaxed. In particular, the assumptions $W_{A}^{1-b} \in S$ and $\mathbb{P}[W_{A}^{1-b} > x] = e^{-o(Q(x))}$ are not needed. Note that the class of distributions $SC$ was used in bounding deviations of process $B$ rather than process $A$. In the case when the deviations of process $B$ are not heavier of those attributed to the CLT, in view of Theorem 8, the square-root insensitivity condition on $W_{A}^{1-b}$ is essentially necessary; this is formalized in the following proposition.

**Proposition 8.** Let $\mathbb{E}e^{\theta \sqrt{N}} < \infty$, $i = 0, 1$ for some $\theta > 0$ and $\mathbb{E}\nu_{i}^{2} < \infty$. If $W_{A}^{1-b}$ is square-root insensitive and for some $a < \phi < 1 - b$

$$\lim_{k \to \infty} \lim_{z \to \infty} \frac{\mathbb{P}[W_{A}^{k} > kz]}{\mathbb{P}[W_{A}^{1-b} > x]} = 0,$$

then as $x \to \infty$

$$\mathbb{P}[W_{A+B} > x] \sim \mathbb{P}[W_{A}^{1-b} > x].$$
CHAPTER 3. MODERATELY HEAVY TAILS

These results extend Propositions 8.2 and 8.3 of [2], where \( A \) is assumed to be an On-Off process with a specific form of the distribution of On periods and \( B \) is exponentially bounded. In particular, Proposition 8.3 in [2] assumes that the tail of the residual distribution of On periods is of the form \( e^{-ax} \) with \( \beta < 1/3 \).

Possible choices of \( A \) include, for instance, an On-Off process with subexponential On periods and a Gaussian process exhibiting long-range dependence, cf. [19]. Next, we specialize our result to the case where \( A \) is a stationary On-Off process. Let On periods be equal in distribution to \( \tau \). Denote by \( r \) and \( \alpha \) the peak and average rate, respectively.

**Corollary 4.** Let \( \mathbb{E}e^{Q(\tau)} < \infty \), \( i = 0, 1 \) for some \( Q \in SC \) and \( \mathbb{E}Q^2_i < \infty \). If \( \tau^{(e)} \in S \) is square-root insensitive, \( \mathbb{P}[\tau^{(e)} > x] = e^{-c(Q(x))}, r > 1 - b > a \) and

\[
\lim_{k \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}[\tau^{(e)} > kx]}{\mathbb{P}[\tau^{(e)} > x]} = 0,
\]

then as \( x \to \infty \)

\[
\mathbb{P}[W_{A+B} > x] \sim \mathbb{P}[W_A^{1-b} > x].
\]

**Proof.** Follows from Theorem 9 and the asymptotics for \( W_A^{1-b} \) (see Theorem 4.3 of [49]). \( \square \)

Before we turn to the proofs of Theorem 9 and Proposition 8, we state two additional preliminary result.

**Lemma 16.** If \( X \in S \) and \( \mathbb{P}[U > x] = o(\mathbb{P}[X > x]) \), then as \( x \to \infty \)

\[
\mathbb{P}[X + U > x, X \leq x] = o(\mathbb{P}[X > x]).
\]
Proof. Corollary 2 of [78] states \( \mathbb{P}[X + U > x] \sim \mathbb{P}[X > x] \) as \( x \to \infty \), which in conjunction with \( \mathbb{P}[X + U > x] = \mathbb{P}[X > x] + \mathbb{P}[X > x, X \leq x] \) concludes the proof.

\[ \square \]

Lemma 17. Let \( \mathbb{E}e^{Q(t)} < \infty, t = 0, 1 \) for some \( Q \in \mathcal{S} \). If \( \mathbb{P}[W_A^{1-b} > x] = e^{-\alpha(Q(x))} \) and for some \( a \prec \phi \prec 1 - b \)

\[ \lim_{l \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}[W_A^{1-b} > lx]}{\mathbb{P}[W_A^{1-b} > x]} = 0, \]

then

\[ \lim_{l \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}[\sup_{t \geq lx} \{A_t + B_t - t\} > x]}{\mathbb{P}[W_A^{1-b} > x]} = 0. \]

Proof. See Section 3.5.

\[ \square \]

We are now ready to present proofs of Theorem 9 and Proposition 8.

Proof of Theorem 9. The proof consists of deriving upper and lower bounds.

Upper bound. Observe that for \( l > 0 \)

\[ \mathbb{P}[W_{A+B} > x] \leq \mathbb{P}\left[ \sup_{0 \leq t \leq lx} \{A_t + B_t - t\} > x \right] + \mathbb{P}\left[ \sup_{t \geq lx} \{A_t + B_t - t\} > x \right]. \quad (3.12) \]

The second term is negligible by Lemma 17 as \( x \to \infty \) for large \( l \). To estimate the first term, we proceed as follows. By using \( \sup_t \{f(t) + g(t)\} \leq \sup_t f(t) + \sup_t g(t) \)
CHAPTER 3. MODERATELY HEAVY TAILS

for any two functions \( f(x) \) and \( g(x) \) one obtains for \( k > 0 \)

\[
P \left[ \sup_{0 \leq t \leq \varepsilon} \{A_t + B_t - t\} > x \right]
\leq P \left[ \sup_{0 \leq t \leq \varepsilon} \{A_t - (1 - b)t\} + \sup_{0 \leq t \leq \varepsilon} \{B_t - bt\} > x \right]
\leq P[W^{1-b}_A > x - k\sqrt{x}] + P[W^{1-b}_A + Y_x > x, W^{1-b}_A \leq x - k\sqrt{x}]
\triangleq f_1 + f_2, \quad (3.13)
\]

where \( Y_x \triangleq \sup_{0 \leq t \leq \varepsilon} \{B_t - bt\} \). Proposition 7 yields an upper bound on \( f_2 \)

\[
f_2 \leq \int_0^{x-k\sqrt{x}} P[Y_x > x - u]dP[W^{1-b}_A \leq u]
\leq C \int_0^{x-k\sqrt{x}} \left( e^{-c(x-u)^2/\varepsilon} + e^{-dx} + lxe^{-Q(x-u)} \right) dP[W^{1-b}_A \leq u]
\triangleq f_{21} + f_{22} + f_{23}. \quad (3.14)
\]

Integration by parts and change of variables \( x = (x - u)/\sqrt{x} \) result in

\[
f_{21} \leq Ce^{-c\varepsilon} + C \frac{k}{\sqrt{x}} \int_0^{x-k\sqrt{x}} \frac{x - u}{\sqrt{x}} e^{-c(x-u)^2/\varepsilon} P[W^{1-b}_A > u] du
= Ce^{-c\varepsilon} + C \frac{k}{\varepsilon} \int_{k\sqrt{x}}^{\sqrt{x}} ze^{-c\varepsilon z^2} P[W^{1-b}_A > x - z\sqrt{x}] dz
\leq Ce^{-c\varepsilon} + C \frac{k}{\varepsilon} P[W^{1-b}_A + Z\sqrt{x} > x, Z > k],
\]

where r.v. \( Z \) is the absolute value of a normal random variable. Combining the

preceding bound with Lemma 11 and Theorem 8 we obtain

\[
\lim_{x \to \infty} \frac{f_{21}}{P[W^{1-b}_A > x]} \leq C \frac{k}{\varepsilon} P[Z > k].
\]

It easily follows that the upper bound for the second term in (3.14) is \( f_{22} \leq Ce^{-dx} \).
CHAPTER 3. MODERATELY HEAVY TAILS

To handle $f_{23}$, define r.v. $U$ such that $\mathbb{P}[U > x] = e^{-cQ(x)}$ for $x \geq x_U$. Then,

$$f_{23} \leq C l x e^{-cQ(k\sqrt{x})} \int_0^{x-k\sqrt{x}} e^{-cQ(x-u)} d\mathbb{P}[W_A^{1-b} \leq u]$$

$$= C l x e^{-cQ(k\sqrt{x})} \mathbb{P}[U + W_A^{1-b} > x, W_A^{1-b} \leq x - k\sqrt{x}]$$

thus, by Lemmas 16 and 11 one obtains $f_{23} = o(\mathbb{P}[W_A^{1-b} > x])$ as $x \to \infty$.

Combining the bounds on $f_{21}$, $f_{22}$ and $f_{23}$ with (3.14), (3.13) and square-root insensitivity of $W_A^{1-b}$ results, after passing $k \to \infty$, in

$$\lim_{x \to \infty} \frac{\mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{A_t + B_t - t\} > x\right]}{\mathbb{P}[W_A^{1-b} > x]} = 1.$$  

Therefore, the proof of the upper bound is concluded recalling (3.12) and Lemma 17.

**Lower bound.** As usual, the lower bound is somewhat easier:

$$\mathbb{P}[W_{A+B} > x] \geq \mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{A_t + B_t - t\} > x\right]$$

$$\geq \mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{A_t - (1-b)t\} + \inf_{0 \leq t \leq l\varepsilon} \{B_t - bt\} > x\right]$$

$$= \mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{A_t - (1-b)t\} - \sup_{0 \leq t \leq l\varepsilon} \{bt - B_t\} > x\right].$$

Hence, for any $k, l > 0$,

$$\mathbb{P}[W_{A+B} > x] \geq \mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{A_t - (1-b)t\} > x + k\sqrt{x}\right] \mathbb{P}\left[\sup_{0 \leq t \leq l\varepsilon} \{bt - B_t\} \leq k\sqrt{x}\right].$$

(3.15)

Note that for $t \geq 0$ and $K_{x} = \max\{n : \sum_{i=1}^{n} \nu_i < x\}$

$$bt - B_t \leq b\nu_0 + b\nu_{K_{x}+1} + \sum_{i=1}^{K_{x}-\nu_0} (b\nu_i - \gamma_i)$$
and, thus, by Lemma 12 and the CLT for maximums [25, Ch. 7] one obtains

\[
\lim_{k \to \infty} \lim_{n \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq n} \{B_t - B_0\} \leq k \frac{\sqrt{x}}{n} \right) = 1.
\]

The proof is now completed by dividing both sides of (3.15) by \(\mathbb{P}[W_A^{1-b} > x + k \sqrt{x}]\), letting \(x \to \infty\), using the square-root insensitivity of \(W_A^{1-b}\), setting first \(k \to \infty\), then \(l \to \infty\) and applying Lemma 17.

