Sojourn time tails in processor-sharing systems

Thomas Stieltjes Institute
for Mathematics

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-1495-3
NUR: 919
Subject headings: queueing theory / communication systems / asymptotics
2000 Mathematics Subject Classification: 60K25, 60F10, 68M20, 90B18, 90B22

Printed by Ponsen & Looijen BV.

Sojourn time tails in processor-sharing systems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op donderdag 5 februari 2009 om 16.00 uur

door

Regina Robertovna Egorova

geboren te Batumi, Georgië

Dit proefschrift is goedgekeurd door de promotor:

prof.dr.ir. S.C. Borst

Copromotor:
dr. A.P. Zwart

Acknowledgements

This thesis describes the research results obtained in the course of my PhD project during the last four years. I gratefully acknowledge the financial support for this project provided by The Netherlands Organization for Scientific Research (NWO).

It is a pleasure to express my gratitude to my supervisors Sem Borst, Onno Boxma and Bert Zwart for their invaluable coaching and guidance. Their scientific advice and continuous support and encouragement were essential for completion of the thesis. I am grateful to Michel Mandjes for excellent lessons in large-deviations theory and the joint work on the material presented in Chapter 4. I thank Lasse Laskelä and Maaike Verloop for many helpful discussions on fluid models. Finally, my warm thanks are due to colleagues at CWI and TU/e for creating such a friendly and stimulating environment.

Regina Egorova
February 2009

Contents

Chapter 1

Introduction

The processor-sharing discipline was originally introduced as a modeling abstraction for the design and performance analysis of the processing unit of a computer system. Under the processor-sharing discipline, all active tasks are assumed to be processed simultaneously, each receiving an equal share of the server capacity. Various extensions of the standard discipline have been developed in order to capture scenarios with heterogeneous service shares and network settings. Over the past several years, the processor-sharing discipline has received renewed attention as a powerful tool in modeling and analyzing dynamic bandwidth sharing among elastic transfers in communication networks like the Internet.

The key property of the processor-sharing discipline is the simultaneous resource sharing among all users present in the system. As a result of the simultaneous processing, small requests can overtake large requests, and are thus protected from experiencing excessive delays. Due to this feature, the processor-sharing discipline is particularly suitable for reducing the adverse impact of the high variability of service requests observed in data networks.

The sojourn time of a customer, i.e. the amount of time a customer spends in the system from his arrival until his service completion, is the most important performance measure for processor-sharing systems. This is a particularly relevant performance measure for modeling data transmissions in the Internet where Quality-of-Service requirements become increasingly stringent. The exact analysis of the sojourn time has however proved to be extremely hard and often impossible due to the fact that knowledge of the residual service times of all the jobs present in the system is required.

In this thesis we study various asymptotic properties of the sojourn time distribution. We are mainly interested in the probability of the sojourn time being extremely large. The advantage of considering the asymptotic behavior is that the analysis often provides insight into the typical scenario for such a long sojourn time to occur. Moreover, the resulting asymptotic formulas can be used for approximate analysis, providing useful estimates in situations when numerical procedures become unreliable. In order to analyze the sojourn time asymptotics, we apply several probabilistic and analytic techniques, such as Laplace transforms, branching arguments,

large-deviations methods and fluid limits.

The remainder of this introductory chapter is organized as follows. In Section 1.1 we provide some basics on queueing systems and discuss how these may be used to model and analyze the performance of communication networks. The basic egalitarian processor-sharing discipline and several of its extensions are discussed in detail in Section 1.2. In Section 1.3 we briefly explain the main concepts and techniques that we have applied in the course of the research. Section 1.4 presents a literature review on the performance analysis of processor-sharing queues and bandwidth-sharing networks. Section 1.5 concludes this chapter with an outline of this monograph.

1.1 Basic queueing concepts

In today's society, telecommunication systems play a crucial role in all aspects of life. Various new applications continue to emerge while both technological capabilities and consumers' demands show continuous growth. Ever since the early 20th century, when public telephony systems first came into service, queueing-theoretic models have been a key technique in the design and performance analysis of telecommunication systems. The pioneering work in queueing theory dates back to Erlang [50]. He developed a model to describe the performance of a telephony system and estimate the fraction of lost calls.

To evaluate the performance of a communication system, various mathematical queueing models may be used. In general, a queueing model describes the operation of a number of servers of finite capacity which are used to provide service to a population of customers. A basic model includes (stochastic) characteristics of the customer arrivals and service requirements, and characteristics of the servers. The terms "servers"and "customers" (the term "jobs" is also often used) may refer to arbitrary objects involved in various sorts of queueing processes; one can think of applications in e.g. public customer service, transportation systems, call centers, inventory systems. In this thesis, we focus on queueing models for the transmission of data files in a network where all transfers simultaneously share a possibly state-dependent transmission rate. Viewing the available bandwidth as the capacity of the server and the individual file transfers as the customers in the system, the above-described scenario can be modeled as a classical processor-sharing system or an extension thereof.

The behavior of a queueing system is analyzed in terms of so-called *performance measures*. Some of the most commonly considered performance measures are the queue length, the workload, the waiting time, the sojourn time, and the throughput. The choice of the relevant performance measure depends on the system in question and the purpose of the analysis. In some situations it is sufficient to gain insight into the average behavior while in other cases it may be critical to obtain the entire probability distribution of the performance measure of interest.

In order to generally characterize queueing models, we distinguish three main components. First of all, the physical structure of the network plays an important

role. By this we mean the amount of available resources, the network's capacity and the connection topology.

Second, the performance of the system depends strongly on the traffic characteristics. The most important elements are the time between two consecutive customer arrivals and the service requirements of the customers. Both the interarrival times and the service requirements are commonly assumed to be sequences of independent identically distributed random variables. Typically interarrival times and service requirements are assumed to be mutually independent. In order to specify the queue in terms of the above-mentioned entities, we use the conventional notation introduced by Kendall [72]. This notation is of the form $A/B/N$ where the first letter refers to the distribution of the interarrival times, the second represents the distribution of the service requirements, and the third stands for the number of servers in the system. The most commonly used distributions are the exponential distribution denoted by M (for memoryless), deterministic denoted by D and the general distribution denoted by G.

The third component which has a significant influence on the behavior of a queueing system is the *service discipline*, which describes the order and the manner in which the customers receive service. There is a wide variety of service disciplines. One of the simplest disciplines is First Come First Served, where the customers are served in the order of arrival. For some systems, disciplines like Last Come First Served, Random Order of Service, etc. can be used as appropriate models. The service discipline may also differentiate among the customers by assigning priorities to specific classes of jobs.

We refer to the textbooks by Asmussen [7], Cohen [37], and Tijms [107] for fundamental models and results in queueing theory.

1.2 Processor-sharing disciplines

The processor-sharing (PS) discipline first became popular by the work of Kleinrock [76, 78], and was originally proposed as an idealization of round-robin scheduling in time-sharing systems. The recent surge of interest in PS queues is motivated by their application in the performance analysis of bandwidth-sharing schemes in the computer communication networks such as the Transmission Control Protocol (TCP) in the Internet, see e.g. Ben Fredj *et al.* [11], Núñez-Queija [87], Roberts and Massoulié [99].

TCP uses an end-to-end flow control protocol which dynamically adjusts the transmission rates in response to the current level of network congestion and is one of the core principles of Internet operation. While individual packets are served one-by-one in a FCFS manner, over somewhat longer time scales TCP ensures that the various transfers are served simultaneously at roughly equal rates. As a result, the service rate of a given transfer fluctuates over time as the total number of active transfers varies when new transfers start or others complete their transmission.

The egalitarian PS (EPS) discipline can be regarded as a basic model which approximates the behavior of a single resource shared in a fair manner. Under EPS

all the capacity of the resource is assumed to be shared equally between all the customers in the system. One of the main limitations of the EPS model is that it does not apply for heterogeneous systems, where jobs may receive different service shares. To model such situations, a number of multi-class extensions of the EPS discipline have been proposed. The main model that allows for *unequal* sharing is Discriminatory Processor Sharing (DPS), where flows of different classes receive service at different rates.

The extension of the PS discipline from a single-node system to a network with multiple shared links gives rise to bandwidth-sharing networks as introduced by Massoulié and Roberts [84], [99]. In such network scenarios, the rate allocation becomes a non-trivial problem, and is commonly assumed to be governed by a utility maximization principle, see Kelly *et al.* [70], Mo and Walrand [85].

We now proceed to discuss in further detail the EPS discipline and some of its extensions mentioned above.

1.2.1 Egalitarian processor sharing

In the EPS queue a single resource is equally shared among all jobs present in the system. In other words, a PS server with capacity c assigns each of $n > 0$ customers present service rate $r = c/n$. Each arriving request is immediately taken into service and continues to receive service at a varying rate which depends on the total number of customers present until it completes, i.e. until the cumulative amount of service received equals the original service requirement.

The PS discipline has several appealing properties. The key feature of the PS discipline is that it prevents small jobs from being excessively delayed by large jobs. At the same time, large jobs receive service continuously and do not experience starvation as in priority systems. This property of the PS discipline is particularly useful in systems with highly diverse service requests, such as data transfers in the Internet.

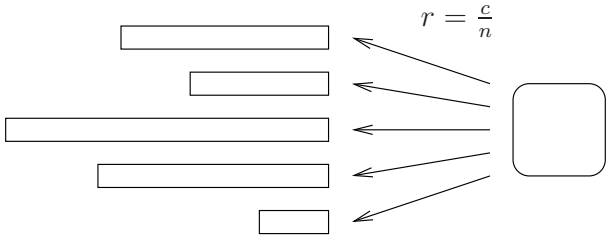

Figure 1.1: Egalitarian processor sharing.

For the PS queue with Poisson arrivals various important properties are known. In the stationary regime, the queue length has a geometric distribution which only depends on the traffic load $\rho = \lambda \mathbf{E}[B] < c$ (Sakata *et al.* [102]),

$$\pi_n = \left(1 - \frac{\rho}{c}\right)\left(\frac{\rho}{c}\right)^n, \qquad n = 0, 1, \ldots, \tag{1.1}$$

where λ denotes the arrival rate and B stands for the service requirement. Thus, the queue length distribution is insensitive in the sense that it only depends on the service requirement distribution through its mean and not any higher-order statistics. By Little's law, this also implies the insensitivity of the mean sojourn time. Furthermore, the expected conditional sojourn time for a particular request is proportional to the size of the request, which indicates the *fairness* of the PS discipline. Given that the job size is equal to τ, the expected sojourn time for this job is

$$\mathbf{E}[V(\tau)] = \frac{\tau}{c-\rho}.$$

For the distribution of the sojourn time, however, there are no simple closed-form expressions available.

It is worth mentioning that the above results on the queue length distribution and mean sojourn time also hold in the EPS model with several traffic classes, see e.g. Cohen [36], Kelly [68]. Suppose that the customers of class i arrive according to a Poisson process with rate λ_i and have service requirements B_i, $i = 1, \ldots, M$. Denote the load of class i by $\rho_i = \lambda_i \mathbf{E}[B_i]$ and the number of customers of class i by Q_i. In the multi-class case, the joint distribution of the number of customers has a simple product form:

$$\mathbf{P}(Q_1 = n_1, ..., Q_M = n_M) = \left(1 - \sum_{i=1}^{M} \frac{\rho_i}{c}\right) \binom{n_1 + ... + n_M}{n_1 ... n_M} \left(\frac{\rho_1}{c}\right)^{n_1} ... \left(\frac{\rho_M}{c}\right)^{n_M}.$$

The classical EPS model assumes a constant service capacity, while in many practical cases the available capacity for data transfers fluctuates dynamically due to the presence of high-priority traffic types with time-varying capacity requirements. For instance, in multi-service communication networks, traffic can be categorized into *streaming* flows (voice, video, etc.) and *elastic* flows (data files, Web pages, etc.), see e.g. Roberts [98]. Streaming flows require strict packet-level delay guarantees for the duration of their connection time, whereas elastic traffic is less sensitive to packet-level delays. One way to meet the Quality-of-Service requirements is by prioritizing streaming traffic. The bandwidth left over by the transmission of streaming traffic is made available to elastic traffic. In this case, the streaming flows 'do not see' the elastic flows, so their performance can be evaluated using traditional queueing models. Assuming fair sharing among elastic flows, the performance experienced by the elastic traffic, on large time scales, can be modeled as a PS system with a service rate (corresponding to the bandwidth left over by the streaming flows) that fluctuates according to some stochastic process (see e.g. Delcoigne *et al.* [40]).

1.2.2 Discriminatory processor sharing

The DPS discipline has gained popularity as a flexible model which allows for differentiation among heterogeneous traffic types. The DPS model was first proposed by Kleinrock [78] under the name Priority Processor Sharing, and is essentially a multi-class extension of the EPS discipline, where the various classes of traffic are

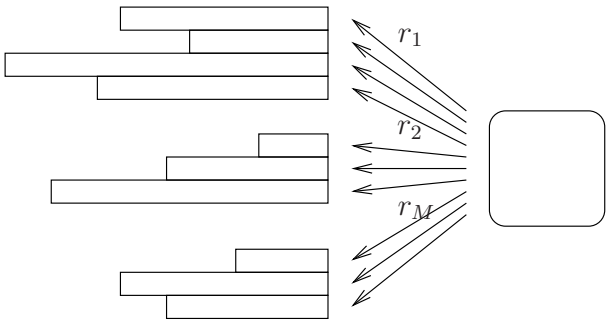

Figure 1.2: Discriminatory processor sharing.

assigned arbitrary positive service weights. The total service capacity is shared among all the present users in proportion to the respective per-class weights, and thus, the per-class service rate depends on the number of users of all the classes currently present in the system.

To give a formal description of the DPS discipline, suppose that there are M customer classes sharing a server of capacity c. All customers present in the system are served simultaneously with rates dependent on a vector of weights $(w_1, \ldots, w_M) > 0$. If there are Q_j customers of class j present in the system, $j = 1, \ldots, M$, each class-k customer is served at rate

$$r_k = \frac{w_k c}{\sum_{j=1}^{M} w_j Q_j}, \qquad k = 1, \ldots, M.$$

Figure 1.2 presents the basic DPS scheme. In case all weight factors are equal, DPS is equivalent to the multi-class EPS discipline. It is worth mentioning that although the DPS discipline has a strong resemblance with the ordinary PS discipline, the analysis of a DPS system is considerably more involved. In particular, the fundamental results for the PS system with Poisson arrivals do not extend to the DPS queue.

Before proceeding to networks of PS queues, we also mention one other related yet different multi-class discipline. In the GPS discipline, the per-class service rate is also governed by pre-assigned weight factors, but in contrast to DPS, the rate only depends on whether a queue is empty or not, and not on its exact length. Each non-empty class now receives a certain guaranteed share of the capacity. We refer to Van Uitert [109] for more details on GPS. We remark that the GPS discipline is different from the Generalized Processor Sharing as considered by Cohen [36]. The latter model, in which the service rate of each customer is determined by an arbitrary positive function of the total queue length, is a more abstract generalization of PS with a state-dependent service rate.

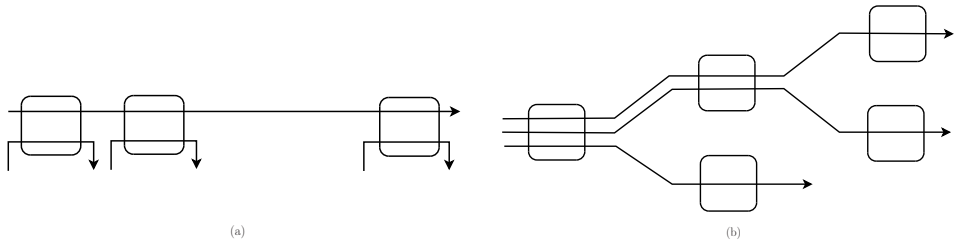

Figure 1.3: Examples of bandwidth-sharing networks: (a) linear network and (b) tree network.

1.2.3 Bandwidth-sharing networks

A further extension of the basic PS model is provided by bandwidth-sharing networks where flows may require simultaneous service from several resources as introduced by Massoulié and Roberts in [84, 99]. More precisely, a network consists of a finite number of links labeled by $j = 1, \ldots, J$. We denote the vector of finite link capacities by $C = (C_1, \ldots, C_J)$. The network is offered traffic from several classes indexed by $i = 1, \ldots, I$. Each class is characterized by a route, i.e., a nonempty subset of $\{1, \ldots, J\}$, which represents the set of links traversed by the traffic from that class. We introduce a $J \times I$ incidence matrix A such that $A_{ji} = 1$ if link j belongs to the route of class i, and $A_{ji} = 0$ otherwise. The distinctive feature of bandwidth-sharing networks is that the flow requires service from all resources on its route simultaneously, which is in contrast to classical queueing networks where a customer visits the nodes sequentially. See Figure 1.3 for an illustration.

In bandwidth-sharing networks, the capacity is allocated to the various traffic classes according to a pre-specified rate allocation policy, while within each class the bandwidth is fairly shared among all competing flows. Such rate allocation policy can be regarded as generalization of a PS discipline from a single node to a network with several shared links. Since the idea was first presented by Kelly *et al.* [70], rate allocation policies based on global network utility optimization principles have been widely used to model various resource-sharing systems and network protocols.

We consider rate allocation policies that maximize a network utility function depending on the current population of active flows. Specifically, for a given number $z = (z_1, \ldots, z_I) \neq (0, \ldots, 0)$ of active flows, the *per-flow* rate allocation $x(z)$ is determined by the solution of the optimization problem:

$$(P) \qquad \begin{array}{ll} \text{maximize} & \sum_{i=1}^{I} z_i U_i(x_i) \\ \text{subject to} & Ax \cdot z \leq C,\ x \geq 0, \end{array}$$

where the utility functions $U_i(\cdot) : \mathbb{R}_+ \to [-\infty, \infty]$ are strictly concave on $(0, \infty)$. By $x \cdot z$ we denote a vector obtained by component-wise multiplication of vectors x and z. With the additional convention that $x_i(z) = 0$ when $z_i = 0$, the rate allocation is

uniquely determined since the above optimization problem is strictly concave.

Often it is more useful to consider the *per-class* rate allocation. These rates can be obtained by multiplying the per-flow rates with the number of flows per class, or directly, by replacing the optimization problem for the per-flow rate allocation $x(z)$ by an equivalent one for the per-class rate allocation $\Lambda_i(z) = x_i(z) \cdot z_i$, $i = 1, \ldots, I$. The rate allocation vector Λ must satisfy the capacity constraints $A\Lambda \leq C$.

The most commonly studied utility-based rate allocation policy is the so-called *(weighted) α-fair* policy introduced by Mo and Walrand [85], where the utility functions $U_i(\cdot)$ are given by

$$U_i(x_i) = \begin{cases} w_i \frac{{x_i}^{1-\alpha}}{1-\alpha}, & \alpha \in (0,\infty)\backslash\{1\}, \\ \\ w_i \log x_i, & \alpha = 1, \end{cases} \tag{1.2}$$

where the weights w_i, $i = 1, \ldots, I$, are some positive constants and α is a fairness coefficient. In a single-link scenario, the α-fair policies actually reduce to the DPS discipline with weights given by $w_i^{\frac{1}{\alpha}}$, and in particular, to EPS in case the weights are equal.

The family of α-fair bandwidth-sharing strategies includes several common fairness concepts as special cases. In particular, the case $\alpha = 1$ and the limiting cases $\alpha \to 0$ and $\alpha \to \infty$ correspond to a rate allocation that is *proportional fair*, achieves *maximum throughput*, and is *max-min fair*, respectively. The special case $\alpha = 2$ corresponds to the bandwidth allocation which achieves *minimal potential delay* [84]. If in addition the weights are chosen to be the reciprocal of the squared round-trip time on the corresponding route, this α-fair bandwidth allocation can be viewed as an appropriate model for the TCP protocol of the Internet, see e.g. Padhye *et al.* [90]. In this context, it is worth noting that the transmission rates in the Internet are not assigned by some centralized control mechanism based on explicit optimization. Instead, the transmission rates are determined through end-to-end congestion control protocols, implemented only in end-user nodes which may be interpreted as solving the utility maximization problem in a distributed fashion (see Kelly [69] for a comprehensive discussion).

In general, the α-fair rate allocation Λ or x as the solution of the optimization problem (P) can not be obtained in explicit form. There are only a few examples of simple network topologies for which a closed-form expression is available, see Bonald and Massoulié [14]. Nevertheless, various useful analytical properties of the rate allocations as function of the number of flows are known (Kelly and Williams [71]).

1.3 Methodology

In this section we briefly sketch the main methods that we apply in this monograph. In Chapters 2 and 3 we use Laplace transform techniques and results for geometric random sums and branching processes in order to obtain the asymptotic behavior of the sojourn time. In Chapters 4 and 6, we apply arguments from

large-deviations theory. In Chapter 5 we analyze the behavior of bandwidth-sharing networks by means of fluid-limit approximations. Below we present the basic ideas of these methods and introduce some preliminaries.

1.3.1 Branching processes and Laplace transforms

In Chapters 2 and 3 we study the sojourn time in a PS system with Poisson arrivals by means of its LST. The LST is particularly useful for asymptotic analysis, e.g., to determine the asymptotic behavior of tail probabilities (see e.g. Widder [113]). The limitations of this approach are that in the first place, it is applicable only to models where an expression for the LST is available in sufficiently explicit form, and second, obtaining the relationship between the LST and the tail probability may be a challenging problem.

In order to simplify the derivation of the LST and the tail probability, we make use of the branching process representation of the sojourn time. The branching process representation and decomposition of the sojourn time into a sum of independent random variables (called *delay elements*), conditioned on the number of customers in the system, was established by Yashkov [116] for the M/G/1 PS queue and later extended by Ott [89]. Grishechkin [56] generalized the method using Crump-Mode-Jagers branching processes and applied it to more general service disciplines and the PS discipline in particular. With this approach, the problem of deriving the LST of the sojourn time reduces to the computation of certain functionals of branching processes which are tractable enough for analysis. Furthermore, the structure of the representation and the properties of the PS discipline allow one to apply powerful asymptotic results for geometric random sums. The latter is discussed in detail in Chapter 2.

The decomposition procedure is as follows. Suppose that a tagged customer with a service requirement τ arrives at time epoch $t = 0$. We consider the dynamics of the system from time 0 until the time epoch when the service of the tagged customer is completed. The first step is to introduce a time scale transformation which allows for the branching process representation. This *time-change* method is widely used in the analysis of PS queues, cf. [87], [116]. With this approach all investigations are performed depending on the amount of service $S(t)$ attained by the tagged customer during the time interval $[0, t]$, rather than the actual time scale. Denote the number of customers in the system (including the tagged customer) at time t on the original time scale by $Q(t)$. The amount of service received by the tagged customer during the time interval $[0, t]$ is then

$$s = S(t) = \int_0^t \frac{1}{Q(u)} \mathrm{d}u. \tag{1.3}$$

Below we use the symbols t and s for time epochs on the original and the transformed time scales, respectively.

We define $V(s) = \inf\{t \geq 0 : S(t) \geq s\}$, that is the time epoch when the attained amount of service reaches level s. In this notation, the sojourn time of the tagged customer is $V(\tau)$. Further, we introduce the process $X(s)$ as the number

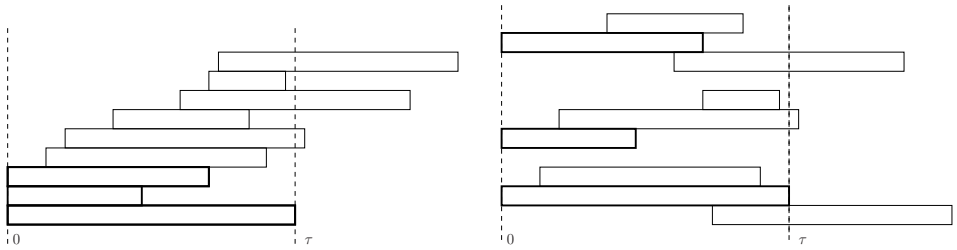

Figure 1.4: Branching process representation.

of customers (including the tagged customer) at the server at the epoch when a cumulative amount of service s is received by the tagged customer. The process $X(s)$ can be defined as $X(s) = Q(V(s))$. Evidently, the sojourn time $V(\tau)$ can be expressed in terms of the process $X(s)$ as

$$V(\tau) = \int_0^\tau X(s)\mathrm{d}s. \tag{1.4}$$

Now we show how to construct a branching process in order to describe the behavior of the PS queue. Consider each active customer in the queue as an individual in a certain population. Each individual has an exponentially distributed life time. During its life time the individual produces children according to a Poisson process with rate λ. The birth of an individual corresponds to the arrival of a new customer and a death corresponds to a service completion. The customers present in the system at the arrival of the tagged customer are called *progenitors* while the new arrivals occurring after $t = 0$ are assumed to be *descendants* of these progenitors. If n progenitors are present in the system then each new arrival is declared with probability $1/n$ to be a descendant of any of these progenitors. The tagged customer is also considered as a progenitor. Each branching process is formed by one progenitor and all its descendants (for more details see [116]). See Figure 1.4 for an illustration. The bars with thick borders represent progenitors and the bars with regular borders represent children. The length of the bars indicates the service requirement. Note that the total birth and death rates in the branching process (after the time change (1.3)) correspond to the arrival and departure rates in the PS queue. Thus, under the described branching construction, the queue length process $X(s)$ is stochastically equivalent to the total size of the population of the branching process at time s.

Let Q denote the number of customers in the system upon arrival of the tagged customer. Define $V_0(\tau)$ as the sum of the ages reached by the tagged customer and its direct descendants up to time τ, and $C_i(\tau)$ as the sum of the ages (attained amount of service) reached by the ith progenitor and its descendants up to time τ, i.e. during the life time of the tagged customer. Yashkov [116] established that the

sojourn time of the tagged customer can be represented as

$$V(\tau) = V_0(\tau) + \sum_{i=1}^{Q} C_i(\tau). \tag{1.5}$$

Notice that the random variable $V_0(\tau)$ is in fact the sojourn time of a customer which arrives into an empty system. In general, we will call the variables $V_0(\tau)$ and $C_i(\tau)$ the *delay elements*.

The essential observation here is that the elements $V_0(\tau)$ and $C_i(\tau)$, $i = 1, 2, ..., Q$, are mutually independent. This is due to the fact that the service requirements and the arrivals of various customers are independent. The elements $C_i(\tau)$ are also identical in distribution. Again, we refer to Yashkov's work [116, 117] for the details behind these results. Another exposition can be found in Chapter 3 of Núñez-Queija [87], where this decomposition result is extended to PS queues with service interruptions.

1.3.2 Large deviations

Large-deviations (LD) theory refers to a very powerful approach which is particularly useful for the analysis of rare-event probabilities in complex queueing systems. The LD approach is in essence a method that transforms the problem of analyzing the stochastic behavior of the system into the problem of optimizing a certain deterministic function. The implications of the method are threefold. First, it provides the exponential decay rate for the probability of a rare event. Second, it enables us to understand the most likely manner in which this rare event occurs, that is to find the most probable sample path that leads to the rare event. Finally, this approach provides foundations for fast and efficient rare-event simulation algorithms.

LD theory is principally concerned with large fluctuations of stochastic objects away from their average behavior, such that the probability of the fluctuations is exponentially small. For a fundamental example of LD results we refer to Cramer's theorem describing the fluctuations of the empirical mean of a sequence of independent identically distributed random variables. The theorem states that the probability of large fluctuations of the empirical mean from the expected value decays exponentially fast as the number of random samples grows large. More details on general LD results can be found in, e.g., Dembo and Zeitouni [41].

An important LD concept is the *sample-path Large-Deviations Principle* (sp-LDP). This principle describes the limiting behavior of a sequence of probability measures in terms of a so-called *rate* function. A formal statement is as follows.

Consider a sequence of stochastic processes $(X_n(t),\ n \in \mathbb{N}, t \geq 0)$. Denote by Ω a space of sample paths, for example the space C of continuous functions. We say that $X_n(\cdot)$ obeys a sp-LDP with rate function I, if

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}(X_n(\cdot) \in S) \leq -\inf_{x\in S} I(x), \qquad \text{for any closed set } S \subset \Omega, \tag{1.6}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(X_n(\cdot) \in T) \geq -\inf_{x\in T} I(x), \qquad \text{for any open set } T \subset \Omega. \tag{1.7}$$

The function $I(\cdot)$ is a non-negative lower-semicontinuous function on Ω. It roughly represents the measure of how likely the occurrence of each sample path is. The minimization of the rate function corresponds to the identification of the most probable sample path.

Inequality (1.6) is the upper bound of the sp-LDP, and (1.7) is the lower bound of the sp-LDP. It is often a challenging task to derive a sp-LDP with useful expressions for I. It is common to first obtain the most probable sample paths in the upper bound and in the lower bound. If the paths coincide, then the most probable path for the rare-event probability has been found.

It is important to mention that the LD method typically applies to rare events that result from a large number of unlikely events that occur at the same time, a so-called conspiracy. For example, in PS systems with time-varying capacity a large sojourn time is the result of both the arrival process generating traffic at a higher rate than usual and the service process offering service at a lower rate than usual. In contrast, when the flow size has a heavy-tailed distribution, in most cases the large size of the request itself is responsible for the large sojourn time, and this scenario can not be captured by the LD method.

1.3.3 Fluid limits

Fluid limits and fluid approximations of stochastic systems have emerged as a key technique for analyzing stability and time-dependent behavior of multi-class stochastic networks. Generally speaking, a fluid approximation represents a functional strong law of large numbers which can be stated for a large class of stochastic systems. This method was first applied to a two-station, two-class network by Rybko and Stolyar [101]. It became popular by the work of Dai [39] who generalized the method and established crucial stability criteria. For an extensive overview on fluid-limit results, the reader may consult the books of Chen and Yao [32] and Whitt [115] and references therein.

In Chapter 5 we will apply fluid models to describe the behavior of the queue length in bandwidth-sharing networks operating in an overload regime. Instead of formulating the fluid-limit approach in a general setting, we provide here a short description of the fluid limits in the context of our model.

We consider a bandwidth-sharing network as in Subsection 1.2.3. Let $\mathcal{R}$ be a sequence of positive real numbers increasing to infinity. With each $r \in \mathcal{R}$ we associate a stochastic model in the following way. We suppose that all the systems have the same vector of link capacities C, incidence matrix A and bandwidth-sharing policy Λ with parameters (α, w). The flow size distribution in the rth system is given by B^r and the arrival rate is λ^r.

Let us now introduce the scaled versions of the stochastic process of interest. For $r \in \mathcal{R}$ and $t \geq 0$, let

$$\overline{Z}^r(t) = \tfrac{1}{r} Z^r(rt),$$

where the vector $Z^r(t)$ denotes the number of active flows at time t in the rth system. The system parameters B^r and λ^r are assumed to converge to the limits B and λ in an appropriate manner. For the sequence of the scaled initial conditions,

we assume that as $r \to \infty$,

$$\overline{Z}^r(0) \to z(0), \qquad \text{almost surely.}$$

A crucial step in the fluid-limit analysis is establishing the tightness of the scaled sequence or, more challenging but more powerful, its convergence. The analysis of the stochastic system then reduces to the derivation and analysis of the deterministic fluid model (in the form of functional equations) that describes the behavior of the limit points.

The fluid-limit models present a convenient tool for establishing the stability of complex queueing systems. It is known that the stability of multi-class queueing networks can not be assured by the usual traffic load conditions and is dependent on the service discipline. Some two-station counterexamples are given for instance, in Bramson [28], Lu and Kumar [80], Rybko and Stolyar [101]. Dai [39] has shown that the queueing network is stable in the sense that the associated Markov process is positive recurrent for any given initial state if the corresponding fluid limit is stable. Based on this result, stability of various priority disciplines was proved. However, it is important to mention that the method of Dai [39] implicitly assumes that the service discipline is a head-of-the-line discipline, and thus, is not applicable to PS type disciplines. The complication is due to the fact that in PS systems (and bandwidth-sharing networks) the number of active customers depends on the arrival rate and on the entire (remaining) flow size distributions of all initial and arriving flows.

1.4 Literature overview

In this section we review several results for the basic PS model and some of its extensions. In Subsections 1.4.1 and 1.4.2 the focus is on the analysis of the sojourn time. References to the literature on other performance measures can be found in e.g. the surveys Altman *et al.* [4], Borst *et al.* [25]. Subsection 1.4.3 gives an overview of the literature on stability and overload behavior of bandwidth-sharing networks.

1.4.1 Egalitarian processor sharing

There exists a vast amount of literature devoted to the derivation of the complete distribution of the (conditional) sojourn time in the egalitarian PS queue. Coffman *et al.* [35] first derived the expression for the LST of the sojourn time conditioned on the service requirement and number of customers upon arrival in the M/M/1 PS queue. Sengupta and Jagerman [106] obtained the LST of the sojourn time conditioned only on the number of customers at the arrival epochs. Yashkov [116] found an analytic expression for the distribution function for the M/G/1 PS queue in terms of a double LST based on the decomposition of the sojourn time into a set of independent random variables. Schassberger [104] developed another approach to derive the LST by considering PS as a limiting case of the round-robin discipline. Using methods similar to Yashkov's, the LST of the conditional sojourn time

was also studied by Grishechkin [56], Ott [89] and Núñez-Queija [87]. Zwart and Boxma [122] derived a new, more explicit expression for the LST involving a series expansion. Van den Berg [13] obtained results for the LST and the moments of the sojourn time by considering the PS queue as a limiting model of the queue with feedback. Using the LST results from [35], Morrison [86] derived an integral representation for the sojourn time probability distribution. Cheung [33] obtained bounds for all moments of the conditional sojourn time in the M/G/1 PS queue based on the LST transform and a novel queue length decomposition approach. For the GI/M/1 PS, Ramaswami [94] derived the LST of the unconditional sojourn time. For a survey on the LST results we refer to [117].

Analytic inversion of these LST's has appeared to be hard, and only partial results are available. The complexity of deriving the complete probability distribution led to an interest in the tail behavior of the sojourn time distribution. Although obtaining the tail behavior seems a more modest goal than obtaining the complete distribution, this task has still proved to be quite challenging and has recently been the subject of extensive research.

Notably, one of the major insights is that there is a fundamental difference between sojourn time asymptotics for heavy-tailed and light-tailed service requirement distributions. A large number of studies have focused on the analysis of the tail of the unconditional sojourn time distribution in case the service time distribution is heavy-tailed. The asymptotic tail behavior of the sojourn time in the M/G/1 PS queue with regularly varying service time distribution was derived in [122] and later generalized in [87] for the case of distributions with intermediately regularly varying tails. The authors established the following asymptotic relationship between the distributions of the sojourn time V and the service requirement B with ρ denoting the traffic load:

$$\mathbf{P}(V > x) \sim \mathbf{P}(B > (1-\rho)x), \tag{1.8}$$

as $x \to \infty$ (for any two real functions $f(\cdot)$ and $g(\cdot)$, $f(x) \sim g(x)$ as $x \to \infty$ denotes that $f(x)/g(x) \to 1$ as $x \to \infty$). This asymptotic equivalence is often referred to as *reduced-load approximation*. The approximation may be heuristically interpreted as follows. Suppose a (tagged) customer with a very large service requirement arrives in the system. During his service, the system behavior may be approximately described as a PS queue with one permanent customer. For such a queue, it is known that the mean service rate received by the permanent (tagged) customer equals $1-\rho$ (cf. [36, 122]). Thus, in order to attain the amount of service B, the customer must spend roughly $B/(1-\rho)$ time units in the system.

It is worth noting that the above heuristics only apply for queues with heavy-tailed service time distributions, so that the customer stays in the system long enough to reach equilibrium behavior. Moreover, the equivalence (1.8) implicitly shows that the most probable scenario for a long sojourn time to occur is due to a large service requirement of the customer itself.

In [65], Jelenković and Momčilović extended the equivalence result to the case when the service time belongs to the class of subexponential distributions with tails heavier than $\mathrm{e}^{-\sqrt{x}}$. Assuming regularly varying distributions, Guillemin *et al.* [62]

proved that the asymptotic equivalence also holds for PS models with admission control and impatience as well as for state-dependent PS models (Generalized Processor Sharing as considered by Cohen [36]). See Borst *et al.* [25] for a survey.

For PS queues with light-tailed service time distributions only a few results are available. The tail asymptotics for the unconditional sojourn time in the M/M/1 PS queue are known, and are of a quite remarkable form:

$$\mathbf{P}(V > x) \sim cx^{-5/6}\mathrm{e}^{-\alpha x^{1/3}}\mathrm{e}^{-\gamma_0 x}, \qquad x \to \infty, \tag{1.9}$$

for positive constants c, α, γ_0. Flatto [54] obtained this asymptotic tail behavior of the waiting time in the M/M/1 Random-Order-of-Service (ROS) queue. Subsequently, Borst *et al.* [22] showed that the waiting-time distribution in the M/M/1 ROS queue, conditioned to be positive, equals the sojourn time distribution in the M/M/1 PS queue.

Mandjes and Zwart [82] analyzed the sojourn time asymptotics in the GI/GI/1 PS queue. Using large-deviations techniques, they derived logarithmic asymptotics for a broad class of light-tailed service time distributions. More precisely, they proved under specific conditions that the sojourn time V obeys

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{\theta\geq 0}(\alpha(\theta) - \theta), \tag{1.10}$$

where $\alpha(s)$ is the so-called (asymptotic) cumulant function of total amount of work fed to the queue, i.e.,

$$\alpha(\theta) = \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{\theta A(0,x)}],$$

with $A(0, x)$ the amount of traffic offered to the system in $(0, x]$.

The overload behavior of a single-server PS system was first analyzed by Jean-Marie and Robert [64], who derived the fluid limit for the number of jobs in the system. They showed that the queue length grows at a linear rate which depends on the entire distribution of the service time in addition to the mean interarrival time. Puha *et al.* [92] studied a single-server overloaded PS system in terms of measure-valued processes. A similar approach was applied to the PS queue with impatient customers in Gromoll *et al.* [58].

There are a few results available for the sojourn time asymptotics in PS queues with time-varying service rate. Assuming the service time distribution to be heavy-tailed, various extensions of the reduced-load approximations (as derived for the situation with constant service rate) were established. Núñez-Queija [87] studied the M/G/1 PS system in which the service rate follows an On-Off process with exponential On-periods. Other versions of the reduced-load approximation for queues with time-varying service rate are given in e.g. Bekker *et al.* [10], Borst *et al.* [24], Guillemin *et al.* [62]. Delcoigne *et al.* [40] evaluated the performance of PS queues in the presence of higher-priority jobs and obtained bounds for the mean sojourn time.

1.4.2 Discriminatory processor sharing

The literature on the sojourn time under the DPS discipline is quite sparse. Despite the rather simple service rate allocation policy which is closely related to the EPS model, the analysis of the DPS system appears to be extremely difficult. Major progress was made by Fayolle *et al.* [52] who obtained for the M/G/1 queue the mean sojourn time conditioned on the service requirement in the form of a set of integro-differential equations. Moreover, they showed that the conditional mean sojourn time under DPS asymptotically coincides with the conditional mean sojourn time under the PS discipline independently of the weights,

$$\mathbf{E}[V_i(\tau)] \sim \frac{\tau}{1-\rho}, \qquad \tau \to \infty.$$

Kim and Kim [74] derived the higher moments of the sojourn time in the M/M/1 queue as a solution to a set of linear equations. Recently, Avrachenkov *et al.* [8] proved that the conditional sojourn times of the various traffic classes are stochastically ordered according to the DPS weights. Rege and Sengupta [95] showed that the sojourn time conditioned on the job size can be decomposed into independent summands.

The sojourn time asymptotics for a general DPS queue with time-varying service rate were analyzed by Borst *et al.* in [26]; the authors proved the reduced-load equivalence in the case when the service rate process does not fluctuate too wildly compared to the service requirement. This result was extended to a wider class of service requirement distributions in [25]. The behavior of the DPS queue under overload conditions was studied by Altman *et al.* in [5]. A heavy-traffic regime was studied in Rege and Sengupta [96], Van Kessel *et al.* [108]. A comprehensive survey on DPS is given in [4] and [25].

1.4.3 Bandwidth-sharing networks

We now present a short overview of the literature on the flow-level analysis of bandwidth-sharing networks as described in Subsection 1.2.3. These networks provide a natural modeling framework for describing the dynamic interaction among competing elastic flows that traverse several links along their source-destination paths. Several studies have focused on the fundamental problem of network stability. Assuming exponential flow size distributions and Poisson arrivals, De Veciana *et al.* [110, 111] proved that weighted max-min and proportional fair bandwidth-sharing strategies achieve stability in such networks (positive recurrence of the associated Markov process) under the nominal condition that no individual link is overloaded. Bonald and Massoulié [14] extended that result to a wide family of weighted α-fair bandwidth-sharing strategies. Massoulié [83] established that the nominal stability condition remains sufficient for the proportional fair strategy with an additional 'routing feature', thus further generalizing the result to phase-type flow size distributions. Bramson [29] showed that the max-min fair strategy guarantees stability under the nominal load condition for general flow size distributions and renewal arrival processes.

The analysis of the flow-level performance of bandwidth-sharing networks appears to be generally difficult, even for the simplest network topologies and exponential flow sizes. Although an α-fair bandwidth-sharing network bears strong resemblance with a single-server PS system, there are two key distinctions that arise in a network scenario: (i) the rate received by a class is no longer constant, but depends on the number of flows of all classes in some intricate fashion; and (ii) the network may show non-work-conserving behavior due to the fact that congestion at other links may prevent a link from utilizing its full capacity, a phenomenon referred to as 'entrainment' by Kelly and Williams [71].

There are a number of results available in the literature for the flow-level performance of networks with *insensitive* rate allocations. Insensitivity is understood in the sense that the distribution of the number of active flows does not depend on the detailed traffic characteristics. The first results are due to Massoulié and Roberts [84] who derived an explicit formula for the distribution of the number of flows in linear networks with proportional fair sharing. The result was extended to a grid network in Bonald and Massoulié [14]. The queue length results in conjunction with Little's law allow to compute the mean sojourn time, see e.g. [14, 17, 18].

Later, Bonald and Proutière [17] proved that the performance of all utility-based policies is sensitive, with the exception of the proportional fair allocation in specific network topologies (namely, homogeneous hypercubes). The authors identified necessary and sufficient conditions for insensitivity in terms of a set of balance equations and introduced an alternative "balanced fairness" allocation policy which is insensitive and Pareto-efficient. Assuming Poisson arrivals, the distribution of the number of flows under an insensitive rate allocation only depends on the traffic intensities and is proportional to

$$\pi(x) = \Phi(x) \prod_{i=1}^{I} \rho_i^{x_i},$$

where $\Phi(\cdot)$ is a so-called *balance function*. Balanced fairness insensitivity can be viewed as a generalization to a network setting of the insensitivity of a single-node PS system with Poisson arrivals.

The difficulty of exact analysis of sensitive rate allocation policies motivated the study of approximations of flow-level performance measures. Assuming Poisson arrival processes and exponential flow sizes, Kelly and Williams [71] studied critical fluid-limit models when the average load on at least one resource is equal to its capacity. Under general distributional assumptions and load conditions, Gromoll and Williams [60, 61] studied the fluid limit for weighted α-fair strategies, and established stability of the fluid limit in some special cases, such as linear and tree topologies. Chiang *et al.* [34] developed a fluid model extending that of [60, 61] to more general network utility maximization policies. They established stability of the fluid model operating under an α-fair policy with α sufficiently small and $1/(1+\alpha)$-approximate stability for arbitrary positive values of α.

1.5 Overview of the thesis

In this first chapter we introduced and discussed various processor-sharing models, namely egalitarian PS, DPS and bandwidth-sharing networks. We briefly described the methods applied in this monograph and provided an overview of the most relevant literature. In the remainder of this monograph we present asymptotic results for the sojourn time distribution in a single-server PS system (Chapters 2–4) and bandwidth-sharing networks (Chapters 5–6).

Before providing a more detailed overview, we like to point out that the main focus in this thesis is on the PS queue where the service time has a "light-tailed" distribution. As mentioned in Subsection 1.4.1, this case has received relatively little attention compared to the case of heavy-tailed distributions. Exact asymptotics (of highly uncommon and interesting form) were only available for the M/M/1 PS queue and were obtained by analytical methods that did not provide insight into the nature of the underlying rare event, cf. (1.9). Even deriving the logarithmic asymptotics has proved to be far from straightforward [82]. The complexity and the scarcity of the available results have triggered our interest in this topic.

In Chapter 2 we consider the M/D/1 queue, and show that the probability $\mathbf{P}(V > x)$ decays exponentially fast as x becomes large. The proof involves a geometric random sum representation of V and a connection with Yule processes, which also enables us to simplify Ott's [89] derivation of the Laplace-Stieltjes transform of V. Numerical experiments show that the asymptotic approximation is highly accurate, even for moderate values of x. Chapter 2 is based on the results published in Egorova *et al.* [49].