Proof of Proposition 8. The proof is identical to the proof of Theorem 9 with \(Q(x) = \theta \sqrt{x}\) except for the derivation of the upper bound for \(f_{23}\). In the following we show that

\(f_{23} = o(\mathbb{P}[W_A^{1-b} > x])\) as \(x \to \infty\). For any \(0 < \delta < 1\) integration by parts yields

\[
f_{23} \leq Clx \int_0^{-k \sqrt{\varepsilon}} e^{-\delta \sqrt{\varepsilon} - u} d\mathbb{P}[W_A^{1-b} \leq u]
\]

\[
\leq Clx e^{-\delta \sqrt{\varepsilon}} + Clx \int_{-k \sqrt{\varepsilon}}^{x - k \sqrt{\varepsilon}} e^{-\delta \sqrt{\varepsilon} - u} d\mathbb{P}[W_A^{1-b} \leq u]
\]

\[
\leq Clx e^{-\delta \sqrt{\varepsilon}} + Clx \int_{-k \sqrt{\varepsilon}}^{x - k \sqrt{\varepsilon}} e^{-\delta \sqrt{\varepsilon} - u} \mathbb{P}[W_A^{1-b} > u] du.
\]

Next, square-root insensitivity yields (see the proof of Lemma 15 for details) that for any \(\varepsilon > 0\) there exists \(x_\varepsilon \geq 1\) such that for all \(x_\varepsilon \leq u \leq x - k \sqrt{\varepsilon}\)

\[
\frac{\mathbb{P}[W_A^{1-b} > u]}{\mathbb{P}[W_A^{1-b} > x - k \sqrt{\varepsilon}]} \leq Ce^{\varepsilon(\sqrt{\varepsilon} - \sqrt{\delta})}.
\]

By using the preceding bound in (3.16) and recalling the concavity of \(Q\) one obtains for \(\delta x \geq x_\varepsilon\)

\[
f_{23} \leq Clx e^{-\delta \sqrt{\varepsilon}} + Clx \mathbb{P}[W_A^{1-b} > x - k \sqrt{\varepsilon}] \int_{-k \sqrt{\varepsilon}}^{x - k \sqrt{\varepsilon}} e^{-\delta \sqrt{\varepsilon} - u + \varepsilon(\sqrt{\varepsilon} - \sqrt{\delta})} du
\]

\[
\leq Clx e^{-\delta \sqrt{\varepsilon}} + Clx^2 \mathbb{P}[W_A^{1-b} > x - k \sqrt{\varepsilon}]
\]

\[
\times \left( e^{-\varepsilon \sqrt{k \sqrt{\varepsilon} + \varepsilon(\sqrt{\varepsilon} - \sqrt{\delta})}} + e^{-\varepsilon \sqrt{\varepsilon(1 - \delta) - \varepsilon(1 - \sqrt{\delta})}} \right).
\]
Clearly, we can chose $\epsilon$ and $\delta$ in the preceding inequality to obtain

$$f_{23} \leq C k x e^{-c \sqrt{z}} + C l x^2 \mathbb{P}[W_A^{1-b} > z - k \sqrt{z}] e^{-c z^{1/4}},$$

which by Lemma 15 and square-root insensitivity yields $f_{23} = o(\mathbb{P}[W_A^{1-b} > z])$ as $x \to \infty$. 

\[\square\]

### 3.4.2 Independent Sampling

In this subsection we illustrate the applicability of the developed methodology by investigating the problem of independent sampling at subexponential times that was recently studied in [7, 35]. The proofs bellow require that $B$ satisfies the CLT and Proposition 1; as a sufficient condition for this we assume that $B$ is regenerative. In general, any process $B$ satisfying the CLT and Proposition 7 is admissible, e.g. certain Gaussian processes as considered in [19]. We would like to stress that the SC framework is used only to estimate the behavior of the sampled process $B$ rather then the sampling time $T$. Note that the deviations of process $B$ are due to the large deviations and CLT effects. When the deviations of $B$ are dominated by those of the CLT, the only condition on $T$ that we require is the square-root insensitivity (see Proposition 9). This, basically necessary, square-root insensitivity condition improves on some of the known results in the literature; this will be further elaborated at the end of this section. When the deviations of $B$ are heavier then those stemming from the CLT, one needs an additional set of conditions on $T$ to ensure that those deviations can be tolerated.
Define the maximum $M_t = \sup_{0 \leq s \leq t} B_s$. Note that $M_t$ inherits the regenerative structure of $B_t$, but has the additional property that its sample paths are non-decreasing. Since $B_t$ has positive drift, heuristically, $B_t$ is not expected to be much smaller than $M_t$. Our theorem below shows that $M_T$ and $B_T$ have similar tail behavior. For convenience, we assume that mean rate $b = 1$.

**Theorem 10.** Let $\mathbb{E}e^{Q(t)} < \infty$, $i = 0, 1$ for some $Q \in SC$ and $\mathbb{E}v^2 < \infty$. If $T \in S$ is square-root insensitive and $\mathbb{P}[T > x] = e^{-c(Q(x))}$, then as $x \to \infty$

\[
\mathbb{P}[B_T > x] \sim \mathbb{P}[M_T > x] \sim \mathbb{P}[T > x].
\]

**Proof.** Since $B_T \leq M_T$, it suffices to provide an upper bound for $\mathbb{P}[M_T > x]$ and a lower bound for $\mathbb{P}[B_T > x]$.

**Upper bound.** Write for $\delta < 1$

\[
\mathbb{P}[M_T > x] \leq \mathbb{P}[T > x - k\sqrt{x}] + \mathbb{P}[M_T > x, \delta x < T \leq x - k\sqrt{x}] + \mathbb{P}[M_{\delta x} > x].
\]

(3.17)

One needs to show that the last two terms are $o(\mathbb{P}[T > x])$ as $x \to \infty$. Note that by Proposition 7 and Lemma 11 as $x \to \infty$

\[
\mathbb{P}[M_{\delta x} > x] \leq C(e^{-c_x} + xe^{-cQ(x)})
\]

\[
\leq Cxe^{-cQ(x)} = o(\mathbb{P}[T > x]).
\]
CHAPTER 3. MODERATELY HEAVY TAILS

To deal with the second term in (3.17), note that, in view of Proposition 7,

\[ \mathbb{P}[M_T > x, \delta x < T \leq x - k\sqrt{x}] \]
\[ = \int_{\delta x}^{x - k\sqrt{x}} \mathbb{P}[M_u > x] d\mathbb{P}[T \leq u] \]
\[ \leq \int_{\delta x}^{x - k\sqrt{x}} \mathbb{P}[\sup_{0 \leq s \leq u} \{B_s - s\} > x - u] d\mathbb{P}[T \leq u] \]
\[ \leq C \int_{\delta x}^{x - k\sqrt{x}} \left( e^{-\frac{(x-u)^2}{u}} + e^{-cu} + ue^{-c(x-u)} \right) d\mathbb{P}[T \leq u] \]
\[ \leq C \int_{\delta x}^{x - k\sqrt{x}} \left( e^{-\frac{(x-u)^2}{u}} + e^{-c\delta x} + xe^{-c(x-u)} \right) d\mathbb{P}[T \leq u]. \]

Now, proceed exactly as in bounding \( f_2 \) in the proof of the upper bound of Theorem 9.

**Lower bound.** Following the steps of the proof of Theorem 3.6 in [7] we write

\[ \mathbb{P}[B_T > x] \geq \int_{x+k\sqrt{x}}^{\infty} \mathbb{P}[B_u > x] d\mathbb{P}[T \leq u] \]
\[ \geq \inf_{u > x + k\sqrt{x}} \mathbb{P}[B_u > x] \mathbb{P}[T > x + k\sqrt{x}]. \]

Note that due to the monotonicity of \((x - u)/\sqrt{u}\) in \( u \) one obtains for \( x > k^2 \)

\[ \inf_{u > x + k\sqrt{x}} \mathbb{P}[B_u > x] \geq \inf_{u > x + k\sqrt{x}} \mathbb{P}\left[ \frac{B_u - u}{\sqrt{u}} > \frac{-k}{\sqrt{1 + k/\sqrt{x}}} \right] \]
\[ \geq \inf_{u > x + k\sqrt{x}} \mathbb{P}\left[ \frac{B_u - u}{\sqrt{u}} > \frac{-k}{2} \right]. \]

Therefore, the square-root insensitivity results in, for an appropriate \( \sigma > 0 \),

\[ \lim_{x \to \infty} \frac{\mathbb{P}[B_T > x]}{\mathbb{P}[T > x]} \geq 1 - \Phi\left( \frac{-k}{2\sigma} \right), \]

where \( \Phi(\cdot) \) is the distribution function of the standard normal r.v. Letting \( k \to \infty \) concludes the proof. \( \square \)
CHAPTER 3. MODERATELY HEAVY TAILS

Proposition 9. If $\mathbb{E}e^{\theta \sqrt{i}} < \infty$, $i = 0, 1$ for some $\theta > 0$, $\mathbb{E} \nu_1^2 < \infty$ and $T$ is square-root insensitive, then as $x \to \infty$

$$\mathbb{P}[B_T > x] \sim \mathbb{P}[M_T > x] \sim \mathbb{P}[T > x].$$

Proof. We follow the same steps as in the proof of Theorem 10. The only difference is that a bound on $C \int_{\delta x}^{cz - k\sqrt{z}} xe^{-c\sqrt{z-u}} d\mathbb{P}[T \leq u]$ is obtained using the same arguments as in bounding $f_{23}$ in the proof of Proposition 8. \qed

The preceding proposition fully generalizes Proposition 3.1 of [7] and shows that Theorems 3.8, 3.10 and 3.11 of [7] hold under less restrictive conditions. In addition, Theorem 10 provides an alternative set of conditions, that may appear easier to verify than those stemming from the extreme value theory used in Theorem 3.6 of [7], under which the sampling result holds. The case of $B_t$ being a counting renewal process has been further examined in [35]. When $B_t$ is such a process, the only condition in Proposition 9 related to $B_t$ is $\mathbb{E} \nu_1^2 < \infty$. An extension of the sampling theorem to the case when the second moment of $\nu_1$ is infinite can be found in Theorem 3.1 of [35].

3.4.3 Busy Period

Investigation of the busy period of the M/G/1 queue with exponentially bounded service distributions has a long history; for recent results see [1] and the references
therein. The first analysis involving the heavy-tailed regularly varying service times has appeared in [30]. The derivation in [30] made use of Karamata Tauberian Theory [13] and the Poisson arrival structure. In [97] this result was generalized for the GI/G/1 queue by developing a sample path technique that exploits the relationship between the busy period and cycle maxima. Furthermore, it was shown in [7] that this result does not hold for distributions that are lighter than $e^{-\sqrt{\alpha}}$.

Here we resolve the question that was left open in [7, 97], by deriving the tail of the busy period distribution for a class of subexponential service times with tails heavier than $e^{-\sqrt{\alpha}}$ but lighter than any polynomial. In addition, our result, in conjunction with [7], shows that the asymptotic behavior of the busy period exhibits a transition in its qualitative behavior depending on the relationship of the service distribution to the Weibull tail $e^{-\sqrt{\alpha}}$.

Without loss of generality we assume that the first (0th) customer arrives to the empty queue at time $t = 0$. Denote by $B_i$ the service requirement of the $i$th customer and by $A_i$ the interarrival time between the $i$th and $(i + 1)$th customers. Random sequences $\{A_i, A_i, i \geq 1\}$ and $\{B_i, B_i, i \geq 0\}$ are respectively i.i.d. and independent of each other. Let $E[A^2] < \infty$ and $N_x$ be a counting process as defined earlier in (3.3) with $A_t$ instead of $J_t$.

The amount of unfinished work in the queue at time $t$ is denoted by $V_t$; for the exact definition of $V_t$ see e.g. [28]. The busy period is a stopping time at which the
queue becomes empty for the first time after \( t = 0 \), i.e.,

\[
P = \inf\{t > 0 : V_t = 0\}.
\]

The traffic load \( \rho \) is equal to \( \mathbb{E}B/\mathbb{E}A < 1 \). Let \( K \) be the number of customers served during the busy period. Note that, since \( \sum_{i=0}^{K-1} B_i = P \), by Wald's lemma \( \mathbb{E}K = \mathbb{E}P/\mathbb{E}B \). The expected number of customers served during the busy period can be also represented as [28, p. 286]

\[
\mathbb{E}K = \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{P}[S_n > 0],
\]

where \( S_n = \sum_{i=1}^{n} (B_i - A_i) \). In the case of the \( M/G/1 \) queue \( \mathbb{E}K = (1 - \rho)^{-1} \).

**Theorem 11.** If \( \mathbb{E}A^2 < \infty \) and \( B \in SC \) with \( \alpha < 1/2 \), then as \( x \to \infty \)

\[
\mathbb{P}[P > x] \sim \mathbb{E}K \mathbb{P}[B > (1 - \rho)x].
\]

**Remark 10.** It is interesting to observe that the asymptotic behavior of the busy period in the \( M/G/\infty \) queue is the same for the whole class of subexponential distributions, irrespective of the relationship of the service distribution to \( e^{-\sqrt{x}} \), as proved in Theorem 3.5 of [49].