In Chapter 3 we investigate the tail behavior of the sojourn time distribution for a service requirement of a given length in an M/G/1 queue. An exponential asymptote is proved for general service times in two special cases: when the traffic load is sufficiently high and when the service requirement is sufficiently small. Furthermore, using the branching process technique we derive exact asymptotics of exponential type for the sojourn time in the M/M/1 queue. We obtain an equation for the asymptotic decay rate and an exact expression for the asymptotic constant. The decay rate is studied in detail and compared to that of other service disciplines. Finally, we investigate the accuracy of the exponential asymptote using numerical methods. This chapter builds upon the analysis of Egorova and Zwart [47] and some basic results presented in Chapter 2.

In Chapter 4 we study the GI/GI/$\cdot$ queue operating under a PS discipline with stochastically varying service rate. The focus is on logarithmic estimates of the tail of the sojourn time distribution, under the assumption that the service time distribution has a light tail. Whereas upper bounds on the decay rate can be derived under fairly general conditions, establishing the corresponding lower bounds requires that the service process satisfies a sample-path large-deviations principle. We show that the class of allowed service processes includes the case where the service rate is modulated by a Markov process. Finally, we extend our results to a similar system operating under the DPS discipline. This chapter presents the results published in Egorova *et al.* [46].

In Chapters 5 and 6 we analyze the behavior of bandwidth-sharing networks as introduced in Subsection 1.2.3. The presented results can be viewed as first steps in the flow-level performance analysis of a network operating under a fair bandwidth-sharing policy. To the best of our knowledge, the sojourn time distribution in such systems has not been studied. The dynamic resource allocation and non-work-conserving behavior make the analysis of the queue length and the sojourn time extremely challenging. In Chapter 5 we analyze an overload regime with the main focus on the queue length growth. While this may appear to be a deviation from the main subject of this monograph, the growth rates of the queue length in an overloaded system are in fact intimately related to the large-deviations behavior of the queue length and sojourn time. Specifically, the most likely way for a large queue or a long delay to occur, commonly entails a scenario where the system temporarily deviates from the normal stochastic laws and behaves as if it experiences overload. In order to estimate the probability of a rare event, it is often convenient to apply a so-called change of measure, a method that allows to transform the system characteristics in such a way that an extremely uncommon phenomenon becomes a more frequent one. Under a particular change of measure the system may exhibit overload behavior. For example, Mandjes and Zwart [82] applied this approach in the single-server case, using a fluid-limit result of Puha *et al.* [92]. Using the overload results from Chapter 5, we perform such a change of measure to derive the sojourn time asymptotics in Chapter 6.

In Chapter 5 we focus on α-fair bandwidth-sharing networks where the load on one or several of the links exceeds the capacity. In order to characterize the overload behavior, we examine the fluid limit, which emerges from a suitably scaled version of the number of flows of the various classes. We derive a functional equation characterizing the fluid limit. The convergence of the scaled number of flows to the fluid limit is proved under the assumption that the fluid limit is strictly positive. Further, we establish the uniqueness of the fluid limit for networks with a tree topology. For the case of a zero initial state and zero-degree homogeneous rate allocation functions, we show that there exists a uniquely determined linear solution to the fluid-limit equation, and obtain a fixed-point equation for the corresponding asymptotic growth rates. The fluid-limit results are illustrated for parking lot, linear and star networks as important special cases. Finally, we discuss extensions to models with user impatience. This chapter is based on results in Borst *et al.* [23], Egorova *et al.* [44, 45].

In Chapter 6 we derive the asymptotics for the sojourn time distribution in a special type of bandwidth-sharing network: a parking lot network. Using large-deviations techniques and the fluid-limit results from Chapter 5, we obtain the logarithmic asymptote under the assumption that flow sizes have a light-tailed distribution. In addition, we derive stochastic bounds for the number of flows and the workload in the system. This chapter is based upon Egorova and Zwart [48].

Chapter 2

Sojourn time asymptotics in the M/D/1 queue

The focus of the present and the next chapter is on the asymptotic behavior of the sojourn time distribution in the classical single-node PS queue. In this chapter we derive exact tail asymptotics for the sojourn time distribution in the PS queue with Poisson arrivals and deterministic service times. Specifically, we assume that customers arrive according to a Poisson process with rate λ at a single server of unit capacity. The service requirement is constant for all customers, denoted by D. Let $\rho = \lambda D$ be the traffic intensity. We assume that $\rho < 1$, so that the system reaches steady state. Our main result is that the tail behavior of the steady-state sojourn time V is of the following form:

$$\mathbf{P}(V > x) \sim \alpha \mathrm{e}^{-\gamma x}, \qquad x \to \infty, \tag{2.1}$$

for some constants α and γ which will be explicitly characterized. Observe that the asymptotic form is fundamentally different from the one for exponential service requirements, cf. (1.9). Note also that the logarithmic asymptotics obtained in [82], which are valid for a broad class of light-tailed distributions, do not extend to distributions with bounded support, such as deterministic service requirements.

Apart from deriving the specific asymptotics, it is of interest to understand *how* large sojourn times take place. In a PS queue, three events may contribute to a large sojourn time of a (tagged) customer: (i) a large service requirement of the tagged customer; (ii) a large number of customers present in the system upon arrival of the tagged customer; (iii) an unusually large number of arrivals after the arrival of the tagged customer. When service requirements are heavy-tailed, event (i) is most likely responsible for a large sojourn time [121]. In [82], the authors show that for a broad class of light-tailed distributions, event (iii) determines the logarithmic asymptotics. Specifically, V becomes large if the traffic load ρ is increased to 1 during the sojourn time of the tagged customer. From the analysis in this chapter, one can infer that the most likely way for the event $\{V > x\}$ for large x to occur not only involves more work feeding into the system between time 0 and x, but also

an increased number of customers at time 0, i.e. the event $\{V > x\}$ occurs by a combination of the events (ii) and (iii) mentioned above.

The analysis in both the present chapter and the next is based on two key ideas. The first cornerstone is the branching method introduced by Yashkov [116]. The branching process representation and decomposition of the sojourn time into a sum of independent random variables (called *delay elements*), conditioned on the number of customers in the system, was established in [116] for the M/G/1 PS queue. The method was further applied in e.g. [87, 89, 95]. For the M/D/1 PS model we make the additional observation that the underlying branching process is a Yule process, which has been treated by, e.g., Ross [100]. We use this connection to obtain a simplified derivation of the Laplace-Stieltjes Transform (LST) of the delay elements associated with V, which also leads to a relatively simple derivation of Ott's result ([89], formula (5.16)) for the LST of V.

The branching process decomposition enables us to represent the sojourn time in terms of a random sum of independent and identically distributed delay elements. Because the number of customers in the system has a geometric distribution, we can apply existing powerful asymptotic results for geometric random sums to obtain the tail behavior of V; this is the second cornerstone of the present analysis.

The remainder of this chapter is organized as follows. In Section 2.1 we provide basic results for geometric random sums. In Section 2.2, we give a closed-form expression for the LST's of the distribution of the delay elements of the branching process decomposition, which is described in Chapter 1. The main result is presented and proved in Section 2.3. In addition, the asymptotic behavior under heavy traffic is considered. It is shown that the large-deviations and heavy-traffic limits are interchangeable. In Section 2.5 we present the results from numerical experiments. We compute the values of $\mathbf{P}(V > x)$ using transform inversion and compare them with the values predicted by (2.1). These experiments demonstrate a remarkable accuracy of the obtained approximation (2.1) even for moderate values of x.

2.1 Preliminaries

This section contains some preliminary results on the branching processes representation and the geometric random sums, which serve as a basis for the analysis in both the present chapter and the next. We assume that customers arrive according to a Poisson process with rate λ at a single PS server with unit capacity. Denote by B the generic service time. We assume that the queue is stable, i.e., that the traffic load in the system is less than one, $\rho = \lambda\,\mathbf{E}[B] < 1$.

In order to obtain the sojourn time tail asymptotics we apply the so-called "tagged-customer" approach. We describe the dynamic behavior of the system on the time interval between the arrival and the departure of a selected customer. Let us now consider a tagged customer with a service requirement τ (abbreviated as τ-requirement) that arrives into the system at the time epoch $t = 0$. Let $V(\tau)$ be its sojourn time.

Following the branching decomposition procedure as explained in Section 1.3, we

can represent the sojourn time of the tagged customer in a more tractable summation form,

$$V(\tau) = V_0(\tau) + \sum_{i=1}^{Q} C_i(\tau), \tag{2.2}$$

where $C_i(\tau)$ is the amount of service received by a certain progenitor and its descendants during the sojourn time of the tagged customer, $V_0(\tau)$ is equal to the amount of service received by the tagged customer and its direct descendants, and Q is the number of customers in the system at $t = 0$.

For convenience, denote $V_1(\tau) = \sum_{i=1}^{Q} C_i(\tau)$. Since the queue length distribution in a PS queue with Poisson arrivals is known, cf. (1.1), the probability distribution of $V_1(\tau)$ can be written as

$$\mathbf{P}(V_1(\tau) > x) = \mathbf{P}\left(\sum_{i=1}^{Q} C_i(\tau) > x\right) = \sum_{n=0}^{\infty}(1-\rho)\rho^n(1 - F_n(x)), \tag{2.3}$$

where F denotes the cumulative distribution function of $C_i(\tau)$, and $F_n(x)$ is the n-fold convolution of F with itself. The random variable $V_1(\tau)$ is called a geometric random sum and such random sums arise in many applied probability settings, the most prominent one being the M/G/1 FCFS queue and the Cramér-Lundberg risk model. From the results in Kalashnikov and Tsitsiashvili [66], it is known that if the Cramér condition holds, the tail of the distribution of such a sum is asymptotically (as $x \to \infty$) equivalent to an exponential function. In particular, in relation to the delay elements in the M/G/1 PS system, the following theorem holds.

Theorem 2.1.1. *Let the Cramér condition hold, i.e. suppose that there exists a* $\gamma = \gamma(\tau) > 0$ *such that*

$$\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}] = \frac{1}{\rho}. \tag{2.4}$$

(i) If $h(\tau) = \rho\int_0^\infty x\mathrm{e}^{\gamma(\tau)x}dF(x) = \rho\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=\gamma(\tau)} < \infty$, *and* F *is non-lattice, then the asymptotic relation*

$$\mathbf{P}(V_1(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{2.5}$$

holds with

$$\alpha(\tau) = \frac{1-\rho}{h(\tau)\gamma(\tau)}. \tag{2.6}$$

(ii) If $h(\tau) = \infty$, *then*

$$\lim_{x\to\infty} \mathbf{P}(V_1(\tau) > x)\mathrm{e}^{\gamma(\tau)x} = 0. \tag{2.7}$$

The above theorem provides an explicit expression for the tail behavior of the delay element $V_1(\tau)$. With this result, the derivation of the tail asymptotics of the sojourn time $V(\tau)$ reduces to two main tasks: (i) to verify the conditions of the approximation for $V_1(\tau)$ (which for some systems appears to be a challenging problem), (ii) to combine the latter asymptotics with the LST of $V_0(\tau)$.

In the remainder of the chapter we assume the service time to be constant. Hence, in the following sections the shorter notation will be used omitting the superfluous $\tau \equiv D$.

2.2 Laplace-Stieltjes transform of the sojourn time distribution

In this section we derive the LST of the sojourn time in the M/D/1 PS queue. In fact, the explicit formula for the LST of V is well known. It was derived by Ott [89] as a special case of the M/G/1 PS queue:

$$\mathbf{E}[\mathrm{e}^{-sV}] = \frac{(1-\rho)(\lambda+s)^2\mathrm{e}^{-(\lambda+s)D}}{s^2+\lambda(s+s(1-\rho)+\lambda(1-\rho))\mathrm{e}^{-(\lambda+s)D}}. \tag{2.8}$$

However, in this section we will give a new simplified proof of this formula using the branching decomposition and existing results for Yule processes. Some intermediate results provided by this decomposition will be applied in the derivation of the tail asymptotics. In fact, the main goal of this section is to obtain the LST of the delay elements.

The remainder of this section is organized as follows. First we consider the situation when the tagged customer enters an empty system. We derive the LST of the sojourn time of this customer in Subsection 2.2.1. In Subsection 2.2.2 we turn to the general case when there is an arbitrary number of customers in the system upon arrival of the tagged customer and finally we prove Ott's formula (2.8).

2.2.1 Sojourn time of the first customer

In this subsection we derive the LST of the sojourn time of the first customer, i.e. the customer that enters an empty system. Notice that in this situation the above-defined process $\{X(t);\ t \in [0, D]\}$, where t is amount of service received by the first customer, can be identified with a Yule process on $[0, D]$ starting with one ancestor. Recall that a Yule process is a pure birth process in which each individual in the population independently gives birth at constant rate. In our model the births correspond to customer arrivals. Until the service requirement of the first customer is completed, a number of other customers may arrive but none leave the system before that time, since under the PS discipline with constant service requirements customers depart from the system in order of their arrival.

The next proposition gives the LST of the first customer's sojourn time.

Proposition 2.2.1.

$$\mathbf{E}[\mathrm{e}^{-sV_0}] = \frac{\lambda+s}{\lambda+s\mathrm{e}^{(\lambda+s)D}}. \tag{2.9}$$

Proof. The integral representation (1.4) of V_0 can be rewritten as follows:

$$V_0 = D + \sum_{k=1}^{X(D)-1} (D - t_k),$$

where $(t_k, k \geq 1)$ are the arrival times of customers that enter the system during the service of the first customer.

Since $\{X(t);\ t \geq 0\}$ is a Yule process, its marginal distribution is known (see e.g. [100], p. 236). At time t the population size is geometrically distributed with parameter $\mathrm{e}^{-\lambda t}$:

$$\mathbf{P}(X(t) = i) = (1 - \mathrm{e}^{-\lambda t})^{i-1}\mathrm{e}^{-\lambda t}, \quad t \in [0, D]. \tag{2.10}$$

Furthermore (see again [100]), the conditional joint probability density of the arrival times $t_1,\ t_2, \ldots,\ t_n$, given the number of customers, $X(t) = n + 1$, is given by

$$p(s_1, s_2, \ldots, s_n | X(t) = n + 1) = \prod_{i=1}^{n} f(s_i), \quad s_i \leq t, \tag{2.11}$$

where

$$f(x) = \frac{\lambda \mathrm{e}^{-\lambda(t-x)}}{1 - \mathrm{e}^{-\lambda t}}, \quad 0 \leq x \leq t.$$

In order to obtain the expression for the LST of V_0, we condition on the number of customers in the system upon departure of the first customer,

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{-sV_0}] = \mathbf{E}[\mathrm{e}^{-s\int_0^D X(t)dt}] = \mathbf{E}[\mathrm{e}^{-s\left(D+\sum_{k=1}^{X(D)-1}(D-t_k)\right)}] \\ = \sum_{n=0}^{\infty} \mathbf{E}[\mathrm{e}^{-s\left(D+\sum_{k=1}^{X(D)-1}(D-t_k)\right)} | X(D) = n+1]\mathbf{P}(X(D) = n+1),\end{aligned} \tag{2.12}$$

where, due to independence of the t_k, $k = 1, \ldots, n$, the conditional expectation is

$$\mathbf{E}[\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))} | X(D) = n+1] = \prod_{k=1}^{n} \mathbf{E}[\mathrm{e}^{-s(D-t_k)} | X(D) = n+1]\mathrm{e}^{-sD}.$$

Computing the inner term of the above product, we get

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{-s(D-t_k)} | X(D) = n+1] &= \int_0^D \mathrm{e}^{-s(D-x)} \frac{\lambda \mathrm{e}^{-\lambda(D-x)}}{1 - \mathrm{e}^{-\lambda D}} dx \\ &= \frac{\lambda}{\lambda + s} \frac{1 - \mathrm{e}^{-(\lambda+s)D}}{1 - \mathrm{e}^{-\lambda D}}.\end{aligned} \tag{2.13}$$

Hence,

$$\mathbf{E}[\mathrm{e}^{-s(D+\sum_{k=1}^{X(D)-1}(D-t_k))} | X(D) = n+1] = \left(\frac{\lambda}{\lambda + s}\right)^n \left(\frac{1 - \mathrm{e}^{-(\lambda+s)D}}{1 - \mathrm{e}^{-\lambda D}}\right)^n \mathrm{e}^{-sD}.$$

Substituting the latter expression into (2.12) we obtain the LST of the sojourn time,

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{-sV_0}] &= \sum_{n=0}^{\infty} \left(\frac{\lambda}{\lambda + s}\right)^n \left(\frac{1 - \mathrm{e}^{-(\lambda+s)D}}{1 - \mathrm{e}^{-\lambda D}}\right)^n \mathrm{e}^{-sD}(1 - \mathrm{e}^{-\lambda D})^n \mathrm{e}^{-\lambda D} \\ &= \frac{\mathrm{e}^{-(\lambda+s)D}}{1 - \frac{\lambda}{\lambda+s}(1 - \mathrm{e}^{-(\lambda+s)D})}.\end{aligned}$$

Simple rewriting gives (2.9). □

Remark 2.2.1. An analog of the above result is given in Kella *et al.* [67]. The authors consider an M/G/1 queue with an arbitrary symmetric queueing discipline (processor sharing is a special case). Let B be a generic service time, D_1 the time epoch of the first departure from the system. Assume that the system is empty at time 0. Then for any positive s,

$$\mathbf{E}[\mathrm{e}^{-sD_1}] = \frac{\lambda}{\lambda + s\mathbf{E}[\mathrm{e}^{(\lambda+s)B}]}.$$

The expression is related to the LST of V_0 as

$$\mathbf{E}[\mathrm{e}^{-sD_1}] = \frac{\lambda}{\lambda + s}\mathbf{E}[\mathrm{e}^{-sV_0}],$$

which is a natural result, since $D_1 = A_1 + V_0$, where A_1 is the time epoch of the first arrival.

The LST of V_0 is recently studied in Van Leeuwaarden *et al.* [79] by investigating a connection between the M/D/1 PS queue and renewal age processes.

2.2.2 Sojourn time of an arbitrary customer

Let us now turn to the derivation of the LST of the sojourn time of a customer who enters the system and sees a number of customers already in service upon its arrival. Denote its sojourn time by V. Suppose that the number of customers in the system upon its arrival is Q. As before, $X(t)$ is the number of customers at the epoch when an amount of service t is received by the tagged customer, $t \in [0, D]$. Then $X(0) = Q$, $X(0+) = Q + 1$.

Proof of Formula (2.8). Conditioning on the number of customers in the system upon arrival of the tagged customer, we can write the LST as

$$\mathbf{E}[\mathrm{e}^{-sV}] = \sum_{n=0}^{\infty} \mathbf{E}[\mathrm{e}^{-sV}|Q = n]\mathbf{P}(Q = n), \tag{2.14}$$

where $\mathbf{P}(Q = n)$ is given by (1.1).

Now we use a branching decomposition of the sojourn time. If n jobs are present in the system at $t = 0$, then the sojourn time is decomposed into a sum of independent delay elements associated with $n + 1$ progenitors:

$$V|_{(Q=n)} = V_0 + \sum_{i=1}^{n} C_i.$$

With this representation the conditional expectation in (2.14) simplifies to

$$\mathbf{E}[\mathrm{e}^{-sV}|Q = n] = \mathbf{E}[\mathrm{e}^{-sV_0}] \left(\mathbf{E}[\mathrm{e}^{-sC_i}]\right)^n.$$

Let us now derive the transform of the random variable C_i. Let B_i^r be the remaining service requirement of the ith progenitor at the moment of the tagged

arrival. $B_i^r = B^r$, $i = 1, \ldots, Q$, is uniformly distributed on the interval $[0, D]$. Conditioning on B^r, we get

$$\mathbf{E}[\mathrm{e}^{-sC_i}] = \frac{1}{D}\int_0^D \mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t]dt. \tag{2.15}$$

Given $B^r = t$, we can express the conditional expectation $\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t]$ as in the previous section. However, in this situation we must distinguish between the intervals $[0, t]$ and $[t, D]$. Since no departures happen before t, on the interval $[0, t]$ we can apply ordinary Yule process properties as for V_0. On the interval $[t, D]$, we represent the number of customers in the system as a Yule process as well: the Yule process $Y(s),\ s \in [0, D-t]$, which starts from a number of customers at the moment $s = 0$: $Y(0) = X(t) - 1$.

Rewriting the conditional expectation

$$\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t] = \mathbf{E}[\mathrm{e}^{-s\sum_{k=1}^{X(t)}(t-t_k)}\mathrm{e}^{-s\sum_{k=X(t)+1}^{X(D)}(t-t_k)}],$$

and using the memoryless property and (2.10), we have

$$\begin{aligned}
&\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t] = \mathbf{E}[(\mathrm{e}^{-s\sum_{k=1}^{X(t)}(t-t_k)}\mathrm{e}^{-s((X(t)-1)(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}] \\
&= \sum_{m=0}^{\infty}\mathbf{E}[\mathrm{e}^{-s\sum_{k=1}^{m+1}(t-t_k)}|X(t) = m+1] \\
&\qquad \times\mathbf{E}[\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}|X(t) = m+1](1-\mathrm{e}^{-\lambda t})^m\mathrm{e}^{-\lambda t}.
\end{aligned}$$

Applying (2.13) with D replaced by t, we can simplify this expression to obtain

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t] &= \sum_{m=0}^{\infty}\mathrm{e}^{-(\lambda+s)t}\left(\frac{\lambda(1-\mathrm{e}^{-(\lambda+s)t})}{\lambda+s}\right)^m \\
&\times \mathbf{E}[\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}].
\end{aligned}$$

For the expectation term in the right-hand side we perform a computation using the Yule process that starts from m individuals (Ross [100]). If the population starts from i individuals, the population size at epoch t is the sum of i i.i.d. geometric random variables with parameter $\mathrm{e}^{-\lambda t}$. Hence, the population size at epoch t has a negative binomial distribution with parameters i and $\mathrm{e}^{-\lambda t}$. As before the distribution of arrival times t_k is defined by (2.11). Using these facts, we obtain:

$$\mathbf{E}[\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{Y(D-t)-Y(0)}(D-t-t_k))}] =$$

$$\begin{aligned}
&= \sum_{l=0}^{\infty}\mathbf{E}[\mathrm{e}^{-s(m(D-t)+\sum_{k=1}^{l}(D-t-t_k))}|Y(D-t) = l+m]\mathbf{P}(Y(D-t) = l+m) \\
&= \sum_{l=0}^{\infty}\mathrm{e}^{-(s+\lambda)m(D-t)}\left(\frac{\lambda}{\lambda+s}\right)^l(1-\mathrm{e}^{-(\lambda+s)(D-t)})^l\frac{(l+m-1)!}{(m-1)!l!} \\
&= \left(\frac{(\lambda+s)\mathrm{e}^{-(\lambda+s)(D-t)}}{s+\lambda\mathrm{e}^{-(\lambda+s)(D-t)}}\right)^m.
\end{aligned}$$

Thus, substituting the latter into the expression for $\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t]$ we get

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t] &= \sum_{m=0}^{\infty} \mathrm{e}^{-(\lambda+s)t} \left(\frac{\lambda(1-\mathrm{e}^{-(\lambda+s)t})}{\lambda+s}\right)^m \left(\frac{(\lambda+s)\mathrm{e}^{-(\lambda+s)(D-t)}}{s+\lambda\mathrm{e}^{-(\lambda+s)(D-t)}}\right)^m \\ &= \frac{\lambda+s\mathrm{e}^{(\lambda+s)(D-t)}}{\lambda+s\mathrm{e}^{(\lambda+s)D}}, \end{aligned} \tag{2.16}$$

and

$$\mathbf{E}[\mathrm{e}^{-sC_i}] = \frac{1}{D}\int_0^D \mathbf{E}[\mathrm{e}^{-sC_i}|B^r = t]dt = \frac{\rho(\lambda+s)-s+s\mathrm{e}^{(\lambda+s)D}}{D(\lambda+s)(\lambda+s\mathrm{e}^{(\lambda+s)D})}. \tag{2.17}$$

Substituting (2.9) and (2.17) into (2.14), we obtain the sojourn time transform

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{-sV}] &= \frac{(\lambda+s)(1-\rho)}{\lambda+s\mathrm{e}^{(\lambda+s)D}} \sum_{n=0}^{\infty} \rho^n \left(\frac{\rho(\lambda+s)-s+s\mathrm{e}^{(\lambda+s)D}}{D(\lambda+s)(\lambda+s\mathrm{e}^{(\lambda+s)D})}\right)^n \\ &= \frac{(1-\rho)(\lambda+s)^2}{s^2\mathrm{e}^{(\lambda+s)D}+\lambda(s+s(1-\rho)+\lambda(1-\rho))}, \end{aligned}$$

which coincides with Ott's formula (2.8). □

2.3 Tail behavior of the sojourn time

In this section we investigate the behavior of $\mathbf{P}(V > x)$ as $x \to \infty$. The following theorem is the main result of this chapter.

Theorem 2.3.1. *As $x \to \infty$,*

$$\mathbf{P}(V > x) \sim \alpha \mathrm{e}^{-\gamma x}, \tag{2.18}$$

where γ is the real solution of the equation

$$\frac{\lambda D(\lambda-s)+s-s\mathrm{e}^{(\lambda-s)D}}{D(\lambda-s)(\lambda-s\mathrm{e}^{(\lambda-s)D})} = \frac{1}{\rho}, \quad s \geq 0, \tag{2.19}$$

and

$$\alpha = \frac{(1-\rho)(\lambda-\gamma)}{2\lambda(1-\rho)-\gamma\rho(2-\rho)}. \tag{2.20}$$

Our derivation is based on the LST results obtained in the previous section. In particular, we will use the moment generating functions (MGF) of the decomposition random variables V_0 and C_i, appearing in the representation (2.2) of V:

$$\mathbf{E}[\mathrm{e}^{sV_0}] = \frac{\lambda-s}{\lambda-s\mathrm{e}^{(\lambda-s)D}}, \tag{2.21}$$

$$\mathbf{E}[\mathrm{e}^{sC_i}] = \frac{\lambda D(\lambda-s)+s-s\mathrm{e}^{(\lambda-s)D}}{D(\lambda-s)(\lambda-s\mathrm{e}^{(\lambda-s)D})}. \tag{2.22}$$

This section is organized as follows. In Subsection 2.3.1 we analyze the singularities of the above MGF s with respect to s. This enables us to prove Theorem 2.3.1 with a version of the Cramér-Lundberg theorem for geometric random sums. The proof is given in Subsection 2.3.2.

2.3.1 Singularities of the delay element LST's

Before we proceed with the proof of Theorem 2.3.1, we need to characterize the singularities of the MGF s $\mathbf{E}[\mathrm{e}^{sV_0}]$ and $\mathbf{E}[\mathrm{e}^{sC_i}]$. It is sufficient to consider only real values of s, since

$$|\mathbf{E}[\mathrm{e}^{sV_0}]| \leq \mathbf{E}[\mathrm{e}^{Re(s)V_0}], \qquad |\mathbf{E}[\mathrm{e}^{sC_i}]| \leq \mathbf{E}[\mathrm{e}^{Re(s)C_i}].$$

We begin with $\mathbf{E}[\mathrm{e}^{sV_0}]$. Let us consider the denominator of $\mathbf{E}[\mathrm{e}^{sV_0}]$ as a separate function $f(s) = \lambda - s\mathrm{e}^{(\lambda-s)D}$. Obviously, singularities of the MGF can only occur at zeros of the denominator. The trivial zero of $f(s)$ is $s = \lambda$. Notice however, that this is a removable singularity of $\mathbf{E}[\mathrm{e}^{sV_0}]$: using L'Hospital's rule we obtain that

$$\lim_{s\to\lambda} \mathbf{E}[\mathrm{e}^{sV_0}] = \frac{1}{1-\rho}.$$

We now show that there exists another zero of the function $f(s)$. The derivative of $f(s)$ is determined as $f'(s) = (Ds-1)\mathrm{e}^{(\lambda-s)D}$ and $f'(s) = 0$ at $s = \frac{1}{D}$. Furthermore, $f(0) = \lambda$, $f(\infty) = \lambda$, $f(\lambda) = 0$ and by stability, $\lambda < \frac{1}{D}$. Since $f'(s) < 0$ for $s < 1/D$ and $f'(s) > 0$ for $s > 1/D$, we can conclude that there is a unique point $\gamma_0 > \frac{1}{D} > \lambda$ such that $f(\gamma_0) = 0$. An important fact is that this point is a pole of the MGF: $\mathbf{E}[\mathrm{e}^{\gamma_0 V_0}] = \infty$.

To analyze the behavior of $\mathbf{E}[\mathrm{e}^{sC_i}]$, let us first consider the conditional MGF $\mathbf{E}[\mathrm{e}^{sC_i}|B^r = t]$, $t \in [0, D]$, cf. (2.16):

$$\mathbf{E}[\mathrm{e}^{sC_i}|B^r = t] = \frac{\lambda - s\mathrm{e}^{(\lambda-s)(D-t)}}{\lambda - s\mathrm{e}^{(\lambda-s)D}}.$$

We already know the zeros of the denominator: λ and γ_0. Again, λ is a removable singularity, since

$$\lim_{s\to\lambda} \mathbf{E}[\mathrm{e}^{sC_i}|B^r = t] = \frac{1-\rho+\lambda t}{1-\rho}, \qquad t \in [0, D].$$

It remains to check if $\mathbf{E}[\mathrm{e}^{sC_i}|B^r = t]$ has a singularity when $s = \gamma_0$. For this purpose we consider the numerator as a separate function, $f_t(s) = \lambda - s\mathrm{e}^{(\lambda-s)(D-t)}$. As a function of the parameter t, the numerator $f_t(s)$ increases for values $s < \lambda$ and decreases for $s > \lambda$. Since $f_0(\gamma_0) \equiv f(\gamma_0) = 0$ and $\gamma_0 > \lambda$, it follows that $f_t(\gamma_0)$ is strictly negative for any $t > 0$. Hence, γ_0 is a pole: $\mathbf{E}[\mathrm{e}^{\gamma_0 C_i}|B^r = t] = \infty$.

Summarizing this subsection we have

Proposition 2.3.1. *There exists a unique value $\gamma_0 > \lambda$ that satisfies the equation*

$$\lambda - s\mathrm{e}^{(\lambda-s)D} = 0, \tag{2.23}$$

and that is an abscissus of convergence of both $\mathbf{E}[\mathrm{e}^{sV_0}]$ *and* $\mathbf{E}[\mathrm{e}^{sC_i}|B^r = t]$, $\forall t \in [0, D]$, *and consequently, of* $\mathbf{E}[\mathrm{e}^{sC_i}]$.

We are now ready to give a proof of Theorem 2.3.1.

2.3.2 Proof of Theorem 2.3.1

The statement of the theorem follows from two known results. First, we obtain the exponential asymptotics for $V_1(\tau)$ using Theorem 2.1.1. Substituting expression (2.22) for $\mathbf{E}[\mathrm{e}^{sC_i}]$ into Equation (2.4) we obtain

$$\mathbf{E}[\mathrm{e}^{\gamma C_i}] = \frac{\lambda D(\lambda-\gamma) + \gamma - \gamma \mathrm{e}^{(\lambda-\gamma)D}}{D(\lambda-\gamma)(\lambda - \gamma \mathrm{e}^{(\lambda-\gamma)D})} = \frac{1}{\rho}.$$

Since the function $\mathbf{E}[\mathrm{e}^{sC_i}]$ monotonically increases from 1 to ∞ on the interval $[0, \gamma_0)$ (by Proposition 2.3.1), for any nonzero value of ρ there exists a unique real solution γ of Equation (2.4), $\gamma < \gamma_0$. Notice also that this solution γ is an abscissus of convergence of $\mathbf{E}[\mathrm{e}^{sV_1}]$.

The MGF $\mathbf{E}[\mathrm{e}^{sC_i}]$ is finite and differentiable at point $s = \gamma$, $\gamma < \gamma_0$, which implies that $h < \infty$. Finally, F is non-lattice, since $\mathbf{P}(C_i = B^r) > 0$, and B^r has a density. Hence, condition (i) of Theorem 2.1.1 is satisfied and we can determine the coefficient α and the asymptotics for $\mathbf{P}(V_1 > x)$.

Taking the derivative of the MGF, performing some simplifications and using the definition of γ, we obtain,

$$\mathbf{P}(V_1 > x) \sim \frac{(1-\rho)(\lambda - \gamma \mathrm{e}^{(\lambda-\gamma)D})}{2\lambda(1-\rho) - \gamma\rho(2-\rho)} \mathrm{e}^{-\gamma x}, \qquad x \to \infty. \tag{2.24}$$

We can now derive an expression for the tail behavior of the sojourn time V. Since V_1 has an asymptotically exponential tail, $\mathbf{P}(V_1 > x) = \mathbf{P}(\mathrm{e}^{V_1} > y) \sim \alpha y^{-\gamma}$, where $y = \mathrm{e}^x$, and $\mathbf{E}[\mathrm{e}^{(\gamma+\varepsilon)V_0}] < \infty$ for any $0 < \varepsilon < \gamma_0 - \gamma$ we can apply Breiman's theorem (see [30]):

$$\mathbf{P}(V > x) = \mathbf{P}(V_0 + V_1 > x) = \mathbf{P}(\mathrm{e}^{V_0}\mathrm{e}^{V_1} > \mathrm{e}^x) \sim \mathbf{E}[\mathrm{e}^{\gamma V_0}]\mathbf{P}(V_1 > x), \ x \to \infty.$$

Combining this with Equation (2.9) for $\mathbf{E}[\mathrm{e}^{\gamma V_0}]$, we obtain (2.18). □

We close this section with some useful observations.

Remark 2.3.1. An interesting issue, raised in the introduction, is how the number of customers in the system plays a role in the occurrence of a large sojourn time. Mandjes and Zwart [82] have shown that, in PS queues with phase-type service times for example, the initial number of customers is of $\mathrm{o}(x)$ when $V > x$. In this remark, we show that this picture drastically changes when service times are deterministic.

The proof of Theorem 2.3.1 indicates that the realizations of the C_i's in the branching representation (2.2) are sampled from the exponentially tilted density $\mathrm{e}^{\gamma x} d\mathbf{P}(C_i \le x)/\mathbf{E}[\mathrm{e}^{\gamma C_i}]$.

Under this density, the expected value of the C_i is

$$\mathbf{E}[C_i \mathrm{e}^{\gamma C_i}]/\mathbf{E}[\mathrm{e}^{\gamma C_i}] = \rho \mathbf{E}[C_i \mathrm{e}^{\gamma C_i}] =: c(\gamma).$$

Thus, in order for V to be of size x, N should be around $x/c(\gamma)$.

Remark 2.3.2. When $s = \lambda$, the denominator and the numerator of Equation (2.22) are both equal to zero. Using L'Hospital's rule we get

$$\lim_{s \to \lambda} \mathbf{E}[\mathrm{e}^{sC_i}] = \frac{2-\rho}{2(1-\rho)}.$$

Solving the equation

$$\frac{2-\rho}{2(1-\rho)} = \frac{1}{\rho},$$

we obtain that, for $\rho = 2 - \sqrt{2}$, Equation (2.19) has a solution $\gamma = \lambda$.

Since the asymptotic constant α in (2.18) has a removable singularity at this value, the tail behavior of V becomes:

$$\mathbf{P}(V > x) \sim \frac{1-\rho}{\rho(2-\rho)} \mathrm{e}^{-\lambda x} = \frac{1}{2} \mathrm{e}^{-\lambda x}, \qquad x \to \infty. \tag{2.25}$$

Remark 2.3.3. The asymptotic behavior of the first customer sojourn time distribution was also obtained in Kella *et al.* [67] and in Van Leeuwaarden *et al.* [79]. Using different approaches, the authors showed that

$$\mathbf{P}(V_0 > x) \sim \frac{\lambda - \gamma_0}{\lambda(1-\gamma_0 D)} \mathrm{e}^{-\gamma_0 x}, \tag{2.26}$$

where $\gamma_0 \neq \lambda$ is solution of Equation (2.23).

2.4 Implications of Theorem 2.3.1

In this section we discuss a number of implications of our main result. First, we take a look at the relationship between the decay rate in the M/D/1 PS queue and decay rates in queues with FCFS and LCFS disciplines. Secondly, we consider the behavior of the decay rate γ and the pre-factor α in heavy traffic.

2.4.1 Other service disciplines

First we consider the FCFS service discipline. A fundamental result of Stolyar and Ramanan [93] states that FCFS is optimal among all work-conserving disciplines, in the sense that it maximizes the decay rate of the sojourn time distribution. The inequality $\gamma_{FCFS} > \gamma_{PS}$ can also be easily verified by the following argument. Recall ([7], Theorem XIII.5.2) that γ_{FCFS} is a solution of the equation $\rho \mathbf{E}[\mathrm{e}^{sB^r}] = 1$, where B^r is the remaining service time. Using Equation (2.4) and the definition of C_i we get:

$$\mathbf{E}[\mathrm{e}^{\gamma_{FCFS} B^r}] = \mathbf{E}[\mathrm{e}^{\gamma_{PS} C_i}] > \mathbf{E}[\mathrm{e}^{\gamma_{PS} B^r}].$$

The decay rate inequality $\gamma_{FCFS} > \gamma_{PS}$ follows from the monotonicity of the MGF s.

For any work-conserving service discipline the sojourn time is bounded from above by a residual busy period, which implies that the decay rate of the residual busy period gives the lowest possible value. Recently Mandjes and Nuyens [81] showed that this lower bound is attained by the decay rate of the sojourn time in the LCFS and the Foreground-Background (FB) queues. A similar result was obtained in [82] for the GI/G/1 PS queue for a class of light-tailed service time distributions excluding deterministic service requirements. However, in [82] it was shown that in the M/D/1 queue the decay rate under the LCFS discipline and the decay rate in the PS case satisfy the strict inequality $\gamma_{LCFS} < \gamma_{PS}$. Thus, the decay rate of the sojourn time in the M/D/1 PS queue is strictly smaller than the decay rate under FCFS and strictly larger than the one under LCFS,

$$\gamma_{LCFS} < \gamma_{PS} < \gamma_{FCFS}.$$

Table 2.1 shows decay rates for the M/D/1 queue with PS, FCFS and LCFS disciplines. For convenience, we take $D = 1$. The decay rate in the M/D/1 LCFS queue is given by $\gamma_{LCFS} = -\log\rho - (1-\rho)$ (Cox and Smith [38]).

ρ	0.2	0.4	0.6	0.8
PS	1.9227	1.0462	0.5578	0.2331
FCFS	2.6604	1.6188	0.9474	0.4308
LCFS	0.8094	0.3163	0.1108	0.0231

Table 2.1: Asymptotic decay rates for the M/D/1 queue with PS, FCFS and LCFS disciplines.

The small value of γ_{LCFS} for $\rho = 0.8$ is related to the fact that $\gamma_{LCFS} = \mathrm{O}((1-\rho)^2)$ as $\rho \to 1$, as opposed to $\gamma_{FCFS} = \mathrm{O}((1-\rho))$. In the next subsection, we show that in heavy traffic γ_{PS} behaves like γ_{FCFS}.

2.4.2 Heavy traffic

Let us now study the sojourn time of a customer in heavy traffic, i.e. when the traffic intensity $\rho \to 1$.

Proposition 2.4.1. *Let γ and α be defined as in Theorem 2.3.1. Then, as $\rho \to 1$, the decay rate $\gamma \sim \lambda(1-\rho)$ and the coefficient $\alpha \to 1$.*

Proof. Obviously, when the traffic intensity $\rho \to 1$, the decay rate γ converges to zero (see Equation (2.4)). Let us study the behavior of γ near zero in more detail. We expand the left-hand side of (2.4) into a two-term Taylor series: $\mathbf{E}[\mathrm{e}^{\gamma C_i}] = \mathbf{E}[1+\gamma C_i + O(\gamma^2)]$. The second-order term is $O(\gamma^2)$ uniformly in ρ, since the second moment $\mathbf{E}[C_i^2]$ is finite if $\rho = 1$. The smoothness of the MGF $\mathbf{E}[\mathrm{e}^{\gamma C_i}]$ near zero implies that all moments of C_i are finite. To calculate the first moment $\mathbf{E}[C_i]$,

let us take the derivative of the MGF at zero: $\mathbf{E}[C_i] = \frac{2-\rho}{\rho\lambda} \to \frac{1}{\lambda}$, and due to $\rho\mathbf{E}[\mathrm{e}^{\gamma C_i}] = \rho + \gamma\mathbf{E}[C_i] + O(\gamma^2) = 1$, we get that

$$\gamma(1/\lambda + o(1)) = 1 - \rho,$$

implying that $\gamma \sim \lambda(1-\rho)$.

Substitution of the expression for γ into (2.20) gives the behavior of the asymptotic constant α:

$$\alpha = \frac{(1-\rho)(\lambda-\gamma)}{2\lambda(1-\rho) - \gamma\rho(2-\rho)} \sim \frac{(1-\rho)(\lambda-\lambda(1-\rho))}{2\lambda(1-\rho) - \lambda(1-\rho)\rho(2-\rho)} \to 1.$$

□

Remark 2.4.1. The above heavy-traffic behavior is related to a result of Yashkov [118]. He derived a heavy-traffic limit result for the sojourn time in the M/G/1 PS queue conditioned on the service requirement. Replacing s in (2.8) by $(1-\rho)s$ and taking the limit $\rho \to 1$ we have:

$$\lim_{\rho\to 1} \mathbf{E}[\mathrm{e}^{-(1-\rho)sV}] = \frac{\lambda}{\lambda+s}. \tag{2.27}$$

Since the limiting value is the LST of the exponential distribution with parameter λ, we obtain the heavy-traffic approximation

$$\mathbf{P}(V > x) \approx \mathrm{e}^{-\lambda(1-\rho)x}. \tag{2.28}$$

Hence, summarizing Proposition 2.4.1 and Remark 2.4.1,

$$\lim_{\rho\to 1}\lim_{x\to\infty} \frac{\mathbf{P}((1-\rho)V > x)}{\alpha\mathrm{e}^{-\gamma x/(1-\rho)}} = \lim_{x\to\infty}\lim_{\rho\to 1} \frac{\mathbf{P}((1-\rho)V > x)}{\alpha\mathrm{e}^{-\gamma x/(1-\rho)}} = 1.$$

This suggests that the asymptotics given in Theorem 2.3.1 provide a good approximation for the sojourn time tail behavior if ρ is close to 1. The numerical results in the next section confirm this.

2.5 Numerical results

In this section we present some numerical results. In particular, we compare the behavior of the sojourn time tail computed numerically from Ott's formula (2.8) with the asymptotics we have obtained. In Ott's formula the sojourn time distribution is expressed in terms of its LST.

The inversion of LST s was considered to be numerically challenging for a long time. However, nowadays there are a number of reliable and effective inversion methods which allow for computing probabilities and other quantities without any complication. We will compute the sojourn time distribution using the inversion

algorithm of Den Iseger [42] and will perform a cross-check with the algorithm proposed by Abate and Whitt [2]. Both methods are known to provide high accuracy, and indeed produce similar results. Since the sojourn time distribution has a jump at point D, we will apply the modified Den Iseger algorithm for functions with discontinuities.

Table 2.2 shows computational results for various arrival rates and service requirements normalized to $D = 1$. For each value of ρ, the first column shows, for different values of x, the approximation (2.18) for $\mathbf{P}(V > x)$. The second column presents the estimates derived with the Den Iseger inversion algorithm.

	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
x	asympt.	LST inv.	asympt.	LST inv.	asympt.	LST inv.
5	1,0935-02	1,0936-02	9,0507-02	9,0507-02	3,6738-01	3,6738-01
10	5,8474-05	5,8474-05	5,5656-03	5,5656-03	1,1452-01	1,1452-01
15	3,1267-07	3,1267-07	3,4225-04	3,4225-04	3,5699-02	3,5699-02
20	1,6718-09	1,6718-09	2,1046-05	2,1046-05	1,1128-02	1,1128-02
25	8,9398-12	8,9398-12	1,2942-06	1,2942-06	3,4689-03	3,4689-03
30	4,7802-14	4,8072-14	7,9587-08	7,9587-08	1,0813-03	1,0813-03
35	2,5560-16	2,1127-16	4,8941-09	4,8941-09	3,3708-04	3,3708-04
40	1,36677-18	1,2177-18	3,0096-10	3,0096-10	1,0507-04	1,0507-04

Table 2.2: Asymptotic approximation and numerical results.

The results show remarkable accuracy of the asymptotic tail approximation. The numbers obtained with LST inversion and the asymptotic formula differ sometimes less than 10^{-16}, which is in fact the maximum accuracy of the inversion algorithm. Moreover, the asymptotics perform well even for relatively small values of x. Already for $x = 10$ the error is of the order 10^{-13}. Results with similar accuracy of exponential asymptotics in FCFS queues are presented in the paper of Abate *et al.* ([1], Table 1).

x	asympt.	LST inv.	HT (2.28)
10	6,17856022-01	6,17856022-01	6,21885056-01
30	2,19011860-01	2,19011860-01	2,40508463-01
50	7,76332889-02	7,76332889-02	9,30144892-02
70	2,75187267-02	2,75187267-02	3,59725188-02
90	9,75458251-03	9,75458251-03	1,39120487-02

Table 2.3: Asymptotic approximations and numerical results for $\rho = 0.95$.

Table 2.3 presents results for high load, $\rho = 0.95$. As before, the service requirement D is equal to 1. We consider two approximations: the asymptotic approximation (2.18) (first column), and the heavy-traffic asymptotics (2.28) (third column).

The second column shows the results from the numerical inversion. Remarkably, the heavy-traffic asymptotics perform less accurately than (2.18).

Chapter 3

Tail behavior of conditional sojourn times

In this chapter we investigate the tail behavior of the sojourn time distribution for a given service requirement in the M/G/1 PS queue. In order to emphasize this conditioning we will use the notation M/G(τ)/1 for the underlying queue, although we stress that all other customers still have generally distributed service requirements.

The analysis in this chapter is based on the same ideas as in the previous chapter: using the branching process decomposition, we first represent the sojourn time as a geometric random sum of delay elements, and secondly, we make use of existing asymptotic results for such sums. However, the task of verifying the conditions under which these asymptotic results are valid, is significantly more challenging here. To obtain rigorous results in the general setting, we need to make additional assumptions. Assuming that either the traffic load is close to one, or that the service requirement is sufficiently small, we show in Section 3.1 that the asymptotic tail behavior

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{3.1}$$

is valid for generally distributed service times.

Sections 3.2 and 3.3 are devoted to the case of exponential service times, for which no further assumptions are necessary. We obtain an equation (of quite an unusual trigonometric form) for the asymptotic decay rate $\gamma(\tau)$ and an exact (though complicated) expression for the asymptotic constant $\alpha(\tau)$. In Section 3.2, we derive expressions for the delay elements of the sojourn time and in Section 3.3 we formulate the main asymptotic result for the M/M(τ)/1 queue.

Finally, in Sections 3.4 and 3.5 we present some numerical results. First, we analyze the behavior of the decay rate depending on the value of τ and compare it with decay rates for an M/M(τ)/1 system with a different service discipline such as Shortest Remaining Processing Time (SRPT), Foreground-Background (FB), FCFS and LCFS. In Section 3.5, we compare the asymptotic result to exact values of $\mathbf{P}(V(\tau) > x)$, obtained by numerical LST inversion. We also compare the accuracy of the asymptotics and the heavy-traffic approximation. The results show that the exponential asymptotics provide a good approximation.

3.1 Tail behavior in the M/G(τ)/1 queue

In this section we present some results for the sojourn time in a system with a general service requirement distribution. Under the condition that the traffic load is sufficiently high, we prove that the sojourn time tail behaves asymptotically as an exponential function. We also consider the situation when the service requirement of the given customer is close to zero.

In order to obtain the sojourn time tail asymptotics we follow the same approach as in Chapter 2. We consider a tagged customer with a service requirement of length τ which arrives at $t = 0$. Using the branching process decomposition procedure as described in Section 1.3, we represent the sojourn time of the tagged customer as a geometric random sum of independent delay elements, cf. (2.2). Subsequently, we apply existing asymptotic results for such sums. We refer to Sections 1.3 and 2.1 for a detailed discussion.

The further derivations in the present chapter predominantly rely on Theorem 2.1.1. In the following proposition, we prove the Cramér condition (2.4) for the case where the traffic intensity is sufficiently large.

Proposition 3.1.1. *Let $h(\tau) = \rho\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=\gamma(\tau)}$. For any value of τ there exists a $\rho(\tau) < 1$ such that for all $\rho > \rho(\tau)$, there exists a solution $\gamma(\tau)$ of Equation (2.4) with $h(\tau) < \infty$.*

Proof. Due to the convexity of the MGF, it suffices to show that for any fixed value of τ there exists a sufficiently large $\rho < 1$ such that there exists an $\bar{s}$ for which $\frac{1}{\rho} < \mathbf{E}[\mathrm{e}^{\bar{s}C_i(\tau)}] < \infty$. Observe that $C_i(\tau)$ is stochastically dominated by the busy period P_τ in a system with services defined as $\min(B, \tau)$ given that the first customer in the busy period has a service requirement τ. Therefore, the inequality $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] \leq \mathbf{E}[\mathrm{e}^{sP_\tau}]$ holds. Due to Theorem 7.1 in [3], P_τ has a decay rate $\hat{s}(\tau, \lambda)$, defined as the solution of the equation $\lambda(d/ds)(\mathbf{E}[\mathrm{e}^{s\min(B,\tau)}]) = 1$, and since $\mathbf{P}(P_\tau > x) \sim const \cdot\ x^{-3/2}\mathrm{e}^{-\hat{s}(\tau,\lambda)x}$, we deduce $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)P_\tau}] < \infty$. Hence, $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)C_i(\tau)}] < \infty$.

To bound the MGF of $C_i(\tau)$ from below, notice that for any τ, $C_i(\tau) \geq \min(B^r, \tau)$, where B^r is the residual service time. Hence, $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)C_i(\tau)}] \geq \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)\min(B^r,\tau)}]$. If $\mathbf{P}(B > \tau) > 0$, then $\mathbf{E}[\min(B, \tau)] < \mathbf{E}[B]$ and the modified queue is still stable. Hence $\hat{s}(\tau, \frac{1}{\mathbf{E}[B]}) > 0$, and

$$\begin{aligned}\lim_{\lambda\to 1/\mathbf{E}[B]} \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)C_i(\tau)}] &\geq \lim_{\lambda\to 1/\mathbf{E}[B]} \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\lambda)\min(B^r,\tau)}] \\ &= \mathbf{E}[\mathrm{e}^{\hat{s}(\tau,1/\mathbf{E}[B])\min(B^r,\tau)}] > 1.\end{aligned}$$

Thus, choosing $\rho > \frac{1}{\mathbf{E}[\mathrm{e}^{\hat{s}(\tau,\frac{1}{\mathbf{E}[B]})\min(B^r,\tau)}]}$, we can find a solution of Equation (2.4). □

The following theorem is a straightforward consequence of the above proposition.

Theorem 3.1.1. *For any value of τ there exists a $\rho(\tau)$ such that for all $\rho > \rho(\tau)$ we have*

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty, \tag{3.2}$$

where $\gamma(\tau)$ is the solution of Equation (2.4) and the constant $\alpha(\tau)$ is given by

$$\alpha(\tau) = \frac{1-\rho}{h(\tau)\gamma(\tau)}\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]. \tag{3.3}$$

Proof. By Proposition 3.1.1, for all $\rho > \rho(\tau)$, there exists a solution $\gamma(\tau)$ of (2.4) and $h(\tau) < \infty$. Further, the distribution function F of the delay element $C_i(\tau)$ is non-lattice, since $\mathbf{P}(C_i(\tau) = B^r) > 0$, and the residual service time B^r has a density. Hence, the conditions of part (i) of Theorem 2.1.1 are satisfied, and we obtain that as $x \to \infty$,

$$\mathbf{P}(V_1 > x) \sim \alpha_1(\tau)\mathrm{e}^{-\gamma(\tau)x},$$

where $\alpha_1(\tau)$ is given by (2.6).

The condition $\mathbf{P}(B > \tau) > 0$ implies that $\mathbf{P}(B^r > \tau) > 0$. Since we consider all elements only on the interval $[0, V(\tau)]$, the elements $C_i(\tau)$ and $V_0(\tau)$ coincide (in distribution) if $B^r > \tau$, and

$$\infty > \frac{\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}]}{\mathbf{P}(B^r > \tau)} \geq \frac{\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}\mathbf{1}(B^r > \tau)]}{\mathbf{P}(B^r > \tau)} = \mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}|B^r > \tau] = \mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}].$$

Applying Breiman's theorem [30] under the weaker condition $\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}] < \infty$ (see [43]), we obtain that

$$\mathbf{P}(V(\tau) > x) = \mathbf{P}(\mathrm{e}^{V_0(\tau)}\mathrm{e}^{V_1(\tau)} > \mathrm{e}^x) \sim \mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]\mathbf{P}(V_1(\tau) > x), \qquad x \to \infty,$$

which completes the proof. □

Using a similar approach, we can prove exponential asymptotics for the sojourn time of a customer with a very small service requirement.

Theorem 3.1.2. *For sufficiently small values of τ,*

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty,$$

where $\gamma(\tau)$ is a solution of Equation (2.4) and the constant $\alpha(\tau)$ is given by (3.3).

Proof. The elements $C_i(\tau)$ can be bounded from above by the delay element $C_i^D(\tau)$ in the M/D/1 PS system with service requirements of size τ. The results in Chapter 2 for the decay rate in the M/D/1 PS queue imply that there exists an $\hat{s}(\tau) > 0$ such that $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i^D(\tau)}] = 1/\rho_D = 1/(\lambda\tau)$. Further, the same argument as in the proof of Proposition 3.1.1 is applicable. However in this case λ, $\mathbf{E}[B]$ and ρ are fixed and the parameter τ is varying: $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i(\tau)}] > \mathbf{E}[\mathrm{e}^{\hat{s}(\tau)\min(B^r,\tau)}] > \mathrm{e}^{\hat{s}(\tau)\tau}\mathbf{P}(B^r > \tau)$. The equation for the decay rate $\hat{s}(\tau)$ (see Equation (2.19)) is

$$\frac{\lambda\tau(\lambda - s) + s - s\mathrm{e}^{(\lambda-s)\tau}}{(\lambda - s)(\lambda - s\mathrm{e}^{(\lambda-s)\tau})} = \frac{1}{\lambda}.$$

Taking $s = c\tau$ and letting $\tau \downarrow 0$ we see that $\liminf_{\tau\downarrow 0} \hat{s}(\tau)\tau \geq c$ for any c, and consequently, $\lim_{\tau\downarrow 0} \hat{s}(\tau)\tau = \infty$. Hence, the decay rate $\hat{s}(\tau)$ is increasing faster than

linear in $1/\tau$ when τ becomes small. Thus, we can conclude that for any $\rho \in (0,1)$ there exists a τ_0 such that $\mathbf{E}[\mathrm{e}^{\hat{s}(\tau)C_i(\tau)}] > 1/\rho$ holds for all $\tau < \tau_0$, which by convexity of the MGF implies the existence of a solution of (2.4) for all $\tau < \tau_0$. The statement of the theorem then follows by the same argument as in Theorem 3.1.1. □

In the following sections, we focus on the behavior of the sojourn time in the M/M(τ)/1 queue.

3.2 The delay elements for exponential service times

The goal of this section is to obtain the LST of the delay elements in the M/M(τ)/1 queue using the approach presented in Yashkov [116], where the general expression for the LST of the sojourn time of a τ-requirement in the M/G/1 queue is derived. The LST of the sojourn time itself is of less importance for our tail behavior investigation; it has been derived in Coffman *et al.* [35].

Define $\varphi(s,\tau) = \mathbf{E}[\mathrm{e}^{-sC_i(\tau)}]$ and $\delta(s,\tau) = \mathbf{E}[\mathrm{e}^{-sV_0(\tau)}]$ as the LST's of the random variables $C_i(\tau)$ and $V_0(\tau)$, respectively.

Theorem 3.2.1. *The delay elements of the sojourn time in the M/M(τ)/1 PS queue have LST's given by the expressions:*

$$\delta(s,\tau) = \frac{2g(s)\mathrm{e}^{-(\lambda+s-\mu)\frac{\tau}{2}}}{(\mu-\lambda+s)\,(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}) + g(s)(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)})} \tag{3.4}$$

and

$$\varphi(s,\tau) = \frac{(\mu-\lambda-s)(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}) + g(s)(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)})}{(\mu-\lambda+s)(\mathrm{e}^{1/2\tau g(s)} - \mathrm{e}^{-1/2\tau g(s)}) + g(s)(\mathrm{e}^{1/2\tau g(s)} + \mathrm{e}^{-1/2\,\tau g(s)})}, \tag{3.5}$$

where $g(s) = \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}$.

Proof. In order to derive the LST's of the delay elements we follow Yashkov [116]. Under the condition that the number of customers in the system upon arrival of the tagged customer is n and the remaining service time of the ith progenitor at $t = 0$ is x_i, the sojourn time $V(\tau)$ can be represented as

$$V(\tau) = V_0(\tau) + \sum_{i=1}^{n} C_i(x_i,\tau).$$

Since the random variables $V_0(\tau)$ and $C_i(x_i,\tau)$ are independent, we can write

$$\mathbf{E}[\mathrm{e}^{-sV(\tau)}|n,x_1,...,x_n] = \delta(s,\tau)\prod_{i=1}^{n}\varphi(s,x_i,\tau).$$

Unconditioning, we obtain that the LST of the sojourn time is

$$\begin{aligned} v(s,\tau) &= (1-\rho)\delta(s,\tau)\left[1-\rho\int_{x=0}^{\infty}\varphi(s,x,\tau)\frac{(1-B(x))}{\mathbf{E}[B]}dx\right]^{-1} \\ &= (1-\rho)\frac{\delta(s,\tau)}{1-\rho\varphi(s,\tau)}, \end{aligned}$$

where $\varphi(s,\tau)$ is the LST of the delay element $C_i(\tau)$.

We now proceed to derive the expressions for $\delta(s,\tau)$ and $\varphi(s,\tau)$. Due to Equations (3.9) and (3.14) in [116] we have

$$\varphi(s,x,\tau) = \begin{cases} \delta(s,\tau)/\delta(s,\tau-x), & x<\tau, \\ \delta(s,\tau), & x\geq\tau. \end{cases}$$

Using Formula (3.16) of [116] we obtain that

$$\delta(s,\tau) = \mathrm{e}^{-(s+\lambda)\tau}\psi(s,\tau)^{-1},$$

where the Laplace transform $\tilde{\psi}(q,s)$ of the function $\psi(s,\tau)$ given by

$$\tilde{\psi}(q,s) = \int_0^{\infty}\mathrm{e}^{-q\tau}\psi(s,\tau)d\tau,$$

is a solution of the following equation (see Equations (3.18)–(3.19) in [116])

$$q\tilde{\psi}(q,s) - 1 + \lambda\tilde{\psi}(q,s)\beta(q+s+\lambda) + \frac{\lambda(1-\beta(q+s+\lambda))}{q+s+\lambda} = 0,$$

where $\beta(\cdot)$ is the LST of the service time. Substituting $\beta(s) = \frac{\mu}{\mu+s}$, we obtain

$$\tilde{\psi}(q,s) = \frac{q+s+\mu}{q^2+(\mu+\lambda+s)q+\lambda\mu}.$$

To derive an expression for $\psi(s,\tau)$ we must invert the LST $\tilde{\psi}(q,s)$ with respect to q. This can be easily done using partial-fraction decomposition of the latter expression. That will lead us to the LST of a sum of two exponential functions. As a result we get

$$\psi(s,\tau) = \frac{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}{g(s)}, \tag{3.6}$$

where $A=(\mu+s-\lambda+g(s))/2$, $B=(\mu+s+\lambda-g(s))/2$, $C=(-\mu-s+\lambda+g(s))/2$, $D=(\mu+s+\lambda+g(s))/2$, and $g(s)=\sqrt{(\mu+\lambda+s)^2-4\lambda\mu}$.

Knowing $\psi(s,\tau)$ we can determine the LSTs $\delta(s,\tau)$ and $\varphi(s,x,\tau)$:

$$\delta(s,\tau) = \mathrm{e}^{-(s+\lambda)\tau}\frac{g(s)}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}},$$

$$\varphi(s,x,\tau) = \begin{cases} \mathrm{e}^{-(s+\lambda)x}\frac{A\mathrm{e}^{-B(\tau-x)}+C\mathrm{e}^{-D(\tau-x)}}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}, & x<\tau, \\ e^{-(s+\lambda)\tau}\frac{g(s)}{A\mathrm{e}^{-B\tau}+C\mathrm{e}^{-D\tau}}, & x\geq\tau. \end{cases}$$

Expression (3.4) for the LST $\delta(s,\tau)$ follows in a straightforward manner. In order to derive the LST $\varphi(s,\tau)$ of the delay element $C_i(\tau)$, we integrate with respect to the residual service time x. After some simplifications we obtain Formula (3.5). □

In order to investigate the sojourn time tail behavior we will need the MGF's of the delay elements rather than the LST's. The results of the previous section yield that the MGF of the delay element $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ is

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)(\mathrm{e}^{\frac{1}{2}\tau f(s)} - \mathrm{e}^{-\frac{1}{2}\tau f(s)}) + f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)} + \mathrm{e}^{-\frac{1}{2}\tau f(s)})}{(\mu-\lambda-s)(\mathrm{e}^{\frac{1}{2}\tau f(s)} - \mathrm{e}^{-\frac{1}{2}\tau f(s)}) + f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)} + \mathrm{e}^{-\frac{1}{2}\tau f(s)})}, \tag{3.7}$$

where $f(s) = g(-s)$ (Theorem 3.2.1),

$$f(s) = \sqrt{(\mu+\lambda-s)^2 - 4\,\lambda\,\mu}.$$

Let us study the function $f(s)$ in more detail.

The expression under the square root is a quadratic function with zeros at $s_l = \lambda+\mu-2\sqrt{\lambda\mu} \equiv \mu(1-\sqrt{\rho})^2$ and $s_r = \lambda+\mu+2\sqrt{\lambda\mu} \equiv \mu(1+\sqrt{\rho})^2$. The function is negative on the interval

$$s \in (\lambda+\mu-2\sqrt{\lambda\mu}, \lambda+\mu+2\sqrt{\lambda\mu})$$

and positive otherwise.

Taking into account the fact that the function $f(s)$ is purely imaginary inside the interval $[s_l, s_r]$, we can rewrite the MGF in two forms depending on the sign of the radicand.

Corollary 3.2.1.

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)\sin[\frac{1}{2}\tau f_2(s)] + f_2(s)\cos[\frac{1}{2}\tau f_2(s)]}{(\mu-\lambda-s)\sin[\frac{1}{2}\tau f_2(s)] + f_2(s)\cos[\frac{1}{2}\tau f_2(s)]}, \qquad \text{if } s \in [s_l, s_r],$$

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu-\lambda+s)\sinh[\frac{1}{2}\tau f_1(s)] + f_1(s)\cosh[\frac{1}{2}\tau f_1(s)]}{(\mu-\lambda-s)\sinh[\frac{1}{2}\tau f_1(s)] + f_1(s)\cosh[\frac{1}{2}\tau f_1(s)]}, \qquad \text{otherwise,}$$

where $f_1(s) = \sqrt{(\mu+\lambda-s)^2 - 4\,\lambda\,\mu}$ *and* $f_2(s) = \sqrt{-(\mu+\lambda-s)^2 + 4\,\lambda\,\mu}$.

The next result is useful to analyze the MGF around the point s_r.

Corollary 3.2.2. *For all values of* $s > 0$ *where* $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ *is finite,*

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{\sum_{n=0}^{\infty} \frac{(\frac{\tau}{2})^{2n}}{(2n)!} d(s)^n \left[\frac{(\mu-\lambda+s)}{(2n+1)}\frac{\tau}{2} + 1\right]}{\sum_{n=0}^{\infty} \frac{(\frac{\tau}{2})^{2n}}{(2n)!} d(s)^n \left[\frac{(\mu-\lambda-s)}{(2n+1)}\frac{\tau}{2} + 1\right]}, \tag{3.8}$$

where $d(s) = (\mu+\lambda-s)^2 - 4\,\lambda\,\mu$.

Proof. The representation follows from the Taylor expansion for the exponential function:

$$\frac{\mathrm{e}^x - \mathrm{e}^{-x}}{2} = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}, \qquad \frac{\mathrm{e}^x + \mathrm{e}^{-x}}{2} = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}.$$

Using Equation (3.5) we can now rewrite the MGF $\varphi(s,\tau)$ as

$$\varphi(-s,\tau) = \frac{\sum_{n=0}^{\infty}\left[(\mu-\lambda+s)\frac{(\frac{\tau}{2})^{2n+1}}{(2n+1)!}f(s)^{2n+1} + \frac{(\frac{\tau}{2})^{2n}}{(2n)!}f(s)^{2n+1}\right]}{\sum_{n=0}^{\infty}\left[(\mu-\lambda-s)\frac{(\frac{\tau}{2})^{2n+1}}{(2n+1)!}f(s)^{2n+1} + \frac{(\frac{\tau}{2})^{2n}}{(2n)!}f(s)^{2n+1}\right]}.$$

Dividing both numerator and denominator by $f(s)$, we get only even powers under the sum

$$\varphi(-s,\tau) = \frac{\sum_{n=0}^{\infty}\frac{(\frac{\tau}{2})^{2n}}{(2n)!}f(s)^{2n}\left[\frac{(\mu-\lambda+s)}{(2n+1)}\frac{\tau}{2}+1\right]}{\sum_{n=0}^{\infty}\frac{(\frac{\tau}{2})^{2n}}{(2n)!}f(s)^{2n}\left[\frac{(\mu-\lambda-s)}{(2n+1)}\frac{\tau}{2}+1\right]},$$

which yields the desired representation as $f(s) = \sqrt{d(s)}$. □

3.3 Tail behavior in the M/M(τ)/1 queue

In this section we present our main result.

Theorem 3.3.1. *Define* $\tau_0 = \frac{1}{\sqrt{\lambda\mu}}\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}$. *For all* $\tau > 0$,

$$\mathbf{P}(V(\tau) > x) \sim \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}, \qquad x \to \infty. \tag{3.9}$$

(i) If $\tau \neq \tau_0$, $\gamma(\tau) > 0$ *is the solution of*

$$\tan\left(\frac{\tau}{2}\sqrt{-(\lambda+\mu-s)^2+4\lambda\mu}\right) = \frac{\sqrt{-(\lambda+\mu-s)^2+4\lambda\mu}}{\lambda-\mu+s\frac{1+\rho}{1-\rho}}, \qquad \text{if } \tau > \tau_0, \tag{3.10}$$

or

$$\tanh\left(\frac{\tau}{2}\sqrt{(\lambda+\mu-s)^2-4\lambda\mu}\right) = \frac{\sqrt{(\lambda+\mu-s)^2-4\lambda\mu}}{\lambda-\mu+s\frac{1+\rho}{1-\rho}}, \qquad \text{if } \tau < \tau_0, \tag{3.11}$$

and

$$\alpha(\tau) = \frac{2(1-\rho)}{\gamma(\tau)}\frac{[(\lambda+\mu-\gamma(\tau))^2-4\lambda\mu]e^{-(-\gamma(\tau)+\lambda-\mu)\frac{\tau}{2}}}{K(\tau)}, \tag{3.12}$$

with

$$\begin{aligned} K(\tau) &= (1+\rho) \\ &\times\left[f(\gamma(\tau))(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} - \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}}) + \gamma(\tau)\tfrac{\tau}{2}(\lambda+\mu-\gamma(\tau))(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} + \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})\right] \\ &-(1-\rho)(\lambda+\mu-\gamma(\tau)) \\ &\times\left[(\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} + \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})(1+\tfrac{(\mu-\lambda)\tau}{2}) + (\mathrm{e}^{f(\gamma(\tau))\frac{\tau}{2}} - \mathrm{e}^{-f(\gamma(\tau))\frac{\tau}{2}})f(\gamma(\tau))\tfrac{\tau}{2}\right] \end{aligned} \tag{3.13}$$

and $f(s) = \sqrt{(\mu+\lambda-s)^2 - 4\,\lambda\,\mu}$.

(ii) If $\tau = \tau_0$, *the decay rate and the asymptotic constant are given by*

$$\gamma(\tau_0) = (\sqrt{\mu}+\sqrt{\lambda})^2, \tag{3.14}$$

$$\alpha(\tau_0) = \frac{12\,(\sqrt{\mu}+\sqrt{\lambda})\sqrt{\rho}\,\lambda\,\mathrm{e}^{1-\frac{1}{\sqrt{\rho}}}}{6\,(\mu\sqrt{\mu}+\lambda\sqrt{\lambda})+(\sqrt{\mu}-\sqrt{\lambda})^3}. \tag{3.15}$$

If the conditions stated in Theorem 2.1.1 hold in the case of exponential service times, the statement of the above theorem follows almost immediately. We will now show that the Cramér condition indeed holds, i.e. that there exists a positive solution to the equation $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{1}{\rho}$.

3.3.1 Cramér condition

Let us first determine some useful thresholds that will play an essential role in our proof.

Proposition 3.3.1. *If* $\tau < \tau_0 = \frac{1}{\sqrt{\lambda\mu}}\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}$, *then the solution* $\gamma(\tau)$ *of Equation (2.4), if it exists, is larger than* $s_r = (\sqrt{\mu}+\sqrt{\lambda})^2$, *and if* $\tau > \tau_0$, *a solution must be inside the interval* $[s_l, s_r] = [(\sqrt{\mu}-\sqrt{\lambda})^2, (\sqrt{\mu}+\sqrt{\lambda})^2]$.

Proof. We claim that the solution $\gamma(\tau)$ of Equation (2.4), if it exists, is always larger than the threshold $s_l = \lambda+\mu-2\sqrt{\lambda\mu}$. Let γ_0 be the leftmost pole of the MGF $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$. Since the MGF is increasing in s on $[0,\gamma_0]$ we only need to show that

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_l} < \frac{1}{\rho}. \tag{3.16}$$

The value of the MGF at s_l is

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_l} = \frac{1+\tau\mu-\tau\sqrt{\lambda\,\mu}}{1-\tau\lambda+\tau\sqrt{\lambda\,\mu}}. \tag{3.17}$$

Thus, inequality (3.16) simplifies to $\lambda+\tau\lambda(\mu-\sqrt{\lambda\,\mu}) < \mu+\tau\mu(\sqrt{\lambda\,\mu}-\lambda)$. Due to the stability assumption it is sufficient to show that $\lambda(\mu-\sqrt{\lambda\,\mu}) < \mu(\sqrt{\lambda\,\mu}-\lambda)$. Notice that this is equivalent to $\lambda+\mu-2\sqrt{\lambda\mu} > 0$ and, hence, the claim is true.

Let us now check the behavior of the MGF at the right boundary $s_r = \lambda+\mu+2\sqrt{\lambda\mu}$. We compare the value of the MGF with $1/\rho$:

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]|_{s=s_r} = \frac{1+\tau\mu+\tau\sqrt{\lambda\,\mu}}{1-\tau\lambda-\tau\sqrt{\lambda\,\mu}} = \frac{1}{\rho}. \tag{3.18}$$

This yields that the MGF at $s = s_r$ is equal to $1/\rho$ if

$$\tau_0 = \frac{\mu-\lambda}{\sqrt{\lambda\mu}(\lambda+\mu+2\sqrt{\lambda\mu})} = \frac{1}{\sqrt{\lambda\mu}}\frac{(\sqrt{\mu}-\sqrt{\lambda})(\sqrt{\mu}+\sqrt{\lambda})}{(\sqrt{\mu}+\sqrt{\lambda})^2} = \frac{1}{\sqrt{\lambda\mu}}\,\frac{1-\sqrt{\rho}}{1+\sqrt{\rho}}.$$

The statement of the proposition follows from the monotonicity of the MGF with respect to both s and τ. □

In the next proposition we prove the existence of the decay rate $\gamma(\tau)$.

Proposition 3.3.2. *For any τ there exists a solution of Equation (2.4).*

Proof. Let us first assume that $\tau > \tau_0$. Hence, a solution of Equation (2.4) can only be inside the interval $[s_l, s_r]$. On this interval Equation (2.4) takes the following form

$$\frac{(\mu-\lambda+s)\sin[\frac{1}{2}f(s)\tau]+f(s)\cos[\frac{1}{2}f(s)\tau]}{(\mu-\lambda-s)\sin[\frac{1}{2}f(s)\tau]+f(s)\cos[\frac{1}{2}f(s)\tau]}=\frac{1}{\rho}, \tag{3.19}$$

where $f(s)=f_2(s)=\sqrt{-(\mu+\lambda-s)^2+4\,\lambda\,\mu}$. After a simple computation we obtain that this equation is equivalent to

$$\tan\left(\frac{\tau}{2}f(s)\right)=\frac{f(s)}{\lambda-\mu+s\frac{1+\rho}{1-\rho}}. \tag{3.20}$$

Let us consider the left-hand side (denoted by F_L) and the right-hand side (denoted by F_R) of the latter equation in more detail. Depending on the value of τ, the qualitative behavior of F_L changes. We will determine the intervals for τ on which F_L behaves differently and prove the Cramér condition on each interval.

The function F_R is independent of τ. As a function of s, F_R has a pole at $s^*=(\mu-\lambda)\frac{1-\rho}{1+\rho}$. On the interval $[s_l, s^*]$, F_R is decreasing from 0 to $-\infty$, and on $[s^*, s_r]$ it is decreasing from $+\infty$ to 0.

Let us now study the behavior of F_L as a function of s and τ. The tangent has infinite jumps when its argument is equal to $\frac{\pi}{2}+\pi k,\ k\in\mathbb{N}$. We are only interested in the first jump, $k=0$. Note that, due to symmetry of $f(s)$ around $s_0=\lambda+\mu$, F_L is also symmetric as a function of s on the interval $[s_l, s_r]$ with respect to the center of the interval, $s_0=\lambda+\mu$.

The first jump of F_L occurs when

$$\frac{\tau}{2}f(s')=\frac{\pi}{2},$$

that is when

$$s'=\lambda+\mu-\sqrt{4\lambda\mu-\frac{\pi^2}{\tau^2}}.$$

We will consider two cases separately: (1) - when F_L has an infinite jump inside the interval $[s_l, s_r]$, (2) - when it does not have such a jump. We derive the conditions and values of τ for which these situations can occur.

(1-a) First suppose that F_L has an infinite jump before the infinite jump of F_R, that is $s'<s^*$. That is equivalent to

$$s'=\lambda+\mu-\sqrt{4\lambda\mu-\frac{\pi^2}{\tau^2}}<(\mu-\lambda)\frac{1-\rho}{1+\rho}=s^*,$$

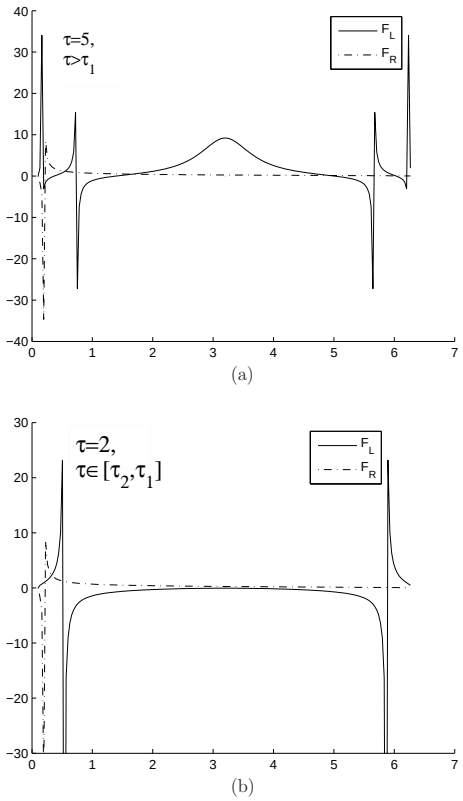

Figure 3.1: Functions F_L and F_R under different conditions on τ, $\lambda = 1.2$, $\mu = 2$.

and hence,

$$\tau > \frac{\pi}{2\sqrt{\lambda\mu}} \frac{\mu + \lambda}{\mu - \lambda} := \tau_1.$$

Thus, for any $\tau > \tau_1$ the function F_L jumps before F_R. Notice that F_R is negative up to s^* and F_L is positive up to s' and negative after s' increasing from $-\infty$. Hence we can conclude that under this condition on τ there is always a solution of the equation $F_L = F_R$. That means that there is a solution $\gamma(\tau)$ of the Equation (2.4) and it is located inside the interval $[s_l, s']$.

Consider now a different situation. Suppose that F_L has no infinite jumps inside the interval $[s_l, s_r]$. This is equivalent to the statement

$$\frac{\tau}{2} f(s) < \frac{\pi}{2},$$

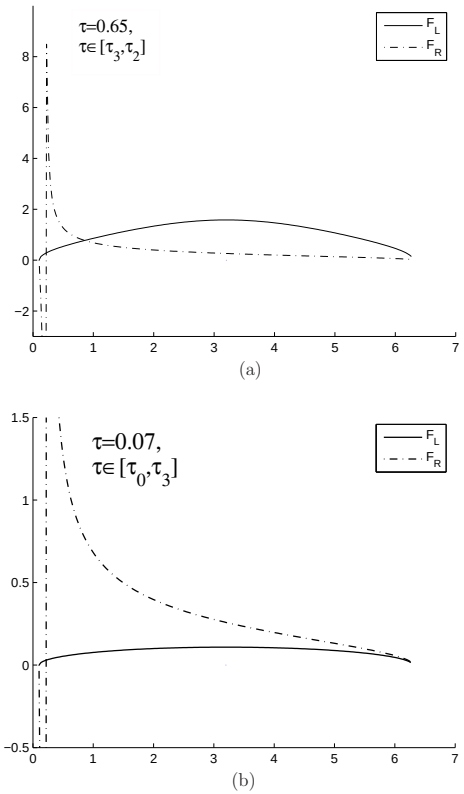

Figure 3.2: Functions F_L and F_R under different conditions on τ, $\lambda = 1.2$, $\mu = 2$.

for all $s \in [s_l, s_r]$, i.e.

$$\tau < \min_{s\in[s_l,s_r]} \frac{\pi}{f(s)} = \frac{\pi}{2\sqrt{\lambda\mu}} := \tau_2.$$

(1-b) Consequently, for any $\tau \in [\tau_2, \tau_1]$ (see Figure 3.1 (b)) there is a jump of F_L in the interval $[s^*, \lambda+\mu)$ (before $\lambda+\mu$ since F_L is symmetric). Due to the properties of both functions for these τ there is always a point $\gamma(\tau)$ at which F_L and F_R intersect, $\gamma(\tau) \in [s^*, \lambda + \mu)$.

(2) Thus, for any $\tau \in [\tau_0, \tau_2]$ the function F_L has no jumps in $[s_l, s_r]$. Comparing the values of F_L and F_R at the center of the interval there are two cases possible in this situation (see Figure 3.2 (a,b)): (a) $F_R|_{s=\lambda+\mu} < F_L|_{s=\lambda+\mu}$ and (b) $F_R|_{s=\lambda+\mu} > F_L|_{s=\lambda+\mu}$.

The values of the functions at this point are:

$$F_L|_{s=\lambda+\mu} = \tan(\tau\sqrt{\lambda\mu}),$$

$$F_R|_{s=\lambda+\mu} = \frac{\mu-\lambda}{2\sqrt{\lambda\mu}}.$$

(2-a) Consider the first case. Let us derive conditions under which this event may occur. Due to the monotonicity of the tangent, the inequality

$$F_L|_{s=\lambda+\mu} = \tan(\tau\sqrt{\lambda\mu}) > \frac{\mu-\lambda}{2\sqrt{\lambda\mu}} = F_R|_{s=\lambda+\mu}$$

reduces to

$$\tau > \frac{1}{\sqrt{\lambda\mu}} \arctan\left(\frac{\mu-\lambda}{2\sqrt{\lambda\mu}}\right) := \tau_3.$$

Hence, for all $\tau \in [\tau_3, \tau_2]$ the value of F_R at the center point is lower than the value of F_L. Observe that F_R is decreasing on $[s^*, s_r]$ and F_L is increasing on $[s_l, \lambda+\mu]$. Therefore, these two functions must intersect at a point $\gamma(\tau)$ on the interval $[s^*, \lambda+\mu]$.

(2-b) Consider now the second case (Figure 3.2 (b)): $F_L|_{s=\lambda+\mu} < F_R|_{s=\lambda+\mu}$. It is easy to check that the derivatives ${F_L}'$ and ${F_R}'$ are equal to infinity when $s = s_r$. For $\tau \in [\tau_0, \tau_3]$ it is impossible for F_L and F_R to intersect before the point $\lambda+\mu$. So we now consider $s \in [\lambda+\mu, s_r]$. For such s and τ both F_R and F_L are decreasing as functions of s and

$$F_R|_{s=s_r} = F_L|_{s=s_r} = 0,$$

$$F_R|_{s=\lambda+\mu} > F_L|_{s=\lambda+\mu}.$$

These functions can intersect if and only if in some neighborhood of the point s_r the decrease of F_L is faster than the decrease of F_R, that is if and only if ${F_L}' < {F_R}'$.

The derivatives are given by

$${F_L}' = \frac{\tau(\lambda+\mu-s)}{2f(s)\cos^2(\frac{\tau}{2}f(s))},$$

$${F_R}' = \frac{4(\lambda-\mu)\lambda s\mu}{f(2\lambda\mu-\mu^2+s\mu-\lambda^2+\lambda s)^2}.$$

Thus, we have

$${F_L}' = \frac{\tau(\lambda+\mu-s)}{2f(s)\cos^2(\frac{\tau}{2}f(s))} < \frac{4(\lambda-\mu)\lambda s\mu}{f(s)(2\lambda\mu-\mu^2+s\mu-\lambda^2+\lambda s)^2} = {F_R}',$$

$$\frac{\tau}{\cos^2(\frac{\tau}{2}f(s))} > \frac{8(\mu-\lambda)\lambda s\mu}{f(s)(s-\lambda-\mu)(2\lambda\mu-\mu^2+s\mu-\lambda^2+\lambda s)^2}.$$

When $s \to s_r$, $\cos(\frac{\tau}{2}f(s))$ converges to one from below. Hence, the right-hand side of the latter inequality is larger than or equal to τ, while in this case $\tau > \tau_0 = \frac{\mu-\lambda}{\sqrt{\lambda\mu}(\lambda+\mu+2\sqrt{\lambda\mu})}$. Notice that when $s \to s_r$ the left-hand side of the inequality converges to τ_0. Hence, the inequality holds for all s close enough to s_r.

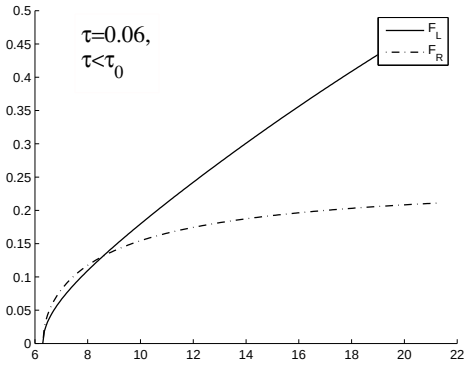

Figure 3.3: Functions F_L and F_R under conditions $\tau < \tau_0$, $\lambda = 1.2$, $\mu = 2$.

Thus, we have considered Equation (2.4) in four possible cases under the condition that $\tau > \tau_0$ and have shown that in all these cases there is a solution of Equation (2.4) and it lies inside the interval $[s_l, s_r]$.

(3) The only case left to consider is when $\tau < \tau_0$. For such values of τ, Equation (2.4) takes the form:

$$\mathbf{E}[\mathrm{e}^{sC_i(\tau)}] = \frac{(\mu - \lambda + s)\sinh[\frac{1}{2}\tau f(s)] + f(s)\cosh[\frac{1}{2}\tau f(s)]}{(\mu - \lambda - s)\sinh[\frac{1}{2}\tau f(s)] + f(s)\cosh[\frac{1}{2}\tau f(s)]}, \tag{3.21}$$

or equivalently,

$$\tanh\left(\frac{\tau}{2}f(s)\right) = \frac{f(s)}{\lambda - \mu + s\frac{1+\rho}{1-\rho}}, \qquad s \in [s_r, \infty), \tag{3.22}$$

where now $f(s) = f_1(s) = \sqrt{(\lambda + \mu - s)^2 - 4\lambda\mu}$.

A useful observation is that when $s \to \infty$, the left-hand side G_L converges to one and the right-hand side G_R converges to $\frac{1-\rho}{1+\rho}$, which is less than one for all $\rho > 0$. The derivatives of both functions are infinite at the point $s = s_r$ and both functions are strictly increasing for $s > s_r$ (see Figure 3.3). To prove the inequality we will use the same technique as in the previous case. We will show that there is a neighborhood of s_r in which the derivatives satisfy $G'_L < G'_R$, that is

$$\frac{\tau}{\cosh^2(\frac{\tau}{2}\,f(s))} < \frac{8\,(\mu - \lambda)\,\lambda\,s\mu}{f(s)(s - \lambda - \mu)\,(2\,\lambda\,\mu - \mu^2 + s\mu - \lambda^2 + \lambda\,s)^2}.$$

Notice that for $s \to s_r$ the function $\cosh(\frac{\tau}{2}\,f(s))$ converges to one from above, and so the left-hand side of the inequality is less than or equal to τ, which is in this case less than τ_0. The inequality follows from the observation that the right-hand side converges to τ_0 when $s \to s_r$.

Thus, we have shown that for all $\tau > 0$ there exists a solution of Equation (2.4).□

Now we are ready to complete the proof of Theorem 3.3.1.

3.3.2 Proof of Theorem 3.3.1.

(i) Due to Propositions 3.3.1 and 3.3.2 we know that the Cramér condition is satisfied. The decay rate $\gamma(\tau)$ is a solution of the following equation:

$$\frac{(\mu-\lambda+s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}-\mathrm{e}^{-\frac{1}{2}\tau f(s)})+f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}+\mathrm{e}^{-\frac{1}{2}\tau f(s)})}{(\mu-\lambda-s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}-\mathrm{e}^{-\frac{1}{2}\tau f(s)})+f(s)(\mathrm{e}^{\frac{1}{2}\tau f(s)}+\mathrm{e}^{-\frac{1}{2}\tau f(s)})}=\frac{1}{\rho}, \tag{3.23}$$

where $f(s)=\sqrt{(\mu+\lambda-s)^2-4\,\lambda\,\mu}$.

Since both the numerator and the denominator in (3.8) for $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ are continuous functions, the MGF becomes infinite only when the denominator equals zero. The fact that $\mathbf{E}[\mathrm{e}^{\gamma(\tau)C_i(\tau)}]=1/\rho$ implies that the denominator is non-zero at $s=\gamma(\tau)$, and hence, due to continuity, is non-zero in some neighborhood of it. From this we conclude that there is a neighborhood of $\gamma(\tau)$ in which the MGF stays finite, and consequently, $\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\big|_{s=\gamma(\tau)}<\infty$. The distribution of $C_i(\tau)$ is non-lattice, since $\mathbf{P}(C_i(\tau)=B^r)>0$, and the residual service time B^r has a density. Thus, the conditions (i) of Theorem 2.1.1 are satisfied and we can conclude that as $x\to\infty$, the probability $\mathbf{P}(V_1(\tau)>x)$ decays exponentially fast with the asymptotic decay rate $\gamma(\tau)$ and the asymptotic constant determined by Equation (2.6).

The fact that the MGF $\mathbf{E}[\mathrm{e}^{sV_0(\tau)}]$ has the same abscissa of convergence as $\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]$ (since B has unbounded support), implies that $\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]$ is finite for any τ in some neighborhood of $\gamma(\tau)$. Applying Breiman's theorem [30], we obtain

$$\mathbf{P}(V(\tau)>x)\sim\frac{1-\rho}{h(\tau)\gamma(\tau)}\mathbf{E}[\mathrm{e}^{\gamma(\tau)V_0(\tau)}]\mathrm{e}^{-\gamma(\tau)x},\qquad x\to\infty. \tag{3.24}$$

Let us now compute the prefactor. We first need to determine the derivative of the MGF of $C_i(\tau)$ at $\gamma(\tau)$. For convenience we use the following notation. Let us denote the denominator in Equation (3.7) by $D(s,\tau)$ and the numerator by $N(s,\tau)$. The exponents e^+ and e^- denote $\mathrm{e}^{f(s)\tau/2}$ and $\mathrm{e}^{-f(s)\tau/2}$ respectively. Then

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\bigg|_{s=\gamma(\tau)}=\left[\frac{N'(s,\tau)D(s,\tau)-N(s,\tau)D'(s,\tau)}{D^2(s,\tau)}\right]\bigg|_{s=\gamma(\tau)}$$

$$=\left[\frac{N'(s,\tau)}{D(s,\tau)}-\frac{D'(s,\tau)}{\rho\cdot D(s,\tau)}\right]\bigg|_{s=\gamma(\tau)}=\frac{\rho N'(s,\tau)-D'(s,\tau)}{\rho D(s,\tau)}\bigg|_{s=\gamma(\tau)},$$

where the derivatives are taken with respect to s,

$$N'(s,\tau)=(\mu-\lambda+s)(\mathrm{e}^+\!+\mathrm{e}^-)\frac{\tau}{2}f'(s)+(\mathrm{e}^+\!-\mathrm{e}^-)+f'(s)\left((\mathrm{e}^+\!+\mathrm{e}^-)+f(s)(\mathrm{e}^+\!-\mathrm{e}^-)\frac{\tau}{2}\right),$$

$D'(s,\tau) = (\mu-\lambda-s)(\mathrm{e}^{+}+\mathrm{e}^{-})\frac{\tau}{2}f'(s)-(\mathrm{e}^{+}-\mathrm{e}^{-})+f'(s)\left((\mathrm{e}^{+}+\mathrm{e}^{-})+f(s)(\mathrm{e}^{+}-\mathrm{e}^{-})\frac{\tau}{2}\right)$,

and $f'(s) = (\lambda+\mu-s)/f(s)$. Hence,

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\Big|_{s=\gamma(\tau)} =$$

$$= \frac{1}{\rho D(\gamma(\tau),\tau)f(\gamma(\tau))}\left[(1+\rho)\left[(f(\gamma(\tau))(\mathrm{e}^{+}-\mathrm{e}^{-})+\gamma(\tau)\frac{\tau}{2}(\lambda+\mu-\gamma(\tau)))(\mathrm{e}^{+}+\mathrm{e}^{-})\right]\right.$$
$$\left.-(1-\rho)(\lambda+\mu-\gamma(\tau))\left[(\mathrm{e}^{+}+\mathrm{e}^{-})(1+(\mu-\lambda)\frac{\tau}{2})+(\mathrm{e}^{+}-\mathrm{e}^{-})f(\gamma(\tau))\frac{\tau}{2}\right]\right].$$

Let us denote the expression between brackets by $K(\tau)$,

$$\frac{d}{ds}\mathbf{E}[\mathrm{e}^{sC_i(\tau)}]\Big|_{s=\gamma(\tau)} = \frac{1}{\rho D(\gamma(\tau),\tau)f(\gamma(\tau))}K(\tau).$$

Since $\mathbf{E}[\mathrm{e}^{sV_0(\tau)}] \equiv \delta(-s,\tau) = \frac{2f(s)\mathrm{e}^{-(-s+\lambda-\mu)\tau/2}}{D(s,\tau)}$ by Equation (3.4), we obtain from Equation (3.24):

$$\alpha(\tau) = \frac{1-\rho}{\gamma(\tau)K(\tau)}2f^2(\gamma(\tau))\mathrm{e}^{-(-\gamma(\tau)+\lambda-\mu)\tau/2},$$

which gives Equation (3.12).