**Proof.** The proof of the lower bound was earlier given in [97]. Thus, it remains to prove the upper bound. Denote by \( S \) the cycle maximum, i.e., \( S = \sup\{V_t, 0 \leq t \leq \)
$P$. Then, following the approach in [97], for some $0 < \delta < 1/2 - \alpha$

\[
\mathbb{P}[P > x] = \mathbb{P}[S > (1 - \rho)(x - x^{1/2+\delta})] + \mathbb{P}[P > x, S \leq (1 - \rho)(x - x^{1/2+\delta})] \\
\leq \mathbb{P}[S > (1 - \rho)(x - x^{1/2+\delta})] \\
+ \mathbb{P}\left[\sum_{i=0}^{N_x} B_i > x, \bigwedge_{i=0}^{N_x} B_i \leq (1 - \rho)(x - x^{1/2+\delta})\right], \quad (3.18)
\]

where the second inequality follows from the facts that \( \{S \leq x\} \) implies \( \{B_i \leq x\} \)

for all $0 \leq i \leq N_P$, and \( \{P > x\} \) implies, by work conservation, \( \{\sum_{i=0}^{N_x} B_i > x\} \).

Next, for $B \in S^*$ the distribution of the cycle maximum $S$ is shown [5] to satisfy

(see also [6]), \( \mathbb{P}[S > x] \sim \mathbb{E}K\mathbb{P}[B > x] \) as $x \to \infty$. Hence, using this fact and

Lemma 11 (ii), the first term in (3.18) satisfies

\[
\lim_{x \to \infty} \frac{\mathbb{P}[S > (1 - \rho)(x - x^{1/2+\delta})]}{\mathbb{P}[B > (1 - \rho)x]} \leq \mathbb{E}K.
\]

Thus, to complete the proof, one needs to show that the second term in (3.18) is

$O(\mathbb{P}[B > (1 - \rho)x])$ as $x \to \infty$. However, that is immediate from Proposition 6. \( \Box \)

### 3.4.4 Processor Sharing

In this subsection we study a processor sharing queue that represents a baseline model of efficient and fair network resource sharing algorithms, e.g. TCP flow control protocol and Web server job scheduling algorithms. Our main result extends the asymptotic reduced load equivalence relationship between the job sizes and their waiting times, derived in [72, 96] for polynomial tails, to a class of subexponential distributions with tails heavier than $e^{-\sqrt{x}}$. In particular, this extension covers the
practically important case of jobs with lognormal distributions that were recently empirically measured in [63, 64].

We start with the basic theory of the M/G/1 processor sharing queue. Customers arrive to the queue of unit capacity according to a Poisson process with rate $\lambda$. Service requirements of customers are i.i.d. r.v.s equal in distribution to $B$. Upon its arrival a customer joins the queue and starts receiving service immediately. The customers are served according to the processor sharing scheduling discipline. Namely, if there are $n$ customers present in the queue, then each of the customers receives service at rate $1/n$. Once a customer receives service equal to its service requirement it departs from the system. The queue is assumed to be stable, i.e., the load $\rho \overset{\Delta}{=} \lambda \mathbb{E}B$ of the system satisfies $\rho < 1$.

The distribution of the number of customers $L$ in the queue in the stationary regime is known to be geometric [54, 83] and depends on $B$ only through $\mathbb{E}B$ (insensitivity property), i.e., for $n = 0, 1, 2, \ldots$

$$\mathbb{P}[L = n] = (1 - \rho)\rho^n.$$ 

Furthermore, the stationary remaining service requirements of the customers present in the queue are i.i.d. random variables equal to $B^{(e)}$ in distribution [90, p. 387].

In the case of regularly varying distributions Zwart and Boxma [96] established the asymptotic relationship between the tails of the waiting time in the M/G/1 PS queue and the customer service requirement. The main result from [96] is derived by means of Tauberian theorem that requires regularly varying service distribution
with non-integer exponent. By using sample path arguments, Núñez-Queija [72] (see also [73]) generalized it to distributions with intermediately regularly varying tails. The result in [72] does not cover the technical case of $\mathbb{E}B^2 = \infty$ and $\mathbb{E}B^c < \infty$ for all $\zeta \in (0,1)$. In Section 3.5 we provide a proof, using a completely different approach, that does not require this minor condition.

**Theorem 12 (Zwart & Boxma [96], Núñez-Queija [72]).** If $B \in \mathcal{IR}$ and $\mathbb{E}B^\alpha < \infty$ for some $\alpha > 1$, then $\mathbb{P}[V > x] \sim \mathbb{P}[B > (1-\rho)x]$ as $x \to \infty$.

In the remaining part of the subsection we extend the preceding theorem to the class $SC$. This main results of this subsection are stated in Theorem 13 and Proposition 10. The proofs are based on an identity that will be described in the following paragraph.

Let $B_i$ and $V_i$ be the job size and waiting time of the customer arriving at time $T_i$, respectively. The sequence of arrival times $\{T_i\}_{i=1}^\infty$ is assumed to be Poisson. Without loss of generality, in view of PASTA property [90], we set $T_0 = 0$. Waiting time of a customer is defined as an amount of time between its arrival and departure, also referred to as sojourn time in the queueing literature. For the customer arriving at time $T_0$ define function $R_0(t) \equiv R_{B_0}(t)$ for $t \geq 0$ as the amount of work that remains to be completed at time $t$. The waiting time satisfies the following min-plus equality which stems from the features of processor sharing

$$V_0 = B_0 + \sum_{i=1}^L B_i^{(e)} \land B_0 + \sum_{i=1}^{N_0} B_i \land R_0(T_i), \quad (3.19)$$
where $L$ is the number of customers in the system just before time $t = 0$, $N_t$ denotes the number of Poisson arrivals in $(0, t)$ and $x \wedge y \equiv \min(x, y)$; the number of customers in the system $L$ and their residual work $B^{(e)}_t$ are independent, e.g. see [90, p. 387]. The identity follows from the fact that in the PS queue any two customers present in the system for some interval of time receive equal amounts of service during that interval, irrespective of other departures and arrivals. In particular, a customer $i$, $1 \leq i \leq L$, present in the system just before $t = 0$ receives $B_0 \wedge B^{(e)}_i$ amount of service during $(0, V_0)$. Similarly, any customer arriving at $T_i \in (0, V_0)$ obtains $B_i \wedge R_0(T_i)$ service in interval $(0, V_0)$. Clearly, 0th customer receives its full requirement $B_0$ by time $V_0$. Therefore, by summing up the services that each customer present in the queue during $(0, V_0)$ receives, one derives (3.19). A related expression to (3.19) can be found in [91, eq. (3.4)] (see also Theorem 5.3.2 in [72]).

**Theorem 13.** Let $B \in SC$ with $\alpha < 1/2$ and

$$\lim_{x \to \infty} \frac{\mathbb{P}[B^{(e)} > x]}{x \mathbb{P}[B > x]} < \infty. \quad (3.20)$$

Then, as $x \to \infty$

$$\mathbb{P}[V > x] \sim \mathbb{P}[B > (1 - \rho)x].$$

**Remark 11.** The condition (3.20) is not very limiting since it is implied if $x^{1+\delta}\mathbb{P}[B > x]$ is eventually monotonically decreasing in $x$ for some $\delta > 0$. 
CHAPTER 3. MODERATELY HEAVY TAILS

The condition $\alpha < 1/2$ in Theorem 13 is crucial. As the proposition below shows, the result does not extend to the whole class of subexponential distributions. An intuitive explanation of this criticality arises from fluctuations induced by the CLT. Informally, the last sum in (3.19) is approximately equal to $\rho V_0 + O(\sqrt{V_0})$ for large $V_0$. Therefore, for the result to hold, the distribution of $V_0$, or equivalently $B_0$, has to be immune to these fluctuations, which translates to $\alpha < 1/2$.

**Proposition 10.** If $\mathbb{P}[B > x] = e^{-x^\alpha}$, $\alpha > 1/2$, then as $x \to \infty$

$$
\mathbb{P}[B > x] = o(\mathbb{P}[V(1 - \rho) > x]).
$$

*Proof.* Presented at the end of this subsection. \qed

*Remark 12.* The proposition implies the result earlier obtained in [7], that the busy period $P$ in the M/G/1 queue satisfies $\mathbb{P}[B > x] = o(\mathbb{P}[P(1 - \rho) > x])$ as $x \to \infty$, when $\mathbb{P}[B > x] = e^{-x^\alpha}$, $\alpha > 1/2$.

For any sequence of i.i.d. r.v.s $\{X_i\}$ we use $W_{X \wedge Y}^\phi$ to denote the stationary workload in a queue with Poisson arrivals of rate $\lambda$, capacity $\phi$ and job sizes equal to $\{X_i \wedge Y\}$; let $W_{X}^\phi = W_{X \wedge \infty}^\phi$. The following lemma estimates the stationary workload in a queue with truncated service requirements.

**Lemma 18.** Assume $\phi > \rho$. 

(i) If $\mathbb{E}B^{1+\delta} < \infty$ for some $\delta > 0$, then for any $\alpha$ there exists $\epsilon > 0$ such that, as $x \to \infty$

$$\mathbb{P} \left[ W_{B_{L\triangle x}}^\phi > x \right] = o(x^{-\alpha}).$$

(ii) If $B \in SC$ and

$$\lim_{x \to \infty} \frac{\mathbb{P}[B^{(c)} > x]}{x \mathbb{P}[B > x]} < \infty,$$

then for any integer $k \geq 1$ there exists $1 > \epsilon > 0$ such that, as $x \to \infty$

$$\mathbb{P}[W_{B_{L\triangle x}}^\phi > x] = o \left( (\mathbb{P}[B > x])^k \right).$$

Proof. Presented in Section 3.5.

The last preparatory result states that asymptotically the long waiting time of a customer cannot be caused by the customers present in the queue upon its arrival.

Proposition 11. Let either (i) $X \in D \cap L$ and $\mathbb{E}X < \infty$ or (ii) $X \in S^\ast$ and

$$\lim_{x \to \infty} \frac{\mathbb{P}[X^{(c)} > x]}{x \mathbb{P}[X > x]} < \infty.$$

If $\{X_i^{(c)}\}_{i=1}^\infty$ are i.i.d. random variables equal in distribution to $X^{(c)}$ and independent of $N$ with $\mathbb{E} \left[ (1+\epsilon)^N \right] < \infty$ for some $\epsilon > 0$, then as $x \to \infty$

$$\mathbb{P} \left[ X + \sum_{i=1}^N X_i^{(c)} \wedge X > x \right] \sim \mathbb{P}[X > x].$$

Proof. See Section 3.5.
Proof of Theorem 13. Expression (3.19) for the waiting time of the 0th customer renders

\[ V_0 - \sum_{i=1}^{N_{V_0}} B_t \wedge R_0(T_i) = B_0 + \sum_{i=1}^{L} B_0 \wedge B_t^{(c)} \triangleq \hat{B}_0, \quad (3.21) \]

where \( \hat{B}_0 \) is introduced for notational convenience. Next, for \( \xi > \max\{1/2, \beta\} \) and \( \delta > 0 \), write \( \mathbb{P}[V_0(1 - \rho) > x] = f_1(x) + f_2(x) + f_3(x) \) where

\[ f_1(x) = \mathbb{P}[V_0(1 - \rho) > x, \hat{B}_0 \leq \xi x], \]
\[ f_2(x) = \mathbb{P}\left[V_0(1 - \rho) > x, \hat{B}_0 \in (\xi x, x - x^{1/2+\delta}]\right], \]
\[ f_3(x) = \mathbb{P}[V_0(1 - \rho) > x, \hat{B}_0 > x - x^{1/2+\delta}]. \]