(ii) The expression for the delay rate is given in Proposition 3.3.1, $\gamma(\tau_0) = s_r$. Let us now compute the prefactor $\alpha(\tau)$.

As before, $N(s,\tau)$ and $D(s,\tau)$ denote the numerator and the denominator in Equation (3.5). Denote by $N_1(s,\tau)$ the numerator and by $D_1(s,\tau)$ the denominator in Equation (3.8). Observe that $N(s,\tau) = f(s)\cdot N_1(s,\tau)$ and $D(s,\tau) = f(s)\cdot D_1(s,\tau)$.

The constant $\alpha(\tau)$ is determined (Equations (3.4) and (3.24)) as

$$\alpha(\tau) = \frac{1-\rho}{\gamma(\tau)\frac{d}{ds}\varphi(s,\tau)|_{s=\gamma(\tau)}}\frac{2f(\gamma(\tau))\,\mathrm{e}^{-(-\gamma(\tau)+\lambda-\mu)\tau/2}}{D(s,\tau)}.$$

Since $\frac{d}{ds}\varphi(s,\tau) = \frac{N_1'(s,\tau)D_1(s,\tau)-D_1'(s,\tau)N_1(s,\tau)}{D_1^2(s,\tau)}$, $\frac{N_1(\gamma(\tau),\tau)}{D_1(\gamma(\tau),\tau)} = 1/\rho$, and $D(s,\tau) = f(s)\cdot D_1(s,\tau)$, we obtain

$$\alpha(\tau) = \frac{2(1-\rho)\,\mathrm{e}^{-(-\gamma(\tau)+\lambda-\mu)\tau/2}\,f(\gamma(\tau))}{D(\gamma(\tau),\tau)\,\gamma(\tau)}\frac{D_1^2(\gamma(\tau),\tau)}{N_1'(\gamma(\tau),\tau)\,D_1(\gamma(\tau),\tau)-D_1'(s_r,\tau)\,N_1(s_r,\tau)}$$
$$= \frac{2\,(1-\rho)\,\mathrm{e}^{-(-\gamma(\tau)+\lambda-\mu)\tau/2}}{\gamma(\tau)}\frac{\rho}{\rho N_1'(\gamma(\tau),\tau)-D_1'(\gamma(\tau),\tau)}, \qquad (3.25)$$

where $N_1'(s,\tau)$ and $D_1'(s,\tau)$ are the derivatives of $N_1(s,\tau)$ and $D_1(s,\tau)$ with respect to s.

Let us now compute these derivatives:

$$N_1'(s,\tau) = \sum_{n=1}^{\infty} \frac{(\frac{\tau}{2})^{2n} d(s)^{n-1}}{(2n-1)!}(\mu+\lambda-s)\left[\frac{(\mu-\lambda+s)\frac{\tau}{2}}{(2n+1)}+1\right] + \sum_{n=0}^{\infty}\frac{(\frac{\tau}{2})^{2n}}{(2n)!}\left[\frac{\tau}{2}d(s)^n}{2n+1}\right],$$

$$D_1'(s,\tau) = \sum_{n=1}^{\infty} \frac{(\frac{\tau}{2})^{2n} d(s)^{n-1}}{(2n-1)!}(\mu+\lambda-s)\left[\frac{(\mu-\lambda-s)\frac{\tau}{2}}{(2n+1)}+1\right] - \sum_{n=0}^{\infty}\frac{(\frac{\tau}{2})^{2n}}{(2n)!}\left[\frac{\tau}{2}d(s)^n}{2n+1}\right].$$

In particular, when $\tau = \tau_0$, we see that $\gamma(\tau_0) = s_r$ and $d(s_r) = 0$. Consequently, the expressions simplify to

$$\begin{aligned} N_1'(s_r,\tau_0) &= \frac{(\frac{\tau_0}{2})^2}{2} 2(\mu+\lambda-s_r)\left[\frac{(\mu-\lambda+s_r)}{3}\frac{\tau_0}{2}+1\right] + \frac{\tau_0}{2} \\ &= \frac{(\sqrt{\mu}-\sqrt{\lambda})(6\lambda-(\sqrt{\mu}-\sqrt{\lambda})^2)}{6\lambda\sqrt{\mu}(\sqrt{\mu}+\sqrt{\lambda})^2}, \end{aligned}$$

$$\begin{aligned} D_1'(s_r,\tau_0) &= \frac{(\frac{\tau_0}{2})^2}{2} 2(\mu+\lambda-s_r)\left[\frac{(\mu-\lambda-s_r)}{3}\frac{\tau_0}{2}+1\right] - \frac{\tau_0}{2} \\ &= \frac{-(\sqrt{\mu}-\sqrt{\lambda})(6\mu+(\sqrt{\mu}-\sqrt{\lambda})^2)}{6\mu\sqrt{\lambda}(\sqrt{\mu}+\sqrt{\lambda})^2}. \end{aligned}$$

The difference $\rho N_1'(s_r,\tau_0) - D_1'(s_r,\tau_0)$ equals

$$\rho N_1'(s_r,\tau_0) - D_1'(s_r,\tau_0) = \frac{(\sqrt{\mu}-\sqrt{\lambda})(6\,(\mu\sqrt{\mu}+\lambda\sqrt{\lambda})+(\sqrt{\mu}-\sqrt{\lambda})^3)}{6\,\mu\,\sqrt{\lambda\mu}\,(\sqrt{\mu}+\sqrt{\lambda})^2}.$$

Substitution of τ_0, s_r, $\rho N_1'(s_r,\tau_0) - D_1'(s_r,\tau_0)$ into (3.25) gives

$$\alpha(\tau_0) = \frac{12\,(1-\rho)\,\lambda\,\sqrt{\lambda\mu}\,\mathrm{e}^{1-\frac{1}{\sqrt{\rho}}}}{(\sqrt{\mu}-\sqrt{\lambda})(6\,(\mu\sqrt{\mu}+\lambda\sqrt{\lambda})+(\sqrt{\mu}-\sqrt{\lambda})^3)}.$$

Obviously, the computed value is a strictly positive finite number. Further simplification leads to (3.15). This completes the proof. □

3.4 Other service disciplines

In this section we investigate the behavior of the decay rate $\gamma(\tau)$ by solving Equations (3.10) and (3.11) numerically. Furthermore, we perform a comparison of the PS decay rate with the decay rates in the M/M(τ)/1 queue under different service disciplines: in particular, the Shortest Remaining Processing Time (SRPT) and the Foreground-Background (FB) disciplines.

The decay rate of the conditional sojourn time $V(\tau)$ under the SRPT and FB disciplines has been studied in Nuyens and Zwart [88] and Mandjes and Nuyens [81],

respectively. For the SRPT discipline, Nuyens and Zwart [88] have shown that the decay rate of the conditional sojourn time $V_{SRPT}(\tau) = [V_{SRPT}|B = \tau]$ coincides with the decay rate of the residual busy period $\gamma^p_{SRPT}(\tau)$ in the queue with service time $B^\tau_{SRPT} = B\mathbf{1}(B < \tau)$. Mandjes and Nuyens [81] have derived a similar result for the FB discipline. They proved that if the generic service time has an exponential moment then the sojourn time $V_{FB}(\tau)$ has the same decay rate $\gamma^p_{FB}(\tau)$ as the residual busy period in the queue with service time $B^\tau_{FB} = \min(B, \tau)$. It is known that the decay rate of the busy period can be determined as

$$\gamma^p(\tau) = -\kappa(\theta_0),$$

where $\kappa(s) = \lambda(\mathbf{E}[\mathrm{e}^{sB^\tau}] - 1) - s$, and $\theta_0 > 0$ is a solution of the equation $\kappa'(\theta_0) = 0$ (or equivalently $\lambda(\mathbf{E}[\mathrm{e}^{sB^\tau}])'_s = 1$).

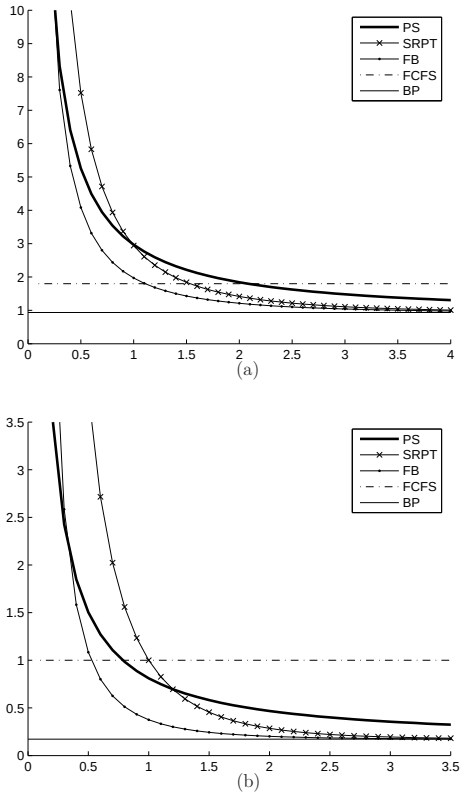

Figure 3.4: Decay rate as a function of τ in the M/M(τ)/1 queue with the PS, SRPT, FB and FCFS service disciplines, $\mu = 2$: (a) $\lambda = 0.2$, (b) $\lambda = 1$.

Figure 3.4 presents the decay rate $\gamma(\tau)$ as a function of τ for the above-mentioned disciplines. The generic service time is exponential with parameter $\mu = 2$. Figure

3.4 (a) shows the decay rates under very low traffic load, $\rho = 0.1$, and Figure 3.4 (b) is for $\rho = 0.5$. In the figures, the horizontal lines show the decay rate in the M/M/1 FCFS queue (dash-dotted line) and the decay rate of the busy period (solid line referred to as BP). The decay rate of the M/M/1 FCFS queue is equal to $\gamma_{FCFS} = \mu - \lambda$ and the decay rate of the busy period is $\gamma^p = (\sqrt{\mu} - \sqrt{\lambda})^2$.

Figure 3.5 shows the decay rates when the traffic intensity is reasonably high, (a) $\rho = 0.9$, (b) $\rho = 0.95$. From the figures we clearly see that when the service requirement τ becomes larger, the decay rates for all disciplines decrease and converge to the decay rate of the busy period γ^p. Thus, the sojourn time of a customer with a large service requirement behaves approximately like the residual busy period.

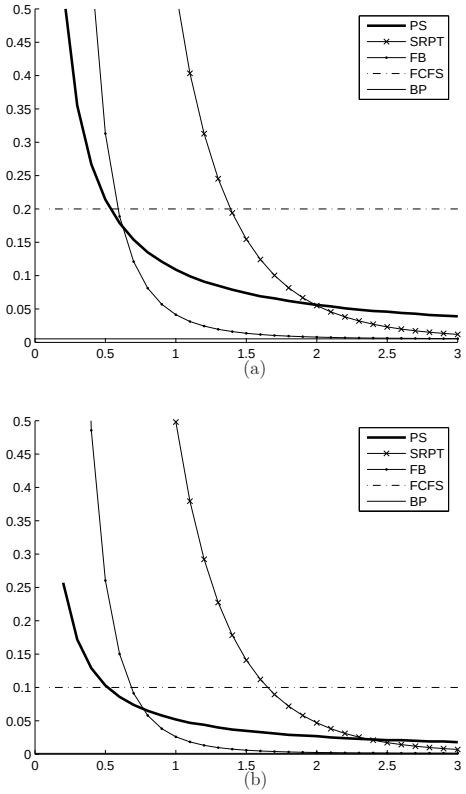

Figure 3.5: Decay rate as a function of τ in the M/M(τ)/1 queue with the PS, SRPT, FB and FCFS service disciplines, $\mu = 2$: (a) $\lambda = 1.8$, (b) $\lambda = 1.9$.

All graphs show that for moderate values of τ the largest decay rate is achieved by SRPT. For larger service requirements the FCFS discipline provides the largest decay rate. Thus, there is a critical value of τ such that for the smaller requirements

SRPT has the largest decay rate and for larger ones FCFS. Analytic results in [88] and our simulations show that in the M/M/1 queue for the majority of the customers (at least 85%) SRPT provides a larger decay rate in comparison to FCFS. Interestingly, for the unconditional sojourn time, the large-deviations results imply on the contrary that large sojourn times are more likely under SRPT than under FCFS. If the decay rate is used as performance measure, PS does not appear to be the optimal discipline for service requirements of any size. See Figures 3.4 (b) and 3.5 (a,b).

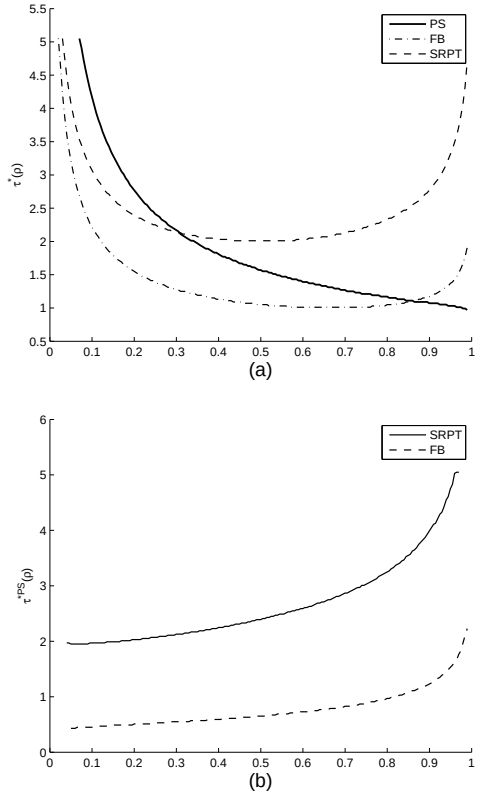

Figure 3.6: Decay rate τ^* as a function of ρ in the M/M(τ)/1 queue: (a) - intersection with FCFS decay rate, (b) - intersection with PS decay rate

However, in Figure 3.4 (a) we see a somewhat different picture. For a certain range of service requirements, not too long and not too short, the decay rate for the PS discipline is the largest. This may be explained as follows. In this case the traffic load is very low implying a small number of customers in the system. Hence, the customers with moderate service requirements receive a sufficiently high service

rate and are protected from being delayed by larger or smaller requirements as in FCFS and SRPT queues, respectively.

Let us introduce τ^*_{PS} as the value of τ at which the PS decay rate $\gamma(\tau)$ is equal to the FCFS decay rate, i.e. the value at which $\gamma(\tau)$ crosses level $\mu-\lambda$. Define similarly τ^*_{SRPT} and τ^*_{FB}. Figure 3.6 shows the behavior of τ^*_{PS} and τ^*_{SRPT} as a function of the traffic load ρ. As we can see, for traffic load $\rho < 0.3$ the PS decay rate reaches the value $\mu - \lambda$ later than the SRPT decay rate. This means that for such ρ, there exists a positive ε_ρ, such that in the interval $[\tau^*_{PS} - \varepsilon_\rho, \tau^*_{PS}]$ the PS discipline has the largest decay rate (compared to FCFS and SRPT). Comparing the PS decay rate to the FB one (see Figure 3.6 (a)), the decay rate shows similar behavior. In this case the threshold load is $\rho < 0.86$.

Figure 3.6 (b) shows τ^{*PS}, the value of τ at which the decay rates $\gamma_{SRPT}(\tau)$ under the SRPT discipline and $\gamma_{FB}(\tau)$ under the FB discipline are equal to the decay rate $\gamma(\tau)$ under PS. As we see, the higher the value of ρ, the narrower is the range of the service requirements for which the PS discipline provides the largest decay rate.

Let us now summarize the results. In the PS queue, as well as in SRPT and FB, the decay rate $\gamma(\tau)$ of the conditional sojourn time decreases and converges to the decay rate of the busy period as $\tau \to \infty$. From the large-deviations point of view, in most cases (except when the traffic load is quite low) the decay rate under the PS discipline is not optimal for any service requirement τ. For larger service requirements, FCFS has the highest decay rate, and for smaller service requirements, SRPT performs the best. For the unconditional sojourn time in the M/M/1 queue however, it is known [88] that the decay rate under the PS discipline coincides with the decay rate under SRPT (and FB as well) and is strictly smaller than the one under FCFS. The decay rate under PS, SRPT and FB is given by the decay rate of the residual busy period.

3.5 Numerical results

Finally, we will study the accuracy of the exponential approximation (3.9) of the sojourn time in the M/M(τ)/1 queue:

$$\mathbf{P}(V(\tau) > x) \approx \alpha(\tau)\mathrm{e}^{-\gamma(\tau)x}.$$

The exponential asymptotics are compared to exact values of $\mathbf{P}(V(\tau) > x)$ computed by numerical LST inversion. We will use the inversion algorithm of Abate and Whitt [2]. In this method the probability distribution function is presented as an infinite sum of complex-valued terms. For the summation of this infinite series the classical Euler summation method is applied. This method is known to provide high accuracy.

Table 3.1 shows the numerical results for various service requirements τ. For simplicity we normalize the generic service time, $\mu = 1$, and take arrival rate $\lambda = 0.9$. For $\tau = 0.8$ and $\tau = 2$ the first column shows the probability $\mathbf{P}(V(\tau) > x)$ obtained by numerical inversion. The second column shows the exponential asymptotics

	$\tau = 0.8$			$\tau = 2$	
x	LST inv.	asympt.	x	LST inv.	asympt.
5	5.49-01	5.77-01	10	6.34-01	7.25-01
10	2.82-01	2.96-01	100	4.72-03	5.41-03
20	7.41-02	7.79-02	150	3.10-04	3.56-04
40	5.14-03	5.39-03	200	2.04-05	2.34-05
80	2.47-05	2.59-05	250	1.33-06	1.54-06
100	1.70-06	1.79-06	300	9.43-08	1.01-07
120	1.24-07	1.24-07	310	5.78-08	5.89-08

Table 3.1: Comparison of the exponential asymptotics to results of numerical inversion.

derived in Theorem 3.3.1. The numbers show reasonably good accuracy of the asymptotic tail approximation. The relative error is on average about 5-10%.

	$\tau = 0.8$				$\tau = 2$		
x	LST inv.	asympt.	HT	x	LST inv.	asympt.	HT
10	5.42-01	5.56-01	5.35-01	10	8.04-01	8.59-01	7.79-01
20	2.84-01	2.92-01	2.87-01	100	7.70-02	8.20-02	8.21-02
50	4.10-02	4.22-02	4.39-02	200	5.67-03	6.03-03	6.74-03
100	1.63-03	1.68-03	1.93-03	300	4.18-04	4.44-04	5.53-04
150	6.45-05	6.66-05	8.48-05	400	3.08-05	3.26-05	4.54-05
200	2.54-06	2.65-06	3.73-06	500	2.26-06	2.40-06	3.73-06
240	1.96-07	2.01-07	3.06-07	600	1.73-07	1.76-07	3.06-07
250	1.05-07	1.05-07	1.64-07	640	6.24-08	6.21-08	1.13-07

Table 3.2: Comparison of the exponential asymptotics to results of numerical inversion and heavy-traffic asymptotics.

Table 3.2 shows results for heavy traffic, in particular $\rho = 0.95$. In addition to the results from numerical inversion and asymptotics, the table presents results of the heavy-traffic approximation. From the results in [105, 118], it is known that under heavy traffic the sojourn time distribution in the M/G/1 PS queue behaves as

$$\mathbf{P}(V(\tau) > x) \approx \mathrm{e}^{-\frac{(1-\rho)x}{\tau}}, \qquad x \to \infty. \tag{3.26}$$

These values are presented in the columns labeled HT.

The accuracy of the asymptotic approximation (3.9) is better for higher traffic load. It is also much more accurate than the heavy-traffic approximation for larger x, while for small x the heavy-traffic approximation performs better.

Chapter 4

Sojourn time tails in queues with varying service rate

In the previous chapters we analyzed the sojourn time behavior in PS queues with constant server capacity. We obtained the exact asymptotics for the tail of the probability distribution of the sojourn time. In the present chapter we consider a more general situation and assume that the capacity of the server varies in time. Such models can be regarded as an appropriate flow-level approximation for modeling the elastic data transfers in integrated communication networks with a mixture of elastic and streaming traffic. We refer to Section 1.2 for further background.

In the present chapter we study the asymptotic properties of the sojourn time distribution of the elastic flows. The main goal is to generalize the result of Mandjes and Zwart [82] to a setting in which the available service capacity varies according to some stochastic process. Mandjes and Zwart [82] derived the logarithmic asymptotics of the sojourn time in the GI/GI/1-PS queue with *constant* service capacity (see (1.10)), under technical assumptions which guarantee that the tail distribution of the service time is not too light and not too heavy. We extend the logarithmic asymptotics in [82] by constructing lower and upper bounds, which asymptotically coincide. The upper bounds can be established under rather general conditions, whereas the lower bound requires that the service process obeys a sample-path large-deviations principle. Again the service requirements should be from a light-tailed distribution (but not too light).

As a special case, we study service processes that have a so-called Markov-fluid structure. Under the additional assumption of the arrival process being Poisson, we derive for these service processes an explicit upper bound on the tail probability of the sojourn time, rather than just an upper bound on the exponential decay rate.

Our proofs predominantly rely on large-deviations tools, such as the classical Chernoff bound, as well as the application of sample-path large-deviations principles. An important role, however, is also played by the insight that, for overloaded PS systems, the queue length increases roughly at a linear rate. As a by-product, the proofs show that the sojourn time asymptotics resemble busy-period asymptotics (in the sense that their exponential decay rates coincide). Although our results are an

extension of the results in [82], we have actually simplified the proofs; in particular, we have eliminated the need to use detailed fluid-limit results for overloaded PS queues, as used in [82]. To obtain our results for the Markov-fluid case we use a change-of-measure argument. We twist the distributions of the arrival and service processes in such a way that the tagged customer sees a critically loaded system.

Finally, our methods allow us to obtain an extension of the result to the DPS discipline. As for the single-class case, we allow the service rate to be random, but note that the obtained asymptotic results are also new for the standard DPS queue with a fixed service rate. More specifically, we show that the decay rate of the sojourn time is weight-independent (and hence the same for customers of any class).

The organization of this chapter is as follows: The model is described in Section 4.1. In Section 4.2 we present our main results on the logarithmic asymptotics for the case with general service rate. In addition, we consider the special case in which the service rate varies according to a Markov-fluid process. The proofs can be found in Sections 4.3 and 4.4. In Section 4.5 we generalize the result to the DPS queue.

4.1 Model description and preliminaries

In this section we introduce the necessary notation and state some preliminary results.

Let A_n, $n \in \mathbb{N}$, be the time between the $(n-1)$-st and n-th arrival after time zero. To emphasize that an arrival occurred in the past, we also use the notation A_{-n}, $n \in \mathbb{N}$, for the time between the $(n-1)$-st and n-th arrival before time zero. Furthermore, let B_n, $n \in \mathbb{Z}$, be the service requirement of the nth customer; recall that B_0 corresponds to the tagged customer. We assume that $(A_n)_n$ and $(B_n)_n$ are mutually independent sequences, each consisting of i.i.d. random variables. We introduce the random walks $S_n^A = A_1 + \ldots + A_n$ and $S_n^B = B_1 + \ldots + B_n$, and similarly, with respect to events in the past, $S_{-n}^A = A_{-n} + \ldots + A_{-1}$, $S_{-n}^B = B_{-n} + \ldots + B_{-1}$. We denote the random variable corresponding to a generic interarrival time (service time) by A (B, respectively).

We set

$$N(t) := \max\{n \in \mathbb{N} : S_n^A \leq t\}$$

representing the number of arrivals in the time interval $(0,t]$. Denote by $A(0,t)$, $t > 0$, the total amount of work fed into the queue in the time interval $(0,t]$, i.e.,

$$A(0,t) = \sum_{i=1}^{N(t)} B_i.$$

Analogously, $C(t_1,t_2)$ is defined as the total service provided in the time interval $(t_1,t_2]$ with $t_2 > t_1$,

$$C(t_1,t_2) = \int_{t_1}^{t_2} R(u)du,$$

where $R(u)$ denotes the (random, non-negative) service rate available at time u. Later we also consider the system in the past, i.e., before time zero; then we use the notation $A(-t,0)$ for the total amount of work fed into the system on $[-t,0)$. Note that the tagged arrival which occurred at time 0 is included in neither $A(0,t)$ nor $A(-t,0)$. The cumulative arrival and service processes are assumed to be independent of each other.

Throughout the chapter we assume the cumulative service process to satisfy the following conditions:

1. the cumulative service process has *stationary increments*, i.e., the distribution of $C(t_1+\delta, t_2+\delta)$ does not depend on δ;
2. the *service rate* $R(\cdot)$ *is bounded from above*, i.e. there exists $r_{\max}$ such that $R(u) \le r_{\max}$ for all u;
3. the asymptotic cumulant function of $C(0,x)$ exists:

$$c(s) := \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sC(0,x)}].$$

Furthermore, the system is assumed to be stable, i.e. the long-run average work offered to the system, say α, is smaller than the average offered service, say c, where

$$\alpha := \lim_{t\to\infty} \frac{\mathbf{E}A(0,t)}{t}, \qquad c := \lim_{t\to\infty} \frac{\mathbf{E}C(0,t)}{t}.$$

Define the MGF s $\Phi_B(s) := \mathbf{E}[\mathrm{e}^{sB}]$ and $\Phi_A(s) := \mathbf{E}[\mathrm{e}^{sA}]$. Since both $\Phi_A(\cdot)$ and $\Phi_B(\cdot)$ are strictly increasing and strictly convex functions, the inverse functions $\Phi_A^{\leftarrow}(\cdot)$ and $\Phi_B^{\leftarrow}(\cdot)$ are well-defined. We assume that either A or B does not have a deterministic distribution. An important result is that the cumulant function of the amount of work fed to the system can be expressed explicitly in terms of the moment generating functions of A and B.

Lemma 4.1.1. *For $s \ge 0$, the asymptotic cumulant function $\alpha(s)$ of $A(0,x)$, $x>0$, is given by*

$$\alpha(s) := \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sA(0,x)}] = -\Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(s)}\right). \tag{4.1}$$

If either A or B is non-deterministic, then $\alpha(\cdot)$ is strictly convex.

The result of Lemma 4.1.1, as stated by Whitt [114], was proved in [82].

In the sequel, we separately consider the special case in which the service process is given by a Markov-fluid process. Such a process can be described as follows. Consider a continuous-time Markov process on a finite state space $\{1,2,\ldots,d\}$. The transition rate matrix is denoted by $Q = (q_{ij})_{i,j=1,2,\ldots,d}$, where $q_{ij} \ge 0$ $(i \ne j)$ and $q_{ii} = -\sum_{j\ne i} q_{ij}$. We assume that the Markov process is irreducible, and π denotes its steady-state distribution. When the Markov process is in state i, the server provides service at constant rate $r_i \ge 0$. Let R be the diagonal matrix with

coefficients r_i on the diagonal. Denote the mean rate by $c = \sum_{i=1}^{d} r_i \pi_i$. We denote this class of processes by $\mathbb{M}\mathrm{f}(Q, R)$; if the service process is of this type, we write $C(\cdot,\cdot) \in \mathbb{M}\mathrm{f}(Q, R)$. Results from Kesidis *et al.* [73] yield the following standard properties.

Property 4.1.1. *Let $C(\cdot,\cdot) \in \mathbb{M}\mathrm{f}(Q, R)$. Then the following statements hold:*

1. *The MGF of the service available in an interval of length x is given by*

$$\mathbf{E}[\mathrm{e}^{sC(0,x)}] = \pi \mathrm{e}^{(Q+sR)x}\mathbf{1},$$

where $\mathbf{1}$ is the all-one vector of dimension d.

Denote by $c_1(s), \ldots, c_d(s)$ the eigenvalues of the matrix $Q + sR$. Hence, the MGF can be represented as, for appropriate numbers $m_1, \ldots, m_d$,

$$\mathbf{E}[\mathrm{e}^{sC(0,x)}] = \sum_{i=1}^{d} m_i \mathrm{e}^{c_i(s)x}.$$

2. *For all real s there exists a limiting MGF:*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sC(0,x)}] = c(s).$$

Moreover, $c(s) = \max\{c_1(s), \ldots, c_d(s)\}$, i.e., $c(s)$ is the largest real eigenvalue of $Q + sR$; the corresponding eigenvector is componentwise positive.

3. *There exists a finite K such that*

$$\mathbf{E}[\mathrm{e}^{sC(0,x)}] \le K \mathrm{e}^{c(s)x}.$$

For instance, $K = \sum_{i=1}^{d} m_i$.

4.2 Main results

In this section we present the main results of the chapter. We focus on the sojourn time V of a tagged customer (with service requirement B_0), which we assume to arrive at time 0. We characterize the logarithmic asymptotic behavior of the tail probability $\mathbf{P}(V > x)$ as $x \to \infty$, under the assumption that the service requirement has a light-tailed distribution.

To put things in perspective, we first recall the asymptotic behavior of the sojourn time distribution in a PS queue with *constant* (rather than fluctuating) service capacity. Mandjes and Zwart [82] derived the following logarithmic estimates under the assumption that the service requirement distribution has a light tail.

Theorem 4.2.1. *([82]) Consider the GI/GI/1 PS queue with unit service rate. If there exists a solution $\nu^* > 0$ to the equation $\alpha'(s) = 1$, and for each constant $c > 0$*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(B > c \log x) = 0,$$

then

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s\geq 0}(\alpha(s) - s) = \alpha(\nu^*) - \nu^*. \tag{4.2}$$

Our main goal is to derive a generalization of the above result for a queue with *varying* service rate. Under similar assumptions on the arrival and service requirement processes, and in addition certain assumptions on the service process, we can prove the following extension of (4.2):

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s\geq 0}(\alpha(s) + c(-s)). \tag{4.3}$$

Despite the simple form, the proof of the above result is quite technical. The proof consists of two parts, derivation of an upper bound (i.e., (4.3) with "="replaced by "$\leq$") and derivation of a lower bound (i.e., (4.3) with "="replaced by "$\geq$") which asymptotically coincide.

The proof of the upper bound is essentially based on classical Chernoff-bound arguments, and applies without imposing additional conditions on the service process. The proof of the lower bound, however, is substantially harder. There we first truncate the service requirement distribution (and then let the truncation threshold increase to ∞), so that we enforce linearly bounded queue length growth. Thus, the problem is reduced to finding the corresponding busy-period asymptotics. The derivation of these busy-period asymptotics requires an additional assumption on the service process: we require the service process to obey a so-called sample-path large-deviations principle (more precisely: only the large-deviations *lower* bound is required here).

In the following subsections we will present results for the system with general service process, but also (more explicit) results for the case the service process is Markov fluid. The proofs are deferred to Sections 4.3 and 4.4.

4.2.1 Upper bound

We first present the asymptotic upper bound for the sojourn time distribution in a GI/GI/$\cdot$ system with a generally distributed service process. We need to make the following assumption.

Assumption 4.2.1. *There exists a $\nu > 0$ such that $\alpha(\nu) + c(-\nu) < 0$.*

This assumption ensures that the service requirements are light-tailed and that the system is stable. To be more precise, what the assumptions states is that in some neighborhood to the right of the origin the cumulant functions stay finite. This implies (due to Lemma 4.1.1) that there exists a neighborhood of the origin in which the MGF $\Phi_B(\cdot)$ is well-defined (as an aside, note that this implies that B is light-tailed). Since the function $g(s) = \alpha(s) + c(-s)$ is strictly convex and equals 0 at $s = 0$, the assumption implies that $g(\cdot)$ has a negative derivative at $s = 0$, $\alpha - c < 0$, and hence, the system is stable.

Due to strict convexity of the cumulant function, we can define $\omega^* > 0$ such that

$$\omega^* = \arg\inf_{s\geq 0}(\alpha(s) + c(-s)).$$

Since $\alpha(s) + c(-s)$ equals zero at $s = 0$ and has a strictly negative derivative at $s = 0$, we also have $\alpha(\omega^*) + c(-\omega^*) < 0$.

The next theorem gives the logarithmic upper bound for $\mathbf{P}(V > x)$ in terms of the cumulant functions.

Theorem 4.2.2. *If Assumption 4.2.1 is satisfied, then*

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) \leq \alpha(\omega^*) + c(-\omega^*). \tag{4.4}$$

Besides the general upper bound on the exponential decay rate, as presented in Theorem 4.2.2, we have a tighter result (namely bounds on the probability $\mathbf{P}(V > x)$ itself, uniformly in x) for an important special case. This result requires an additional assumption; it implies Assumption 4.2.1 and existence of ω^*.

Assumption 4.2.2. *There exists a solution $\nu^* > 0$ to $\alpha(\nu^*) + c(-\nu^*) = 0$.*

As a special case we consider Poisson arrivals (rather than renewal arrivals; the arrival process is thus a compound Poisson process) and Markov-fluid service. We remark that the constant K, as used in Theorem 4.2.3, will be explicitly given in the proof of the result.

Theorem 4.2.3. *Suppose the arrival process is given by a compound Poisson process (with rate λ) and the service process is in $\mathbb{M}\mathrm{f}(Q, R)$. Then, under Assumption 4.2.2,*

$$\mathbf{P}(V > x) \leq K\mathrm{e}^{(a(\omega^*)+c(-\omega^*))x}, \tag{4.5}$$

uniformly in $x \geq 0$, and $\alpha(\omega^*) = \lambda(\Phi_B(\omega^*) - 1)$.

4.2.2 Lower bound

Let us now turn to the results for the lower bound on $\mathbf{P}(V > x)$. Here we need the following assumption.

Assumption 4.2.3. *For each constant $c > 0$, we have*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(B > c \log x) = 0.$$

It is readily checked that this assumption is satisfied by most distributions of interest, such as phase-type, Gamma, Weibull distributions, etc. However, it is noted that it is violated by distributions with *extremely light* tails. For instance, the assumption does not hold for service times for which $\mathbf{P}(B > x)$ is of the form $\exp(-\mathrm{e}^x)$, and by service requirement distributions with bounded support (including deterministic service requirements).

The derivation of the lower bound is considerably more involved than the corresponding upper bound. Importantly, it requires extra structure of the process $C(\cdot, \cdot)$, namely that the process $C(\cdot, \cdot)$ must satisfy the lower bound of a *sample-path large-deviations principle*.

Definition 4.2.1. *Denote by $\mathcal{AC}$ the space of all absolutely continuous functions (see e.g. [41]), i.e.,*

$$\mathcal{AC} = \left\{ f \in C([0,1]) : \begin{array}{l} \text{if } \sum_{l=1}^{k} |t_l - s_l| \to 0,\ s_l \le t_l \le s_{l+1} < t_{l+1}, \\ \text{then } \sum_{l=1}^{k} |f(t_l) - f(s_l)| \to 0 \end{array} \right\}.$$

Define the space $\Omega := \{f \in [0,1] \to R,\ f \in \mathcal{AC}, f(0) = 0\}$.

Let the process $Z_x(\cdot)$ be given through

$$Z_x(u) := \frac{1}{x} \int_0^{ux} c(s)ds = \frac{1}{x}\, C(0, ux).$$

The process $Z_x(\cdot)$ obeys a sample-path large-deviations principle *(sp-LDP) if for all $S \subset \Omega$:*

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}(Z_x(\cdot) \in S) \le - \inf_{f\in\overline{S}} \int_0^1 \Lambda(f'(t))dt, \tag{4.6}$$

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}(Z_x(\cdot) \in S) \ge - \inf_{f\in S^o} \int_0^1 \Lambda(f'(t))dt, \tag{4.7}$$

where $\Lambda(t) := \sup_{s\in\mathbb{R}}(st - c(s))$, $\overline{S}$ is the closure and S^o is the interior of set S. We say that (4.6) *is the upper bound of the sp-LDP, and* (4.7) *is the lower bound of the sp-LDP.*

Assumption 4.2.4. *The process $Z_x(\cdot)$, defined through $Z_x(u) := C(0, ux)/x$, satisfies the lower bound of the sp-LDP* (4.7).

The next theorem presents the main result of the present chapter; its upper bound was already stated in Theorem 4.2.2.

Theorem 4.2.4. *If Assumptions 4.2.1, 4.2.3 and 4.2.4 are satisfied, then*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) = \alpha(\omega^*) + c(-\omega^*).$$

Although, to our best knowledge, no sp-LDP was established for a Markov-fluid process, we were still able to prove the corresponding logarithmic lower bound.

Theorem 4.2.5. *If Assumptions 4.2.1 and 4.2.3 are satisfied and $C(\cdot,\cdot) \in \mathbb{M}\mathrm{f}(Q, R)$, then*

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) \ge a(\omega^*) + c(-\omega^*). \tag{4.8}$$

Thus, combining the results in Theorems 4.2.3 and 4.2.5, we conclude that if the service process is of Markov-fluid type, the logarithmic asymptote (4.3) holds under Assumptions 4.2.1 and 4.2.3.

Remark 4.2.1. In this chapter we assume that the tagged customer (with service time B_0) and customers arriving into the system after time 0 (with generic service time B) have the same service requirement distribution. However, this assumption is not necessary as will become clear from our proofs. If the distributions of B_0 and B are different, the result still holds if just B_0 satisfies Assumption 4.2.3; it is not necessary that B satisfies this assumption.

Remark 4.2.2. Our results allow us to compare the performance of systems with varying service rate and with constant rate (where the mean service rate is the same in both systems). It is a quite typical phenomenon that performance improves if a random process is replaced by a deterministic process with the same mean.

Therefore, we now consider the GI/GI/1 PS system with fixed service rate c (recall that this is the mean service rate of the system considered in this chapter). Applying Jensen's inequality we obtain that

$$c(s) = \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sC(0,x)}] \geq \lim_{x\to\infty} \frac{1}{x} \log \mathrm{e}^{\mathbf{E}[sC(0,x)]} = \lim_{x\to\infty} \frac{1}{x} \mathbf{E}[sC(0,x)] = sc.$$

Hence,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) = \inf_{s\geq 0}(\alpha(s) + c(-s)) \geq \inf_{s\geq 0}(\alpha(s) - sc),$$

where the latter is the exponential decay rate in the system with the constant service rate c. If the function $c(-s)$ is strictly convex, it can be shown that the above inequality is strict. Thus, we conclude that, informally speaking, the random service rate increases the probability of a long sojourn time.

We now provide the proofs of the results presented above.

4.3 Proof of the upper bound

We start by proving the upper bound.

Proof of Theorem 4.2.2. The event $\{V > x\}$ implies that the queue does not empty before time x. Evidently, as we assume the system to be in steady state (with respect to the arrival process), the workload present at time 0, say W, can be identified with the FCFS waiting time. In other words, W has the representation $W = \sup_{t\geq 0}(A(-t,0) - C(-t,0))$. Hence, we can write

$$\begin{aligned} \mathbf{P}(V > x) &\leq \mathbf{P}(W + B_0 + A(0,x) - C(0,x) > 0) \\ &= \mathbf{P}\left(\sup_{t>0}(A(-t,0) - C(-t,0)) + B_0 + A(0,x) - C(0,x) > 0\right). \end{aligned} \tag{4.9}$$

Now note that the process $A(0,x)$ jumps at the arrival epochs and is constant in between, while we assumed the process $C(0,x)$ to be non-decreasing. Hence, the difference $A(0,x) - C(0,x)$ has positive jumps at arrival epochs and is non-increasing in between. Therefore, the supremum can only be attained at arrival epochs. This yields that expression in the right-hand side of (4.9) is equivalent to

$$\mathbf{P}\left(\sup_{n\in\mathbb{N}}\left(A(-S^A_{-n},0) - C(-S^A_{-n},0)\right) + B_0 + A(0,x) - C(0,x) > 0\right).$$

Remark that the quantities $A(-S^A_{-n},0)$ and $A(0,x)$ are independent. Now applying

the standard union bound, this expression is further bounded by

$$\begin{aligned}&\sum_{n=1}^{\infty}\mathbf{P}\left(A(-S^A_{-n},0)-C(-S^A_{-n},0)+B_0+A(0,x)-C(0,x)>0\right)\\&\quad=\quad\sum_{n=1}^{\infty}\mathbf{P}\left(A(-S^A_{-n},0)+B_0+A(0,x)-C(-S^A_{-n},x)>0\right),\end{aligned}$$

where we recall that $-S^A_{-n}$ denotes the time of the nth arrival in the past. Now we can apply the Chernoff bound to (each term in) the last expression, so that we arrive at

$$\begin{aligned}\mathbf{P}(V>x)\quad&\leq\quad\sum_{n=1}^{\infty}\mathbf{E}[\mathrm{e}^{\omega^*(A(-S^A_{-n},0)+B_0+A(0,x)-C(-S^A_{-n},x))}]\\&=\quad\sum_{n=1}^{\infty}\int_0^{\infty}\mathbf{E}[\mathrm{e}^{\omega^*(A(-S^A_{-n},0)+B_0+A(0,x)-C(S^A_{-n},x))}|S^A_{-n}=y]d\mathbf{P}(S^A_{-n}\leq y)\\&=\quad\sum_{n=1}^{\infty}\int_0^{\infty}(\mathbf{E}[\mathrm{e}^{\omega^*B}])^{n+1}\mathbf{E}[\mathrm{e}^{\omega^*A(0,x)}]\mathbf{E}[\mathrm{e}^{-\omega^*C(-y,x)}]d\mathbf{P}(S^A_{-n}\leq y),\end{aligned}$$

where in the last equality $A(-S^A_{-n},0)$ is interpreted as the sum of n service requirements. Now applying the definition of the cumulant function $c(\cdot)$, we obtain that for any $\varepsilon>0$ for x large enough the expression in the previous display is bounded from above by

$$\sum_{n=1}^{\infty}\int_0^{\infty}\left(\mathbf{E}[\mathrm{e}^{\omega^*B}]\right)^{n+1}\mathrm{e}^{(\alpha(\omega^*)+\varepsilon)x}\mathrm{e}^{(c(-\omega^*)+\varepsilon)(x+y)}d\mathbf{P}(S^A_{-n}\leq y).$$

Evaluating the integral and using the definition of S^A_{-n}, we see that the last expression equals

$$\begin{aligned}&\sum_{n=1}^{\infty}\left(\mathbf{E}[\mathrm{e}^{(c(-\omega^*)+\varepsilon)A}]\right)^n\mathrm{e}^{(\alpha(\omega^*)+c(-\omega^*)+2\varepsilon)x}(\mathbf{E}[\mathrm{e}^{\omega^*B}])^{n+1}\\&\quad=\quad\mathbf{E}[\mathrm{e}^{\omega^*B}]\mathrm{e}^{(\alpha(\omega^*)+c(-\omega^*)+2\varepsilon)x}\sum_{n=1}^{\infty}\left(\Phi_B(\omega^*)\Phi_A(c(-\omega^*+\varepsilon))\right)^n.\end{aligned}$$

Now observe that the summation over n does not depend on x; we therefore now verify whether this sum is finite. Note that (apply Lemma 4.1.1)

$$\alpha(\omega^*)+c(-\omega^*)=-\Phi_A^{\leftarrow}\left(\frac{1}{\Phi_B(\omega^*)}\right)+c(-\omega^*)<0.$$

Hence, due to continuity of the MGF s, we see that for ε small enough the product under the sum is less than one, and hence the geometric series is converging. Furthermore, $\mathbf{E}[\mathrm{e}^{\omega^*B}]<\infty$. Thus, we conclude that $\mathbf{P}(V>x)$ can be bounded from above by

$$\mathbf{P}(V>x)\leq M\mathrm{e}^{(\alpha(\omega^*)+c(-\omega^*)+2\varepsilon)x},$$

where $M < \infty$ is some positive constant. Taking logarithms, dividing by x, letting $x \to \infty$ and $\varepsilon \downarrow 0$, we obtain

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V > x) \leq \alpha(\omega^*) + c(-\omega^*).$$

This proves the upper bound. □

We now turn to the proof of Theorem 4.2.3. Let us first state the basic result for the workload distribution which is useful for our proof. Denote by $X(t)$ the state of the underlying Markov process at time t; $X(t) \in \{1, 2, \ldots, d\}$.

Proposition 4.3.1. *If $C(\cdot,\cdot) \in \mathbb{M}\mathrm{f}(Q, R)$ and Assumption 4.2.2 is satisfied, then there exists a constant $K > 0$ such that for any initial state of the service process $X(0) = i$, $i \in \{1, 2, \ldots, d\}$, uniformly in x it holds that*

$$\mathbf{P}\left(\sup_{t\geq 0} A(-t, 0) - C(-t, 0) > x | X(0) = i\right) \leq K\mathrm{e}^{-\nu^* x}. \tag{4.10}$$

Proof of Proposition 4.3.1. We present a proof that is based on a change-of-measure argument; there are several alternative approaches possible. This change of measure is such that the event $\{W > x\}$ becomes more likely than under the old measure. We introduce a process

$$T(x) := \inf\{t : A(-t, 0) - C(-t, 0) > x\}.$$

Then we can write

$$\mathbf{P}(W > x) = \mathbf{P}(T(x) < \infty).$$

Let us first twist the interarrival time and service requirement distributions. Define a new probability measure $\mathbf{P}_\omega$ for $\omega > 0$ such that

$$\mathbf{P}_\omega(A \in dx) = \mathbf{P}(A \in dx)\mathrm{e}^{-\alpha(\omega)x}/\Phi_A(-\alpha(\omega)),$$

$$\mathbf{P}_\omega(B \in dx) = \mathbf{P}(B \in dx)\mathrm{e}^{\omega x}/\Phi_B(\omega).$$

In order to construct the change of measure for the service process, let us first define the largest real eigenvalue of the matrix $Q + \omega R$, which coincides with $c(\omega)$, where the corresponding right eigenvector $(v_1, \ldots, v_d)^{\mathrm{T}}$ is component-wise positive, see Property 4.1.1(2). Note that the eigenvector also depends on ω, but for compactness we suppress this. With the new probability measure we associate the modified Markov process with transition matrix Q^* defined as (for $i \neq j$)

$$q^*_{ij} = q_{ij}v_j/v_i,$$

$$q^*_{ii} = q_{ii} + r_i\omega + c(-\omega).$$

It is not hard to verify that these rates indeed constitute a generator matrix (use that $c(\omega)$ is eigenvalue of $Q + \omega R$).

We have the following fundamental identity

$$\mathbf{P}(W > x) = \mathbf{E}_\omega[L_{T(x)}\mathbf{1}\{T(x) < \infty\}], \tag{4.11}$$

see e.g. Theorem XIII.3.2 in [7]; here $\mathbf{E}_\omega$ denotes the expectation under the new measure $\mathbf{P}_\omega$, and $L \equiv L_{T(x)}$ is the likelihood ratio process stopped at $T(x)$, which we specify below.

In this proof we take the parameter ω (the 'exponential twist') to be equal to ν^*. Suppose that in $[-T(x), 0)$ there were n arrivals; denote a_i, b_i, $i = 1, \ldots, n$, the interarrival times and corresponding service requirements. Also suppose that there were m transitions of the Markov process governing the service process; let, in the time interval $[-T(x), 0)$, the Markov process $X(\cdot)$ visit states $i_0, i_1, \ldots, i_m$. Define by t_{i_j}, $j = 1, \ldots, m$, the time which the service process spends in state i_j. Then, considering the likelihood ratio $L_{T(x)}$ stopped at time $T(x)$, we can write

$$\begin{aligned} L_{T(x)} &= \frac{v_{i_0}}{v_{i_m}} \times \left(\mathrm{e}^{\nu^* \sum_{j=1}^m r_{i_j} t_{i_j} + c(-\nu^*) \sum_{j=1}^m t_{i_j}}\right) \times \\ &\quad \left(\mathrm{e}^{\alpha(\nu^*) \sum_{i=1}^n a_i}\right) \times \left(\mathrm{e}^{-\nu^* \sum_{i=1}^n b_i}\right) \times (\Phi_A(-\alpha(\nu^*))\Phi_B(\nu^*))^n \,. \end{aligned}$$

As $-T(x)$ corresponds to an arrival epoch, we have that $\sum a_i = T(x)$, $\sum b_i = A(0, T(x))$. Also, recall from Lemma 4.1.1 that $\Phi_A(-\alpha(\nu^*))\Phi_B(\nu^*) = 1$. Recall the new measure was chosen so that the event $\{T(x) < \infty\}$ occurs with probability 1. We thus find

$$L_{T(x)} \leq \frac{v_{i_0}}{v_{i_m}} \times \left(\mathrm{e}^{-\nu^*(A(0,T(x)) - C(0,T(x)))}\right) \times \left(\mathrm{e}^{\alpha(\nu^*)T(x) + c(-\nu^*)\sum_{j=1}^m t_{i_j}}\right).$$

Taking into account that $\{\mathbf{1}\{T(x) < \infty\} = 1\}$ implies $A(-T(x), 0) - C(-T(x), 0) > x$, in conjunction with $\alpha(\nu^*) = -c(-\nu^*)$, we have identified a $K > 0$ such that

$$L_{T(x)}\mathbf{1}\{T(x) < \infty\} \leq K\mathrm{e}^{-\nu^* x}.$$

We conclude that the identity (4.11) implies that indeed $\mathbf{P}(W > x) \leq K\mathrm{e}^{-\nu^* x}$, irrespective of the value of $X(0) = i$. □

Proof of Theorem 4.2.3. Since the event $\{V > x\}$ implies that the queue does not empty before time x, we obtain by using the Chernoff bound

$$\begin{aligned} \mathbf{P}(V > x) &\leq \mathbf{P}(W + B_0 + A(0, x) - C(0, x) > 0) \leq \mathbf{E}[\mathrm{e}^{\omega^*(W + B_0 + A(0,x) - C(0,x))}] \\ &= \mathbf{E}[\mathbf{E}[\mathrm{e}^{\omega^*(W + B_0 + A(0,x) - C(0,x))}]|X(0)]. \end{aligned}$$

Conditioning on the state of the Markov process at time 0 provides the independence between the workload process and the arrival and service process after time 0. Therefore, the last expression in the previous display is equal to

$$\begin{aligned} &\mathbf{E}[\mathrm{e}^{\omega^* B_0}]\mathbf{E}\left[\mathbf{E}[\mathrm{e}^{\omega^* W}|X(0)]\mathbf{E}[\mathrm{e}^{\omega^*(A(0,x) - C(0,x))}|X(0)]\right] \\ &= \mathbf{E}[\mathrm{e}^{\omega^* B_0}] \sum_{i=1}^d \mathbf{E}[\mathrm{e}^{\omega^* W}|X(0) = i]\mathbf{E}[\mathrm{e}^{\omega^*(A(0,x) - C(0,x))}|X(0) = i]\pi_i, \end{aligned}$$

where we recall that π is the equilibrium distribution of $X(\cdot)$. Since $\alpha(s)+c(-s)$ equals zero at $s=0$, and has a strictly negative derivative at $s=0$, it follows that $\omega^* < \nu^*$. Then, Proposition 4.3.1 implies that there is a K_1 such that

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{\omega^* W}|X(0)] &= \int_0^\infty \mathbf{P}(\mathrm{e}^{\omega^* W} > x|X(0))dx = \int_0^\infty \mathbf{P}(W > (\log x)/\omega^*|X(0))dx \\ &\leq 1 + \int_1^\infty \mathbf{P}(W > (\log x)/\omega^*|X(0))dx \leq 1 + \int_1^\infty K_1 \mathrm{e}^{-(\nu^*/\omega^*)\log x}dx \\ &< 1 + K_1 \int_1^\infty x^{-\nu^*/\omega^*}dx =: K_2 < \infty.\end{aligned}$$

Consequently,

$$\mathbf{P}(V > x) \leq K_2 \cdot \mathbf{E}[\mathrm{e}^{\omega^* B_0}]\mathbf{E}[\mathrm{e}^{\omega^* A(0,x)}]\mathbf{E}[\mathrm{e}^{-\omega^* C(0,x)}]. \tag{4.12}$$

Note that due to Assumption 4.2.2, $\mathbf{E}[\mathrm{e}^{\omega^* B}] < \infty$. Since the process $A(0,x)$ is a compound Poisson process we have

$$\mathbf{E}[\mathrm{e}^{\omega^* A(0,x)}] = \mathrm{e}^{\alpha(\omega^*)x} = \mathrm{e}^{\lambda x(\Phi_B(\omega^*)-1)}.$$

Due to Property 4.1.1(3), there exists a $K_3 < \infty$ such that

$$\mathbf{E}[\mathrm{e}^{\omega^* C(0,x)}] \leq K_3 \mathrm{e}^{c(\omega^*)x}.$$

Combining this with (4.12), we have identified a $K > 0$ such that, uniformly in $x \geq 0$, $\mathbf{P}(V > x) \leq K\mathrm{e}^{(\alpha(\omega^*)+c(-\omega^*))x}$, where $\alpha(\omega^*) = \lambda(\Phi_B(\omega^*) - 1)$, as desired. □

4.4 Proof of the lower bound

We now proceed with proving the lower bound results.

Proof of Theorem 4.2.4. Our proof consists of five steps: (i) we truncate the service requirement distribution to find a lower bound on $\mathbf{P}(V > x)$ which, by virtue of Assumption 4.2.3, reduces the problem to finding a lower bound on a related busy-period problem for the system with truncated service requirements; (ii) next, we show that long busy periods are due to large deviations of both the arrival process and the service process; (iii) after that, we analyze the large deviations of the arrival process, and pay special attention to the technicality of dealing with the truncated service requirements; (iv) we then invoke the sp-LPD lower bound (Assumption 4.2.4) to analyze the large deviations of the service process; (v) finally, we combine all results to establish the stated.

Step (i). We truncate the service requirement distribution, by introducing a new stochastic process $A_k(0,x)$, $k > 0$, as follows:

$$A_k(0,x) := \sum_{i=1}^{N(x)} B_i \mathbf{1}\{B_i < k\}.$$

By definition of the PS queue with varying service capacity,

$$\mathbf{P}(V > x) = \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1+Q(u)} dC(0,u)\right),$$

where $Q(u)$ is the number of customers in the system at time u *excluding the tagged customer*.

If we have $A_k(0,u) - C(0,u) > \varepsilon u$, then also $A_k(0,u) > \varepsilon u$, and as all service requirements are at most of size k, we find a linear lower bound on the number of customers present at time u: $Q(u) \geq \varepsilon u/k$. We thus obtain

$$\begin{aligned}\mathbf{P}(V > x) &\geq \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1+Q(u)} dC(0,u),\ A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)\right) \\ &\geq \mathbf{P}\left(B_0 > \int_0^x \frac{1}{1+\varepsilon u/k} dC(0,u) \,\middle|\, A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)\right) \\ &\quad \times \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)).\end{aligned}$$

By applying integration by parts and standard calculus,

$$\begin{aligned}&\int_0^x \frac{1}{1+\varepsilon u/k} dC(0,u) = \frac{C(0,x)}{1+\varepsilon x/k} + \frac{\varepsilon}{k}\int_0^x C(0,u)\frac{1}{(1+\varepsilon u/k)^2} du \\ &\leq \frac{C(0,x)}{1+\varepsilon x/k} + \frac{\varepsilon}{k} r_{\max} \int_0^x \frac{u}{(1+\varepsilon u/k)^2} du \\ &\leq \frac{r_{\max} x}{1+\varepsilon x/k} + \frac{r_{\max} k}{\varepsilon}\left(\frac{1}{1+\varepsilon x/k} - 1 + \log\left(1+\frac{\varepsilon}{k}x\right)\right) = \frac{r_{\max} k}{\varepsilon}\log\left(1+\frac{\varepsilon}{k}x\right).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbf{P}(V > x) &\geq \mathbf{P}\left(B_0 > \frac{r_{\max} k}{\varepsilon}\log\left(1+\frac{\varepsilon}{k}x\right) \,\middle|\, A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)\right) \\ &\times \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)).\end{aligned} \tag{4.13}$$

Now observe that in the first probability in the right-hand side of the previous display, the value of B_0 does not depend on the condition, so that we finally arrive at the lower bound

$$\mathbf{P}\left(B_0 > \frac{k r_{\max}}{\varepsilon}\log\left(1+\frac{\varepsilon}{k}x\right)\right) \times \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\ u \in (0,x)). \tag{4.14}$$

Due to Assumption 4.2.3 we conclude that the first probability in (4.14) asymptotically behaves as $\mathrm{e}^{o(x)}$. Therefore, we are left with analyzing the second probability, which could be interpreted as the probability of a busy period exceeding x in the system with truncated service requirements and a service rate perturbed by ε.

Step (ii). We bound the second factor in (4.14) as follows:

$$\mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\ u \in [0,x]) \geq \mathbf{P}_1(x) \cdot \mathbf{P}_2(x);$$

here $\mathbf{P}_1(x) := \mathbf{P}(A_k(0,u) - bu > 0,\ u \in (0,x))$, $\mathbf{P}_2(x) := \mathbf{P}(C(0,u) < (b-\varepsilon)u,\ u \in (0,x))$, and $b < c$ is any fixed number. We have thus decomposed the probability of a long busy period into a large deviation of the arrival process and a large deviation of the service process; the intuitive explanation is that the occurrence of a long busy period is the result of both the arrival process generating traffic at a *higher* rate than usual and the service process offering service at a *lower* rate than usual. We emphasize that the value of b is free now, but in Step (v) we choose an appropriate value. We now deal with each of the probabilities separately; in Step (iii) we analyze $\mathbf{P}_1(x)$, and in Step (iv) $\mathbf{P}_2(x)$.

Step (iii). Consider $\mathbf{P}_1(x)$. Denote by P_k the busy period in the system with truncated service requirement (at threshold k) and constant service rate b. In [91] the asymptotics for large busy periods in this system were derived; it is readily checked that the corresponding conditions apply for truncated service requirements. We thus find

$$\begin{aligned}&\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}(A_k(0,u) - bu > 0,\ u \in (0,x)) \\ &\quad = \liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}(P_k > x) = \inf_{s\geq 0}(\alpha_k(s) - bs) = \gamma_b^k < 0,\end{aligned}$$

where

$$\alpha_k(s) := \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sA_k(0,x)}].$$

We now show that $\gamma_b^k \to \gamma_b := \inf_{s\geq 0}(\alpha(s) - bs)$ as $k \to \infty$. To this end, define $f_k(s) := \alpha_k(s) - bs$. Clearly, $f_k(s) \to f(s) = \alpha(s) - bs$ pointwise as $k \to \infty$ and $f_k(s)$ is increasing in k. Consequently, we have that the limit of γ_b^k for $k \to \infty$ exists and that

$$\gamma_b^* := \lim_{k\to\infty} \gamma_b^k = \lim_{k\to\infty} \inf_{s\geq 0} f_k(s) \leq \inf_{s\geq 0} f(s) = \gamma_b.$$

It remains to be shown that the reverse inequality holds. For this we follow an argument similar to the proof of Proposition 2.2 in Nuyens and Zwart [88].

Note that the function $f_k(\cdot)$ is continuous in s. Moreover, it is non-decreasing in k, and thus so is $\gamma_b^k \equiv \inf_{s\geq 0} f_k(s)$. Clearly, $\inf_{s\geq 0} f_k(s) \leq f_k(0) \leq f(0) = 0$, and hence $\gamma_b^* \equiv \lim_{k\to\infty} \inf_{s\geq 0} f_k(s) \leq 0$.

Now denote by B^k the service requirement truncated at k. Take k_0 such that $\mathbf{P}(B^k > bA) > 0$ for $k > k_0$. Then there exist δ, $\eta > 0$ such that $\mathbf{P}(B^k - bA \geq \delta) \geq \eta > 0$ for $k > k_0$. Hence, for $k > k_0$,

$$\Phi_{B^k}(s)\Phi_A(-bs) = \mathbf{E}[\mathrm{e}^{sB^k}]\mathbf{E}[\mathrm{e}^{-sbA}] = \mathbf{E}[\mathrm{e}^{s(B^k - bA)}] \geq \eta \mathrm{e}^{s\delta},$$

and consequently, for s large enough,

$$\Phi_A(-bs) \geq \frac{1}{\Phi_{B^k}(s)}.$$

Since $\Phi_A^{-1}(-s)$ is increasing in s, we find that for s and k large enough,

$$\alpha_k(s) - bs = -\Phi_A^{-1}\left(\frac{1}{\Phi_{B^k}(s)}\right) - bs \geq -\Phi_A^{-1}(\Phi_A(-bs)) - bs = 0,$$

and $\gamma_b^* > -\infty$. Therefore, the level sets $L_k = \{s \geq 0 : f_k(s) \leq \gamma_b^*\}$ are non-empty, compact sets that are nested with respect to k, which implies that there exists at least one point, say s_0, in their intersection. By definition of s_0, we have $f_k(s_0) \leq \gamma_b^*$ for every k. Since f_k converges pointwise, we find

$$\gamma_b = \inf_{s\geq 0} f(s) \leq f(s_0) = \lim_{k\to\infty} f_k(s_0) \leq \gamma_b^*.$$

Thus, we conclude that $\gamma_b^k \to \gamma_b$ as $k \to \infty$, and

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}_1(x) = \inf_{s\geq 0}(\alpha(s) - bs).$$

Step (iv). We now analyze the asymptotic behavior of $\mathbf{P}_2(x)$. First observe that we can rewrite $\mathbf{P}_2(x)$ as follows:

$$\begin{aligned}\mathbf{P}_2(x) &= \mathbf{P}(C(0,u) < (b-\varepsilon)u,\, u \in (0,x)) = \mathbf{P}(C(0,ux) < (b-\varepsilon)ux,\, u \in (0,1)) \\ &= \mathbf{P}\left(\frac{1}{x}\, C(0,ux) < (b-\varepsilon)u,\, u \in (0,1)\right) = \mathbf{P}\left(\frac{1}{x}\, C(0,\cdot\, x) \in S\right),\end{aligned}$$

where $S := \{f \in \Omega : f(u) < (b-\varepsilon)u,\, u \in (0,1)\}$. As we assumed that $C(0,\cdot\, x)/x$ obeys the lower bound of the sp-LDP (Assumption 4.2.4) we have

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}\left(\frac{1}{x}\, C(0,ux) \in S\right) \geq -\inf_{f\in S^o} I(f) =: -I^*, \qquad I(f) := \int_0^1 \Lambda(f'(t))dt,$$

where we recall that $\Lambda(t) = \sup_{s\in\mathbb{R}}(st - c(s))$. Since the infimum of $I(f)$ over all $f \in S^o$ is not larger than $I(f^*)$ for any particular $f^* \in S^o$, taking $f^*(u) := (b-\bar{\varepsilon})u$ with $\bar{\varepsilon} := \varepsilon(1+\delta)$ for some small $\delta > 0$, we obtain the lower bound

$$-I^* \geq -\sup_{s\in\mathbb{R}}((b-\bar{\varepsilon})s - c(s)).$$

Observe that since the constant b is chosen such that $b < c$, the supremum is attained for $s \leq 0$. Hence, we may write

$$\begin{aligned}\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}_2(x) &\geq -I^* \geq -\sup_{s\leq 0}((b-\bar{\varepsilon})s - c(s)) \\ &= -\sup_{s\geq 0}(-(b-\bar{\varepsilon})s - c(-s)) = \inf_{s\geq 0}((b-\bar{\varepsilon})s + c(-s)).\end{aligned}$$

Step (v). By combining the results for $\mathbf{P}_1(x)$ and $\mathbf{P}_2(x)$ we find that, for any $b < c$,

$$\begin{aligned}&\liminf_{x\to\infty} \frac{1}{x} \log \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x)) \\ &\qquad \geq \inf_{s\geq 0}(\alpha(s) - bs) + \inf_{s\geq 0}((b-\bar{\varepsilon})s + c(-s)). \qquad (4.15)\end{aligned}$$

Take $\varepsilon > 0$ sufficiently small and note that $\log \Phi_B(\cdot)$ is convex, $(\log \Phi_B(\cdot))' = \Phi'_B(\cdot)/\Phi_B(\cdot)$ is increasing and, due to Assumption 4.2.2, is finite and continuous in a neighborhood of ω^*. Similar arguments yield that $\Phi'_A(\cdot)/\Phi_A(\cdot)$ is an increasing, finite and continuous function as well, and $\alpha(\cdot)$ is continuous and increasing. Thus, there exists an $\varepsilon > 0$ for which there is $\omega = \omega_\varepsilon$ such that $\Phi_B(\omega_\varepsilon) < \infty$, $\Phi'_B(\omega_\varepsilon) < \infty$, and $\alpha'(\omega_\varepsilon) - c'(-\omega_\varepsilon) = \bar{\varepsilon}$. Since $\alpha(\cdot) + c(-\cdot)$ is a strictly convex function (this follows from the fact that $\alpha(\cdot)$ is strictly convex and $c(-\cdot)$ is convex), $\alpha'(\cdot) - c'(-\cdot)$ is increasing and hence, ω_ε is the unique solution. The continuity properties imply that $\lim_{\varepsilon \to 0} \omega_\varepsilon = \omega^*$.

Let us now take $b := \alpha'(\omega_\varepsilon)$ in (4.15). Note that this choice satisfies the requirement $b < c$: since the cumulant function $c(\cdot)$ is a convex function, its derivative is increasing, and consequently, for $\bar{\varepsilon}$ small, $b = c'(-\omega_\varepsilon) + \bar{\varepsilon} < c'(0) = c$.

Now consider the first optimization in (4.15): $\inf_{s \geq 0}(\alpha(s) - \alpha'(\omega_\varepsilon)s)$. It is readily checked that its first-order condition is $\alpha'(s) = \alpha'(\omega_\varepsilon)$, which is obviously met for $s = \omega_\varepsilon$ (and there is at most one solution, so ω_ε is the unique minimizer). The first-order condition for the second optimization in (4.15) is then $\alpha'(\omega_\varepsilon) - c'(-s) = \bar{\varepsilon}$, which is by definition solved for $s = \omega_\varepsilon$. We conclude that

$$\inf_{s \geq 0}(\alpha(s) - bs) + \inf_{s \geq 0}((b - \bar{\varepsilon})s + c(-s)) = \inf_{s \geq 0}(\alpha(s) + c(-s) - \bar{\varepsilon}s).$$

Now let $\varepsilon \to 0, \delta \to 0$ (and hence also $\bar{\varepsilon} \to 0$). Due to continuity we have that $\omega_\varepsilon \to \omega^*$, and consequently,

$$\liminf_{x \to \infty} \frac{1}{x} \log \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x)) \geq \alpha(\omega^*) + c(-\omega^*).$$

This completes the proof. □

Proof of Theorem 4.2.5. The proof strongly resembles that of Theorem 4.2.4. We leave it to the reader to check that only the argumentation in Step (iv) needs to be modified. This step relies on the validity of the lower bound of the sp-LDP, and to our best knowledge, an sp-LDP for the processes in $\mathbb{M}\mathrm{f}(Q,R)$ is not available from the literature. Therefore we need a different approach to analyze the large deviation $\mathbf{P}_2(x)$ of the service process $C(\cdot,\cdot)$. The main idea of this modification is to apply results of Chang [31] for Markov-type processes in discrete time. For that we need to cast our model into Chang's framework. This is done as follows.

Consider, as before, $\mathbf{P}_2(x) = \mathbf{P}(C(0,u) < (b-\epsilon)u, u \in (0,x))$. For any fixed $M < x$ and $C_M < (b-\varepsilon)M$,

$$\mathbf{P}_2(x) \geq \mathbf{P}(C(0,u) < (b-\epsilon)u, u \in (0,x), C(0,M) < C_M, X(M) = j),$$

as the event in the right-hand side is fully contained in that of the left-hand side. Now consider separately the intervals $(0,M]$ and (M,x). By using the conditional independence and a straightforward time-shift, we have that the previous probability is not smaller than

$$\mathbf{P}(C(0,u) < (b-\epsilon)u, u \in (0,M), C(0,M) < C_M, X(M) = j) \times \bar{\mathbf{P}}_2(x), \text{ where}$$

$$\bar{\mathbf{P}}_2(x) := \mathbf{P}(C(0,u) < (b-\epsilon)u + (b-\varepsilon)M - C_M, u \in (0, x-M) \mid X(0) = j).$$

Observe that the former probability is constant in x; therefore we need to concentrate just on $\bar{\mathbf{P}}_2(x)$. Now the fact that the service rate is bounded by $r_{\max}$ entails

$$C(0,u) \leq C\left(0, \left\lfloor \frac{u}{\delta} \right\rfloor \delta\right) + r_{\max}\delta,$$

for any δ. As a consequence, $\bar{\mathbf{P}}_2(x)$ majorizes

$$\mathbf{P}\left(C(0,i\delta) + r_{\max}\delta < (b-\epsilon)i\delta + (b-\varepsilon)M - C_M, i = 0, \ldots, \left\lceil \frac{x-M}{\delta} \right\rceil \mid X(0) = j\right).$$

Let us take $\delta < ((b-\varepsilon)M - C_M)/r_{\max}$. Then the probability in the previous display is not smaller than

$$\mathbf{P}\left(C(0,i\delta) \leq (b-\epsilon)i\delta, i = 0, \ldots, \left\lceil \frac{x-M}{\delta} \right\rceil \mid X(0) = j\right).$$

Now it can be verified that $C(0,i\delta)$ is a discrete-time process fitting in the framework of the sp-LDP of Chang [31]. Applying the sp-LDP lower bound on the last probability, it is straightforward to prove that the decay rate (in x) of the latter probability is indeed

$$-\sup_{s \geq 0}((b-\epsilon)s - c(s)),$$

as desired. Proceeding with Step (v) as before completes the proof. □

4.5 Extension to Discriminatory Processor Sharing

We now consider the extension of our analysis to the GI/GI/· queue with varying service rate operating under DPS. The proof indicates that essentially the same argumentation can be used as in the case of PS (as dealt with in the previous sections).

Suppose that there are M customer classes sharing the available capacity. The aggregate arrival process is assumed to be a renewal process as considered in Section 4.2. An arriving customer is of type k with probability $p_k, k = 1, \ldots, M$. All customers present in the system are served simultaneously with rates controlled by a vector of weights $(g_1, \ldots, g_M) > 0$. If there are Q_j customers of class j present in the system, $j = 1, \ldots, M$, each class-k customer is served at rate $g_k / \sum_{j=1}^{M} g_j Q_j$ (see also Subsection 1.2.2).

The service times B_n in Section 4.1 denote the unconditional service requirements (for our purposes, we do not need to specify the conditional service requirements distributions). Thus, the asymptotic cumulant generating function of the aggregate arrival process is still given by $\alpha(s)$.

The proofs of the previous section show that the logarithmic sojourn time asymptotics coincide with the logarithmic busy-period asymptotics. The following theorem

states that the same result holds in the DPS queue, regardless of the specific values of the weight factors.

Suppose the tagged customer belongs to class 1. Denote by V_1 its sojourn time, and B_0^1 its size.

Theorem 4.5.1. *If Assumptions 4.2.1, 4.2.3 and 4.2.4 are satisfied, then*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) = \inf_{s>0}(\alpha(s) + c(-s)).$$

Thus, the large-deviations estimate does not change when one assigns different weights to the various customer classes. This may not be surprising since we already obtained the insight that on a large-deviations scale, the behavior of the sojourn time resembles that of the busy period. The decay rate of the latter is obviously weight-independent (as the length of a busy period is the same for all work-conserving service disciplines, such as DPS).

On the one hand, this asymptotic insensitivity might be considered as a negative fact. It says that independent of the particular weights assignment, the DPS discipline does not reduce the likelihood of extremely long sojourn times. Long sojourn times are inevitable, since they are typically caused by the large amount of work brought by customers during the service of the tagged customer. On the other hand, the insensitivity property may be regarded as a positive result, because it implies that preferential treatment of classes with large weights does not carry the penalty of increasing the occurrence of long sojourn times for classes with smaller weights.

Proof of Theorem 4.5.1. The proof of the upper bound uses the same arguments as for the single-class PS queue, which we will not repeat here. The proof of the lower bound is similar to that of Theorem 4.2.4. We truncate the work process by accepting only customers with the service requirements of size smaller than k into the system and proceed in a similar fashion as before. The only extra step involves the minimal weight $g_{\min} = \min_{j=1,\ldots,M} g_j$,

$$\begin{aligned}
&\mathbf{P}(V_1 > x) \\
&\quad \geq \quad \mathbf{P}\left(B_0^1 > \int_0^x \frac{g_1 dC(0,u)}{1+\sum_{j=1}^M g_j Q_j(u)} \middle| A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x) \right) \\
&\qquad\quad \times \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x)) \\
&\quad \geq \quad \mathbf{P}\left(B_0^1 > \int_0^x \frac{g_1 dC(0,u)}{1+g_{\min}\sum_{j=1}^M Q_j(u)} \middle| A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x) \right) \\
&\qquad\quad \times \mathbf{P}(A_k(0,u) - C(0,u) > \varepsilon u,\, u \in (0,x)).
\end{aligned}$$

Since the service requirements are not larger than k, under the above condition, the total number of customers can be bounded from below in terms of the workload as

$$\sum_{j=1}^M Q_j(u) \geq \frac{\varepsilon u}{k}.$$

It is now straightforward to verify that the first probability behaves as $e^{\sigma(x)}$ when $x \to \infty$. The second probability gives the desired decay rate. For details see Theorem 4.2.4. □

Chapter 5

Fluid limits for bandwidth-sharing networks in overload

In Chapters 2-4 we evaluated the flow-level performance of elastic data transfers on a single bottleneck link by means of the classical processor-sharing model. In the remainder of the thesis we turn our attention to networks of processor-sharing links as considered by Massoulié and Roberts [84, 99]. Such bandwidth-sharing networks provide a natural extension for modeling the dynamic interaction among competing elastic flows that traverse several links along their source-destination paths. For a detailed description of bandwidth-sharing networks we refer to Section 1.2.3.

In the present chapter we focus on bandwidth-sharing networks where the load on one or several of the links exceeds the capacity. Obviously, with adequate provisioning, a network should not experience overload, or even approach overload, in normal operating conditions. However, even in a properly dimensioned system with a low typical load, the actual traffic volume may substantially fluctuate over time and exhibit transient surges, see also Bonald and Roberts [20]. Furthermore, an understanding of the overload behavior plays a crucial role in analyzing the performance in terms of long transfer delays or low flow throughputs as caused by large queue build-ups. The most likely way for such rare events to occur, commonly entails a scenario where the system temporarily appears to deviate from the normal stochastic laws and behaves as if it experiences overload, see for instance Anantharam [6].

As discussed earlier in Section 1.2.3, the intricate rate allocation and the non-work-conserving behavior of the network not only render the flow-level performance largely intractable, but also complicate the analysis of the overload behavior. For example, even on links with excess capacity, the workloads may grow because of the non-work-conserving behavior mentioned above. In addition, while the total number of flows must grow in overload conditions, the exact nature of the growth patterns of the various classes is far from clear, and may even potentially involve oscillatory effects in certain cases as observed in Bramson [28] and Lu and Kumar [80] for example.

In order to characterize the growth dynamics, we examine the fluid limit, which

emerges when the flow dynamics are scaled in both space and time. Gromoll and Williams [60] provided a characterization of the limit points of the scaled sequence in terms of measure-valued processes. We propose a related but slightly different fluid model, and derive a functional equation characterizing the fluid limit. We show that any strictly positive solution must be unique, which in particular implies the convergence of the scaled number of flows to the fluid limit for nonzero initial states when the load is sufficiently high. In addition, we establish the uniqueness of the fluid limit for tree networks. For the case of a zero initial state and zero-degree homogeneous rate allocation functions, we show that there exists a linear solution to the fluid-limit equation, and obtain a fixed-point equation for the corresponding asymptotic growth rates. It is proved that the solution to the fixed-point equation is also a solution to a related strictly concave optimization problem, and hence exists and is unique. Finally, we discuss extensions to models with user impatience, which has a particularly pronounced impact in overload conditions, see also Bonald and Roberts [20].

The remainder of the chapter is organized as follows. In Section 5.1 we present a detailed description of the network, the bandwidth-sharing strategy, and the flow dynamics. In Section 5.2 we introduce the fluid scaling, define the notion of a fluid-model solution, and prove that any limit point of the scaled sequence satisfies the fluid-model equation. In Section 5.3 we prove uniqueness of the strictly positive fluid-model solution. In Section 5.4 we focus on the case of a zero initial state. In Section 5.5 we prove the uniqueness of the fluid limit for a network with a tree topology. In Section 5.6 we study the fluid limits in a parking lot network. In Section 5.7 we focus on networks with a linear topology. Section 5.8 is devoted to the special case of a star network. In Section 5.9 we elaborate on the numerical experiments that we conducted to illustrate and support the analytical findings. We discuss extensions to models with user impatience in Section 5.10. Proofs of the results in Sections 5.2–5.5 are presented in Appendices 5.A–5.E.

5.1 Model description

In this section we present a detailed model description. For compatibility, we adhere to the notation used in [34, 60] to the extent possible.

Network model

We consider a bandwidth-sharing network as described in Section 1.2.3. The network consists of a finite number of links labeled by $j = 1, \ldots, J$ and is offered traffic from several classes indexed by $i = 1, \ldots, I$. Denote by $C = (C_1, \ldots, C_J)$ the vector of link capacities. Each class is characterized by a route, i.e., a nonempty set of links traversed by the traffic from that class. Let A be a $J \times I$ incidence matrix such that $A_{ji} = 1$ if link j belongs to the route of class i, and $A_{ji} = 0$ otherwise. For now, we do not make any specific assumptions on the topology of the network or the structure of the route sets.

Rate allocation policy

Denote by $x(z) = (x_1(z), \ldots, x_I(z))$ the vector of rates received by each individual flow of the various classes as function of the population of active flows $z = (z_1, \ldots, z_I)$. Denote by $\Lambda(z) = (\Lambda_1(z), \ldots, \Lambda_I(z))$, with $\Lambda_i(z) = z_i x_i(z)$, the vector of aggregate rates allocated to the various classes. A class-i flow that is continuously active throughout the time interval $[s,t]$, receives a cumulative amount of service

$$S_i(s,t) = \int_s^t x_i(Z(u))du,$$

with $Z(t) = (Z_1(t), \ldots, Z_I(t))$ representing the population of active flows at time t. We further introduce $T(t) = (T_1(t), \ldots, T_I(t))$, with $T_i(s,t) = T_i(t) - T_i(s)$ representing the aggregate cumulative amount of service received by class-i flows during the time interval $[s,t]$,

$$T_i(s,t) = \int_s^t \Lambda_i(Z(u))du.$$

The bandwidth sharing among competing flows is governed by a utility-maximizing strategy. Specifically, for a given population $z = (z_1, \ldots, z_I) \neq (0, \ldots, 0)$ of active flows, the rate allocation vector $x(z)$ is determined as the solution of the optimization problem (P) defined in Section 1.2.3. With the additional convention that $x_i(z) = 0$ when $z_i = 0$, the rate allocation is uniquely determined since the optimization problem is strictly concave.

The family of α-fair policies as described in Section 1.2.3 are the most commonly studied utility-maximizing policies. Recall that the per-flow rate allocation vector $x(z)$ for an α-fair policy is the solution of the optimization problem (cf. (1.2))

$$(P') \qquad \begin{array}{ll} \text{maximize} & \sum_{i=1}^I w_i z_i U_i(x_i) \\ \text{subject to} & A\Lambda \leq C,\ \Lambda \geq 0, \end{array}$$

with

$$U_i(x_i) = \begin{cases} \frac{x_i^{1-\alpha}}{1-\alpha}, & \alpha \in (0,\infty) \setminus \{1\}, \\ \log x_i, & \alpha = 1. \end{cases}$$

In the present chapter we consider rate allocation policies that satisfy the following assumption.

Assumption 5.1.1. *The utility functions $U_i(\cdot)$ are such that the per-class rate allocation vector $\Lambda(z) = z \cdot x(z)$ is (i) a continuous function of z on $\mathbb{R}^I_{++} = (0,\infty)^I$, and (ii) zero-degree homogeneous, i.e., $\Lambda(az) = \Lambda(z)$ for any scalar $a > 0$, $z \geq 0$.*

Kelly and Williams [71] established that α-fair utility functions satisfy property (i) of Assumption 5.1.1, and Chiang *et al.* [34] extended this property to the case when the parameter α varies among the different classes. Furthermore, in the present chapter we prove the Lipschitz continuity of the rate allocation function on

the set $\mathbb{R}^I_{++}$ for a large class of utility functions. A sufficient condition for property (ii) to be satisfied is that $U_i(\kappa x_i) = V(\kappa)U_i(x_i)$ for all $i = 1, \ldots, I$ and some function $V(\cdot)$. This is a natural property, implying that the relative utilities are scale-invariant, which is satisfied by α-fair utility functions as long as the parameter α is common to all classes.

Flow dynamics

The traffic of the various classes consists of elastic file transfers. A flow is a continuous transfer of a file through the links along the route associated with its class. The duration of the flow thus depends on its size and the simultaneous service rate it receives on all the links along its route.

Class-i flows arrive according to a delayed renewal process of rate λ_i. Let A_{ik} be the arrival epoch of the kth class-i flow. Define $E_i(t) = \max\{k : A_{ik} < t\}$ as the number of class-i flows that arrive during the time interval $(0, t]$. Let B_{ik} be the size of the kth class-i flow, i.e., the total amount of service required to complete the transfer. The random variables $B_{i1}, B_{i2}, \ldots$ are independent and identically distributed copies of a generic random variable B_i with mean $1/\mu_i$. Denote by $\rho = (\rho_1, \ldots, \rho_I)$ with $\rho_i := \lambda_i/\mu_i$ the vector of traffic intensities.

Let $\overline{B}_{il}$ be the residual size of the lth initial class-i flow at time 0. We assume that for each class i the initial number of flows and the initial workload are finite, i.e., $Z_i(0) < \infty$ and $\sum_{l=1}^{Z_i(0)} \overline{B}_{il} < \infty$.

The residual size at time t of the lth initial flow and the kth arriving flow of class i, assuming $A_{ik} \leq t$, are given by $\overline{B}_{il}(t) = \left(\overline{B}_{il} - S_i(0,t)\right)^+$ and $B_{ik}(t) = \left(B_{ik} - S_i(A_{ik}, t)\right)^+$, respectively.

The number of active class-i flows at time t may be related to the arrival times, service requirements, and received service amounts as:

$$\begin{aligned} Z_i(t) &= \sum_{l=1}^{Z_i(0)} \mathbf{1}\left(\overline{B}_{il}(t) > 0\right) + \sum_{k=1}^{E_i(t)} \mathbf{1}\left(B_{ik}(t) > 0\right) \\ &= \sum_{l=1}^{Z_i(0)} \mathbf{1}\left(\overline{B}_{il} > S_i(0,t)\right) + \sum_{k=1}^{E_i(t)} \mathbf{1}\left(B_{ik} > S_i(A_{ik}, t)\right). \end{aligned} \tag{5.1}$$

Likewise, the aggregate cumulative amount of service received by class-i flows during the time interval $[0, t]$ may be expressed as:

$$T_i(t) = \sum_{l=1}^{Z_i(0)} \min\left(\overline{B}_{il}, S_i(0,t)\right) + \sum_{k=1}^{E_i(t)} \min\left(B_{ik}, S_i(A_{ik}, t)\right). \tag{5.2}$$

Load conditions

In the present paper we focus on an overload scenario where the stability condition $A\rho < C$ is violated for at least one of the links. While the total number of flows must grow in such a scenario, the exact nature of the growth patterns of the various classes is not so evident. In the next sections we will examine the growth dynamics in terms of the fluid limit.

5.2 Fluid model

In this section we introduce the fluid limit, which arises from a scaled sequence of the stochastic processes $(Z(t), T(t); t \geq 0)$, and derive a characterization of the fluid limit in the form of a set of integral equations.

Following Gromoll and Williams [60], we consider a sequence of models as described in the previous section associated with a sequence $r \in \mathcal{R}$ of positive numbers increasing to infinity. The numbers $r \in \mathcal{R}$ are attached as superscripts to the corresponding model parameters and stochastic processes. Each model has the same arrival rates and network characteristics in terms of the vector of link capacities C and the incidence matrix A. The flow size distributions are allowed to vary with r but converge: $B_i^r \stackrel{d}{\to} B_i$ and $\mu_i^r \to \mu_i$ as $r \to \infty$. The scaled initial conditions are assumed to converge as well: $\overline{Z}^r(0) \to z(0)$ as $r \to \infty$. We consider the behavior of the system on the law-of-large-numbers scale, or fluid scale, and define the scaled processes

$$\begin{aligned} \overline{Z}^r(t) &= \tfrac{1}{r} Z^r(t), \\ \overline{S}^r(s,t) &= S^r(rs, rt), \\ \overline{T}^r(t) &= \tfrac{1}{r} T^r(rt). \end{aligned} \tag{5.3}$$

Since the rate allocation vector is a zero-degree homogeneous function, the scaled amount of service received by a class-i flow that is continuously active during the time interval $[s,t]$ is

$$\overline{S}_i^r(s,t) = \int_s^t \frac{\Lambda_i(\overline{Z}^r(u))}{\overline{Z}_i^r(u)} du.$$

Gromoll and Williams [60] established that the sequence of scaled processes $(\overline{Z}^r(t), \overline{T}^r(t); t \geq 0)$ is tight. Moreover, they provided a characterization of the limit points in terms of a so-called state descriptor, a Borel measure that contains information on the residual sizes of all active flows.

We propose a related but different fluid model, and derive a functional equation that is satisfied by the limit points of the scaled sequence $(\overline{Z}^r(t), \overline{T}^r(t); t \geq 0)$. In order to define the fluid model, we first introduce a slightly modified version of the rate allocation functions which may be interpreted as the service rates on the fluid scale. The service rate $R_i(z)$ received by class i is defined as follows: $R_i(z) \equiv \Lambda_i(z)$ if $z_i > 0$, where $\Lambda_i(z) = z_i x_i(z)$ and $x(z)$ is the solution of the optimization problem (P); and $R_i(z) \equiv \rho_i$ if $z_i = 0$. The above distinction reflects the fact that at the fluid scale, $z_i = 0$ requires that class i receives service at rate ρ_i, rather than 0.

We define

$$x_i^*(z) = \begin{cases} x_i(z), & \text{if } z_i > 0, \\ \infty, & \text{if } z_i = 0. \end{cases} \tag{5.4}$$

In the remainder of the chapter we consider

$$S_i(s,t) = \int_s^t x_i^*(z(u))du$$

as the cumulative amount of service received at the fluid scale by a class-i flow that is continuously active throughout the time interval $[s,t]$.

Further we introduce $\tau_i(s,t) = \tau_i(t) - \tau_i(s)$ representing the aggregate cumulative amount of service received by class i at the fluid scale during the time interval $[s,t]$. Gromoll and Williams [60] derived

$$\tau_i(s,t) = \int_s^t \left(\Lambda_i(z(u))\mathbf{1}(z_i(u) > 0) + \rho_i \mathbf{1}(z_i(u) = 0)\right) du = \int_s^t R_i(z(u))du.$$

Definition 5.2.1. *A nonnegative continuous function $z(\cdot)$ is a* fluid-model solution *if it satisfies the functional equation*

$$z_i(t) = z_i(0)\mathbf{P}\left(\overline{B}_i > S_i(0,t)\right) + \lambda_i \int_0^t \mathbf{P}\left(B_i > S_i(s,t)\right) ds. \tag{5.5}$$

Moreover,

$$\tau_i(t) = z_i(0)\mathbf{E}[\min(\overline{B}_i, S_i(0,t))] + \lambda_i \int_0^t \mathbf{E}[\min(B_i, S_i(s,t))]ds, \tag{5.6}$$

and

$$\sum_{i=1}^{I} A_{ji}\tau_i(s,t) \leq C_j(t-s) \tag{5.7}$$

for all $t \geq s \geq 0$.

Define $\mathcal{M}(C) = \{z \in \mathbb{R}_+^I : AR(z) \leq C\}$. An important implication of Inequality (5.7) is that a fluid-model solution $z(\cdot) \in \mathcal{M}(C)$ almost everywhere (see also [60]).