In what follows, we examine the asymptotic behavior of the three terms. We start with \( f_1(x) \). Observe that, since \( \hat{B}_0 \geq B_0 \geq R_0(t) \), (3.21) implies

\[ V_0(1 - \rho - \delta) \leq \hat{B}_0 + \sup_{t \geq 0} \left\{ \sum_{i=1}^{N_t} B_t \wedge \hat{B}_0 - (\rho + \delta)t \right\} \]
\[ \overset{d}{=} \hat{B}_0 + W_{\Delta \hat{B}_0}^{+, \delta}, \]

where \( \overset{d}{=} \) denotes the equality in distribution. Therefore, for \( \delta_\rho \equiv \delta/(1 - \rho) \)

\[ f_1(x) \leq \mathbb{P}\left[\hat{B}_0 + W_{\Delta \hat{B}_0}^{+, \delta} > (1 - \delta_\rho)x, \hat{B}_0 \leq \xi x\right] \]
\[ \leq \mathbb{P}\left[W_{\Delta \hat{B}_0}^{+, \delta} > (1 - 2\delta_\rho)x\right] + \int_{\delta_\rho x}^{\xi x} \mathbb{P}\left[W_B^{+, \delta} > (1 - \delta_\rho)x - u\right] d\mathbb{P}[\hat{B}_0 \leq u]. \]

If \( \delta_\rho \) (i.e. \( \delta \)) is chosen small enough, the first term in the preceding sum, by Lemma 18 (ii), is upper bounded by \( C(\mathbb{P}[B > (1 - 2\delta_\rho)x])^2 = o(\mathbb{P}[B > x]) \), where the last equality follows from Lemma 11 (i). The bound for the second term is as
follows. By Pakes’ asymptotic result for the workload of a stable M/G/1 queue [75] and assumption (3.20)

\[
f_1(x) \leq C x \int_{\delta_x}^{\xi_x} e^{-Q((1-\delta_x)x-u)} dP[\hat{B}_0 \leq u] + o(P[B > x])
\]

\[
= o(P[B > x]),
\]

as \( x \to \infty \), where the last equality follows from Lemma 21.

Bounding \( f_2(x) \) requires the most work. Introduce a continuous function \( H : \mathbb{R}_+ \to \mathbb{R}_+ \) defined as

\[
H(t) \overset{\triangle}{=} \sup_{0 \leq u \leq t} \left\{ u - \sum_{i=1}^{N^a} B_i \right\}, \quad t > 0.
\]

(3.22)

The function \( H \) is nondecreasing and, hence, it is possible to define a right-continuous inverse \( H^- (x) = \inf\{ t > 0 : H(t) > x \} \). From Figure 3.1, due to the memoryless property of exponential distribution, it is clear that \( H(t) \) increases linearly at rate 1 over exponential intervals of parameter \( \lambda \) and then stays constant for the amounts of time that are equal in distribution to the busy period \( P \) of the original M/G/1 queue. Thus, \( H^- \) can be written in the following form

\[
H^-(t) = t + \sum_{i=1}^{N^a} P_i,
\]

(3.23)

where r.v.s \( \{P_i\}_{i=1}^{\infty} \) are i.i.d. copies of \( P \). Note that \( H^-(t) \) is the busy period of an M/G/1 queue with initial workload equal to \( t \).

From (3.21), \( V_0 \) can be interpreted as the first time \( t \) that \( t - \sum_{i=1}^{N^a} B_0 \wedge R_0(t) = \).
CHAPTER 3. MODERATELY HEAVY TAILS

Figure 3.1: A typical sample path realization of function $H$; $T_n$-s are the Poisson arrival points.

$\hat{B}_0$, implying

$$V_0 \leq \inf \left\{ t > 0 : t - \sum_{i=1}^{N_t} B_i > \hat{B}_0 \right\}$$

$$= \inf \{ t > 0 : H(t) > \hat{B}_0 \} = H^-(\hat{B}_0). \tag{3.24}$$

By using (3.23) and (3.24), $f_2(x)$ can be upper bounded by

$$f_2(x) \leq \int_{x \to \infty} \int_{\xi_x} \left( \int_{\xi_x}^{x} \mathbb{P} \left[ \sum_{i=1}^{N_t} P_i - \lambda u \mathbb{E} P > \frac{x-u}{1-\rho} \right] d\mathbb{P}[\hat{B}_0 \leq u] \right)$$

$$\leq C \int_{x \to \infty} \left( e^{-\frac{(x-u)^2}{2u}} + ue^{-\frac{1}{2}(1-\delta)(x-u)^2} \right) d\mathbb{P}[\hat{B}_0 \leq u],$$

where the second inequality follows from Theorems 11 and 6. Therefore, for all $\delta$ sufficiently small, we obtain by Lemma 20

$$f_2(x) = o(\mathbb{P}[B > x]).$$

The bound for $f_3(x)$ is straightforward by Lemma 11 and Proposition 11

$$\lim_{x \to \infty} \frac{f_3(x)}{\mathbb{P}[B > x]} \leq \lim_{x \to \infty} \frac{\mathbb{P}[B > x - x^{1/2+\delta}]}{\mathbb{P}[B > x]} = 1.$$
Combination of bounds for $f_1$, $f_2$ and $f_3$ yields the upper bound. The lower bound is a corollary of Lemma 19.

Denote by $V^{(x)}$ the waiting time of a customer conditional on the fact that its service requirement is equal to $x$. In the same fashion, let $R^{(x)}(t)$ be the conditional amount of service to be completed at time $t$.

Remark 13. Waiting time $V_0$ can be represented as sampling at subexponential time $B_0$ of the monotonically increasing process

$$V^{(x)} = x + \sum_{i=1}^{L} B_i^{(x)} \land x + \sum_{i=1}^{N_{V^{(x)}}} B_i \land R^{(x)}(T_i),$$

for which one could potentially use results obtained in [7, 35]. However, the major difficulty in carrying out this approach is that $V^{(x)}$ is implicitly defined, i.e. understanding $V^{(x)}$ requires the knowledge of $V^{(x)}$ and $R^{(x)}$. This is similar to the situation that arises in the analysis of the busy period, as pointed out in [7].

The following lemma provides a general lower bound on the waiting time. The proof is based on the Central Limit Theorem, and, therefore, the second moment of the service requirement is assumed.

Lemma 19. If $\mathbb{E}B^2 < \infty$ and $\mathbb{P}[B > x] \sim \mathbb{P}[B > x + x^{1/2+\delta}]$ as $x \to \infty$ for some $\delta > 0$, then

$$\lim_{x \to \infty} \frac{\mathbb{P}[V > x]}{\mathbb{P}[B > (1-\rho)x]} \geq 1.$$
CHAPTER 3. MODERATELY HEAVY TAILS

Proof. Consider the stationary M/G/1 PS queue at \( t < 0 \), as described at the beginning of this section. Next, assume that at time \( t = 0 \) a special customer with infinite service requirement arrives. Let \( Z(t) \), \( t \geq 0 \), be the total unfinished work in the new PS systems of all but the infinite customer. It is known that \( Z(t) \) converges in distribution to an a.s. finite random variable \( Z \) [90, p. 339]. In what follows we exploit the fact that at time \( t = V_0 \) the remaining unfinished work in the original PS queue is equal to \( Z(V_0) \). Thus, (3.19) and (3.22) render

\[
H(V_0) \geq V_0 - \sum_{i=1}^{N_{V_0}} B_i \\
\geq B_0 - \left( \sum_{i=1}^{N_{V_0}} B_i - \sum_{i=1}^{N_{V_0}} B_i \land R_0(T_i) \right) \\
\geq B_0 - \left( \sum_{i=1}^{N_{V_0}} B_i - \sum_{i=1}^{N_{V_0}} B_i \land R_0(T_i) + \sum_{i=1}^{L} (B_i^{(e)} - B_i^{(e)} \land B_0) \right) \\
= B_0 - Z(V_0). \tag{3.25}
\]

Next, from (3.25) one obtains

\[
\mathbb{P}((1 - \rho)V_0 > x) \\
\geq \mathbb{P}\left[(1 - \rho)H^{-}(B_0 - Z(V_0)) > x, B_0 > x + k\sqrt{x}, Z(V_0) \leq \sqrt{x}\right] \\
\geq \mathbb{P}\left[(1 - \rho)H^{-}(x + (k - 1)\sqrt{x}) > x, B_0 > x + k\sqrt{x}, Z(V_0) \leq \sqrt{x}\right],
\]

where \( k > 1 \), and the last inequality is due to the monotonicity of \( H^{-} \). Then, by
independence of $H^-(x)$ and $Z(x)$ from $B_0$,

\[
\mathbb{P}[(1 - \rho)V_0 > x] \\
\geq \int_{x + k\sqrt{x}}^\infty \mathbb{P}\left[(1 - \rho)H^-(x + (k - 1)\sqrt{x}) > x, \ Z(V^{(y)}) \leq \sqrt{x}\right] d\mathbb{P}[B \leq y] \\
\geq \left(\mathbb{P}[(1 - \rho)H^-(x + (k - 1)\sqrt{x}) > x] - \sup\limits_{y \geq x + k\sqrt{x}} \mathbb{P}[Z(V^{(y)}) > \sqrt{x}]\right) \\
\times \mathbb{P}[B > x + k\sqrt{x}] ;
\]

the second inequality follows from the union bound. Observe that, since $Z(t)$ converges in distribution to a.s. finite $Z$, the supremum in the preceding inequality tends to 0 as $x \to \infty$. Therefore,

\[
\lim\limits_{x \to \infty} \frac{\mathbb{P}[(1 - \rho)V_0 > x]}{\mathbb{P}[B > x]} \geq \lim\limits_{x \to \infty} \mathbb{P}[(1 - \rho)H^-(x + (k - 1)\sqrt{x}) > x].
\]

Next, it is known that $\mathbb{E}P^2 < \infty$ if and only if $\mathbb{E}B^2 < \infty$ [1]. Thus, by (3.23), the process $H^-(t)$ as a function of $t$ satisfies the Central Limit Theorem, yielding

\[
\lim\limits_{x \to \infty} \frac{\mathbb{P}[(1 - \rho)V_0 > x]}{\mathbb{P}[B > x]} \geq \lim\limits_{k \to \infty} \lim\limits_{x \to \infty} \mathbb{P}[(1 - \rho)H^-(x + (k - 1)\sqrt{x}) > x] = 1,
\]

which concludes the proof. \qed

Proof of Proposition 10. The proof is a minor modification of the proof of Lemma 19.
CHAPTER 3. MODERATELY HEAVY TAILS

Equation (3.25) leads to

\[ \mathbb{P}[(1 - \rho)V_0 > x] \]
\[ \geq \mathbb{P} \left[ (1 - \rho)H^{-}(B_0 - Z(V_0)) > x, \ B_0 > x - \sqrt{x}, \ Z(V_0) \leq \sqrt{x} \right] \]
\[ \geq \mathbb{P} \left[ (1 - \rho)H^{-}(x - 2\sqrt{x}) > x, \ B_0 > x - \sqrt{x}, \ Z(V_0) \leq \sqrt{x} \right] \]
\[ \geq \left( \mathbb{P}[(1 - \rho)H^{-}(x - 2\sqrt{x}) > x] - \sup_{y \geq x - \sqrt{x}} \mathbb{P}[Z(V^{(y)}) > \sqrt{x}] \right) \mathbb{P}[B > x - \sqrt{x}]. \]

Thus,

\[ \lim_{\alpha \to -\infty} \frac{\mathbb{P}[(1 - \rho)V > x]}{\mathbb{P}[B > x]} \geq \lim_{\alpha \to -\infty} \mathbb{P}[(1 - \rho)H^{-}(x - 2\sqrt{x}) > x] \lim_{\alpha \to -\infty} \frac{\mathbb{P}[B > x - \sqrt{x}]}{\mathbb{P}[B > x]} \]
\[ \geq c \lim_{\alpha \to -\infty} e^{x_\alpha - (x_\alpha - \sqrt{x})^\alpha} = \infty, \]

since \( \alpha > 1/2 \); the existence of \( c > 0 \) follows from the Central Limit Theorem. \( \square \)