The main result of the present section is the following characterization of the limit points of the scaled sequence $(\overline{Z}^r(t), \overline{T}^r(t); t \geq 0)$.

Theorem 5.2.1. *The limit point of any convergent subsequence of $(\overline{Z}^r(t), \overline{T}^r(t); t \geq 0)$ is almost surely a solution of the fluid-model Equations (5.5)–(5.7).*

To prove the above theorem, we apply the fluid scaling to the set of Equations (5.1)–(5.2) satisfied by the pre-limit processes. Taking subsequently $r \to \infty$, we deduce Equations (5.5)–(5.6). Similarly, Inequality (5.7) follows from the fact that the pre-limit cumulative unused capacity $U(s,t) = C(t-s) - AT(s,t)$ is nonnegative for any $s < t$. The proof is presented in Appendix 5.A.

Remark 5.2.1. In case of exponential flow sizes, the fluid-model Equations (5.5)–(5.6) take a simpler form:

$$z_i(t) = z_i(0) + \lambda_i t - \mu_i \tau_i(t), \tag{5.8}$$

or equivalently, for almost every $t \geq 0$,

$$z_i'(t) = \lambda_i - \mu_i R_i(z(t)). \tag{5.9}$$

We refer to Kelly and Williams [71] for a detailed discussion of this model.

Remark 5.2.2. It is worth observing that Chiang *et al.* [34] considered a slightly different scaling, commonly referred to as 'large-capacity scaling', where the arrival rates of the various classes and link capacities are scaled by r. This may be interpreted as a slightly different way of scaling time, and yields the same fluid-limit equation, but has the advantage that the rate allocation vector is not required to be zero-degree homogeneous.

5.3 Uniqueness of fluid-model solutions

In this section we establish the uniqueness of the fluid-limit solution in two scenarios of interest.

In preparation for the proof of uniqueness, we first state two important auxiliary results. For fixed $0 < \delta < M$, define $\mathcal{Z} := \{z \in \mathbb{R}^I : \delta \leq z_i \leq M, i = 1, \ldots, I\} = (\delta, M)^I$.

Proposition 5.3.1. *Assume that the utility functions $U_i(\cdot)$ are twice differentiable on $\mathbb{R}^I_{++} = (0, \infty)^I$. Then the rate allocation vector $\Lambda(\cdot)$ is Lipschitz continuous on the set $\mathcal{Z}$.*

Proposition 5.3.2. *Assume that the utility functions $U_i(\cdot)$ are twice differentiable on $\mathbb{R}^I_{++}$. Then any fluid-limit solution that is strictly positive must be unique.*

The proofs of the above propositions are provided in Appendices 5.B and 5.C.

5.3.1 Per-class overload conditions

We first prove convergence of the scaled sequence in situations where each individual class is overloaded. Denote by $C_i^{\min} = \min\{C_j : A_{ij} = 1\}$ the minimum link capacity along the route of class i.

Theorem 5.3.1. *Assume that the utility functions $U_i(\cdot)$ are twice differentiable on $\mathbb{R}^I_{++}$. If $z(0) > 0$ and $\rho_i > C_i^{\min}$ for all $i = 1, \ldots, I$, then the scaled sequence $(\overline{Z}^r(t); t \geq 0)$ converges almost surely to a solution of the fluid-model Equation (5.5).*

Proof. Since the sequence $(\overline{Z}^r(t); t \geq 0)$ is tight, it suffices to show that any limit point is unique in order to establish the convergence. To the contrary, suppose that there are two different limit points, $z(t)$, $h(t)$, with $z(0) = h(0)$.

It is easily verified that the residual flow sizes $B_{ik}(t)$ and $\bar{B}_{il}(t)$ at time t are larger than the corresponding quantities in an isolated single-server PS system with service capacity $C_i^{\min}$ and class-i traffic only. In particular, the number $Z_i(t)$ of class-i flows at time t is larger than in the isolated PS system. The results in [64] imply that in

case $\rho_i > C_i^{\min}$ the latter number grows at a strictly positive rate. It follows that $z_i(t), h_i(t) \geq z_i(0) > 0$ for all $t \geq 0$. In addition, Theorem 5.2.1 shows that $z(t)$ and $h(t)$ are almost surely solutions of the fluid-model Equation (5.5). Proposition 5.3.2 then implies that $z(t) = h(t)$, contradicting the initial supposition. □

5.3.2 Fluid-model solution with permanent flows

Theorem 5.3.2. *Let the utility functions $U_i(\cdot)$ be twice differentiable on $\mathbb{R}^I_{++}$. The fluid-model equations*

$$z_i^\varepsilon(t) = \varepsilon_i + \lambda_i \int_0^t \mathbf{P}(B_i > S_i^\varepsilon(s,t))ds, \tag{5.10}$$

$$\tau_i^\varepsilon(t) = \int_0^t R_i(z^\varepsilon(u))du = \varepsilon_i S_i^\varepsilon(0,t) + \lambda_i \int_0^t \mathbf{E}[\min(B_i, S_i^\varepsilon(s,t))]ds, \tag{5.11}$$

have a unique solution $z^\varepsilon(t)$, $z^\varepsilon(t) \in \mathcal{M}(C)$.

Proof. The proof is based on the derivations in Appendix 5.C. The idea is to consider the time interval $[0, t']$, for some suitably chosen $t' > 0$, and then show that Equation (5.10) has a unique solution for all $t \in [0, t']$. Applying an induction argument we extend the proof to the entire time line.

In order to prove the existence of a solution of the above equations, we construct a mapping $\Psi : C_b^I[0,t'] \to C_b^I[0,t']$, for some fixed $t' > 0$:

$$\Psi_i^\varepsilon(z,t) = \varepsilon_i + \lambda_i \int_0^t \mathbf{P}(B_i > S_i^z(s,t))ds, \qquad i = 1, \ldots, I,\ t \in [0,t'].$$

We use superscript z to emphasize that $S^z(\cdot,\cdot)$ is determined for vector z. The set $C_b^I[0,t']$ is the set of continuous bounded functions on the interval $[0,t']$. We let $z^0(\cdot) = \varepsilon$. We recursively define the sequence of functions $z^n(\cdot) = \Psi^\varepsilon(z^{n-1},\cdot)$, for each $n \geq 1$. Observe that $z^n(\cdot)$ is a continuous function and $z^n(t) \in [\varepsilon, \varepsilon + \lambda t']$, for all $t \in [0,t']$.

Consider the distance between two successive functions. By definition,

$$||z^{n+1} - z^n|| = ||\Psi^\varepsilon(z^n,\cdot) - \Psi^\varepsilon(z^{n-1},\cdot)||,$$

where the norm is defined as $||f|| = \sup_{t\in[0,t'],i\in\mathcal{I}} |f_i(t)|$, $f \in C^I[0,t']$. Since $z^n(\cdot)$ is bounded away from zero, invoking Inequalities (5.64)–(5.66) in the proof of Proposition 5.3.2 (Appendix 5.C), we obtain that

$$||\Psi^\varepsilon(z^n,\cdot) - \Psi^\varepsilon(z^{n-1},\cdot)|| \leq \frac{1}{4}||z^n - z^{n-1}||,$$

if $t' = \min_{i\in\mathcal{I}} \left(\frac{c_i}{4\lambda_i\gamma_i}\right)$, where $c_i = \min_{t\in[0,t']} x_i(z(t))$ and γ_i is a Lipschitz constant of $x_i(\cdot)$ (see Appendix 5.B for an explicit expression). Thus, we derive for all $m > n$,

$$||z^m - z^n|| \leq 2\left(\frac{1}{4}\right)^n ||z^1 - z^0||.$$

The above inequality shows that z^n is a Cauchy sequence. Hence, by completeness of $C_b^I[0,t']$, z^n converges as $n \to \infty$.

Let $z^*(\cdot) = \lim_{n\to\infty} z^n(\cdot)$. We now show that the function $z^*(\cdot)$ is a solution of the equation $z^*(t) = \Psi^\varepsilon(z^*,t)$. By construction,

$$||z^{n+1} - \Psi^\varepsilon(z^*,\cdot)|| = ||\Psi^\varepsilon(z^n,\cdot) - \Psi^\varepsilon(z^*,\cdot)|| \leq \frac{1}{4}||z^n - z^*||.$$

This implies that $z^n \to \Psi^\varepsilon(z^*,\cdot)$ as $n \to \infty$. Since $z^n \to z^*$ and the limit is unique, we deduce $z^* = \Psi(z^*,\cdot)$. Thus, for any $t \in [0,t']$ there exists a solution of Equation (5.10). Uniqueness of the solution of the equation $z^* = \Psi^\varepsilon(z^*,\cdot)$, follows by the argument in Appendix 5.C.

Suppose now Equation (5.10) has a unique solution z^* on the time interval $[0,kt']$. The next step is to consider the interval $[kt',(k+1)t']$. We introduce a sequence of functions z^n, $n \geq 0$, such that $z^n(t) = z^*(t)$ if $t \leq kt'$, and $z^0(kt'+t) = z^*(kt')$, $z^n(kt'+t) = \Psi^\varepsilon(z^{n-1}, kt'+t)$ if $t \in [0,t']$. The mapping Ψ^ε can now be written for any $t \in [0,t']$ as

$$\begin{aligned}\Psi^\varepsilon(z,kt'+t) &= \varepsilon_i + \lambda_i \int_0^{kt'} \mathbf{P}\left(B_i > S_i^z(s,kt') + S_i^z(kt',kt'+t)\right) ds \\ &+ \lambda_i \int_{kt'}^{kt'+t} \mathbf{P}\left(B_i > S_i^z(s,kt'+t)\right) ds.\end{aligned}$$

Noting that $S^{z^n}(s,kt') = S^{z^m}(s,kt')$ for any m,n, and applying Inequalities (5.64)–(5.66) to each integral, we obtain

$$||\Psi^\varepsilon(z^n,\cdot) - \Psi^\varepsilon(z^{n-1},\cdot)|| \leq \frac{1}{2}||z^n - z^{n-1}||,$$

which by the above argument implies existence and uniqueness of the solution of Equation (5.10) on the interval $[0,(k+1)t']$ and hence, on the entire time line. Furthermore, since function z^* is a solution of Equation (5.10) and $z^* \geq \varepsilon$, it trivially follows by definition of the rate allocation that $z^* \in \mathcal{M}(C)$. □

5.4 Fluid-model solution with zero initial state

In this section our focus is on the case of a zero initial state. In this important special case, the fluid model equations admit a linear solution.

The next theorem states that there exists exactly one linear solution of the fluid-limit Equations (5.5)–(5.7).

Theorem 5.4.1. *Assume that $z(0) = 0$ and that $\Lambda(az) = \Lambda(z)$ for any scalar $a > 0$ and vector $z \in \mathbb{R}_+^I$. Then the fluid-limit Equations (5.5)–(5.7) admit a linear solution*

$$z(t) \equiv mt,$$

where the vector $m = (m_1, \ldots, m_I)$ *forms the unique solution in the set* $\mathcal{M}(C)$ *of the fixed-point equation*

$$R_i(m) = \rho_i \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)}B_i^*}\right], \qquad i = 1, \ldots, I, \tag{5.12}$$

and B_i^* *represents a residual class-i flow size, i.e., a random variable with density* $\mu_i \mathbf{P}(B_i > x)$ *and Laplace-Stieltjes Transform (LST)* $\mathbf{E}\left[\mathrm{e}^{-xB_i^*}\right] = \mu_i(1-\mathbf{E}\left[\mathrm{e}^{-xB_i}\right])/x$.

The above theorem holds for arbitrary network topologies and arbitrary flow size distributions. In a single-link scenario, i.e., $J = 1$, it reduces to known results for single-server processor-sharing type systems. In particular, in the single-class case, i.e., $I = 1$, we have, dropping the class index, $R(m) = 1$. The fixed-point Equation (5.12) specializes to

$$1 = \rho\, \mathbf{E}[\mathrm{e}^{-mB^*}],$$

which corresponds to the result in [64]. In the multi-class case, we have $R_i(m) = w_i m_i / \sum_{k=1}^I w_k m_k$, and Equation (5.12) takes the form

$$\frac{w_i m_i}{\sum_{k=1}^I w_k m_k} = \rho_i\, \mathbf{E}\left[\mathrm{e}^{-w_i^{-1}\sum_{k=1}^I w_k m_k B_i^*}\right], \qquad i = 1, \ldots, I,$$

which agrees with the fixed-point equation in [5] for overloaded discriminatory processor-sharing queues.

The remainder of the section is organized as follows. We first provide a heuristic interpretation of the fixed-point Equation (5.12). Next, we give the proof of Theorem 5.4.1. We then proceed to discuss some qualitative properties of the asymptotic growth rates.

5.4.1 Heuristic interpretation

The fixed-point Equation (5.12) may be heuristically derived in a similar way as explained by Jean-Marie [63]; we are not aware of an article where this derivation has been published.

Suppose that $\frac{Z(r)}{r} \to m$ (or equivalently, $\frac{Z^r(t)}{r} \to mt$) almost surely as $r \to \infty$ for some vector $m = (m_1, \ldots, m_I)$. Then, for large t, a class-i flow will receive service at a rate of approximately $\frac{R_i(mt)}{m_i t}$. Let a_i^n be the arrival epoch of the n-th class-i flow. Then the size B_i^n of that flow and its sojourn time V_i^n may be related as:

$$B_i^n = \int_{a_i^n}^{a_i^n + V_i^n} \frac{R_i(mu)}{m_i u} du.$$

Since the rate allocation function is zero-degree homogeneous, i.e., $R_i(mu) = R_i(m)$, it follows that

$$\frac{m_i}{R_i(m)} B_i^n = \int_{a_i^n}^{a_i^n + V_i^n} \frac{1}{u} du = \log(a_i^n + V_i^n) - \log a_i^n.$$

Taking the exponent on both sides, we obtain

$$V_i^n = a_i^n \left(\mathrm{e}^{\frac{m_i}{R_i(m)} B_i^n} - 1 \right).$$

The number of active class-i flows at time t may then be expressed as $Z_i(t) = \#\{n : a_i^n + V_i^n \geq t, a_i^n \leq t\} = \#\{n : t \geq a_i^n \geq t\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^n}\}$. Because $a_i^n \approx n/\lambda_i$ for large n, we have

$$Z_i(t) = \#\{n : t \geq n/\lambda_i \geq t\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^n}\} \approx \lambda_i t \left(1 - \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i}\right]\right).$$

Dividing both sides by t and letting t tend to infinity, we deduce

$$m_i = \lambda_i \left(1 - \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i}\right]\right) = \frac{\rho_i m_i}{R_i(m)} \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^*}\right], \tag{5.13}$$

which is equivalent to Equation (5.12).

Remark 5.4.1. In the case of exponential flow sizes, Equation (5.12) specializes to

$$m_i = \lambda_i - \mu_i R_i(m), \qquad i = 1, \ldots, I, \tag{5.14}$$

which makes sense, since $\mu_i R_i(m)$ is indeed the departure rate of class-i flows. This is also consistent with the convention $R_i(m) = \rho_i$ when $m_i = 0$.

5.4.2 Proof of Theorem 5.4.1

The statement of Theorem 5.4.1 follows from the next two lemmas.

Lemma 5.4.1. *If the rate allocation function is zero-degree homogeneous, then Equations (5.5)–(5.7) admit a linear solution given by*

$$z_i(t) = m_i t, \qquad i = 1, \ldots, I,$$

where m_i is a solution of Equation (5.12) in the set $\mathcal{M}(C)$.

Proof. Since the rate allocation policy is zero-degree homogeneous, the fact that $m \in \mathcal{M}(C)$, i.e. $\sum_{i=1}^{I} A_{ji} R_i(m) \leq C_j$, implies $z(t) = mt \in \mathcal{M}(C)$ for almost every $t \geq 0$. Consequently, Equation (5.7) is satisfied.

Suppose $z_i(t) = m_i t$, where m_i is some constant. Substituting this into Equa-

tion (5.5) and using the fact that $R(mu) = R(m)$, we obtain that

$$\begin{aligned}
z_i(t) &= m_i t \\
&= \lambda_i \int_0^t \mathbf{P}\left(B_i > \int_s^t \frac{R_i(m_i u)}{m_i u} du\right) ds \\
&= \lambda_i \int_0^t \mathbf{P}\left(B_i > \frac{R_i(m)}{m_i} \int_s^t \frac{1}{u} du\right) ds \\
&= \lambda_i \int_0^t \mathbf{P}\left(-B_i \frac{m_i}{R_i(m)} < \log \frac{s}{t}\right) ds \\
&= \lambda_i t \int_0^1 \mathbf{P}\left(-B_i \frac{m_i}{R_i(m)} < \log u\right) du \\
&= \lambda_i t \int_0^1 \mathbf{P}\left(\mathrm{e}^{-B_i \frac{m_i}{R_i(m)}} < u\right) du \\
&= \lambda_i t \left(1 - \mathbf{E}\left[\mathrm{e}^{-B_i \frac{m_i}{R_i(m)}}\right]\right) \\
&= \frac{m_i \rho_i t}{R_i(m)} \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^*}\right],
\end{aligned}$$

which yields that m is a solution of Equation (5.12). If we assume $z_i(t) = m_i t$, so that $R_i(z(t)) = R_i(mt)$, then substituting into Equation (5.6), and using the fact that $R(mu) = R(m)$, we obtain

$$\begin{aligned}
\tau_i(t) &= R_i(m) t \\
&= \lambda_i \int_0^t \mathbf{E}\left[\min\left(B_i, \int_s^t \frac{R_i(mu)}{m_i u} du\right)\right] ds \\
&= \lambda_i \int_0^t \mathbf{E}\left[\min\left(B_i, \frac{R_i(m)}{m_i} \int_s^t \frac{1}{u} du\right)\right] ds \\
&= \frac{\lambda_i R_i(m)}{m_i} \int_0^t \mathbf{E}\left[\min\left(\frac{m_i}{R_i(m)} B_i, -\log \frac{s}{t}\right)\right] ds \\
&= \frac{\lambda_i t R_i(m)}{m_i} \int_0^1 \mathbf{E}\left[\min\left(\frac{m_i}{R_i(m)} B_i, -\log u\right)\right] du \\
&= \frac{\lambda_i t R_i(m)}{m_i} \int_0^1 \int_0^\infty \mathbf{P}\left(\min\left(\frac{m_i}{R_i(m)} B_i, -\log u\right) > v\right) dv du \\
&= \frac{\lambda_i t R_i(m)}{m_i} \int_0^1 \int_0^{-\log u} \mathbf{P}\left(\frac{m_i}{R_i(m)} B_i > v\right) dv du \\
&= \rho_i t \int_0^1 \mathbf{P}\left(\frac{m_i}{R_i(m)} B_i^* < -\log u\right) du \\
&= \rho_i t \int_0^1 \mathbf{P}\left(\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^*} > u\right) du \\
&= \rho_i t \mathbf{E}\left[\mathrm{e}^{-\frac{m_i}{R_i(m)} B_i^*}\right],
\end{aligned}$$

which also yields that m is a solution of Equation (5.12). □

Lemma 5.4.2. *Equation (5.12) has a unique solution $m = (m_1, \ldots, m_I)$ in the set $\mathcal{M}(C)$.*

The proof of the above lemma is presented in Appendix 5.D. The idea of the proof is to show that the rate allocation vector $R(m)$ associated with a solution m in the set $\mathcal{M}(C)$ of the fixed-point Equation (5.12) is also a solution of a related strictly concave optimization problem, and hence exists and is unique. Uniqueness of $R(m)$ together with Equation (5.12) then implies uniqueness of the vector m. Besides proving uniqueness, the latter relationship also provides a way for actually computing the asymptotic growth rates m_i, since the vector $R(m)$ can be calculated by solving a concave programming problem

$$(Q) \qquad \begin{array}{ll} \text{maximize} & G(R) = \sum_{i=1}^{I} G_i(R_i) \\ \text{subject to} & AR \leq C,\ R \leq \rho,\ R \geq 0, \end{array} \tag{5.15}$$

where the function $G_i : [0, \rho_i] \to \mathbb{R}$ is determined by its derivative

$$G_i'(x) = U_i' \left(\frac{1}{\beta_i^{-1}\left(\frac{x}{\rho_i}\right)} \right),$$

and $\beta_i^{-1}(\cdot)$ is the inverse of the LST $\beta_i(y) = \mathbf{E}\left[\mathrm{e}^{-yB_i^*}\right]$. The details on the above construction can be found in Appendix 5.D.

We conclude this section with some observations about qualitative properties of the growth rates.

Remark 5.4.2. If the arrival rates and the flow sizes of all classes are scaled by common factors $K > 0$ and $1/K$, respectively, thus keeping the traffic intensities constant, then the asymptotic growth rates scale by K. This makes sense as the scaling simply amounts to a change of time scale.

Remark 5.4.3. Suppose we focus on a particular class and, dropping the class index, examine the impact of the variability of the flow size B on the asymptotic growth rate m for a fixed mean flow size $\mathbf{E}[B]$ and service rate R. It may be deduced from the fixed-point Equation (5.13) that the growth rate is non-increasing in the variability of the flow size in the sense of the LST ordering. A random variable X is said to be larger or more variable than a random variable Y in the LST ordering if $\mathbf{E}\left[\mathrm{e}^{-sX}\right] \geq \mathbf{E}\left[\mathrm{e}^{-sY}\right]$ for all $s \geq 0$. Note that the LST ordering is implied by the more common convex ordering, which provides a measure for the degree of variability of a distribution. This monotonicity property does not directly extend to a network setting where the service rate R depends on the growth rate m.

Remark 5.4.4. If $m_i = 0$, then the asymptotic growth rates m_j, $j \neq i$, are identical to those in a corresponding system with both class-i traffic and all links j with $A_{ji} > 0$ removed, which makes sense as none of these links are bottlenecks if $m_i = 0$.

Remark 5.4.5. Suppose we consider a sequence of systems where the arrival rate of a particular class i in the k-th system is scaled in such a manner that $\lim_{k\to\infty} \lambda_i^{(k)} = 0$, while the flow size may be scaled in an arbitrary way, so that it is not necessarily the case that $\lim_{k\to\infty} \rho_i^{(k)} = 0$. It may then be deduced that $\lim_{k\to\infty} m_i^{(k)} = 0$, and thus, in view of Remark 5.4.4, $m_j^{(k)} \to m_j$, $j \neq i$, with m_j representing the asymptotic growth rate of class j in a corresponding system with class i removed.

5.5 Uniqueness of the fluid-model solution for tree networks

We now proceed to show convergence of the scaled sequence in so-called tree networks. Tree networks are practically useful as a model for communication networks which exhibit a certain hierarchical structure such as access networks consisting of several multiplexing stages [17]. A tree network has a central link (usually referred to as the root of the tree) which belongs to all routes. The key property is that a tree can be decomposed into a set of subtrees, which represent a tree structure in itself. Figure 5.1 (a) shows one example of a tree topology.

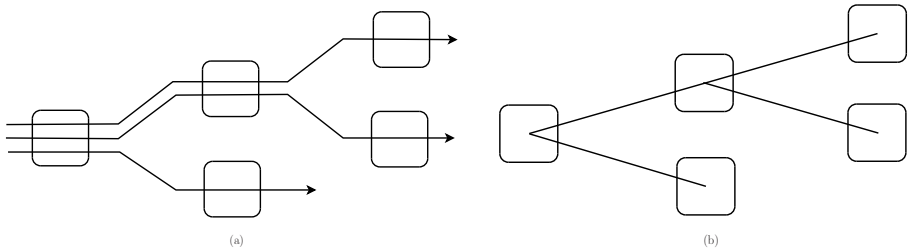

Figure 5.1: Example of tree network: (a) tree network (b) graph representation.

In the context of bandwidth-sharing networks, it will be convenient to define a tree network in terms of links and classes or routes, consisting of subsets of links, as represented by the incidence matrix. Note that in graph theory a tree is also defined as an acyclic network, but the network is described in terms of a set of vertices (or nodes) and a set of edges (or node pairs), rather than routes. The two notions may be formally related as follows. If we take a tree network in the graph-theoretic sense, pick an arbitrary vertex as root, and consider a collection of vertex paths with the root as common end point, then we obtain a tree network in our setting, with what we refer to as links somewhat confusingly corresponding to the vertices rather than the edges of the graph. See Figure 5.1 (b) for a graph representation of a tree network depicted in Figure 5.1 (a).

We build upon the following representation of a tree network.

Definition 5.5.1. *A bandwidth-sharing network with J links and I traffic classes has a tree topology if its incidence matrix A can be represented in the following manner:*

1. *if* $J = 1$, A *is a* $1 \times I$ *unit vector,*

$$A = (1, 1, \ldots, 1),$$

2. *if* $J > 1$, *there exists an integer* $m > 0$ *such that*

$$A = \begin{pmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 \\ A^{(1)} & & & & & & \\ & A^{(2)} & & & & & \\ & & \cdot & & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & A^{(m)} & B \end{pmatrix} \tag{5.16}$$

where B *is a* $(J-1) \times N$ *zero matrix,* $N \geq 0$, *and each* $A^{(k)}$ *is an incidence matrix of a tree network.*

This representation reflects that the network contains a single link which belongs to the routes of all classes and which is connected to m (second-level) links which in their turn constitute m disjoint subtrees. It may also contain N classes which traverse the root link only.

The incidence matrix corresponding to the network presented in Figure 5.1 can be written as

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{5.17}$$

where the subtree matrices are given by

$$A^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \text{ and } A^{(2)} = (1). \tag{5.18}$$

Tree networks have a few useful properties.

Proposition 5.5.1. *Tree networks operating under a weighted* α*-fair policy are monotone, i.e. if* $z \leq \hat{z}$, $z_i > 0$, *then* $x_i(z) \geq x_i(\hat{z})$.

In plain words, a network is monotone if adding a flow of any class reduces the rates allocated to all flows, or equivalently, removing a flow increases the rates of all flows. Whether or not the monotonicity property is satisfied depends on both the network topology and the rate allocation policy. Observe that if the network is monotone in the above sense, the per-flow rate allocation on a fluid scale satisfies a similar ordering property. For any $z \leq \hat{z}$, $x^*(z) \geq x^*(\hat{z})$. For all positive components of the vector z the inequality follows from monotonicity. For zero components the inequality holds trivially, since $x_i^*(z) = \infty$ if $z_i = 0$.

Proposition 5.5.2. *Tree networks are rate-preserving, i.e., if* $z \leq \hat{z}$, $z, \hat{z} \in \mathcal{M}(C)$, *then* $\sum_{i=1}^{I} R_i(z) \leq \sum_{i=1}^{I} R_i(\hat{z})$.

In a rate-preserving network adding a flow of any class increases the aggregate rate, or equivalently, removing a flow reduces the aggregate rate. Note that the rate preservation property trivially holds in (work-conserving) single-node systems.

The proofs of the above propositions are presented in Appendix 5.E.

Our goal is to prove uniqueness of fluid-model solutions for tree networks with a zero initial state. This in conjunction with the tightness result of [60] would imply the convergence of the scaled sequence. We need to make an additional assumption on the flow size distribution.

Assumption 5.5.1. *The flow size distribution has a bounded hazard rate, i.e., there exists an* $M \in (0, \infty)$ *such that*

$$\frac{f_{B_i}(x)}{\mathbf{P}(B_i > x)} < M, \qquad \forall x \geq 0. \tag{5.19}$$

This assumption is satisfied by a large class of distributions, including phase-type, log-normal, Pareto distributions, etc.

The main result of this section is the following theorem.

Theorem 5.5.1. *Consider a tree network. Let* $m = (m_1, \ldots, m_I) \in \mathcal{M}(C)$ *be the unique solution of the fixed-point Equation (5.12). Assume* $m_i > 0$ *for all* $i = 1, \ldots, I$. *Suppose* $\overline{Z}^r(0) \to 0$ *as* $r \to \infty$. *Then the scaled sequence* $(\overline{Z}^r(t); t \geq 0)$ *converges almost surely to the unique solution*

$$z(t) = mt$$

of the fluid-model equations

$$z_i(t) = \lambda_i \int_0^t \mathbf{P}(B_i > S_i(s,t))ds \tag{5.20}$$

and

$$\tau_i(t) = \int_0^t R_i(z(u))du = \lambda_i \int_0^t \mathbf{E}[\min(B_i, S_i(s,t))]ds. \tag{5.21}$$

In preparation for the proof of the above theorem, we first state two important auxiliary results.

Lemma 5.5.1. *Consider a monotone network. Assume the flow size distribution has a bounded density. Let* $z(t)$, $z(t) \in \mathcal{M}(C)$ *a.e., satisfy Equations (5.20)–(5.21) and let* $z^{\varepsilon}(t)$ *be a solution of the fluid-model Equations (5.10)–(5.11). Then, for all* $t \geq 0$,

$$z(t) \leq z^{\varepsilon}(t). \tag{5.22}$$

Proof. Let us construct a sequence of functions $z^{\varepsilon,n}(\cdot)$, $n > 0$, in the following manner. Let $z^{\varepsilon,0}(t) = z(t)$ be a solution of Equation (5.20). Lemma 5.4.1 implies that there exists at least one such solution. For a fixed $t \geq 0$, introduce the function $\Psi^{\varepsilon}(\cdot, t) :\to [\varepsilon, \infty)$,

$$\Psi_i^{\varepsilon}(z,t) = \varepsilon_i + \lambda_i \int_0^t \mathbf{P}\left(B_i > S_i(s,t)\right) ds,$$

where $S_i(s,t) = \int_0^t x_i^*(z(u))du$. Define

$$z^{\varepsilon,n}(t) = \Psi^{\varepsilon}(z^{\varepsilon,n-1}, t), \qquad n \geq 1.$$

We show that the sequence is non-decreasing by induction. Clearly,

$$z_i^{\varepsilon,1}(t) = \Psi_i^{\varepsilon}(z,t) = \varepsilon_i + z_i(t) \geq z_i^{\varepsilon,0}(t).$$

Suppose $z^{\varepsilon,n}(t) \geq z^{\varepsilon,n-1}(t)$. Since the network is monotone, this implies $x^*(z^{\varepsilon,n}(t)) \leq x^*(z^{\varepsilon,n-1}(t))$. Hence,

$$z_i^{\varepsilon,n+1}(t) = \varepsilon_i + \lambda_i \int_0^t \mathbf{P}\left(B_i > \int_s^t x_i^*(z^{\varepsilon,n}(u))du\right) ds$$

$$\geq \varepsilon_i + \lambda_i \int_0^t \mathbf{P}\left(B_i > \int_s^t x_i^*(z^{\varepsilon,n-1}(u))du\right) ds = z^{\varepsilon,n}(t).$$

Since $z^{\varepsilon,n}(t)$ is non-decreasing in n and bounded from above on any finite time interval, there exists a function $z^{\varepsilon,*}(t)$ such that

$$\lim_{n\to\infty} z^{\varepsilon,n}(t) = z^{\varepsilon,*}(t).$$

Let us now show that the function $z^{\varepsilon,*}(t)$ is continuous in t. Fix $h > 0$. Then we have

$$|z^{\varepsilon,*}(t+h) - z^{\varepsilon,*}(t)| = \lim_{n\to\infty} |z^{\varepsilon,n}(t+h) - z^{\varepsilon,n}(t)|$$

$$\leq \lambda_i \int_t^{t+h} \mathbf{P}(B_i > S_i(s,t+h))ds + \lambda_i \int_0^t \mathbf{P}(S_i(s,t) < B_i < S_i(s,t+h))ds. \quad (5.23)$$

The first term is bounded from above by $\lambda_i h$. Consider now the second term. Since the flow size distribution has a bounded density, there exists an $M \in (0,\infty)$ such that $f_{B_i}(u) \leq M$, for all $u \geq 0$, $i = 1, \ldots, I$. Hence,

$$\int_0^t \mathbf{P}(S_i(s,t) < B_i < S_i(s,t+h))ds = \int_0^t \int_{S_i(s,t)}^{S_i(s,t+h)} f_{B_i}(u)duds$$

$$\leq M \int_0^t (S_i(s,t+h) - S_i(s,t))ds = MS_i(t,t+h)t.$$

From $z_i^{\varepsilon,*}(t) \geq \varepsilon_i$, by monotonicity we derive $x_i^*(z^{\varepsilon,*}(t)) \leq \frac{C_i^{min}}{\varepsilon_i}$, $C_i^{min} = \min\{C_j : A_{ji} > 0\}$. Consequently,

$$\int_0^t \mathbf{P}(S_i(s,t) < B_i < S_i(s,t+h))ds \leq \frac{MC_i^{min}}{\varepsilon_i} ht.$$

Thus, as $h \to 0$, the right-hand side of (5.23) tends to zero, yielding continuity of the function $z^{\varepsilon,*}(t)$.

We now show that $z^{\varepsilon,*}(t)$ satisfies Equations (5.10)–(5.11). Since the sequence $z^{\varepsilon,n}(t)$ is bounded away from zero and is non-decreasing in n, $x^*(z^{\varepsilon,n}(t))$ is continuous and non-increasing in n. Then, $S_i^{\varepsilon,n}(s,t) \to S_i^{\varepsilon,*}(s,t)$ by monotone convergence. Hence, $\Psi^\varepsilon(z^{\varepsilon,n},t) \to \Psi^\varepsilon(z^{\varepsilon,*},t)$, implying

$$z^{\varepsilon,*}(t) = \Psi^\varepsilon(z^{\varepsilon,*},t). \tag{5.24}$$

Now since the function $z^{\varepsilon,*}(t)$ is continuous and satisfies the fluid-model Equation (5.10), Theorem 5.3.2 yields $z^{\varepsilon,*}(t) \equiv z^\varepsilon(t)$, a unique solution of Equation (5.10). Since the sequence $z^{\varepsilon,n}(t)$ is non-decreasing and for all n, $z^{\varepsilon,n}(t) \geq z(t)$, we deduce that for all $t > 0$, $z(t) \leq z^\varepsilon(t)$. □

Lemma 5.5.2. *Consider a tree network. Let B_i satisfy Assumption 5.5.1. Let $z(t)$, $z(t) \in \mathcal{M}(C)$ a.e., satisfy Equations (5.20)–(5.21) and let $z^\varepsilon(t)$ be a solution to the fluid-model Equations (5.10)–(5.11). Let $m = (m_1, \ldots, m_I) \in \mathcal{M}(C)$ be the unique solution of Equation (5.12). Assume $m_i > 0$ for all $i = 1, \ldots, I$. Then, for all $t \geq 0$, $i = 1, \ldots, I$,*

$$z_i^\varepsilon(t) - z_i(t) \leq \kappa(\varepsilon, t), \tag{5.25}$$

where

$$\kappa(\varepsilon,t) = K \sum_{i=1}^I \left(\frac{\varepsilon_i}{m_i} \left(1 + \max\left(\log\left(\frac{m_i t}{\varepsilon_i} \right), 0 \right) \right) \right), \tag{5.26}$$

for some constant $K \in (0, \infty)$.

Proof. Since the tree network is monotone (Proposition 5.5.1), Lemma 5.5.1 implies

$$z_i(t) \leq z_i^\varepsilon(t), \qquad t \geq 0.$$

This in particular yields that $\sum_{i=1}^I R_i(z(t)) \leq \sum_{i=1}^I R_i(z^\varepsilon(t))$ for almost every $t \geq 0$ because the network is rate-preserving by Proposition 5.5.2. Hence,

$$\sum_{i=1}^I \tau_i(t) \leq \sum_{i=1}^I \tau_i^\varepsilon(t) \tag{5.27}$$

for all $t \geq 0$.

Monotonicity yields $x^*(z(t)) \geq x^*(z^\varepsilon(t))$ and consequently, $S_i(s,t) \geq S_i^\varepsilon(s,t)$ for all $s \in [0,t]$, $t \geq 0$. For compactness, denote $\psi_i(t) = \int_0^t \mathbf{P}\left(S_i^\varepsilon(s,t) < B_i < S_i(s,t)\right) ds$. The fluid-limit equations yield

$$z_i^\varepsilon(t) - z_i(t) = \varepsilon_i + \lambda_i \int_0^t \mathbf{P}\left(S_i^\varepsilon(s,t) < B_i < S_i(s,t)\right) ds = \varepsilon_i + \lambda_i \psi_i(t), \tag{5.28}$$

and

$$\tau_i(t) - \tau_i^\varepsilon(t) = -\varepsilon_i S_i^\varepsilon(0,t) + \lambda_i \int_0^t \left(\mathbf{E}[\min(B_i, S_i(s,t))] - \mathbf{E}[\min(B_i, S_i^\varepsilon(s,t))]\right) ds. \tag{5.29}$$

Consequently, due to Assumption 5.5.1,

$$\mathbf{E}[\min(\overline{B}_i, S_i(s,t))] - \mathbf{E}[\min(\overline{B}_i, S_i^\varepsilon(s,t))] = \int_{S_i^\varepsilon(s,t)}^{S_i(s,t)} \mathbf{P}(B_i > x) dx$$

$$\geq \frac{1}{M} \int_{S_i^\varepsilon(s,t)}^{S_i(s,t)} f_{B_i}(x) dx = \frac{1}{M} \mathbf{P}\left(S_i^\varepsilon(s,t) < B_i < S_i(s,t)\right).$$

Hence,

$$\tau_i(t) - \tau_i^\varepsilon(t) \geq -\varepsilon_i S_i^\varepsilon(0,t) + \frac{\lambda_i}{M} \psi_i(t).$$

Let us now consider the term $\varepsilon_i S_i^\varepsilon(0,t)$ in more detail,

$$\varepsilon_i S_i^\varepsilon(0,t) = \varepsilon_i \int_0^t x_i^*(z^\varepsilon(u)) du.$$

By Lemma 5.5.1, $z^\varepsilon(t)$ is bounded from below by any solution $z(t)$ of Equations (5.20)–(5.21), and in particular, by $z(t) = mt$, where $m > 0$ is a solution of Equation (5.12). Moreover, $z_i^\varepsilon(t) \geq \varepsilon_i$. Thus, using the fact that $z_i^\varepsilon(t) \geq \max(m_i t, \varepsilon_i)$, $m_i > 0$, we derive

$$\begin{aligned} \varepsilon_i S_i^\varepsilon(0,t) &\leq \varepsilon_i \int_0^t \frac{C_i^{min}}{\max(m_i u, \varepsilon_i)} du = \varepsilon_i C_i^{min} \left(\int_0^{\frac{\varepsilon_i}{m_i}} \frac{1}{\varepsilon_i} du + \int_{\frac{\varepsilon_i}{m_i}}^t \frac{1}{m_i u} du \right) \\ &= \frac{\varepsilon_i C_i^{min}}{m_i} \left(1 + \log\left(\frac{m_i t}{\varepsilon_i}\right)\right), \end{aligned}$$

if $t > \frac{\varepsilon_i}{m_i}$, and $\varepsilon_i S_i^\varepsilon(0,t) \leq C_i^{min} t \leq C_i^{min} \frac{\varepsilon_i}{m_i}$, otherwise. Invoking (5.27), we obtain

$$\sum_{i=1}^I (\tau_i(t) - \tau_i^\varepsilon(t)) \geq \sum_{i=1}^I \left(-\frac{\varepsilon_i C_i^{min}}{m_i} \left(1 + \max\left(\log\left(\frac{m_i t}{\varepsilon_i}\right), 0\right)\right) + \frac{\lambda_i}{M} \psi_i(t) \right),$$

$$\sum_{i=1}^I \frac{\lambda_i}{M} \psi_i(t) \leq \sum_{i=1}^I \left(\frac{\varepsilon_i C_i^{min}}{m_i} \left(1 + \max\left(\log\left(\frac{m_i t}{\varepsilon_i}\right), 0\right)\right)\right),$$

and hence, for all $i = 1, \ldots, I$,

$$\frac{\lambda_i}{M} \psi_i(t) \leq \sum_{i=1}^I \left(\frac{\varepsilon_i C_i^{min}}{m_i} \left(1 + \max\left(\log\left(\frac{m_i t}{\varepsilon_i}\right), 0\right)\right)\right) := \hat{\kappa}(\varepsilon, t).$$

Substituting this into (5.28), we find

$$z_i^\varepsilon(t) - z_i(t) = \varepsilon_i + \lambda_i \psi_i(t) \leq \varepsilon_i + M\hat{\kappa}(\varepsilon, t).$$

□

We are now in a position to prove Theorem 5.5.1.

Proof of Theorem 5.5.1. Since the sequence $(\overline{Z}^r(t); t \geq 0)$ is tight, and Theorem 5.2.1 shows that any limit point is almost surely a solution of the fluid-limit Equations (5.20)–(5.21), it suffices to show the latter equation has a unique solution. To the contrary, suppose that there are two different fluid-limit solutions $z(t)$ and $h(t)$, with $z(0) = h(0) = 0$. Then there exist i, t and $\delta > 0$ such that $|z_i(t) - h_i(t)| > \delta$. Because of the symmetry, we may assume $z_i(t) - h_i(t) > \delta$.

Let $z_i^\varepsilon(t)$ be the solution of the fluid-model Equations (5.10)–(5.11). Lemma 5.5.1 implies $z(t), h(t) \leq z^\varepsilon(t)$ for all $t \geq 0$. Moreover, it follows from Lemma 5.5.2 that

$$z_i(t), h_i(t) \geq z_i^\varepsilon(t) - \kappa(\varepsilon, t).$$

Thus, we derive

$$z_i(t) - h_i(t) = z_i(t) - z_i^\varepsilon(t) + z_i^\varepsilon(t) - h_i(t) \leq \kappa(\varepsilon, t),$$

with $\kappa(\varepsilon, t)$ as in (5.26). Note that $\kappa(\varepsilon, t)$ tends to zero when $\varepsilon \to 0$. Taking $\varepsilon > 0$ sufficiently small so that $\kappa(\varepsilon, t) < \delta$ then yields a contradiction. □

Remark 5.5.1. In the present chapter we established uniqueness of the fluid limit for networks with a tree topology operating under a weighted α-fair policy. However, our derivations show that under the assumptions of Theorem 5.5.1 the fluid limit is unique for any monotone and rate-preserving bandwidth-sharing network.

5.6 Fluid limits in the two-link parking lot

In the present section our focus is on a so-called parking lot network. We consider a two-link network with link capacities $c_1 = 1$, $c_2 = c < 1$. Class-1 flows require service from link 1 only, while class-2 flows demand capacity on both links simultaneously. See Figure 5.2 for an illustration.

The name of the network topology is motivated by parking lots which consist of several parking areas connected by a single exit route [103]. The visitors with the cars parked in the first lot only need to traverse one segment of the exit link, the visitors parked in the second parking lot need to traverse two segments, etc.

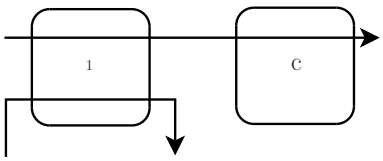

Figure 5.2: 2-link parking lot network.

This network presents one particular example of a tree topology. The rate allocations under any unweighted α-fair policy (Appendix 5.E) are given by

$$\begin{aligned} \Lambda_1(z) &= \max\left(1-c, \tfrac{z_1}{z_1+z_2}\right), \quad z_1 > 0, \\ \Lambda_2(z) &= \min\left(c, \tfrac{z_2}{z_1+z_2}\right). \end{aligned} \tag{5.30}$$

In this section we assume that the load on the root link exceeds the capacity and the load of class 1 exceeds the minimum guaranteed service rate, i.e.,

$$\rho_1 + \rho_2 > 1, \qquad \rho_1 > 1 - c. \tag{5.31}$$

Our interest in these specific load conditions is related to the large-deviations analysis of the parking lot network in Chapter 6. In order to derive the logarithmic asymptotics for the sojourn time, we perform a change of measure which induces overload of the network as described above.

Proposition 5.6.1. *Under Assumption (5.31),* $\mathcal{M}(C) = \{z \in \mathbb{R}_+^2 : R_1(z)+R_2(z) \leq 1,\ R_2(z) \leq c\} \subset (0,\infty)^2$.

Proof. Suppose to the contrary that there exists $z \geq 0$ in $\mathcal{M}(C)$, such that $z_i = 0$ for some $i \in \{1,2\}$. We consider three cases.

First let $z_1 = 0, z_2 = 0$. By definition of $R(z)$ we have $R_1(z) = \rho_1,\ R_2(z) = \rho_2$. Since $\rho_1 + \rho_2 > 1$, $(0,0) \notin \mathcal{M}(C)$.

Suppose now $z_1 > 0,\ z_2 = 0$. In this case, $R_1 = 1,\ R_2 = \rho_2$, implying $(z_1, 0) \notin \mathcal{M}(C)$. Observe that $(z_1, 0) \notin \mathcal{M}(C)$ under any load conditions.