3.5 Proofs

This section contains the proofs of more technical results. First, we supply proofs of large deviation bounds including Lemma 12, Theorem 6, Proposition 6, Lemma 14 and state Lemma 20, Lemma 21. Then, in the second subsection we provide the proof of the PS result from Theorem 12 for the IR case. In addition we prove a useful queueing bound from Lemma 18. Finally, the last subsection is the proof of Proposition 11.
CHAPTER 3. MODERATELY HEAVY TAILS

3.5.1 Large deviation bounds

Proof of Lemma 12. The definition of \( N_z \) and Markov's inequality yield for \( s > 0 \)
\[
P[N_u - \lambda u > x] = P \left[ \sum_{i=1}^{\lfloor \lambda u + x \rfloor} X_i < u \right] \leq e^{su} \left( E e^{-sJ} \right)^{\lambda u + x - 2}. \tag{3.26}
\]

Next we estimate \( E e^{-sJ} \) as follows
\[
E e^{-sJ} = E[e^{-sJ} 1\{sJ \leq 1\}] + E[e^{-sJ} 1\{sJ > 1\}] \leq 1 - sEJ + s^2EJ^2 + e^{-1}P[sJ > 1],
\]
since \( e^{-x} \leq 1 - x + x^2 \) for \( 0 \leq x \leq 1 \). The preceding inequality leads to
\[
E[e^{-sJ}] = 1 - sEJ + s^2(1 + e^{-1})EJ^2
\]
which substituted in (3.26) results in for \( s = x/(2\lambda Cu) \)
\[
P[N_u - \lambda u > x] \leq e^{su} e^{-(sEJ + s^2C)(\lambda u + x - 2)} \leq Ce^{-cs^2}.
\]

Proof of Theorem 6. (i) For any max(1/2, \( \beta \)) < \( \gamma \) < 1 the union bound gives
\[
P \left[ \sum_{i=1}^{N_u} X_i - E X E N_u > x \right] \leq P \left[ \sum_{i=1}^{N_u} X_i 1\{X_i \leq \gamma x\} - E X E N_u > x \right] + \lambda u P[X > \gamma x].
\]

We point out that, by assumption \( P[X > x] \leq Cx e^{-Q(x)} \) and (3.1),
\[
\lambda u P[X > \gamma x] \leq C x e^{-\gamma Q(x)} \leq C e^{-\frac{1}{2}Q(x)},
\]
CHAPTER 3. MODERATELY HEAVY TAILS

from which one concludes that the first statement of the lemma holds if for all $x$ and $u$

$$\mathbb{P} \left[ \sum_{i=1}^{N_u} X_i \mathbb{1}_{\{X_i \leq \gamma x\}} - \mathbb{E} X \mathbb{E} N_u > x \right] \leq C \left( e^{-c \frac{x^2}{u}} + e^{-\frac{1}{2} Q(x)} \right).$$  \hspace{1cm} (3.27)

In the proof of this statement we restrict our attention to $u \leq \eta x^2$ for some $\eta > 0$, since for any $\eta > 0$ and $u > \eta x^2$ the bound holds trivially if $C$ is chosen large enough, i.e. $Ce^{-c/\eta} > 1$. Next, let

$$\frac{1}{\gamma x} \leq s \leq \frac{Q(x)}{x}. \hspace{1cm} (3.28)$$

Then, Markov’s inequality yields

$$\mathbb{P} \left[ \sum_{i=1}^{N_u} X_i \mathbb{1}_{\{X_i \leq \gamma x\}} - \lambda u \mathbb{E} X > x \right] \leq e^{-s(x + \lambda u \mathbb{E} X)} e^{\lambda u (e^{sX[X \leq \gamma x]} - 1)}. \hspace{1cm} (3.29)$$

We start with estimating the moment generating function of $X \mathbb{1}_{\{X \leq \gamma x\}}$

$$\mathbb{E} e^{sX \mathbb{1}_{\{X \leq \gamma x\}}} = \int_{0}^{1/s} e^{sy} d\mathbb{P}[X \leq y] + \int_{1/s}^{\gamma x} e^{sy} d\mathbb{P}[X \leq y] + \mathbb{P}[X > \gamma x]. \hspace{1cm} (3.30)$$

The last term, by Markov’s inequality, can be upper bounded as

$$\mathbb{P}[X > \gamma x] \leq \frac{\mathbb{E} X^2}{\gamma^2 s^2 x^2} s^2 \leq \mathbb{E} X^2 s^2; \hspace{1cm} (3.31)$$

recall that (3.1) implies $\mathbb{E} X^2 < \infty$. Inequality $e^x \leq 1 + x + x^2$ on $[0, 1]$, gives rise to

$$\int_{0}^{1/s} e^{sy} d\mathbb{P}[X \leq y] \leq \int_{0}^{1/s} (1 + sy + s^2 y^2) d\mathbb{P}[X \leq y] \leq 1 + s\mathbb{E} X + s^2 \mathbb{E} X^2. \hspace{1cm} (3.32)$$
CHAPTER 3. MODERATELY HEAVY TAILS

Next, concavity of $Q(y)$ renders for $1/s \leq y \leq \gamma x$

$$sy - Q(y) \leq \max \{ s\gamma x - Q(\gamma x), 1 - Q(1/s) \}$$

and, hence, integration by parts and Markov’s inequality yield

$$\int_{1/s}^{\gamma x} e^{sy} dP[X \leq y] \leq eP[X > 1/s] + Csx \int_{1/s}^{\gamma x} e^{sy-Q(y)}dy$$

$$\leq s^2 eEX^2 + Csx^2 \left( e^{s\gamma x-Q(\gamma x)} + e^{1-Q(1/s)} \right)$$

$$\leq Cs^2 \left( 1 + x^3 \left( e^{s\gamma x-Q(\gamma x)} + e^{1-Q(1/s)} \right) \right).$$ (3.33)

The expression in brackets in (3.33) is an increasing function in $s$ that achieves its maximum for $s = Q(x)/x$ (see (3.28)). Then, by (3.2), $\gamma Q(x) - Q(\gamma x) \leq -(1 - \alpha)(1 - \gamma)Q(x)$ and, by Lemma 11 (i) $Q(x/Q(x)) \geq Q(\sqrt{x})$; thus, the last two bounds, (3.1) and (3.33) yield

$$\int_{1/s}^{\gamma x} e^{sy} dP[X \leq y] \leq Cs^2. \quad \text{(3.34)}$$

Hence, combining bounds (3.30), (3.31), (3.32) and (3.34) we derive

$$Ee^{sX1(X \leq \gamma x)} \leq 1 + sEX + Cs^2,$$ (3.35)

where $C^*$ is a constant. Substituting this estimate for $Ee^{sX1(X \leq \gamma x)}$ in (3.29) yields

$$P \left[ \sum_{i=1}^{N_u} X_i 1 \{ X_i \leq \gamma x \} - \lambda uEX > x \right] \leq e^{-sx + \lambda uC^*s^2}. \quad \text{(3.36)}$$

Next, if $u \leq x^2/(2\lambda C^*Q(x))$, then by setting $s = Q(x)/x$ we derive

$$P \left[ \sum_{i=1}^{N_u} X_i 1 \{ X_i \leq \gamma x \} - \lambda uEX > x \right] \leq e^{-\frac{1}{2}Q(x)}. \quad \text{(3.37)}$$
On the other hand, if \( u \geq x^2/(2\lambda C^*Q(x)) \), \( s = x/(2\lambda u C^*) \) \( \leq Q(x)/x \) yields
\[
P \left[ \sum_{i=1}^{N_u} X_i 1\{X_i \leq \gamma x\} - \lambda u \mathbb{E} X > x \right] \leq e^{-\frac{4\lambda^2 \gamma^2}{x^2}}. \tag{3.38}
\]
Since for any value of \( u \) either (3.37) or (3.38) holds we conclude that (3.27) and, therefore, the first statement of the theorem holds.

(ii) Let \( Y_i = X_i 1\{X_i \leq \gamma x\} \) and \( s \) be as in (3.28). As in part (i) we restrict our attention to the case \( u \leq \eta x^2 \) for some \( \eta > 0 \) since otherwise the bound holds trivially by selecting \( Ce^{-c/n} > 1 \). Then for \( \max(1/2, \beta) < \gamma < 1 \)
\[
P \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n \mathbb{E} X \right\} > x \right] \
\leq P \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} Y_i - n \mathbb{E} Y_1 \right\} > x \right] + uP[X > \gamma x] \
\leq P \left[ \max_{1 \leq n \leq u} \exp \left\{ \sum_{i=1}^{n} s(Y_i - \mathbb{E} Y_1) \right\} > e^{sx} \right] + uP[X > \gamma x].
\]
Next, note that \( \exp\{\sum_{i=1}^{n} s(Y_i - \mathbb{E} Y_1)\} \) is a submartingale. Therefore, applying the submartingale inequality (e.g. see Theorem 9.4.1 [25] in or Theorem 35.3 in [12]) in the preceding equation leads to
\[
P \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n \mathbb{E} X \right\} > x \right] \leq e^{-sx} \mathbb{E} \left[ e^{s(Y_1 - \mathbb{E} Y_1)} \right]^u + uP[X > \gamma x] \
\leq e^{-sx - su \mathbb{E} Y_1} (\mathbb{E} e^{sY_1})^u + Cue^{-\frac{1}{2}Q(s)},
\]
where the last inequality is due to Markov's inequality and Lemma 11. By combining the preceding inequality with (3.35) one obtains
\[
P \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n \mathbb{E} X \right\} > x \right] \leq e^{-zx + u C^* s^2} e^{su(\mathbb{E} X - \mathbb{E} Y_1)} + Cue^{-\frac{1}{2}Q(s)}.\]
CHAPTER 3. MODERATELY HEAVY TAILS

Next, by following the same steps of the proof of (i) after (3.36), for $u \leq x^2/(2Q(x)C^*)$ we set $s = Q(x)/x$ to obtain

$$
\mathbb{P} \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n\mathbb{E}X \right\} > x \right] \leq e^{-\frac{1}{2}Q(x)}e^{\frac{2}{\eta}Q(x) (\mathbb{E}X - \mathbb{E}X_i)} + Cu e^{-\frac{1}{4}Q(x)}
$$

$$
\leq Ce^{-\frac{1}{4}Q(x)} + Cu e^{-\frac{1}{4}Q(x)},
$$

and for $\eta x^2 \geq u \geq x^2/(2Q(x)C^*)$ we set $s = x/(2uC^*)$ to get

$$
\mathbb{P} \left[ \max_{1 \leq n \leq u} \left\{ \sum_{i=1}^{n} X_i - n\mathbb{E}X \right\} > x \right] \leq e^{-xs + uC^* x^2} e^{s\mathbb{E}Q(x) - \mathbb{E}X_i} + Cu e^{-\frac{1}{4}Q(x)}
$$

$$
\leq Ce^{-\frac{x^2}{4\eta x^2}} + u e^{-\frac{1}{4}Q(x)}.
$$

Hence, the statement follows.