The remaining case is $z_1 = 0,\ z_2 > 0$. We have $R_1 = \rho_1,\ R_2 = c$. Since $\rho_1 > 1-c$, we have $R_1 + R_2 > 1$, and $(0, z_2) \notin \mathcal{M}(C)$. □

Asymptotic growth rates

Let us now consider a fluid-model solution z, satisfying Equations (5.5)–(5.6) with $z(0) = 0$. We determine the asymptotic growth rates m_i, $i = 1, 2$. As shown in the proof of Lemma 5.4.2, there exist nonnegative Lagrange multipliers p_j associated with the links so that the m_i and the corresponding rate allocations R_i together with the p_j form a solution to the system of equations

$$\begin{cases} m_1 = R_1 p_1, \\ m_2 = R_2(p_1 + p_2), \\ p_1(R_1 + R_2 - 1) = 0, \\ p_2(R_2 - c) = 0, \end{cases} \tag{5.32}$$

in conjunction with the set of fixed-point Equations (5.12). The latter equations in fact allow us to express the m_j's in terms of the R_j's, yielding

$$\begin{cases} \beta_1^{-1}\left(\frac{R_1}{\rho_1}\right) = p_1, \\ \beta_2^{-1}\left(\frac{R_2}{\rho_2}\right) = p_1 + p_2, \\ p_1(R_1 + R_2 - 1) = 0, \\ p_2(R_2 - c) = 0, \end{cases} \tag{5.33}$$

where $\beta_i^{-1}(\cdot)$ is the inverse of the LST $\beta_i(y) = \mathbf{E}\left[\mathrm{e}^{-yB_i^*}\right]$. In total the above system provides four equations for four unknown variables R_i, $i = 1, 2$, and p_j, $j = 1, 2$.

Proposition 5.6.1 implies that under the overload assumptions (5.31) the solution of the system of Equations (5.32) is strictly positive. We distinguish between two scenarios: *(I)* $R_2 < c$, *(II)* $R_2 = c$. It is important to note that the inequality $R_i(m) < \rho_i$ must be satisfied when $m_i > 0$. Hence, due to the positivity of the solution m, scenario *(I)* occurs if and only if $\rho_2 < c$, while scenario *(II)* occurs if and only if $\rho_2 > c$. The solutions may then be represented as

$$(I) \quad R_1 = \tfrac{m_1}{m_1+m_2},\ R_2 = \tfrac{m_2}{m_1+m_2},\ p_1 = m_1 + m_2,\ p_2 = 0, \quad \text{if } \rho_2 < c,$$

$$(II) \qquad R_1 = 1 - c,\ R_2 = c,\ p_1 = \tfrac{m_1}{1-c},\ p_2 = \tfrac{m_2}{c} - \tfrac{m_1}{1-c}, \qquad \text{if } \rho_2 > c.$$

Exponential flow sizes

In order to determine the growth rates explicitly, we specify the flow size distribution in the set of Equations (5.33). Let us assume exponentially distributed flow sizes. In the above scenario *(I)*, the growth rates of both classes may be represented in terms of the single variable $n = m_1 + m_2$ as

$$m_i = \lambda_i - \mu_i \frac{m_i}{n} = \frac{\lambda_i n}{\mu_i + n}, \qquad i = 1, 2. \tag{5.34}$$

Summing the above equations results in

$$n = \frac{\lambda_1 n}{\mu_1 + n} + \frac{\lambda_2 n}{\mu_2 + n},$$

which yields the quadratic equation

$$n^2 + n(\mu_1 + \mu_2 - \lambda_1 - \lambda_2) + \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1 = 0.$$

The latter equation has indeed a unique positive solution since the zero-order constant is non-positive by the assumption $\rho_1 + \rho_2 > 1$.

In scenario *(II)* when $\rho_2 > c$, we have a trivial solution

$$m_1 = \lambda_1 - (1 - c)\mu_1, \qquad m_2 = \lambda_2 - c\mu_2.$$

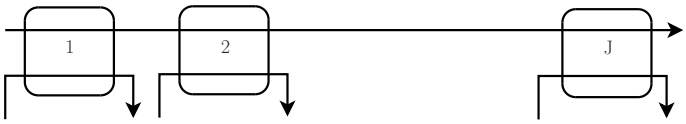

Figure 5.3: Linear network.

5.7 Asymptotic growth rates in linear networks

We now turn to network topologies for which the rate allocation cannot be expected to be monotone. Since we strongly conjecture that an analogue of Theorem 5.5.1 still holds for non-monotone bandwidth-sharing allocations, we consider it worthwhile to investigate properties of the asymptotic growth rates for several practically relevant topologies.

In the present section we focus on the special case of a linear network as illustrated in Figure 5.3. Linear networks provide a useful model for traffic that traverses several links and experiences bandwidth competition from independent cross-traffic. The network consists of links $1, \ldots, J$, each of unit capacity, and is offered traffic from classes $0, 1, \ldots, J$. Class-j flows require service from link j only, $j = 1, \ldots, J$, while class-0 flows demand capacity on all links simultaneously. The rate allocation is governed by the proportional fair policy with unit class weights, i.e., the objective function is given by $G_z(\Lambda) = \sum_{i=0}^{J} z_i \log(\Lambda_i)$. We assume that the load on at least one of the links exceeds the capacity, i.e., $\max_{j=1,\ldots,J} \rho_j > 1 - \rho_0$. The capacity constraints take the form $\Lambda_0 + \Lambda_j \leq 1$ for all $j = 1, \ldots, J$. For now, we allow the flow sizes to have general distributions.

We are interested in determining the asymptotic growth rates m_i of the various classes. As shown in the proof of Lemma 5.4.2, there exist nonnegative coefficients (Lagrange multipliers) p_j associated with the various links so that the m_i and the corresponding rate allocations R_i together with the p_j form a solution to the system of equations

$$\begin{cases} m_0 = R_0 \sum_{j=1}^{J} p_j, \\ m_j = R_j p_j, \qquad j = 1, \ldots, J, \\ p_j(R_0 + R_j - 1) = 0, \qquad j = 1, \ldots, J, \end{cases} \tag{5.35}$$

in conjunction with the set of fixed-point Equations (5.67). The latter equations in fact allow us to express the m_j's in terms of the R_j's, yielding

$$\begin{cases} \beta_0^{-1}\left(\frac{R_0}{\rho_0}\right) = \sum_{j=1}^{J} p_j, \\ \beta_j^{-1}\left(\frac{R_j}{\rho_j}\right) = p_j, \qquad j = 1, \ldots, J, \\ p_j(R_0 + R_j - 1) = 0, \qquad j = 1, \ldots, J. \end{cases} \tag{5.36}$$

In total the above system provides $2J+1$ equations for $2J+1$ unknown variables R_i, $i=0,\ldots,J$, and p_j, $j=1,\ldots,J$.

In order to solve the above system of equations, we consider the nonempty subset $\mathcal{J}_+ := \{j : p_j > 0\}$ of links with strictly positive Lagrange multipliers. (The subset $\mathcal{J}_+$ cannot be empty, since that would imply $\rho_0+\rho_j \le 1$ for all $j=1,\ldots,J$, and contradict the overload assumption.) Observe that $p_j = 0$ means $m_j = 0$. As stated in Property 5.4.4, the growth rates of classes 0 and $j \in \mathcal{J}_+$ thus correspond to those in a scenario with classes $j \notin \mathcal{J}_+$ as well as links $j \notin \mathcal{J}_+$ removed. In particular, when $\mathcal{J}_+ = \{j_+\}$, the growth rates of classes 0 and j_+ are identical to those in a single-node processor-sharing system with classes 0 and j_+.

For compactness, denote $n = \sum_{j\in\mathcal{J}_+} p_j$. Then the solution to the system of Equations (5.35) may be represented as

$$R_0 = \frac{m_0}{n}; \qquad R_j \equiv S_{\mathcal{J}_+} = 1 - \frac{m_0}{n} \ \text{ if } j \in \mathcal{J}_+; \qquad R_j = \rho_j \ \text{ if } j \notin \mathcal{J}_+;$$

$$p_j = \frac{m_j}{R_j}, \ j=1,\ldots,J. \tag{5.37}$$

Summing the last equality in Equation (5.37) over $j \in \mathcal{J}_+$, it follows that $n = m_0 + \sum_{j\in\mathcal{J}_+} m_j$.

What remains is to determine the subset $\mathcal{J}_+$ in terms of the system parameters. Note that $j \in \mathcal{J}_+$ implies $R_0 + R_j = 1$, and thus necessitates $\rho_0+\rho_j \ge 1$. However, the latter inequality is not sufficient for $j \in \mathcal{J}_+$, since it is possible that $m_j = 0$ when other classes at other links sufficiently throttle the service rate of class 0. In order to characterize the subset $\mathcal{J}_+$, observe that $\rho_j \le S_{\mathcal{J}_+}$ for all $j \notin \mathcal{J}_+$ and $\rho_j > S_{\mathcal{J}_+}$ for all $j \in \mathcal{J}_+$. In view of the inherent symmetry, we may assume without loss of generality that the links are indexed such that $\rho_1 \le \rho_2 \le \cdots \le \rho_J$. Denote by σ_j the common service rate obtained by classes $j,\ldots,J$ in a system with links $j,\ldots,J$ and classes 0 and $j,\ldots,J$ only, $\sigma_j = 1 - \Lambda_0(m_j,\ldots,m_J) = 1 - \frac{m_0}{m_0+\sum_{k=j}^{J} m_k}$. Then the subset $\mathcal{J}_+$ is of the form $\{j_+,\ldots,J\}$, with $j_+ := \max\{j : \rho_{j-1} \le \sigma_j\}$. In case $B_0 \equiv B_J$, it is easily verified that $\sigma_J = \lambda_J/(\lambda_0+\lambda_J)$.

The above characterization of the subset $\mathcal{J}_+$ may be interpreted as follows. If $\rho_{j-1} \le \sigma_j$, then competition from classes $j,\ldots,J$ alone against class 0 is sufficient to throttle the rate of class 0 to an extent that what remains available for classes $1,\ldots,j-1$ exceeds their respective loads, and hence $m_1 = \cdots = m_{j-1} = 0$. This scenario occurs when the loads of classes $1,\ldots,j-1$ are relatively low and the loads of classes $j,\ldots,J$ are sufficiently high. Note that this may occur even when $\rho_0+\rho_i > 1$ for some classes $i=1,\ldots,j-1$. Although these classes rely on service from overloaded links, they remain stable thanks to the much stronger competition at other higher-loaded links. In contrast, if $\rho_{j-1} > \sigma_j$, then competition from classes $j,\ldots,L$ alone is not sufficient to provide stability to class $j-1$, and hence $m_{j-1} > 0$.

Remark 5.7.1. For any subset $\mathcal{K} \subseteq \mathcal{L} = \{1,\ldots,J\}$, we may construct a reduced version of the original linear network with similar characteristics but classes $i \in \mathcal{K}$ removed. We attach superscripts $\mathcal{L}$ and $\mathcal{L}\setminus\mathcal{K}$ to the variables associated with the original and reduced version of the network, respectively. It is easily verified that

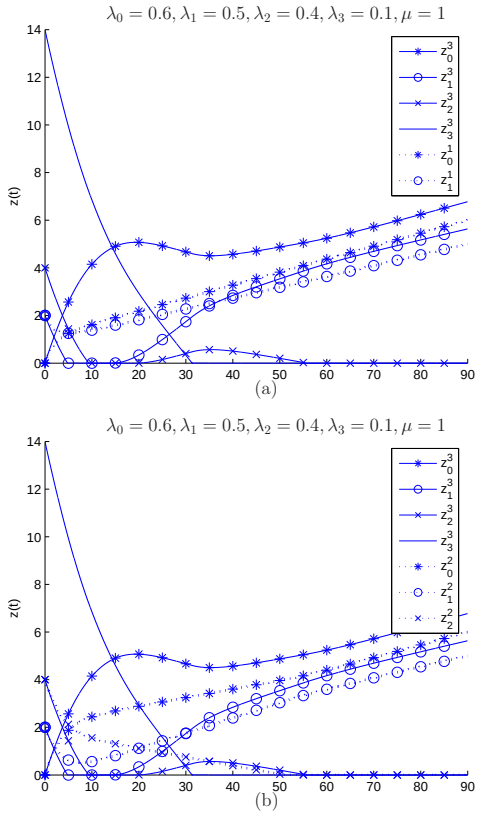

Figure 5.4: Fluid-model solutions for a 3-link linear network and (a) a PS node, (b) a 2-link linear network.

$S^{\mathcal{L}\setminus\mathcal{K}}_{\mathcal{J}^{\mathcal{L}\setminus\mathcal{K}}_+} \leq S^{\mathcal{L}}_{\mathcal{J}^{\mathcal{L}}_+}$ for any subset $\mathcal{K}$, which implies that $m_0^{\mathcal{L}\setminus\mathcal{K}} \leq m_0^{\mathcal{L}}$ and $m_i^{\mathcal{L}\setminus\mathcal{K}} \geq m_i^{\mathcal{L}}$ for all $i \in \mathcal{L}\setminus\mathcal{K}$. In other words, removing competing classes reduces the asymptotic growth rate of class 0 and increases the asymptotic growth rates of the remaining classes, which is intuitively plausible. It might seem natural to expect that a similar monotonicity property holds for the entire fluid-limit trajectories, but that turns out *not* to be the case, as is graphically illustrated in Figure 5.4. This may be explained from the fact that when class 0 becomes smaller by removing competing classes, this actually also has a beneficial effect on the remaining classes.

Two-link network

As an illustrative example, we now elaborate on the case of a two-link (three-class) network, i.e., $J = 2$. In that case we need to distinguish between two scenarios: *(I)* only one Lagrange multiplier is strictly positive; and *(II)* both Lagrange multipliers are strictly positive. Note that it cannot occur that both Lagrange multipliers are

zero, since that would imply $\rho_0 + \rho_i < 1$, $i = 1, 2$, and contradict the overload assumption $\max\{\rho_1, \rho_2\} > 1 - \rho_0$. As noted above, $\rho_0 + \rho_1 > 1$, $\rho_0 + \rho_2 > 1$ is needed for case *(II)* to arise. These two inequalities are however not sufficient for case *(II)* to occur, since it is possible that $m_i = 0$ even when $\rho_0 + \rho_i > 1$, for either $i = 1$ or $i = 2$ (not both). The exact demarcation between cases *(I)* and *(II)* is determined by slightly more involved conditions which will be further discussed below.

For compactness, denote $m := m_0 + m_1 + m_2$ and $n = m_0 + m_i$. The solutions in the above two scenarios may then be represented as

$$(I) \qquad (R_0, R_i, R_{3-i}) = \left(\frac{m_0}{n}, \frac{m_i}{n}, \rho_{3-i}\right), \qquad p_i = n, \qquad p_{3-i} = 0,$$

and

$$(II) \qquad (R_0, R_1, R_2) = \left(\frac{m_0}{m}, \frac{m_1 + m_2}{m}, \frac{m_1 + m_2}{m}\right), \qquad p_1 = p_2 = \frac{m_2}{m_1 + m_2} m.$$

Note that in case *(I)* the growth rates of classes 0 and i are identical to those in a scenario with both link $3 - i$ and class $3 - i$ removed, i.e., a single-node processor-sharing system with classes 0 and i only, cf. Property 5.4.4.

Exponential flow sizes

In order to determine the growth rates explicitly, we need to specify the flow size distributions of the various classes that occur in the set of Equations (5.36). In the case of exponential flow sizes the growth rates of the various classes may be represented in terms of the single variable n as

$$m_0 = \lambda_0 - \mu_0 \frac{m_0}{n} = \frac{\lambda_0 n}{\mu_0 + n}, \tag{5.38}$$

$$m_j = \lambda_j - \mu_j \frac{n - m_0}{n} = \lambda_j - \mu_j \left(1 - \frac{\lambda_0}{\mu_0 + n}\right), \qquad j \in \mathcal{J}_+. \tag{5.39}$$

Summing the above equations results in

$$n = \frac{\lambda_0 n}{\mu_0 + n} + \sum_{j \in \mathcal{J}_+} \left(\lambda_j - \mu_j \left(1 - \frac{\lambda_0}{\mu_0 + n}\right)\right),$$

which yields the quadratic equation

$$n^2 + \nu n + \kappa = 0, \tag{5.40}$$

with

$$\nu := \mu_0 + \sum_{j \in \mathcal{J}_+} \mu_j - \lambda_0 - \sum_{j \in \mathcal{J}_+} \lambda_j,$$

$$\kappa := \sum_{j \in \mathcal{J}_+} \mu_j(\mu_0 - \lambda_0) - \mu_0 \sum_{j \in \mathcal{J}_+} \lambda_j.$$

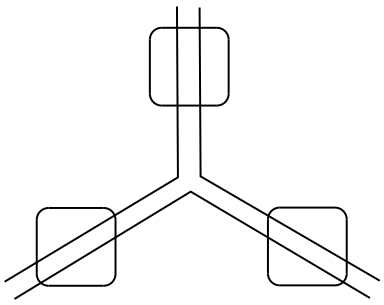

Figure 5.5: Star network with three links.

Substituting the positive solution in Equations (5.38)–(5.39) gives expressions for the asymptotic growth rates. To see that there is indeed a unique positive solution, recall that a quadratic equation of the type (5.40) has a unique positive solution when the zero-order constant is nonpositive, which may be written as

$$\rho_0 + \sum_{j \in \mathcal{J}_+} \rho_j \frac{\mu_j}{\sum_{j \in \mathcal{J}_+} \mu_j} \geq 1. \tag{5.41}$$

Noting that

$$\sum_{j \in \mathcal{J}_+} \rho_j \frac{\mu_j}{\sum_{j \in \mathcal{J}_+} \mu_j} \geq \sum_{j \in \mathcal{J}_+} \min_{k \in \mathcal{J}_+} \rho_k \frac{\mu_j}{\sum_{j \in \mathcal{J}_+} \mu_j} = \min_{k \in \mathcal{J}_+} \rho_k,$$

the inequality (5.41) is seen to hold by virtue of the fact that $\rho_0 + \rho_j \geq 1$, $j \in \mathcal{J}_+$.

5.8 Asymptotic growth rates in star networks

In this section we focus our attention on the special case of a star network. As mentioned earlier, star networks offer a convenient abstraction for scenarios where the core is highly over-provisioned and congestion predominantly occurs at the edge with comparatively low-capacity access links. The network is composed of J links, each of unit capacity, and is offered traffic from $J(J-1)/2$ classes labeled as $\{i, j\}$, $i, j = 1, \ldots, J$, $i \neq j$. The route of class $\{i, j\}$ simply consists of the two links i and j. We assume that the load on at least one of the links exceeds the capacity, i.e., $\max_{j=1,\ldots,J} \sigma_j > 1$, with $\sigma_j := \sum_{k \neq j} \rho_{\{j,k\}}$. The rate allocation is governed by the proportional fair policy with unit class weights, i.e., the objective function is given by $G_z(\Lambda) = \sum_{j \neq k} z_{\{j,k\}} \log(\Lambda_{\{j,k\}})$. The capacity constraints take the form $\sum_{k \neq j} \Lambda_{\{j,k\}} \leq 1$ for all $j = 1, \ldots, J$.

For star networks, the proof of Lemma 5.4.2 shows that the Lagrange multipliers p_j associated with the links in the network and the corresponding rate alloca-

tions $R_{\{j,k\}}$ satisfy the following system of equations

$$\begin{cases} \beta_{\{j,k\}}^{-1}\left(\frac{R_{\{j,k\}}}{\rho_{\{j,k\}}}\right) = R_{\{j,k\}}(p_j + p_k), & j \neq k, \\ p_j(\sum_{k\neq j} R_{\{j,k\}} - 1) = 0, & j = 1, \ldots, J. \end{cases} \tag{5.42}$$

In total the above system provides $J(J+1)/2$ equations for $J(J+1)/2$ unknown variables $R_{\{j,k\}}$, $j \neq k$, and p_j, $j = 1, \ldots, J$. In the case of exponential flow sizes, the set of fixed-point equations takes the explicit form in Equation (5.14). The above system of equations then simplifies to

$$\begin{cases} \lambda_{\{j,k\}} - \mu_{\{j,k\}} R_{\{j,k\}} = R_{\{j,k\}}(p_j + p_k), & j \neq k, \\ p_j(\sum_{k\neq j} R_{\{j,k\}} - 1) = 0, & j = 1, \ldots, J. \end{cases} \tag{5.43}$$

As before, we need to consider the subset of links with strictly positive Lagrange multipliers in order to solve the above system of equations.

Three-link network

As an illustrative example, we now focus on the case of a star network with three links and three classes, which is topologically equivalent to a triangular network as depicted in Figure 5.5. In that case we need to distinguish three scenarios, *(I)*, *(II)* and *(III)*, depending on whether one, two or all three of the Lagrange multipliers are strictly positive, respectively. It cannot occur that all three Lagrange multipliers are zero, since that would imply $\sum_{k\neq j} \rho_{\{j,k\}} < 1$, $j = 1, 2, 3$, and contradict the overload assumption.

With minor abuse of notation, we define $m_i := m_{\{1,2,3\}\setminus\{i\}}$, $R_i := R_{\{1,2,3\}\setminus\{i\}}$, and $\rho_i := \rho_{\{1,2,3\}\setminus\{i\}}$. The above system of Equations (5.42) may then be rewritten as

$$\begin{cases} m_i = R_i(p_j + p_k), & \{i,j,k\} = \{1,2,3\}, \\ p_j(R_i + R_k - 1) = 0, & \{i,j,k\} = \{1,2,3\}. \end{cases} \tag{5.44}$$

For compactness, denote $m := m_1 + m_2 + m_3$, and $n = m_i + m_j$. The solutions in the above three scenarios may be then represented as

$$\begin{aligned} (I) \qquad & (R_i, R_j, R_k) = \left(\tfrac{m_i}{n}, \tfrac{m_j}{n}, \rho_k\right), \\ & p_i = p_j = 0, \quad p_k = n, \\ (II) \qquad & (R_i, R_j, R_k) = \left(\tfrac{m_i}{m}, \tfrac{m_j+m_k}{m}, \tfrac{m_j+m_k}{m}\right), \\ & p_i = 0, \quad p_j = \tfrac{m_j}{m_j+m_k} m, \quad p_k = \tfrac{m_k}{m_j+m_k} m, \\ (III) \qquad & R_1 = R_2 = R_3 = \tfrac{1}{2}, \\ & p_i = \textstyle\sum_{j\neq i} m_j - m_i > 0, \quad i = 1, 2, 3. \end{aligned}$$

The above results reveal an interesting trichotomy in the behavior of the triangular network. In case *(I)* the network behaves as a single-node processor-sharing system with classes i and j only. The conditions for case *(I)* to occur in terms of the system parameters also coincide with the corresponding ones in the linear network with $|\mathcal{J}_+| = 1$. Case *(II)* corresponds to the case of the linear network with $|\mathcal{J}_+| = 2$. The conditions for this case to arise subsume the corresponding ones in the linear network, but include an additional condition that the loads of the three classes should be slightly unbalanced. If the latter condition is violated, i.e., the loads of the three classes are nearly equal, then case *(III)* arises, which has no counterpart in the linear network. In this case, each of the three classes behaves as in an isolated processor-sharing system with capacity $\frac{1}{2}$.

5.9 Numerical results

In this section we discuss the numerical experiments that we performed to corroborate and illustrate the analytical findings. We present simulation results to demonstrate the convergence of the scaled number of flows to the fluid limit. In addition, we examine the impact of the traffic intensities and the variability of the flow sizes on the asymptotic growth rates. Throughout the numerical experiments we focus the attention on a two-link linear network operating under the proportional fair policy with unit class weights.

Exponential flow sizes

We first consider the case of exponential flow sizes, and specifically investigate the impact of the traffic intensity on the asymptotic growth rates.

Figures 5.6–5.7 plot the asymptotic growth rates m_0, m_1, m_2 for exponential flow sizes as function of the traffic intensity. We let all classes have the same mean flow size and let the arrival rates vary. Figures 5.6 (a,b) show the growth rates in a situation where the loads of classes 0 and 2 are fixed and the load of class 1 is varied. The figures reveal natural qualitative trends. As the load of class 1 increases, the competition with class 0 becomes stronger, and as a result both queues grow at a higher rate. The reduced service rate of class 0 in turn leaves more capacity available for class 2, and thus its growth rate decreases, ultimately reaching stability.

Figure 5.7 shows the growth rates in a situation where the loads of classes 1 and 2 are fixed and the load of class 0 is varied. As the load of class 0 increases, both classes 1 and 2 receive less service. Consequently, all three classes build up queues at a higher rate.

In particular, the figures illustrate the stability properties of the linear networks discussed in Section 5.7. Recall that the link overload conditions $\rho_0 + \rho_1 > 1$, $\rho_0 + \rho_2 > 1$ are necessary but not sufficient for the number of flows z_1 or z_2 to grow. For instance, in Figure 5.6 (b) the asymptotic growth m_1 becomes positive only when ρ_1 reaches the value of 0.3. While the first link is already overloaded due to the large number of class-0 flows, the strong competition with class 2 provides sufficient bandwidth for class 1 to remain stable. Note that if $m_1 = 0$ the network

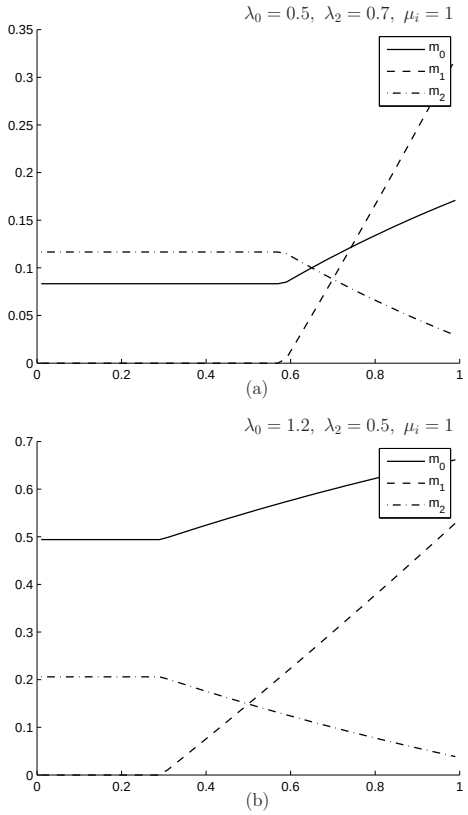

Figure 5.6: Asymptotic growth rates m_0, m_1, m_2 as function of ρ_1.

behaves as a single-node processor-sharing system with classes 0 and 2 only. Thus, the stability condition for class 1 in this case is determined as $\rho_1 < 1 - \Lambda_0 = \frac{\rho_2}{\rho_0+\rho_2}$.

Figure 5.8 plots the value of $Z_i(t)/t$ as function of t obtained by simulation. The horizontal lines represent the asymptotic growth rates m_0, m_1, and m_2 as computed from Equation (5.40). All classes have exponential flow sizes with unit mean. The traffic intensities are $\rho_0 = 1.2$, $\rho_1 = 0.5$, $\rho_2 = 0.7$.

Hyperexponential flow sizes

We now turn to the case of hyperexponential flow sizes, and investigate the impact of the variability of the flow sizes on the asymptotic growth rates by varying the parameter values of the hyperexponential distribution.

The flow sizes of all classes have the same hyperexponential distribution: the flow size is exponential with mean $1/\nu_1$ with probability p, and exponential with mean $1/\nu_2$ otherwise. Moreover, we assume that the contributions to the mean are

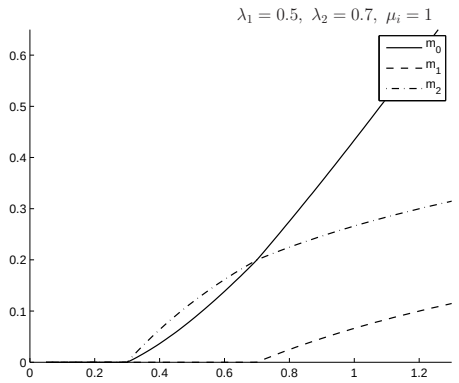

Figure 5.7: Asymptotic growth rates m_0, m_1, m_2 as function of ρ_0.

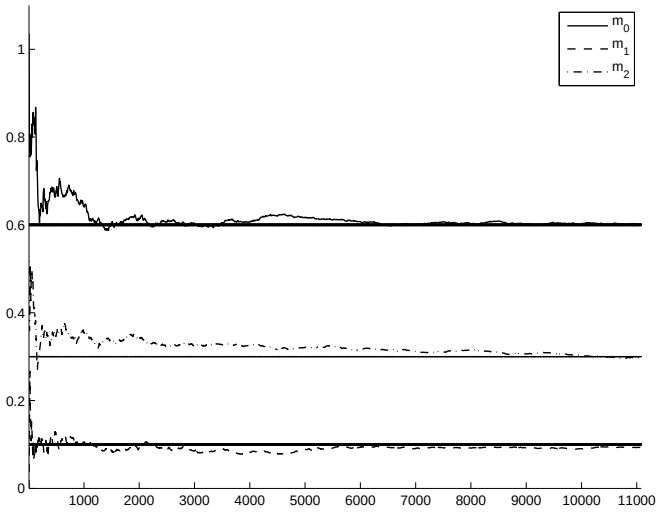

Figure 5.8: Scaled queue length $\frac{Z_i(t)}{t}$ as function of time, $\rho_0 = 1.2$, $\rho_1 = 0.5$, $\rho_2 = 0.7$.

balanced, i.e., $p/\nu_1 = (1-p)/\nu_2$, so that as $p \to 0$, $\nu_1 = \frac{2p}{\mathbf{E}[B]} \to 0$ and $\nu_2 \to \frac{2}{\mathbf{E}[B]}$.

We fix the mean flow size $\mathbf{E}[B] = 2$, and vary the value of p in the interval $[0, 1/2]$. Note that when $p = \frac{1}{2}$, the flow size becomes simply an exponential random variable with mean 2 and squared coefficient of variation $\sigma^2 = 1$. However, as p tends to 0, the squared coefficient of variation grows like $1/p$.

Figures 5.9 (a,b) plot the growth rates m_0, m_1, m_2 as function of the squared coefficient of variation. Two limiting cases are shown as the markers on the graphs which provide lower and upper bounds for the asymptotic growth rates; white markers show the growth rates computed for exponential flow sizes with mean 2 (the case $p = \frac{1}{2}$); the black markers indicate the growth rates obtained for exponential flow sizes with mean 1.

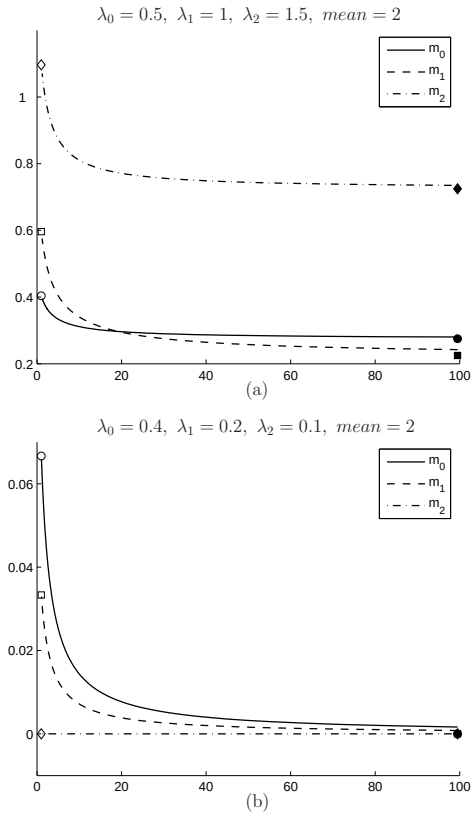

Figure 5.9: Asymptotic growth rates m_0, m_1, m_2 as function of the squared coefficient of variation.

The observation that the growth rates approach those for exponential flow sizes with mean 1 may be explained as follows. Note that a particular class i with arrival rate λ_i and a hyperexponential flow size distribution with parameters μ_{i1}, μ_{i2}, p_{i1} and p_{i2}, with $p_{i1} + p_{i2} = 1$ may be equivalently replaced by two classes with arrival rates $\lambda_{ik} = \lambda_i p_{ik}$ and exponential flow sizes with parameter μ_{ik}, $k = 1, 2$. Now suppose that we consider a regime where the coefficient of variation grows large by letting $p_{i1}, \mu_{i1} \downarrow 0$, with $p_{i1}/\mu_{i1} = c_{i1}$ and $p_{i2}/\mu_{i2} = c_{i2}$ fixed, meaning that $\mu_{i2} \uparrow 1/c_{i2}$. Property 5.4.5 then states that the growth rates of the three origi-

nal classes with hyperexponential flow sizes, in the limit are identical to those in a scenario with three classes with arrival rates λ_{i2} and exponential flow sizes with parameter $1/c_{i2}$.

In the numerical experiments, we have focused the attention on an admittedly simple two-link linear network. Nevertheless, the network model already appears to be sufficiently rich to reveal several interesting features and qualitative properties which also may be expected to occur in more complex scenarios.

5.10 User impatience

In this section we discuss an extension of the model to a scenario with user impatience. Impatient users may abandon the system before completing service. An impatient user is characterized by a positive random *initial lead time* in addition to its flow size. Denote by D_i the initial lead time of class-i flows. An impatient user has a deadline (arrival time plus initial lead time); the user leaves the system either when it completes service or when the deadline expires, whichever occurs first. As mentioned earlier, user impatience has a particularly pronounced impact in overload conditions.

The single-link single-class version of the above model has been studied in [20, 58, 59]. Following similar arguments as in those papers, we propose to approximate the number of flows of each class by the solution $z \in \mathcal{M}(C)$ of the fixed-point equation

$$z_i = \lambda_i \mathbf{E}\left[\min\left(D_i, \frac{z_i}{R_i(z)}B_i\right)\right]. \tag{5.45}$$

This equation may be heuristically explained as follows. Let Z_i^r be the steady-state number of class-i flows in the r-th system, and let $V_i^r(B)$ be the sojourn time of a user that does *not* abandon. Assume that users are relatively patient, i.e., let their initial lead time also be scaled as $D_i r$. Then the actual sojourn time is given by $\min\{V_i^r(B), D_i r\}$, and Little's law implies

$$\mathbf{E}[Z_i^r] = \lambda_i \mathbf{E}[\min(V_i^r(B), D_i r)]. \tag{5.46}$$

Divide both sides of the equality by r. Since we observe the system in steady state at time 0, the number of flows hardly changes over the course of a sojourn time, and by the so-called 'snapshot principle' we conclude that $V_i^r = \frac{z_i}{R_i(z)}B_i + \mathrm{o}(r)$. Noting that $Z_i^r/r \to z_i$ then gives Equation (5.45) after dividing both sides of (5.46) by r and letting $r \to \infty$.

To make these heuristics rigorous, requires us to unify the frameworks of [58] and [60]. In addition, it needs to be shown that the fixed-point Equation (5.45) has in general a unique solution in $\mathcal{M}(C)$. These issues will be further discussed in the following subsection.

5.10.1 Uniqueness

In order to prove that the fixed-point Equation (5.45) has a unique solution z, we construct a new optimization problem which we will use to establish uniqueness of the rate allocation vector $R(z)$ subject to the constraint $AR(z) \leq C$. We first rewrite Equation (5.45) in a more convenient form:

$$R_i(z) = \lambda_i \mathbf{E}\left[\min\left(\frac{R_i(z)}{z_i}D_i, B_i\right)\right] := g_i\left(\frac{R_i(z)}{z_i}\right). \tag{5.47}$$

Since the left-hand side does not contain the term z_i explicitly, we can apply a similar argument as in the proof of Lemma 5.4.2. In this case we construct a convex optimization problem (Q) with functions $G_i(x)$ such that

$$G_i'(x) = U_i'\left(g_i^{-1}(x_i)\right),$$

where $g_i^{-1}(\cdot)$ denotes the inverse of function $g_i(\cdot)$:

$$g_i^{-1}(x_i) = \min\{y_i \geq 0 : \lambda_i \mathbf{E}\left[\min\left(y_i D_i, B_i\right)\right] = x_i\}, \qquad 0 \leq x_i \leq \rho_i.$$

Now observe that given Equation (5.47)

$$G_i'(R_i(z)) = U_i'\left(\frac{R_i(z)}{z_i}\right).$$

Following the argument in the proof of Lemma 5.4.2, we conclude that $R(z)$ satisfying Equation (5.47) obeys the KKT sufficient conditions for the optimization problem (Q), yielding the following lemma.

Lemma 5.10.1. *Equation (5.45) has a unique solution $z = (z_1, \ldots, z_I)$ in the set $\mathcal{M}(C)$.*

5.10.2 Examples

We now present some simple examples illustrating various properties of Equation (5.45). We examine several special cases that allow for explicit calculations. In addition, we investigate what fraction of the flows successfully complete their transfer, that is the probability $\mathbf{P}_i^s$ of the event $\{D_i > \frac{z_i}{R_i(z)}B_i\}$, $i = 1, \ldots, I$.

The following property is a direct consequence of Equation (5.45). Consider the system operating under a zero-degree homogeneous rate allocation policy. Then multiplication of the lead times of all classes with the same arbitrary coefficient leads to an increase of the number of flows by this coefficient and does not affect the success probabilities.

Property 5.10.1. *Consider two systems such that $(B_i^1, D_i^1) \equiv (B_i^2, aD_i^2)$, $i = 1, \ldots, I$, for some $a > 0$ with the same arrival rates and operating under the same zero-degree homogeneous rate allocation policy. Then,*

$$z_i^1 = az_i^2, \qquad \mathbf{P}_i^{s,1} = \mathbf{P}_i^{s,2}, \qquad i = 1, \ldots, I.$$

Completely dependent lead times

We now consider the case of completely dependent lead times, i.e., we assume that $D_i = \theta_i B_i$ for some coefficient $\theta_i > 0$, independent of B_i. The coefficient θ_i may be interpreted as the average service rate expected by a class-i user. In this case, the equations for the number of active flows and success probability are

$$z_i = \rho_i \mathbf{E}\left[\min\left(\theta_i, \frac{z_i}{R_i(z)}\right)\right], \qquad \mathbf{P}_i^s = \mathbf{P}\left(\theta_i > \frac{z_i}{R_i(z)}\right).$$

It was shown in [20, 58, 59] for a single-link scenario, that impatience can have a substantial impact, especially if the initial lead time is a *constant* times the flow size. In that case, the abandonment rate is one hundred percent. To see whether this holds in network scenarios as well, we assume that $D_i = \theta B_i$ for some coefficient $\theta > 0$, and consider the case of a two-link linear network operating under the proportional fair policy with unit class weights. Class-0 flows require both links simultaneously, while flows of classes 1 and 2 use only links 1 and 2, respectively. Then the equations simplify to

$$\begin{aligned} z_0 &= \rho_0 \min(\theta, z_0 + z_1 + z_2), \\ z_i &= \rho_i \min\left(\theta, \tfrac{z_i}{z_1+z_2}(z_0 + z_1 + z_2)\right), \quad i = 1, 2. \end{aligned} \tag{5.48}$$

In addition, the constraints $R_0 + R_i \leq 1$ should hold for a solution of the fixed-point Equation (5.48) to be admissible. If $\rho_2 < \rho_1/(\rho_0 + \rho_1)$ and $\rho_0 + \rho_1 > 1$, then it follows that $z_2 = 0, z_0 = \theta\rho_0, z_1 = \theta\rho_1$ is a feasible solution of the fixed-point Equation (5.48). We conclude that the overall abandonment rate can be lower than one hundred percent in network scenarios, due to the fact that some classes may only traverse links with surplus capacity. It is interesting to observe that when at least one link is overloaded, all class-0 flows leave the system due to impatience.

Now suppose that the coefficients θ_i are *exponentially* distributed with parameters φ_i. Then the solution of Equation (5.45) satisfies

$$\varphi_i \frac{z_i}{\rho_i} = \left(1 - \mathrm{e}^{-\varphi_i \frac{z_i}{R_i(z)}}\right), \qquad \mathbf{P}_i^s = \mathrm{e}^{-\frac{z_i}{R_i(z)}}.$$

Independent lead times

Assume now that the lead times are *independent* of the flow sizes. In this case, Equation (5.45) can be written as

$$z_i = \lambda_i \int_0^\infty \mathbf{P}(D_i > u)\mathbf{P}\left(\frac{z_i}{\Lambda_i(z)} B_i > u\right) du,$$

or

$$\Lambda_i(z) = \lambda_i \int_0^\infty \mathbf{P}\left(D_i > \frac{z_i}{R_i(z)} u\right) \mathbf{P}(B_i > u) du.$$

If $\mathbf{E}[B_i] < \infty$, then this is equivalent to $\mathbf{P}\left(D_i > \frac{z_i}{R_i(z)} B_i^*\right) = \frac{R_i(z)}{\rho_i}$, where B_i^* represents the residual class-i flow size. Note that if B_i has an exponential distribution, then $\mathbf{P}_i^s = R_i(z)/\rho_i$.

In case D_i has an exponential distribution with parameter ν_i (and B_i a general distribution), we see that z is the solution of

$$R_i(z) = \rho_i \beta_i \left(\frac{z_i}{R_i(z)} \nu_i \right), \tag{5.49}$$

where $\beta_i(s) = \mathbf{E}[\mathrm{e}^{-sB_i^*}]$. If in addition the rate allocation function is zero-degree homogeneous and $\nu_i \equiv \nu$ for all $i = 1, \ldots, I$, then we observe that the number of flows in the system with impatience is ν times smaller than the number of flows m in the ordinary system as characterized by Equation (5.12).

Appendix

5.A Proof of Theorem 5.2.1

In this appendix we present the proof of Theorem 5.2.1, which shows that any limit point of the scaled sequence $(\overline{Z}^r(t), \overline{T}^r(t); t \geq 0)$ is almost surely a solution of the fluid-model Equations (5.5)–(5.6).

Proof of Theorem 5.2.1
The starting point is provided by Equations (5.1) and (5.2) expressing $Z_i(t)$ and $T_i(t)$ in terms of the arrival times, service requirements and amounts of service received by class-i flows.

Since these equations are similar in nature, we consider an equation of the generic form

$$F_i[f](t) = \sum_{l=1}^{Z_i(0)} f(\overline{B}_{il}, S_i(0,t)) + \sum_{k=1}^{E_i(t)} f(B_{ik}, S_i(A_{ik}, t)), \tag{5.50}$$

which reduces to (5.1) and (5.2) in case $f(x,y) = \mathbf{1}(x < y)$ and $f(x,y) = \min(x,y)$, respectively. At this point we only assume that function $f(\cdot,\cdot)$ is monotone in the second argument.

Applying the fluid scaling to each term in (5.50), we obtain

$$\overline{F}_i^r[f](t) = \frac{1}{r} \sum_{l=1}^{r\overline{Z}_i^r(0)} f(\overline{B}_{il}, \overline{S}_i^r(0,t)) + \frac{1}{r} \sum_{k=1}^{E_i^r(t)} f(B_{ik}, \overline{S}_i^r(A_{ik}^r, t)) := I_i^r + J_i^r. \tag{5.51}$$

For compactness, the implicit dependence of I_i^r and J_i^r on t and the function $f(\cdot,\cdot)$ will be suppressed where appropriate.

We now proceed to derive $\lim_{r\to\infty} F_i^r[f](t)$, and distinguish two cases, depending on whether $z_i(u) > 0$ for all $u \in [0,t]$ or not.

We first deal with the latter case, and let $\eta_i(t) = \sup(u \in [0,t] : z_i(u) = 0)$. It is useful to distinguish two further cases, depending on whether $\eta_i(t) < t$ or $\eta_i(t) = t$. We start with the former case, and fix $\varepsilon > 0$ such that $\eta_i(t) + \varepsilon < t$.