(iii) The proof of the third statement proceeds along the similar lines as in part (i). Choosing $\gamma < (k + 1)^{-1/(1-\alpha)}$, $s = (k + 1)Q(x)/x$ and using Lemma 11 (i) we verify

$$
s\gamma x - Q(\gamma x) \leq (k + 1)\gamma Q(x) - Q(\gamma x)
$$

$$
\leq ((k + 1)^{1-\alpha} - 1)\gamma^\alpha Q(x) < 0. \quad (3.39)
$$

Next, Markov's inequality yields

$$
\mathbb{P} \left[ \sum_{i=1}^{n} X_i \wedge \gamma x - n\mathbb{E}X > x \right] \leq e^{-s(x + n\mathbb{E}X)} \left( \mathbb{E}e^{s(X \wedge \gamma x)} \right)^n. \quad (3.40)
$$

The moment generating function of $X \wedge \gamma x$ can be bounded as

$$
\mathbb{E}e^{s(X \wedge \gamma x)} = \int_0^{1/s} e^{sy} d\mathbb{P}[X \leq y] + \int_{1/s}^{\gamma x} e^{sy} d\mathbb{P}[X \leq y] + e^{\gamma sx - Q(\gamma x)}
$$

$$
\leq 1 + s\mathbb{E}X + Cs^2,
$$
where the second inequality holds by the same arguments used in obtaining (3.35), and (3.39). Substituting the preceding bound in (3.40) results in
\[
\mathbb{P}
\left[
\sum_{i=1}^{n} X_i \wedge \gamma x - n\mathbb{E}X > x
\right] \leq e^{-\delta(x+n\mathbb{E}X) + n \log(1 + \mathbb{E}X + Cx^2)}
\leq e^{-\delta(x+n\mathbb{E}X) + n(\mathbb{E}X + Cx^2)}
\leq e^{-(k+1)Q(\epsilon)(1-(k+1)CQ(\epsilon)/x)} \leq e^{-kQ(\epsilon)},
\]
for all \( x \) large enough, which renders the second part of Theorem 6.

Proof of Proposition 6. In order to simplify the notation we define \( \rho \triangleq \lambda \mathbb{E}X \) and
\[
f \triangleq \mathbb{P}
\left[
\sum_{i=0}^{N_x} X_i > (\rho + y)x, \bigvee_{i=0}^{N_x} X_i \leq y(x - x^{1/\delta})
\right].
\]
The following straightforward identity represents the basis of our analysis
\[
\sum_{i=0}^{N_x} X_i = \sum_{i=0}^{N_x} X_i \wedge \left(\bigvee_{j=0}^{N_x} X_j\right) = \bigvee_{j=0}^{N_x} \left(\sum_{i=0}^{N_x} X_i \wedge X_j\right).
\]
Using this identity, the union bound, Lemma 12 and Lemma 11 (i)
\[
f \leq \mathbb{P}
\left[
\bigvee_{j=0}^{N_x} \left\{ \sum_{i=0}^{N_x} X_i \wedge X_j \{X_j \leq y(x - x^{1/\delta})\} \right\} > (\rho + y)x
\right]
\leq Cx \mathbb{P}
\left[
\sum_{i=0}^{N_x} X_i \wedge X_0 \{X_0 \leq y(x - x^{1/\delta})\} > (\rho + y)x, N_x \leq Cx
\right]
+ o(\mathbb{P}[X > yx]),
as $x \to \infty$. Next, for $0 < \xi < 1/2$, conditioning on $X_0$ yields as $x \to \infty$

$$f \leq Cx \int_0^{y(x - x_0^{1/2 + \varepsilon})} \mathbb{P} \left[ \sum_{i=1}^{N_x} X_i \land u > (\rho + y)x - u \right] \, d\mathbb{P}[X \leq u] + o(\mathbb{P}[X > yx])$$

$$\leq \int_0^{yx} + \int_{yx}^{y(x - x_0^{1/2 + \varepsilon})} + o(\mathbb{P}[X > yx])$$

$$\triangleq f_1 + f_2 + o(\mathbb{P}[X > yx]). \quad (3.41)$$

Now, we examine the two terms in the preceding expression. For all $1/4 > \varepsilon > 0$, Lemma 12 and 11 (i) yield

$$f_1 \leq Cx \int_0^{yx} \mathbb{P} \left[ \sum_{i=1}^{N_x} X_i \land u > (\rho + y)x - u \right] \, d\mathbb{P}[X \leq u] + \mathbb{P} \left[ N_x > \lambda x + x_0^{1-\varepsilon} \right]$$

$$\leq Cx \int_0^{yx} \mathbb{P} \left[ \sum_{i=1}^{N_x} X_i \land u > (\rho + y)x - u \right] \, d\mathbb{P}[X \leq u] + o(\mathbb{P}[X > yx]),$$

as $x \to \infty$. Then, by Theorem 6 and Lemma 14, we can choose $\varepsilon$ small enough, such that as $x \to \infty$

$$f_1 \leq Cx \int_{yx}^{yx} e^{-Q(yx-u)} \, d\mathbb{P}[X \leq u] + o(\mathbb{P}[X > yx]).$$

and Lemma 21 yields $f_1 = o(\mathbb{P}[X > yx])$ as $x \to \infty$.

At this point, we upper bound the expression under the integral in $f_2$. To ease the notation, for $u$ in the interval of integration we define

$$g(x, u) \triangleq (yx - u)x^{-d/2},$$

and observe that for any $\varepsilon > 0$ there exists $x_\varepsilon$ such that for all $x \geq x_\varepsilon$

$$g(x, u) \leq \varepsilon(yx - u).$$
Then, neglecting the minimums under the sum \( \sum X_i \wedge u \) and using Theorem 6, Lemma 12, Lemma 11 (i) and the preceding inequality, result in, for all values of \( u \) in the interval of integration for \( f_2 \)

\[
P \left[ \sum_{i=1}^{N} X_i \wedge u > (\rho + y)x - u \right] \leq P \left[ \sum_{i=1}^{N} X_i > (\rho + y)x - u \right]
+ P \left[ N_x - \lambda x > \frac{g(x, u)}{EX} \right]
\leq C \left( e^{-\frac{c_{x-u}^2}{2}} + xe^{-\frac{1}{2}Q(x-z-u)} + e^{-c_{x-u}^2} \right).
\]

Hence, the upper bound on \( f_2 \) for sufficiently large \( x \) is as follows

\[
f_2 \leq Cx \int_{y \leq x} \left( e^{-\frac{c_{y-x-U}^2}{2}} + xe^{-\frac{1}{2}Q(y-x-u)} + e^{-c_{x-u}^2} \right) dP[X \leq u]
= \Delta f_{21} + f_{22} + f_{23}.
\]

The first two terms are \( o(P[X > yx]) \) as \( x \to \infty \) by Lemma 21. In bounding \( f_{23} \) we use

\[
f_{23} \leq Cx \int_{y \leq x} e^{-\frac{c_{y-x-U}^2}{2}} dP[X \leq u]
\leq Cxe^{-ax} + Cxe^{-Q(yz)} \int_{y \leq x} e^{Q(yz) - Q(u) - c_{y-x-U}^2} du
= o(P[X > yx]) \quad \text{as} \quad x \to \infty,
\]

by the same arguments as in the proof of Lemma 20, we omit the details. Finally, by replacing the preceding bounds in (3.41) we complete the proof. \( \square \)

**Proof of Lemma 14.** Markov’s inequality yields for \( s > 0 \)

\[
P \left[ \sum_{i=1}^{N} X_i \wedge u - \lambda zEX > yx - u \right] \leq e^{-s((\lambda EX + y)x - u)} \left( Ee^{s(X \wedge u)} \right)^{\lambda z + z^{1} - z}.
\]

(3.42)
Next, for some $1 < \beta < \left( \frac{1 + 2v}{1 - 2v} \right)^{-\alpha}$ (where $\alpha$ is the parameter of $Q$) we set

$$s = \beta \frac{Q(yx - u)}{yx - u}$$

and estimate the expectation in (3.42) as a sum of three terms

$$\mathbb{E}e^{s(x\wedge u)} = \int_0^{1/s} e^{sz} \, d\mathbb{P}[X \leq z] + \int_{1/s}^u e^{sz} \, d\mathbb{P}[X \leq z] + e^{su-Q(u)}$$

$$\leq 1 + s\mathbb{E}X + s^2\mathbb{E}X^2 + \int_{1/s}^u e^{sz} \, d\mathbb{P}[X \leq z] + e^{su-Q(u)}, \quad (3.43)$$

where we used $e^x \leq 1 + x + x^2$ on $[0,1]$ as in the proof of Theorem 6. Now, let $v = \frac{u}{yx}$ and, therefore, by the assumption $\varepsilon \leq v \leq 1/2 - \varepsilon$. Then, note that by Lemma 11 (i)

$$su - Q(u) = \frac{\beta u}{yx - u}Q(yx - u) - Q(u)$$

$$= \frac{\beta v}{1 - v}Q\left( \frac{1 - v}{v} \right) - Q(vyx)$$

$$\leq \left[ \frac{\beta v}{1 - v} \left( \frac{1 - v}{v} \right)^\alpha - 1 \right] Q(vyx) < 0, \quad (3.44)$$

by the choice of $\beta$ and $u$. The last inequality for all $u$ in the assumed interval leads to $e^{su-Q(u)} \leq C_s^2 x^2 e^{-\alpha Q(\varepsilon x)} \leq C_s^2$. On the other hand, integration by parts results in

$$\int_{1/s}^u e^{sz} \, d\mathbb{P}[X \leq z] \leq \mathbb{E}[X > 1/s] + C_s x \int_{1/s}^u e^{z-Q(s)} \, dz$$

$$\leq s^2 \mathbb{E}X^2 + C_s x \left( e^{su-Q(u)} + e^{1-Q(1/s)} \right)$$

$$\leq C \left( 1 + x^3 \left( e^{su-Q(u)} + e^{1-Q(1/s)} \right) \right).$$
Hence, due to (3.44) and $X \in SC$, the right hand side of the preceding inequality is bounded by $Cs^2$. The obtained bounds, in connection with (3.43), yield $E e^{(X \wedge u)} \leq 1 + sE X + C^* s^2$, for some constant $C^*$ and all $u$ in the given interval. Then, by replacing this estimate in (3.42) and using the definition of $s$, we obtain

$$P \left[ \sum_{i=1}^{\lambda x + z^1} X_i \wedge u - \lambda x E X > yx - u \right] \leq e^{-s(yx-u) + sz^1 - \varepsilon E X + z^2 Cs^2}$$

$$\leq Ce^{-Q(yx-u)(\beta - Cz^{-\varepsilon} - Cz^{-1}Q(z))}$$

$$\leq Ce^{-Q(yx-u)},$$

for all $x$ large enough, since $\beta > 1$; this concludes the proof. \qed

**Lemma 20.** Let $Q \in SC$ with $\alpha < 1/2$ and $Ee^{Q(x)} < \infty$. Then, for all $\delta < 1/2 - \alpha$, $\xi > (3/5)^2$ and all sufficiently small $\varepsilon > 0$, as $x \to \infty$

$$\int_{\xi x}^{x - \frac{1}{2} + \delta} \left( e^{-c(x-u)^2 \varepsilon} + xe^{-\frac{1}{2}zQ(x-u)} \right) \, dP[X \leq u] = o(P[X > x]).$$

**Proof.** Let

$$\int_{\xi x}^{x - \frac{1}{2} + \delta} \left( e^{-c(x-u)^2 \varepsilon} + xe^{-\frac{1}{2}zQ(x-u)} \right) \, dP[X \leq u] \triangleq f_1(x) + f_2(x).$$

Integration by parts and Proposition 11 give a bound for $f_1(x)$

$$f_1(x) \leq Ce^{-cz-Q(\xi x)} + C \int_{\xi x}^{x - \frac{1}{2} + \delta} \frac{x^2 - u^2}{u^2} e^{-Q(u)} e^{-c(\xi x-u)^2 \varepsilon} \, du$$

$$\leq Ce^{-cz} + Ce^{-Q(\xi x)} \int_{\xi x}^{x - \frac{1}{2} + \delta} e^{Q(u) - Q(\xi x) - c(\xi x-u)^2 \varepsilon} \, du,$$
CHAPTER 3. MODERATELY HEAVY TAILS

where in the second inequality we used that \((x^2 - u^2)/u^2 = O(1)\) for all \(u\) in the interval of integration. To show that \(f_1(x) = o(\mathbb{P}[X > x])\), it is enough to verify that the exponent in the last integral is upper bounded by \(-cx^{2\delta}\) for the given interval of \(u\). Thus, by assumption (3.2) and Lemma 11 (i), for all large \(x\)

\[
Q(x) - Q(u) - c\frac{(x - u)^2}{x} \leq aQ(x)\frac{x - u}{x} - c\frac{(x - u)^2}{x} \\
\leq Cx^\alpha \frac{x - u}{x} - c\frac{(x - u)^2}{x} \\
\leq Cx^{-(\frac{1}{2} - \alpha) + \delta - cx^{2\delta}}; \quad (3.45)
\]

since for all \(x\) large enough the right-hand side of the second inequality is increasing in \(u\) and \(u \leq x - x^{1/2 + \delta}\). Now, by choosing \(\delta < 1/2 - \alpha\) it follows that (3.45) is upper bounded by \(-cx^{2\delta}\). As far as \(f_2(x)\) is concerned, by discretization of the integral below, we have

\[
f_2(x) \leq Cx \int_{\xi x}^{x - x^{1/2 + \delta}} e^{-\frac{1+\delta}{2}Q(x-u)} d\mathbb{P}[X \leq u] \\
\leq Cx \sum_{j=1}^{\lfloor (1-\xi) x^{1/2 - \delta} \rfloor} e^{-\frac{1+\delta}{2}Q(x^{1/2 + \delta})} e^{-Q(x - (j+1)x^{1/2 + \delta})} \\
\leq Cx^{\frac{1}{2}} \left\{ e^{-\frac{1+\delta}{2}Q(x^{1/2 + \delta}) - Q(x - 2x^{1/2 + \delta})} \sqrt{e^{-\frac{1+\delta}{2}Q((1-\xi)x) - Q(\xi x - 2x^{1/2 + \delta})}} \right\}, \quad (3.46)
\]

where in the last inequality we used the concavity property of \(Q\), i.e. the maximum of all the summands is equal to either the first or the last summand. Thus, Lemma 11 (i) and (ii) imply that the first term in the preceding maximum is \(o(\mathbb{P}[X > x])\); the exponent of the second term is by Lemma 11 (i) bounded by

\[
\frac{1 - \varepsilon}{2} Q((1 - \xi)x) + Q(\xi x - 2x^{\frac{1}{2} + \delta}) \geq Q(x) \left(\frac{1 - \varepsilon}{2} (1 - \xi)^\alpha + \xi^\alpha \right).
\]
CHAPTER 3. MODERATELY HEAVY TAILS

Next, it is easy to verify that for any \( \xi > (3/5)^2 \), we can choose \( \epsilon > 0 \) sufficiently small such that \( \frac{1-\xi}{2}(1-\xi) + \xi > 1 \), and therefore we have \( f_2 = o(\mathbb{P}[X > x]) \) as \( x \to \infty \) and we obtain the statement of the lemma.

\[ \square \]

Lemma 21. Let \( Q \in SC \) and \( \mathbb{E}e^{Q(X)} < \infty \). Then for \( 1 > \xi > 0 \) and sufficiently small \( \epsilon > 0 \), as \( x \to \infty \)

\[ x \int_{\epsilon x}^{\xi x} e^{-Q((1-\epsilon)z-u)} d\mathbb{P}[X \leq x] = o(\mathbb{P}[X > x]). \]

Proof. By discretizing the integral for some \( \Delta > 0 \) one obtains

\[ x \int_{\epsilon x}^{\xi x} e^{-Q((1-\epsilon)z-u)} d\mathbb{P}[X \leq x] \]

\[ \leq x \sum_{j=0}^{\left\lfloor \frac{\epsilon x}{\Delta} \right\rfloor} e^{-Q((1-\epsilon)(\epsilon + j\Delta)z)} d\mathbb{P}[X \leq u] \]

\[ \leq x \sum_{j=0}^{\left\lfloor \frac{\epsilon x}{\Delta} \right\rfloor} e^{-Q((1-2\epsilon-(j+1)\Delta)z)} \mathbb{P}[X > (\epsilon + j\Delta)z] \]

\[ \leq Cx \sum_{j=0}^{\left\lfloor \frac{\epsilon x}{\Delta} \right\rfloor} e^{-Q((1-2\epsilon-(j+1)\Delta)z)-Q((\epsilon + j\Delta)x)}. \]

Next, Lemma 11 (iv) shows that each term in the last sum is \( o(\mathbb{P}[X > x]) \) for sufficiently small \( \epsilon \) and \( \Delta \) and, thus, the lemma holds.

\[ \square \]

3.5.2 Queueing Results

Proof of Theorem 12. Upper bound. Note that \( B \in TR \) implies \( B \in D \cap L \) and recall the definition of \( \hat{B}_0 \) from (3.21). Based on (3.19) and \( \hat{B}_0 \geq B_0 \geq R_0(t) \) for all
CHAPTER 3. MODERATELY HEAVY TAILS

$t > 0$, the waiting time $V_0$ can be upper bounded as

\[
V_0(1 - \rho - \delta) \leq \hat{B}_0 + \sum_{i=1}^{N(V_0)} B_i \wedge \hat{B}_0 - (\rho + \delta)V_0
\leq \hat{B}_0 + \sup_{t \geq 0} \left\{ \sum_{i=1}^{N(t)} B_i \wedge \hat{B}_0 - (\rho + \delta)t \right\}
\]

and, thus, for any positive $\delta < 1 - \rho$

\[
\mathbb{P}[V_0(1 - \rho - \delta) > x] \leq \mathbb{P}[\hat{B}_0 > (1 - \delta)x] + \mathbb{P}[W^{\rho + \delta}_{B \wedge \hat{B}_0} > \delta x]. \tag{3.47}
\]

Next we examine the asymptotic behavior of the second term in (3.47)

\[
\mathbb{P}[W_{B \wedge \hat{B}_0}^{\rho + \delta} > \delta x] \leq \mathbb{P}[\hat{B}_0 > \delta^2 x \mathbb{P}[W_{B}^{\rho + \delta} > \delta x] + \mathbb{P}[W_{B \wedge \delta^2 x}^{\rho + \delta} > \delta x],
\]

which by Lemmas 18, 1 and Proposition 11 for $\delta$ sufficiently small results in

\[
\mathbb{P}[W_{B \wedge \hat{B}_0}^{\rho + \delta} > \delta x] = o(\mathbb{P}[B > x]) \quad \text{as} \quad x \to \infty; \tag{3.48}
\]

we use the fact that for any $B \in \mathcal{BR}$ there exists $\alpha > 0$, such that $\mathbb{P}[B > x] \geq C/x^\alpha$
(see (1.6) in [79]). Substituting (3.48) in (3.47) yields

\[
\lim_{x \to \infty} \frac{\mathbb{P}[V_0(1 - \rho) > x]}{\mathbb{P}[B_0 > x]} \leq \lim_{x \to \infty} \frac{\mathbb{P}[\hat{B}_0 > (1 - \delta)x]}{\mathbb{P}[B_0 > \frac{1 - \rho - \delta}{1 - \rho - \delta} x]}. 
\]

Finally, recall Proposition 11 and then let $\delta \downarrow 0$ to complete the proof of the upper bound.

Lower bound. Recalling (3.25) for $\delta > 0$ results in

\[
\mathbb{P}[(1 - \rho)V_0 > x] \geq \mathbb{P}[(1 - \rho)H^{\omega}(x + \delta x) > x, B_0 > (1 + 2\delta)x, Z(V_0) \leq \delta x],
\]
CHAPTER 3. MODERATELY HEAVY TAILS

where $Z$ is defined in the proof of Lemma 19. Next,

$$\mathbb{P}[(1 - \rho)W > x] \geq \left( \mathbb{P}[(1 - \rho)H^\infty(x + \delta x) > x] - \sup_{y \geq (1 + 2\delta)x} \mathbb{P}[Z(V(y)) > \delta x] \right) \times \mathbb{P}[B > (1 + 2\delta)x]$$

and the result follows from the Law of Large Numbers, convergence of $Z(t)$ to a.s. finite $Z$ and $B \in \mathcal{T}\mathcal{R}$. □

Proof of Lemma 18. The proof is based on the analysis of sums of truncated random variables. Denote by $K$ the number of positive ladder heights in the Pollaczek-Khintchine representation of the stationary workload in an $M/G/1$ queue (see Ch. VII and IX in [4]).

(i) First observe that for all $y < \varepsilon x$

$$\mathbb{P}[(B_i \wedge \varepsilon x)^{(e)} > y] \leq \frac{EB}{E(B \wedge \varepsilon x)} \mathbb{P}[B^{(e)} > y]$$

(3.49)

and introduce a new absolutely continuous random variable $S$ defined by

$$\mathbb{P}[S > y] = \begin{cases} 
(1 + \delta)\mathbb{P}[B^{(e)} > y] \wedge 1 & y < y_0, \\
(1 + \delta)y_0^\beta \mathbb{P}[B^{(e)} > y_0]y^{-\beta} & y \geq y_0,
\end{cases}$$

where $y_0$ is finite. Then, for sufficiently large $x$ and all $y$, $\mathbb{P}[S \wedge \varepsilon x > y] \geq \mathbb{P}[(B \wedge \varepsilon x)^{(e)} > y]$ and, thus,

$$\mathbb{P} \left[ W_{B \wedge \varepsilon x} > x \right] = \mathbb{P} \left[ \sum_{i=1}^{K} (B_i \wedge \varepsilon x)^{(e)} > x \right]$$

$$\leq \mathbb{P} \left[ \sum_{i=1}^{\lfloor k \log \varepsilon x \rfloor} S_i \wedge \varepsilon x > x \right] + \mathbb{P}[K > k \log \varepsilon x],$$

(3.50)
where random variables \(\{S_i\}\) are i.i.d. equal in distribution to \(S\). Setting \(s = [\epsilon^{-1}]\) and using an easy modification of the proof of Theorem 1 of [44] we derive

\[
\mathbb{P} \left[ \sum_{i=1}^{[k \log \epsilon x]} S_i \wedge \epsilon x > x \right] \leq \left( \frac{[k \log \epsilon x]}{s + 1} \right) C x^{-(s+1)\beta} \\
\leq C (\log x)^{s+1} x^{-(s+1)\beta}.
\]

By an appropriate choice of \(s\) (and hence of \(\epsilon\)) one can ensure that \(\alpha < (s + 1)\beta\) and, thus,

\[
\mathbb{P} \left[ \sum_{i=1}^{[k \log \epsilon x]} S_i \wedge \epsilon x > x \right] = o(x^{-\alpha}). \quad (3.51)
\]

Furthermore, \(K\) is geometric and, therefore, a large enough \(k\) ensures \(\mathbb{P}[K > k \log \epsilon x] = o(x^{-\alpha})\). Finally, this bound on \(K\), (3.51) and (3.50) imply the proof of part (i).

(ii) The proof is similar to the proof of part (i). The Pollaczek-Khintchine representation results in

\[
\mathbb{P}[W_{B \wedge \epsilon x}^0 > x] = \mathbb{P} \left[ \sum_{i=1}^{K} (B_i \wedge \epsilon x)^{(e)} > x \right] \\
\leq \mathbb{P} \left[ \sum_{i=1}^{[\epsilon x]} (B_i \wedge \epsilon x)^{(e)} > x \right] + \mathbb{P}[K > \epsilon x].
\]

We point out that by (3.49) and the assumption of the lemma \(\mathbb{P}[(B \wedge \epsilon x)^{(e)} > y] \leq C \epsilon e^{-Q(y)}\). Next, introduce a new random variable defined by \(\mathbb{P}[S > y] = C \epsilon e^{-Q(y)} \wedge 1\) and note that for all \(y \geq 0\)

\[
\mathbb{P}[(B \wedge \epsilon x)^{(e)} > y] \leq \mathbb{P}[S \wedge \epsilon x > y].
\]
Thus, for any $1/k > \Delta > 0$ we can choose $c < \Delta/\mathbb{E}S$, rendering for sufficiently small $\varepsilon$

$$
P \left[ \sum_{i=1}^{\lfloor cx \rfloor} (B_i \land \varepsilon x)^{(c)} > x \right] \leq P \left[ \sum_{i=1}^{\lfloor cx \rfloor} S_i \land \varepsilon x - \lfloor cx \rfloor \mathbb{E}S > (1 - \Delta)x \right]$$

$$
\leq C e^{-(k+1)Q((1-\Delta)x)},
$$

where the last bound follows by Theorem 6. Assumption (3.2), for $\Delta < 1 - \beta$, yields $kQ(x) \leq (k + 1)Q((1 - \Delta)x)$ and, hence, part (ii) holds. \hfill \Box

### 3.5.3 Proof of Proposition 11

The proof is an immediate consequence of the following four lemmas (22 - 25) and the dominant convergence.

**Lemma 22.** Let $\{Y_i\}_{i=1}^{n}$ be independent, a.s. finite random variables and $X \in \mathcal{D} \cap \mathcal{L}$.

Then, for any fixed $n$, as $x \to \infty$

$$
P \left[ X + \sum_{i=1}^{n} X \land Y_i > x \right] \sim P[X > x].
$$

**Proof.** Note that $\{X + \sum_{i=1}^{n} X \land Y_i > x\}$ only if $\{X > \frac{x}{n+1}\}$. Hence, for any $k > 0$ the union bound yields

$$
P \left[ X + \sum_{i=1}^{n} X \land Y_i > x \right] \leq P[X > x - kn] + P \left[ X > \frac{x}{n+1} \right] \sum_{i=1}^{n} P[Y_i > k].$$
CHAPTER 3. MODERATELY HEAVY TAILS

Since \( X \) is both in \( \mathcal{L} \) and \( \mathcal{D} \) one easily obtains from the preceding inequality

\[
\lim_{x \to \infty} \frac{\mathbb{P}[X + \sum_{i=1}^{n} X \wedge Y_i > x]}{\mathbb{P}[X > x]} \leq 1 + C \sum_{i=1}^{n} \mathbb{P}[Y_i > k].
\]

Setting \( k \to \infty \), since \( Y_i < \infty \) a.s., yields the statement of the lemma. \( \square \)

**Lemma 23.** Let \( X \in \mathcal{D} \cap \mathcal{L} \), \( \mathbb{E}X < \infty \) and \( \{X_i^{(e)}\}_{i=1}^{n} \) be i.i.d. random variables equal in distribution to \( X^{(e)} \). Then for any \( \epsilon > 0 \) there exist \( C \) such that for all \( x > 0 \) and \( n \geq 1 \)

\[
\mathbb{P} \left[ X + \sum_{i=1}^{n} X \wedge X_i^{(e)} > x \right] \leq C(1 + \epsilon)^n \mathbb{P}[X > x].
\]

**Proof.** Define \( S_n \triangleq \sum_{i=1}^{n} X_i^{(e)} \). Then

\[
\mathbb{P} \left[ X + \sum_{i=1}^{n} X \wedge X_i^{(e)} > x \right] \leq \mathbb{P} \left[ X > \frac{x}{n+1}, X + S_n > x \right]
\]

\[
\leq \mathbb{P} \left[ X > \frac{x}{n+1} \right] + \mathbb{P}[X + S_n > x, S_n \leq x]. \tag{3.52}
\]

Next, by definition of the class of distributions \( \mathcal{D} \) we have \( s \triangleq \sup_{\mathbb{P}[X > 2x]} \mathbb{P}[X > x] < \infty \) and, hence,

\[
\mathbb{P} \left[ X > \frac{x}{n+1} \right] \leq s^{\log_2(n+1)} \mathbb{P}[X > x]. \tag{3.53}
\]

On the other hand, Theorem 1 and Lemma 4 (i) result in

\[
\mathbb{P}[X + S_n > x, S_n \leq y] = \int_{0}^{x} \mathbb{P}[X > x - y] d\mathbb{P}[S_n \leq y]
\]

\[
\leq C(1 + \epsilon)^n \int_{0}^{x} \mathbb{P}[X > x - y] \mathbb{P}[X > y] dy
\]

\[
\leq C(1 + \epsilon)^n \mathbb{P}[X > x], \tag{3.54}
\]
where in the last inequality we used Definition 5 of $S^*$. Inequality (3.52) in conjunction with (3.53) and (3.54) yields the statement of the lemma. \hfill \Box

**Lemma 24.** Let $X \in S^*$ and

$$
\lim_{x \to \infty} \frac{\mathbb{P}[X^{(e)} > x]}{x \mathbb{P}[X > x]} < \infty.
$$

If \(\{X_i^{(e)}\}_{i=1}^n\) are i.i.d. random variables equal in distribution to $X^{(e)}$, then, for any fixed $n$, as $x \to \infty$

$$
\mathbb{P} \left[ X + \sum_{i=1}^n X \wedge X_i^{(e)} > x \right] \sim \mathbb{P}[X > x].
$$

**Proof.** Define $S_n \overset{\triangle}{=} \sum_{i=1}^n X_i^{(e)}$. Then,

$$
\mathbb{P} \left[ X + \sum_{i=1}^n X \wedge X_i^{(e)} > x \right] = \mathbb{P} \left[ X > \frac{x}{n+1}, X + \sum_{i=1}^n X \wedge X_i^{(e)} > x \right]
$$

$$
\leq \mathbb{P} \left[ X > \frac{x}{n+1} \right] \mathbb{P}[S_n > x - k] + \mathbb{P}[X + S_n > x, S_n \leq x - k]
$$

$$
\overset{\triangle}{=} I_1(x) + I_2(x),
$$

(3.55)

where the second line follows from the independence of $X$ and $S_n$. Next we examine the asymptotic behavior of $I_1(x)$ and $I_2(x)$. First, the assumption of the lemma and Theorem 1 yield

$$
\lim_{x \to \infty} \frac{I_1(x)}{\mathbb{P}[X > x]} \leq \lim_{x \to \infty} \frac{\mathbb{P}[X > \frac{x}{n+1}] \mathbb{P}[X^{(e)} > x]}{\mathbb{P}[X > x]} \lim_{x \to \infty} \frac{\mathbb{P}[S_n > x - k]}{\mathbb{P}[X^{(e)} > x]}
$$

$$
\leq n \lim_{x \to \infty} \frac{\mathbb{P}[X^{(e)} > x]}{x \mathbb{P}[X > x]} \lim_{x \to \infty} x \mathbb{P} \left[ X > \frac{x}{n+1} \right] = 0.
$$

(3.56)
CHAPTER 3. MODERATELY HEAVY TAILS

By definition of the class of distributions $S^*$

$$
\int_0^\infty \frac{\mathbb{P}[X > x - y] \mathbb{P}[X > y]}{\mathbb{P}[X > x]} \, dy \rightarrow 2\mathbb{E}X \quad \text{as} \quad x \rightarrow \infty,
$$

and, therefore, the following double limit holds

$$
\lim_{k \to \infty} \lim_{x \to \infty} \int_k^{x-k} \frac{\mathbb{P}[X > x - y] \mathbb{P}[X > y]}{\mathbb{P}[X > x]} \, dy = 0. \quad (3.57)
$$

The quantity $I_2(x)$ can be upper bounded as

$$
I_2(x) \leq \mathbb{P}[X > x - k] + \int_k^{x-k} \mathbb{P}[X > x-y] \, d\mathbb{P}[S_n \leq y]. \quad (3.58)
$$

By Lemma 4 (ii) for any $\varepsilon > 0$ there exists $k_0$ such that for all $x > k > k_0$

$$
\int_k^{x-k} \mathbb{P}[X > x-y] \, d\mathbb{P}[S_n \leq y] \leq \frac{1+\varepsilon}{\mathbb{E}X} \int_k^{x-k} \mathbb{P}[X > x-y] \mathbb{P}[X > y] \, dy. \quad (3.59)
$$

Finally, substituting (3.59) in (3.58), using (3.57) and recalling $X \in S^* \subset \mathcal{L}$ results in

$$
\lim_{x \to \infty} \frac{I_2(x)}{\mathbb{P}[X > x]} \leq 1.
$$

Combining (3.55) with (3.56) and the preceding limit concludes the proof. \qed

Lemma 25. Let $X \in S^*$ and

$$
\lim_{x \to \infty} \frac{\mathbb{P}[X^{(e)} > x]}{x \mathbb{P}[X > x]} < \infty.
$$

If $\{X_i^{(e)}\}_{i=1}^n$ are i.i.d. random variables equal in distribution to $X^{(e)}$, then for any $\varepsilon > 0$ there exist $C$ such that for all $x \geq 0$ and $n \geq 1$

$$
\mathbb{P} \left[ X + \sum_{i=1}^n X_i^{(e)} > x \right] \leq C(1+\varepsilon)^n \mathbb{P}[X > x].
$$
CHAPTER 3. MODERATELY HEAVY TAILS

Proof. Define $S_n \triangleq \sum_{i=1}^{n} X_i^{(e)}$. Then

$$
P \left[ X + \sum_{i=1}^{n} X_i \wedge X_i^{(e)} > x \right] \leq P \left[ X > \frac{x}{n+1}, X + S_n > x \right]
$$

$$
\leq P \left[ X > \frac{x}{n+1} \right] P[S_n > x] + P[X + S_n > x, S_n \leq x].
$$

(3.60)

Next, Theorem 1 and Lemma 4 yield

$$
P \left[ X > \frac{x}{n+1} \right] P[S_n > x] \leq C(1 + \varepsilon)^n P \left[ X > \frac{x}{n+1} \right] P[X^{(e)} > x]
$$

$$
\leq C(1 + \varepsilon)^n (n+1) \sup_{z \geq 0} \{ z P[X > z] \} x^{-1} P[X^{(e)} > x]
$$

$$
\leq Cn(1 + \varepsilon)^n P[X > x],
$$

(3.61)

where in the last inequality we used the assumption and the fact that $EX < \infty$.

The second term in (3.60) can be bounded in the same way as in (3.54) since $X \in S^*$

$$
P[X + S_n > x, S_n \leq x] \leq C(1 + \varepsilon)^n P[X > x].
$$

Combining (3.60) with (3.61) and the preceding inequality concludes the proof. \(\square\)
Chapter 4

Conclusions

We focus on quantifying benefits of resource sharing in the presence of self-similar/heavy-tailed demand. The thesis evolves around open and closed models of resource sharing.

In the first part of the thesis, we study the classical model of a network switching element, a finite buffer single server queue fed by On-Off traffic sources. The primary performance measures of this model are the loss rate and buffer overflow probability. These quantities indicate the quality of service provided by the system. For the case of heavy-tailed On-periods, explicit asymptotic formulas for the loss rate and buffer overflow probability are derived. The results provide important insight into qualitative tradeoffs between the performance measures and system parameters, the key element for proper dimensioning of the system. Furthermore, we quantify the benefits of buffer sharing and scheduling in the same model.
The second part of the thesis is devoted to resource sharing with feedback control, such as TCP, the predominant transport protocol in the Internet. The processor sharing queue represents a baseline model of ideal flow control, i.e., when a number of users share bandwidth, each user receives an equal share. It is shown that job (file) transmission times admit an easy asymptotic characterization depending on whether the job size has a heavier or lighter tail than the Weibull distribution $e^{-\sqrt{x}}$. In other words, this fundamental model exhibits a phase transition at $e^{-\sqrt{x}}$. Furthermore, the newly developed large deviations approach also provides a mathematical framework for proving related results.
Bibliography