We first consider the term J_i^r, which may be rewritten as

$$J_i^r = \frac{1}{r} \sum_{k=1}^{E_i^r(\eta_i(t)+\varepsilon)} f(B_{ik}, \overline{S}_i^r(A_{ik}^r, t)) + \frac{1}{r} \sum_{k=E_i^r(\eta_i(t)+\varepsilon)+1}^{E_i^r(t)} f(B_{ik}, \overline{S}_i^r(A_{ik}^r, t)) =: J_{1,i}^r + J_{2,i}^r. \tag{5.52}$$

We first determine $\lim_{r\to\infty} J_{2,i}^r$. By definition, $z_i(u) > 0$ for all $[\eta_i(t) + \varepsilon, t]$. Hence, the bounded convergence theorem yields

$$\lim_{r\to\infty} \overline{S}_i^r(u, v) = S_i(u, v)$$

for all $t \geq v \geq u \geq \eta_i(t)+\varepsilon$. Since $\overline{S}_i^r(s, t)$ is decreasing in s and $S(\cdot, t)$ is continuous on $[\eta_i(t) + \varepsilon, t]$, the convergence is uniform on $[\eta_i(t) + \varepsilon, t]$, i.e., for any $\delta > 0$ there exists an r_δ such that

$$\sup_{s\in[\eta_i(t)+\varepsilon,t]} \left| \overline{S}_i^r(s, t) - S_i(s, t) \right| \leq \delta, \qquad \text{for all } r \geq r_\delta. \tag{5.53}$$

We partition the interval $[\eta_i(t)+\varepsilon, t]$ into N subintervals $[t_{j-1}^N, t_j^N]$, $j = 1, \ldots, N$, for some integer $N \geq 1$, in such a way that $\max_{j=0,\ldots,N} \left(t_j^N - t_{j-1}^N\right) \to 0$ as $N \to \infty$. Then,

$$J_{2,i}^r = \frac{1}{r} \sum_{j=1}^{N} \sum_{k=E_i^r\left(t_j^N\right)+1}^{E_i^r\left(t_{j+1}^N\right)} f\left(B_{ik}, \overline{S}_i^r(A_{ik}^r, t)\right).$$

Suppose that $t_{j-1}^N \leq A_{ik}^r \leq t_j^N$ for some $j \in \{1, \ldots, N\}$, some $k \in \{E_i^r(\eta_i(t) + \varepsilon) + 1, \ldots, E_i^r(t)\}$, and some $r > r_\delta$. It then follows from (5.53) that for $r > r_\delta$

$$S_i\left(t_j^N, t\right) - \delta \leq \overline{S}_i^r\left(A_{ik}^r, t\right) \leq S_i\left(t_{j-1}^N, t\right) + \delta. \tag{5.54}$$

If the function $f(\cdot, \cdot)$ is non-decreasing in its second argument, then we derive for $r > r_\delta$

$$f\left(B_{ik}, S_i\left(t_j^N, t\right) - \delta\right) \leq f\left(B_{ik}, \overline{S}_i^r\left(A_{ik}^r, t\right)\right) \leq f\left(B_{ik}, S_i\left(t_{j-1}^N, t\right) + \delta\right), \tag{5.55}$$

which yields

$$\begin{aligned} &\frac{1}{r} \sum_{j=1}^{N} \sum_{k=E_i^r\left(t_{j-1}^N\right)+1}^{E_i^r\left(t_j^N\right)} f\left(B_{ik}, S_i\left(t_j^N, t\right) - \delta\right) \leq J_{2,i}^r \\ &\leq \frac{1}{r} \sum_{j=1}^{N} \sum_{k=E_i^r\left(t_{j-1}^N\right)+1}^{E_i^r\left(t_j^N\right)} f\left(B_{ik}, S_i\left(t_{j-1}^N, t\right) + \delta\right). \end{aligned}$$

Using Lemma 5.1 in [58], we obtain

$$\limsup_{r\to\infty} J_{2,i}^r \leq \lambda_i \sum_{j=1}^{N} \left(t_j^N - t_{j-1}^N\right) \mathbf{E}[f\left(B_i, S_i\left(t_{j-1}^N, t\right) + \delta\right)],$$

$$\liminf_{r\to\infty} J^r_{2,i} \geq \lambda_i \sum_{j=1}^{N} \left(t^N_j - t^N_{j-1}\right) \mathbf{E}[f\left(B_i, S_i\left(t^N_j, t\right) - \delta\right)].$$

For $s \in [\eta_i(t)+\varepsilon, t]$, the bounded convergence theorem implies that

$$\lim_{N\to\infty} \sum_{j=1}^{N} \mathbf{1}_{[t^N_{j-1}, t^N_j)}(s) \mathbf{E}[f(B_i, S_i\left(t^N_{j-1}, t\right) + \delta)] = \mathbf{E}[f\left(B_i, S_i(s,t) + \delta\right)],$$

$$\lim_{N\to\infty} \sum_{j=1}^{N} \mathbf{1}_{[t^N_{j-1}, t^N_j)}(s) \mathbf{E}[f(B_i, S_i\left(t^N_j, t\right) - \delta)] = \mathbf{E}[f\left(B_i, S_i(s,t) - \delta\right)].$$

Letting $N \to \infty$, we deduce

$$\limsup_{r\to\infty} J^r_{2,i} \leq \lambda_i \int_{\eta_i(t)+\varepsilon}^{t} \mathbf{E}[f\left(B_i, S_i(s,t) + \delta\right)] ds,$$

$$\liminf_{r\to\infty} J^r_{2,i} \geq \lambda_i \int_{\eta_i(t)+\varepsilon}^{t} \mathbf{E}[f\left(B_i, S_i(s,t) - \delta\right)] ds.$$

Passing $\delta \downarrow 0$ and $\varepsilon \downarrow 0$, we obtain because of continuity,

$$\lim_{r\to\infty} J^r_{2,i} = \lambda_i \int_{\eta_i(t)}^{t} \mathbf{E}[f(B_i, S_i(s,t))] ds. \tag{5.56}$$

If the function $f(\cdot,\cdot)$ is non-increasing in its second argument, then the inequalities in (5.55) reverse, but yield the same limit.

We now determine $\lim_{r\to\infty} J^r_{1,i}$ and $\lim_{r\to\infty} I^r_i$. Fatou's lemma and the fact that $\eta_i(t) < t$ imply

$$\liminf_{r\to\infty} \overline{S}^r_i(0,t) \geq \int_0^t \liminf_{r\to\infty} \frac{\Lambda_i(\overline{Z}^r_i(u))}{\overline{Z}^r_i(u)} du = S_i(0,t) = \infty. \tag{5.57}$$

We partition the interval $[0, \eta_i(t)]$ into M subintervals $[s^M_{j-1}, s^M_j]$, $j = 1, \ldots, M$, for some integer $M \geq 1$, in such a way that $\max_{j=1,\ldots,M}\left(s^M_j - s^M_{j-1}\right) \to 0$ as $N \to \infty$. Suppose $s^M_{j-1} \leq A^r_{ik} \leq s^M_j$ for some $j \in \{1, \ldots, M\}$, and some $k \in \{1, \ldots, E^r_i(\eta_i(t))\}$. It then follows from (5.57) that

$$\liminf_{r\to\infty} \overline{S}^r_i(A^r_{ik}, t) \geq \liminf_{r\to\infty} \overline{S}^r_i(s^M_j, t) \geq S_i(s^M_j, t) = \infty. \tag{5.58}$$

We now turn our attention to the specific functions of interest $f_1(\cdot,\cdot) = \mathbf{1}(x > y)$ and $f_2(\cdot,\cdot) = \min(x, y)$.

We first consider $\lim_{r\to\infty} J^r_{1,i}[f_2](t)$. For $t \in [0, \eta_i(t) + \varepsilon]$, we have

$$J^r_{1,i}[f_2](t) \leq \frac{1}{r} \sum_{k=1}^{E^r_i(\eta_i(t)+\varepsilon)} B_{ik},$$

$$J^r_{1,i}[f_2](t) \geq \frac{1}{r} \sum_{k=1}^{E^r_i(\eta_i(t))} \min(B_{ik}, \overline{S}^r_i(A^r_{ik}, t)) \geq \frac{1}{r} \sum_{k=1}^{E^r_i(\eta_i(t))} \min(B_{ik}, \overline{S}^r_i(s^M_j, t)).$$

It follows from (5.58) that for any L_1, $\overline{S}^r_i(s^M_j, t) > L_1$ for r sufficiently large. Applying Lemma 5.1 in [58], and letting $L_1 \to \infty$, we obtain

$$\rho_i \eta_i(t) \leq J^r_{1,i}[f_2](t) \leq \rho_i(\eta_i(t) + \varepsilon).$$

Passing $\varepsilon \to 0$, we derive

$$\lim_{r\to\infty} J^r_{1,i}[f_2](t) = \rho_i \eta_i(t) = \int_0^{\eta_i(t)} \mathbf{E}[\min(B_i, S_i(s,t))]ds. \tag{5.59}$$

We now derive $\lim_{r\to\infty} I^r_i[f_2](t)$. It follows from (5.57) that for any L_2, for r sufficiently large

$$\frac{1}{r} \sum_{l=1}^{r\overline{Z}^r_i(0)} \min(\overline{B}_{il}, L_2) \leq I^r_i[f_2](t) = \frac{1}{r} \sum_{l=1}^{r\overline{Z}^r_i(0)} \min(\overline{B}_{il}, \overline{S}^r_i(0,t)) \leq \frac{1}{r} \sum_{l=1}^{r\overline{Z}^r_i(0)} \overline{B}_{il}.$$

Multiplying and dividing $I^r_i[f_2](t)$ by $Z^r_i(0)$, and letting $L_2 \to \infty$, we deduce

$$\lim_{r\to\infty} I^r_i[f_2](t) = z_i(0)\mathbf{E}[\overline{B}_i] = z_i(0)\mathbf{E}[\min(\overline{B}_i, S_i(0,t))]. \tag{5.60}$$

Taking the sum of (5.59), (5.56) and (5.60) yields the right-hand side of (5.6). The limit on the left-hand side is $\lim_{r\to\infty} \overline{T}^r_i(t) = T_i(t)$. This proves (5.6) in case $\eta_i(t) < t$.

We now move to $\lim_{r\to\infty} J_{1,i}[f_1](t)$:

$$0 \leq J^r_{1,i}[f_1](t) \leq \frac{1}{r} \sum_{k=1}^{E^r_i(\eta_i(t))} \mathbf{1}(B_{ik} > \overline{S}^r_i(A^r_{ik}, t)) + \frac{1}{r}\left(E^r_i(\eta_i(t) + \varepsilon) - E^r_i(\eta_i(t))\right).$$

The first term on the right-hand side tends to 0 by (5.58), while the second term converges to $\lambda_i \varepsilon$ according to Lemma 5.1 in [58].

Passing $\varepsilon \to 0$, we obtain

$$\lim_{r\to\infty} J_{1,i}[f_1](t) = 0 = \lambda_i \int_0^{\eta_i(t)} \mathbf{P}(B_i > S_i(s,t))ds. \tag{5.61}$$

The term $I^r_i[f_1](t)$ follows from (5.57):

$$\lim_{r\to\infty} \frac{1}{r} \sum_{l=1}^{r\overline{Z}^r_i(0)} \mathbf{1}\left(\overline{B}_{il} > \overline{S}^r_i(0,t)\right) = 0 = z_i(0)\mathbf{P}\left(\overline{B}_i > S_i(0,t)\right). \tag{5.62}$$

Taking the sum of (5.61), (5.56) and (5.62), yields the right-hand side of (5.5). The limit on the left-hand side is $\lim_{r\to\infty} \overline{Z}^r_i(t) = z_i(t)$. This proves (5.5) in case $\eta_i(t) < t$.

In case $\eta_i(t) = t$, Equations (5.5) and (5.6) immediately follow from the fact that $\overline{S}_i^r(s,t) \to \infty$ for any $s \in [0,t]$.

It remains to treat the case when $z_i(u) > 0$ for all $u \in [0,t]$. Then $\overline{S}_i^r(u,v)$ converges uniformly to $S_i(u,v)$ for any $u, v \in [0,t]$, while the expression for $J_i^r(t)$ follows by the same argument as used for $J_{2,i}^r$ on the interval $[\eta_i(t)+\varepsilon, t]$.

For any $\varepsilon > 0$, there exists an r_ε such that $\overline{S}_i^r(0) \in (S_i(0,t)-\varepsilon, S_i(0,t)+\varepsilon)$ for all $r > r_\varepsilon$. Multiplying and dividing I_i^r by $Z_i^r(0)$, we deduce in case $f(x,y) = \min(x,y)$,

$$\limsup_{r\to\infty} \frac{1}{r} \sum_{l=1}^{r\overline{Z}_i^r(0)} f(\overline{B}_{il}, \overline{S}_i(0)) \le z_i(0)\mathbf{E}[f(\overline{B}_i, S_i(0,t)-\varepsilon)],$$

$$\liminf_{r\to\infty} \frac{1}{r} \sum_{l=1}^{r\overline{Z}_i^r(0)} f(\overline{B}_{il}, \overline{S}_i^r(0,t)) \ge z_i(0)\mathbf{E}[f(\overline{B}_i, S_i(0,t)+\varepsilon)];$$

in case $f(x,y) = \mathbf{1}(x < y)$ the reverse inequalities apply. Letting $\varepsilon \to 0$, we find that in both cases

$$\lim_{r\to\infty} I_i^r = z(0)\mathbf{E}[f(\overline{B}_i, S_i(0,t))].$$

This completes the proof. □

5.B Proof of Proposition 5.3.1

In this appendix we provide the proof of Proposition 5.3.1, which shows that the rate allocation function $\Lambda(z)$ is Lipschitz continuous on the set $\mathcal{Z}$.

We first introduce some useful notation and a definition which will play a critical role in the proof. For $x \in \mathbb{R}^I$, let $||x|| = \max_{i=1,\dots,I} |x_i|$.

Definition 5.B.1. *Consider the optimization problem*

$$\min_{x\in\Phi} f(x),$$

where $f : X \to \mathbb{R}$. Let S be a nonempty subset of the feasible set Φ such that $f(x) = f_0$ for all $x \in S$ and some $f_0 \in \mathbb{R}$. We say that the second-order growth condition *is satisfied on S if there exists a constant $\nu > 0$ and a neighborhood N of S such that, for all $x \in N \cap \Phi$, the following inequality holds:*

$$f(x) \ge f_0 + \nu[\mathrm{dist}\,(x,S)]^2, \tag{5.63}$$

where $\mathrm{dist}\,(x,S) = \min_{y\in S} ||x-y||$.

In particular, if $S = \{x_0\}$ is a singleton, then condition (5.63) takes the form

$$f(x) \ge f_0 + \nu||x-x_0||^2,$$

for any x in a feasible neighborhood of x_0.

Proof of Proposition 5.3.1
(i) We first prove Lipschitz continuity of the per-flow rate allocation. Let $\hat{x}(z) = (\hat{x}_1(z), \ldots, \hat{x}_I(z))$ be the optimal solution of problem (P) defined in Section 1.2.3 as function of $z = (z_1, \ldots, z_I)$. Introduce set $\mathcal{X}_z = \{x \in \mathbb{R}^I : x = \hat{x}(z), z \in \mathcal{Z}\}$.

Denote by $f_z(\cdot)$ the objective function $G_z(\cdot)$ taken with negative sign,

$$f_z(x) = -\sum_{i=1}^{I} z_i U_i(x_i).$$

Consider the difference function $f_z(\cdot) - f_y(\cdot)$. The derivative of the difference function with respect to x_i is $(y_i - z_i)U_i'(x_i)$. Define $c_i^* = \min_{z \in \mathcal{Z}} \hat{x}_i(z)$. Note that c_i^* is well-defined and strictly positive, as the rate $x_i(z)$ is a strictly positive continuous function on $\mathbb{R}^I_{++}$, and the set $\mathcal{Z}$ is closed and bounded. Combined with the fact that the derivative $U_i'(\cdot)$ is decreasing (because of concavity), this yields Lipschitz continuity of $f_z(\cdot) - f_y(\cdot)$ on $\mathcal{X}_z$, with the Lipschitz constant $\kappa = I||U'(c^*)||||z - y||$, with $U'(c^*) = (U_1(c_1^*), \ldots, U_I(c_I^*))$.

We now proceed to verify that the second-order growth condition is satisfied by the objective function $f_z(\cdot)$ at the optimal point $\hat{x}(z)$, i.e., that there exists a constant $\nu > 0$ such that for any x in a neighborhood of $\hat{x}(z)$ holds

$$f_z(x) - f_z(\hat{x}(z)) \geq \nu ||x - \hat{x}(z)||^2.$$

A sufficient condition for the second-order growth at a particular point is the positivity of the second derivative of the Lagrangian around this point (Theorem 3.63 in [21]). In the present problem (with linear optimization constraints) the second derivatives of the Lagrangian coincide with the second derivatives of the objective function $-z_i U_i''(x_i)$, $i = 1, \ldots, I$, which are indeed positive for any x_i. The constant ν can be determined using a Taylor series expansion around $\hat{x}(z)$. The optimality of $\hat{x}(z)$ implies

$$f_z(x) - f_z(\hat{x}) = \sum_{i=1}^{I} (z_i U_i(x_i) - z_i U_i(\hat{x}_i(z))) \geq -\frac{1}{2} \sum_{i=1}^{I} z_i U_i''(\hat{x}_i(z))(x_i - \hat{x}_i(z))^2.$$

Since $-U_i''(\cdot)$ is strictly positive and $\hat{x}(z)$ is continuous and bounded on $\mathcal{Z}$, there exists a $\sigma > 0$ such that $-U_i''(\hat{x}_i(z)) \geq \sigma$, which yields

$$f_z(x) - f_z(\hat{x}(z)) \geq \frac{\delta}{2} \sigma \sum_{i=1}^{I} (x_i - \hat{x}_i(z))^2 \geq \frac{\delta}{2} \sigma ||x - \hat{x}(z)||^2,$$

and hence the second-order growth constant is $\nu = \frac{\delta}{2}\sigma$.

Because the difference function is Lipschitz continuous with constant κ and $f_z(\cdot)$ satisfies the second-order growth condition with constant ν, Proposition 4.32 in [21] implies that the optimal solution of problem (P) satisfies

$$||\hat{x}(z) - \hat{x}(y)|| \leq \frac{\kappa}{\nu} = \frac{I||U'(c^*)||}{\nu} ||z - y||,$$

for all $y, z \in \mathcal{Z}$, i.e., the per-flow rate allocation is Lipschitz continuous on $\mathcal{Z}$ with constant $\gamma = \frac{I||U'(c^*)||}{\nu}$.

(ii) The Lipschitz continuity of the per-class rate allocation easily follows from that of the per-flow rate allocation.

Specifically, part (i) of the proof gives that

$$\left|\left|\frac{\Lambda(z)}{z} - \frac{\Lambda(y)}{y}\right|\right| \leq \gamma||z-y||, \qquad \gamma > 0.$$

The left-hand side is bounded from below by

$$\left|\left|\frac{1}{z}(\Lambda(z) - \Lambda(y))\right|\right| - \left|\left|\Lambda(y)\left(\frac{1}{y} - \frac{1}{z}\right)\right|\right|.$$

Since the norm is the maximum norm, the latter is greater than or equal to

$$\frac{1}{||z||}||\Lambda(z) - \Lambda(y)|| - ||\Lambda(y)||\left|\left|\frac{1}{z} - \frac{1}{y}\right|\right| \geq \frac{1}{M}||\Lambda(z) - \Lambda(y)|| - C\frac{1}{\delta^2}||z-y||,$$

which implies $||\Lambda(z) - \Lambda(y)|| \leq M\left(\gamma + \frac{C}{\delta^2}\right)||z-y||$. □

5.C Proof of Proposition 5.3.2

In this appendix we present the proof of Proposition 5.3.2, which shows that any fluid-limit solution that is strictly positive, must be unique in case the utility functions $U_i(\cdot)$ are twice differentiable on $\mathbb{R}^I_{++}$.

Proof of Proposition 5.3.2
Suppose, contrary to the statement of the proposition, that there are two different strictly positive solutions $z(t)$ and $h(t)$ of Equation (5.5). Let $t_0 = \inf(t : z(t) \neq h(t))$. The idea of the proof is to consider the time interval $[t_0, t_0 + t']$, for some suitably chosen $t' > 0$, and then show that $z(t) = h(t)$ for all $t \in [t_0, t_0 + t']$, yielding a contradiction with the definition of t_0.

For notational convenience, we assume that $t_0 = 0$ and consider the time interval $[0, t']$ (which does not cause any loss of generality, since we could equivalently introduce a shifted time variable $t' = t - t_0$). At this point, t' is an arbitrary positive constant. The appropriate value of this constant will be given below. By definition of t_0, we have $\varepsilon = \sup_{t\in[0,t']}||z(t) - h(t)|| > 0$. Define the function $H_i(x) = z_i(0)\mathbf{P}\left(\overline{B}_i > x\right)$. Note that $H_i(S(0,t))$ corresponds to the first term in the fluid-limit Equation (5.5) representing the number of initial class-i flows that are still active at time t.

Using simple estimates, we have for each class $i = 1, \ldots, I$,

$$|z_i(t) - h_i(t)| \leq \left|H_i(S_i(0,t)) - H_i(\tilde{S}_i(0,t))\right|$$

$$+\lambda_i \int_0^t \left| \mathbf{P}\left(B_i > S_i(s,t)\right) - \mathbf{P}\left(B_i > \tilde{S}_i(s,t)\right) \right| ds := A_1 + A_2,$$

with $\tilde{S}_i(s,t) = \int_s^t \hat{x}_i(h(u))du$.

We first derive an upper bound for the term A_1.

By definition of the residual flow size, we have $|H_i(x) - H_i(x')| \leq z_i(0)\mu_i|x - x'|$ for any $x, x' \in \mathbb{R}_+$. Thus, the function $H_i(\cdot)$, $i = 1, \ldots, I$, is Lipschitz continuous with Lipschitz constant $L_i = z_i(0)\mu_i$.

In order to bound the difference $|S_i(0,t) - \tilde{S}_i(0,t)|$, we first consider the integrands in the definition of the functions $S_i(\cdot,\cdot)$ and $\tilde{S}_i(\cdot,\cdot)$. A fluid-limit solution is bounded from above on the interval $[0,t']$ by

$$\sup_{i=1,\ldots,I, u\in[0,t']} z_i(u) \leq ||z_0|| + t' \sum_{i=1}^{I} \lambda_i.$$

Thus, Proposition 5.3.1 implies that the rate allocation functions are Lipschitz continuous with some constant γ:

$$|\hat{x}_i(z(u)) - \hat{x}_i(h(u))| \leq \gamma ||z(u) - h(u)|| \leq \gamma\varepsilon. \tag{5.64}$$

It follows that

$$\left| S_i(0,t) - \tilde{S}_i(0,t) \right| \leq \int_0^t |\hat{x}_i(z(u)) - \hat{x}_i(h(u))|\, du \leq \gamma\varepsilon t \leq \gamma\varepsilon t', \tag{5.65}$$

and hence,

$$A_1 \leq L_i \gamma \varepsilon t'.$$

We apply similar arguments to obtain an upper bound for the term A_2.

Since

$$\tilde{S}_i(s,t) \leq S_i(s,t) + \gamma\varepsilon(t-s),$$

we have

$$A_2 \leq \lambda_i \int_0^t \mathbf{P}\left(S_i(s,t) < B_i < S_i(s,t) + \gamma\varepsilon(t-s)\right) ds.$$

Denote by $S_i^{-1}(r)$, $r > 0$, the inverse of $S_i(0,t)$, i.e., $S_i^{-1}(r) = \inf\{s : S_i(0,s) \geq r\}$. Because $z(t)$ is strictly positive for all t, $S_i(0,s)$ is strictly increasing in s, implying $S_i^{-1}(r)$ is well-defined and finite for all r.

For $u, v > 0$, consider now the integral

$$\int_0^t \mathbf{P}\left(S_i(t) - S_i(s) + u < B_i < S_i(t) - S_i(s) + v\right) ds$$

$$= \int_0^{S_i(t)} \mathbf{P}\left(r + u < B_i < r + v\right) \frac{1}{x_i^*(z(S_i^{-1}(r)))} dr.$$

Define $c_i = \inf_{r\in[0,S_i(t)]} \hat{x}_i(z(S_i^{-1}(r)))$. Note that c_i is strictly positive and well-defined since $S_i^{-1}(\cdot)$ is finite and $z(t)$ and $x_i^*(z)$ are continuous functions. Thus, the above integral is less than or equal to

$$\begin{aligned}
&\int_0^{S_i(t)} \mathbf{P}\left(r+u<B_i<r+v\right)\frac{1}{c_i}dr\\
\leq\ & \frac{1}{c_i}\int_0^{\infty} \mathbf{P}\left(r+u<B_i<r+v\right)dr\\
=\ & \frac{1}{c_i}\int_0^{\infty}\int_{r+u}^{r+v} d\mathbf{P}\left(B\leq w\right)dr\\
=\ & \frac{1}{c_i}\int_0^{\infty}\int_{r-v}^{r-u} drd\mathbf{P}\left(B\leq w\right)\\
\leq\ & \frac{1}{c_i}|u-v|, \qquad (5.66)
\end{aligned}$$

which implies

$$A_2 \leq \frac{\lambda_i}{c_i}\gamma\varepsilon t'.$$

Summing the bounds for A_1 and A_2, we obtain

$$|z_i(t)-h_i(t)| \leq A_1+A_2 \leq \left(L_i+\frac{\lambda_i}{c_i}\right)\gamma\varepsilon t'.$$

Taking

$$t' = \min_{i=1,\ldots,I}\frac{1}{2(L_i+\frac{\lambda_i}{c_i})\gamma},$$

we have that for any $i=1,\ldots,I$, $|z_i(t)-h_i(t)|\leq\frac{\varepsilon}{2}$, and hence $||z(t)-h(t)||\leq\frac{\varepsilon}{2}$. This contradicts the original supposition that $\sup_{t\in[0,t']}||z(t)-h(t)||=\varepsilon$, and implies that $z(t)=h(t)$ for all $t\in[0,t']$. □

5.D Proof of Lemma 5.4.2

In this section we provide the proof of Lemma 5.4.2, which shows that Equation (5.12) has a unique solution $m=(m_1,\ldots,m_I)$ in the set $\mathcal{M}(C)$.

Proof of Lemma 5.4.2
A crucial role in the proof is played by a related optimization problem. To formulate this optimization problem, we rewrite the fixed-point Equation (5.12) in the equivalent form

$$\frac{m_i}{R_i(m)} = \beta_i^{-1}\left(\frac{R_i(m)}{\rho_i}\right), \qquad i=1,\ldots,I, \qquad (5.67)$$

where $\beta_i^{-1}(\cdot)$ is the inverse of the LST $\beta_i(y)=\mathbf{E}\left[\mathrm{e}^{-yB_i^*}\right]$, $\beta_i^{-1}(\beta_i(y))=y$. We will establish uniqueness of the rate allocation vector $R(m)=(R_1(m),\ldots,R_I(m))$

subject to the constraint $AR(m) \leq C$. By (5.67), uniqueness of solution $R(m)$ would imply uniqueness of the growth rate vector $m = (m_1, \ldots, m_I)$.

Observe that the right-hand side only depends on m through $R_i(m)$. This motivates us to introduce the function $G_i : [0, \rho_i] \to \mathbb{R}$ with derivative $G_i'(x) = U_i'\left(\frac{1}{\beta_i^{-1}\left(\frac{x}{\rho_i}\right)}\right)$. Since $\beta_i(x)$ is strictly decreasing in x, its inverse is strictly decreasing in x as well. Because of the concavity of $U_i(\cdot)$, we thus conclude that $G_i(\cdot)$ is strictly concave.

Now consider the optimization problem Q as defined in (5.15). This optimization problem is strictly concave, and hence has a unique solution $R = (R_1^*, \ldots, R_I^*)$ (see for instance [27]).

We proceed to show that the rate allocation vector $R(m) = (R_1(m), \ldots, R_I(m))$ satisfying Equation (5.67) is the unique solution to the optimization problem (Q).

First recall that $R_i(m) = \Lambda_i(m)$ when $m_i > 0$ while $R_i(m) = \rho_i$ when $m_i = 0$, and that $AR(m) \leq C$ so that $R(m)$ is a feasible solution.

Also, the rate allocation vector $\Lambda(m)$ is a solution to the optimization problem:

$$(P') \qquad \begin{array}{ll} \text{maximize} & \sum_{i=1}^I m_i U_i\left(\frac{\Lambda_i}{m_i}\right) \\ \text{subject to} & A\Lambda \leq C,\ \Lambda \geq 0. \end{array}$$

Let us consider the Karush-Kuhn-Tucker (KKT) necessary conditions [9] for problem (P'). As $\Lambda(m)$ is an optimal solution, there exist Lagrange multipliers $p_j(m) \geq 0$ such that

$$U_i'\left(\frac{\Lambda_i(m)}{m_i}\right) = \sum_{j=1}^J A_{ji} p_j(m), \qquad \text{if } m_i > 0, \tag{5.68}$$

$$p_j(m)\left(\sum_{i=1}^I A_{ji}\Lambda_i(m) - C_j\right) = 0, \qquad j = 1, \ldots, J. \tag{5.69}$$

Let us further consider the KKT sufficient conditions for problem (Q). A feasible solution $R^* = (R_1^*, \ldots, R_I^*)$ is a global optimum if there exist Lagrange multipliers $p_j^*, q_i^* \geq 0$ such that

$$U_i'\left(\frac{1}{\beta_i^{-1}\left(\frac{R_i^*}{\rho_i}\right)}\right) = \sum_{j=1}^J A_{ji} p_j^* + q_i^*, \qquad i = 1, \ldots, I, \tag{5.70}$$

$$q_i^*\left(R_i^* - \rho_i\right) = 0, \qquad i = 1, \ldots, I, \tag{5.71}$$

$$p_j^*\left(\sum_{i=1}^I A_{ji} R_i^* - C_j\right) = 0, \qquad j = 1, \ldots, J. \tag{5.72}$$

Note that $\sum_{i=1}^{I} A_{ji}R_i(m) = \sum_{i:m_i>0} A_{ji}R_i(m)+\sum_{i:m_i=0} A_{ji}\rho_i = \sum_{i:m_i>0} A_{ji}\Lambda_i(m)+\sum_{i:m_i=0} A_{ji}\Lambda_i(m) + \sum_{i:m_i=0} A_{ji}\rho_i = \sum_{i=1}^{I} A_{ji}\Lambda_i(m) + \sum_{i:m_i=0} A_{ji}\rho_i$. Further using that $\sum_{i=1}^{I} A_{ji}R_i(m) \le C_j$, Equation (5.69) yields

$$p_j(m)\left(\sum_{i=1}^{I} A_{ji}R_i(m) - C_j\right) = 0, \qquad j = 1,\dots,J. \tag{5.73}$$

In addition, if $m_i = 0$, then for any j with $A_{ji} = 1$, we have the strict inequality

$$\sum_{i=1}^{I} A_{ji}\Lambda_i(m) < C_j,$$

and thus Equation (5.69) forces $p_j(m) = 0$. We deduce that

$$\sum_{j=1}^{J} A_{ji}p_j(m) = 0, \qquad \text{if } m_i = 0. \tag{5.74}$$

Further define $q_i(m) = 0$ when $m_i > 0$ and $q_i(m) = U_i'(\infty) \ge 0$ when $m_i = 0$, so that

$$q_i(m)\,(R_i(m) - \rho_i) = 0, \qquad i = 1,\dots,I. \tag{5.75}$$

Using the fixed-point Equation (5.67), the definition of $q_i(m)$, and Equations (5.68) and (5.74), we obtain

$$U_i'\left(\frac{1}{\beta_i^{-1}\left(\frac{R_i(m)}{\rho_i}\right)}\right) = \sum_{j=1}^{J} A_{ji}p_j(m) + q_i(m), \qquad i = 1,\dots,I. \tag{5.76}$$

Equations (5.73), (5.75) and (5.76) yield that $R(m)$, together with $p_j(m)$, $q_i(m)$, satisfies the KKT sufficient conditions (5.70)–(5.72) for problem (Q), and hence is a global optimum. □

5.E Properties of tree networks

Consider a tree network. By Definition 5.5.1 we can partition the network into m subtrees which are connected to the root link. The vectors of rate allocations and the capacity constraints in the k-th subtree, $k = 1,\dots,m$, are given by $R^{(k)}$ and $C^{(k)}$, respectively.

Proof of Proposition 5.5.1
We first prove that the weighted α-fair rate allocation in a tree network may be obtained using a weighted version of the so-called water-filling procedure. In this procedure, the allocation to each of the class-i flows is continuously increased at rate $w_i^{1/\alpha}$ until it is no longer feasible to do so, i.e., until the capacity of one of the links

along the route of class i is exhausted. The class-i flows then collectively drop out, and their rate allocation remains frozen from that point onward. The procedure continues until eventually all flows have dropped out.

The water-filling procedure may be formally described as follows. We first introduce some convenient terminology. A link is said to saturate once its capacity is exhausted by the aggregate rate of the flows traversing it. A link is called locked out when all the flows traversing it have dropped out. In particular, a link gets locked out once it saturates. Note that in a tree network a link gets locked out as soon as one of its upstream links saturates, but it may also get locked out if all the flows traversing it have dropped out due to saturation of one or several of its downstream links. Link j is said to be downstream from link j' (and link j' is upstream from j) if the (unique) path from link j to the root link contains link j', with the convention that a link is neither upstream nor downstream from itself. Let $\mathcal{D}_j$ be the set of links that are downstream from link j.

We now define some additional useful notation. Let $z = (z_1, \ldots, z_I)$ be the population vector, with z_i the number of class-i flows. Denote by $r^{(k)}$ the k-th 'time epoch' at which a link or a group of links, represented by the set $\mathcal{F}^{(k)}$, saturate, and denote by $\mathcal{G}^{(k)}$ the set of classes that drop out at that point. Define $\mathcal{I}^{(k+1)} := \mathcal{I}^{(k)} \cup \mathcal{G}^{(k+1)}$, with $\mathcal{I}^{(0)} = \emptyset$, as the set of classes that have dropped out by time $r^{(k+1)}$. Define $\mathcal{J}^{(k+1)} := \mathcal{J}^{(k)} \cup \mathcal{F}^{(k+1)}$, with $\mathcal{J}^{(0)} = \emptyset$, as the set of links that have saturated by time $r^{(k+1)}$. Then $\mathcal{G}^{(k+1)} = \{i \notin \mathcal{I}^{(k)} : \sum_{j \in \mathcal{F}^{(k)}} A_{ij} \geq 1\}$ and $\mathcal{I}^{(k)} = \{i : \sum_{j \in \mathcal{J}^{(k)}} A_{ij} \geq 1\}$. Define $\mathcal{L}^{(k)} := \{j : \sum_{i \notin \mathcal{I}^{(k)}} A_{ij} z_i = 0\}$ as the set of links that have been locked out by time $r^{(k)}$. Define $\mathcal{H}^{(k+1)} := \mathcal{L}^{(k+1)} \setminus \mathcal{L}^{(k)}$ as the set of links that become locked out at time $r^{(k+1)}$.

Define $C_j^{(k)}$ as the residual capacity of link j at time $r^{(k)}$, with $r^{(0)} = 0$ and $C_j^{(0)} = C_j$. As long as $\mathcal{L}^{(k)} \neq \mathcal{J}$, we recursively compute

$$\Delta_j^{(k+1)} = \frac{C_j^{(k)}}{\sum_{i \notin \mathcal{I}^{(k)}} A_{ij} z_i w_i^{1/\alpha}},$$

for all $j \notin \mathcal{L}^{(k)}$,

$$\Delta^{(k+1)} = \min_{j \notin \mathcal{L}^{(k)}} \Delta_j^{(k+1)},$$

$$\mathcal{G}^{(k+1)} = \arg \min_{j \notin \mathcal{L}^{(k)}} \Delta_j^{(k+1)}.$$

Also, for all $j \notin \mathcal{L}^{(k)}$,

$$\begin{aligned} C_j^{(k+1)} &= C_j^{(k)} - \sum_{i \notin \mathcal{I}^{(k)}} A_{ij} z_i w_i^{1/\alpha} \Delta^{(k)} \\ &= C_j - \sum_{l=1}^{k} r^{(l)} \sum_{i \in \mathcal{G}^{(l)}} A_{ij} z_i w_i^{1/\alpha} - r^{(k+1)} \sum_{i \notin \mathcal{I}^{(k)}} A_{ij} z_i w_i^{1/\alpha} \end{aligned}$$

and for all $j \in \mathcal{F}^{(k+1)}$

$$r^{(k+1)} = r^{(k)} + \Delta^{(k+1)} = \frac{C_j - \sum_{l=1}^{k} r^{(l)} \sum_{i \in \mathcal{G}^{(l)}} A_{ij} z_i w_i^{1/\alpha}}{\sum_{i \notin \mathcal{I}^{(k)}} A_{ij} z_i w_i^{1/\alpha}}.$$

Denote by $K := \min\{k : \mathcal{L}^{(k)} = \mathcal{J}\}$ the number of iterations, and note that $K \leq J$ since at least one link must saturate at each iteration.

For later use, define the set of classes that have dropped out by time t as $\mathcal{I}(t) = \mathcal{I}^{(k^*(t))}$, the set of links that have saturated by time t as $\mathcal{J}(t) = \mathcal{J}^{(k^*(t))}$, and the set of links that have been locked out by time t as $\mathcal{L}(t) = \mathcal{L}^{(k^*(t))}$, with $k^*(t) := \max\{k : r^{(k)} \leq t\}$ representing the number of iterations up to time t. Thus $\mathcal{L}(t) = \mathcal{L}^{(k)}$, $\mathcal{I}(t) = \mathcal{I}^{(k)}$ for $t \in [r^{(k)}, r^{(k+1)})$, $k = 1, \ldots, K$, and $\mathcal{L}(t) = \mathcal{J}$, $\mathcal{I}(t) = \mathcal{I}$ for $t \geq r^{(K)}$.

The 'time' that class $i \in \mathcal{G}^{(k)}$ drops out is $t_i = r^{(k)}$, and the time that link $j \in \mathcal{H}^{(k)}$ gets locked out is $s_j = r^{(k)}$. Note that

$$s_j = \sup\{t : j \notin \mathcal{L}(t)\} = \max_{i:A_{ij}=1} t_i, \tag{5.77}$$

and

$$t_i = \sup\{t : i \notin \mathcal{I}(t)\} = \min_{j:A_{ij}=1} s_j.$$

The rate allocation of each class-i flow is

$$x_i(z) = t_i w_i^{1/\alpha}. \tag{5.78}$$

Also, denote by $x(u)$, with

$$x_i(u) = \sup\{t \in [0, u] : i \notin \mathcal{I}(u)\} w_i^{1/\alpha}$$

the rate allocation vector at time u.

For convenience, we henceforth assume that only a single link saturates at a time, i.e., $\mathcal{F}^{(k)} = \{j^{(k)}\}$, say, for all $k = 1, \ldots, K$, but the arguments below may be easily extended to the case where several links may saturate simultaneously.

For $k, l = 1, \ldots, K$, let the 0–1 variable $D_{kl} = \mathbf{1}_{\{j^{(k)} \in \mathcal{D}_{j^{(l)}}\}}$ indicate whether link $j^{(k)}$ is downstream from link $j^{(l)}$ or not. Observe that $D_{kl} = 0$ when $k > l$: link $j^{(k)}$ cannot be upstream from link $j^{(l)}$ since once a link saturates all its downstream links are locked out and can no longer saturate at a later stage. Formally, $k > l$ implies $j^{(k)} \notin \mathcal{L}^{(l)}$, while $\mathcal{D}_{j^{(l)}} \subseteq \mathcal{L}^{(l)}$, and hence $j^{(k)} \notin \mathcal{D}_{j^{(l)}}$, i.e., $D_{kl} = 0$.

Recursively define

$$p_{j^{(K)}} = (r^{(K)})^{-\alpha},$$

$$p_{j^{(K-1)}} = (r^{(K-1)})^{-\alpha} - D_{K-1,K} p_{j^{(K)}},$$

up to

$$p_{j^{(1)}} = (r^{(1)})^{-\alpha} - D_{1,2} p_{j^{(2)}} - \cdots - D_{1,K} p_{j^{(K)}},$$

i.e.,

$$p_{j^{(k)}} = (r^{(k)})^{-\alpha} - \sum_{l=k+1}^{K} D_{k,l} p_{j^{(l)}},$$

$k = 1, \ldots, K$.

If $\sum_{l=k+1}^{K} D_{k,l} \geq 1$, then define $I_k = 1$ and $m_k := \min\{l : D_{k,l} = 1\} > k$. Note that $D_{k,l} = 0$ for $l = k+1, \ldots, m_k - 1$, and $D_{k,l} = D_{m_k,l} = 1$ for $l = m_k + 1, \ldots, K$ because of transitivity: if link $j^{(m_k)}$ is downstream from link $j^{(l)}$, then link $j^{(k)}$ must be downstream from it as well. Formally, if $m \in \mathcal{D}_l$ and $k \in \mathcal{D}_m$, then $k \in \mathcal{D}_l$.

Thus we may write

$$\begin{aligned}
p_{j^{(k)}} &= (r^{(k)})^{-\alpha} - \sum_{l=k+1}^{K} D_{k,l} p_{j^{(l)}} \\
&= (r^{(k)})^{-\alpha} - I_k (p_{j^{(m_k)}} + \sum_{l=m_k+1}^{K} D_{k,l} p_{j^{(l)}}) \\
&= (r^{(k)})^{-\alpha} - I_k (p_{j^{(m_k)}} + \sum_{l=m_k+1}^{K} D_{m_k,l} p_{j^{(l)}}) \\
&= (r^{(k)})^{-\alpha} - I_k (r^{(m_k)})^{-\alpha} \\
&> 0.
\end{aligned}$$

Also, define $p_j = 0$ for all $j \neq j^{(1)}, \ldots, j^{(K)}$.

By construction,

$$p_j (C_j - \sum_{i=1}^{I} A_{ij} z_i x_i(z)) = 0$$

for all $j \in \mathcal{J}$.

Note that if $i \in \mathcal{G}_k$, then $i \notin \mathcal{I}^{(k-1)}$, i.e., $\sum_{j \in \mathcal{J}^{(k-1)}} A_{ij} = 0$, implying that $A_{ij^{(l)}} = 0$ for all $l = 1, \ldots, k-1$. Also, by definition, $A_{ij^{(k)}} = 1$, and $A_{ij^{(l)}} = D_{j^{(k)},j^{(l)}}$ for $l = k+1, \ldots, K$. Thus we obtain

$$\begin{aligned}
w_i U'(x_i(z)) &= w_i (w_i^{1/\alpha} r^{(k)})^{-\alpha} = (r^{(k)})^{-\alpha} = p_{j^{(k)}} + \sum_{l=k+1}^{K} D_{k,l} p_{j^{(l)}} \\
&= \sum_{l=1}^{K} A_{ij^{(l)}} p_{j^{(l)}} = \sum_{j=1}^{J} A_{ij} p_j
\end{aligned}$$

for all $i \in \mathcal{G}_k$, $k = 1, \ldots, K$.

In conclusion, the rate allocation produced by the water-filling procedure satisfies the KKT conditions for the weighted α-fair utility maximization problem.

We now proceed to show the stated monotonicity property, and attach the two population vectors y and z, with $y \leq z$, as subscripts to the various sets and variables

as defined above. We will prove that $\mathcal{L}_y(t) \subseteq \mathcal{L}_z(t)$ for all $t \geq 0$, i.e., if any link is locked out for y, then it is locked out for z as well. In view of (5.77), this is equivalent to $s_{j,y} \geq s_{j,z}$ for all $j \in \mathcal{J}$, which in turn is equivalent to $t_{i,y} \geq t_{i,z}$ for all $i \in \mathcal{I}$, and hence (5.78) yields $x_i(y) \geq x_i(z)$ for all $i \in \mathcal{I}$.

In order to establish the above property, we will use induction on the number of links J. In case $J = 1$, the water-filling procedure terminates after $K = 1$ iteration, and $t_i = s_1 = r^{(1)}$ for all $i \in \mathcal{I}$, with

$$r_y^{(1)} = \frac{C^0}{\sum_{i=1}^I y_i} \geq \frac{C^0}{\sum_{i=1}^I z_i} = r_z^{(1)},$$

so $s_{1,y} \geq s_{1,z}$ and $t_{i,y} \geq t_{i,z}$ for all $i \in \mathcal{I}$.

Now assume that the above property holds for any tree network with at most $J \geq 1$ links. We will consider a tree network with $J + 1$ links, with the root link labeled as 1, and show that the property holds as well.

Suppose that were not the case, and let u be the first time epoch at which the property is violated, i.e., $\mathcal{L}_y(t) \subseteq \mathcal{L}_z(t)$ for all $t \in [0, u)$, but it is *not* the case that $\mathcal{L}_y(u) \subseteq \mathcal{L}_z(u)$. In other words, there must exist some link $j^* \in \mathcal{L}_y(u) \setminus \mathcal{L}_y(u^-)$, $j^* \notin \mathcal{L}_z(u)$, which gets locked out for y at time u, but is not yet locked out for z. In fact, we may impose $j^* \in \mathcal{J}_y(u) \setminus \mathcal{J}_y(u^-)$ as some link must saturate in order for any link to get locked out.

Now consider a reduced network that is comprised of the links indexed by the set $\mathcal{J}_y(u^-) \cup \mathcal{J}_z(u^-) \cup \{j^*\}$ that have saturated for either y or z up to time u, along with link j^*. If the water-filling procedure is executed in this network, it will take identical actions up to time u, and thus link j^* will saturate at time u in case of y, but not in case of z. If the number of links in the reduced network were J or less, then the latter event cannot occur by virtue of the induction hypothesis. Thus the reduced network must contain $J+1$ links, i.e., $|\mathcal{J}_y(u^-) \cup \mathcal{J}_z(u^-) \cup \{j^*\}| = J+1$, which implies $|\mathcal{J}_z(u^-) \cup \mathcal{J}_y(u^-)| \geq J$. Since $\mathcal{J}_y(u^-), \mathcal{J}_z(u^-) \subseteq \mathcal{L}_z(u^-)$, and $j^* \notin \mathcal{L}_z(u^-)$, it follows that $\mathcal{L}_z(u^-) = \mathcal{J}$ and $j^* = 1$, i.e., the root link saturates at time u for y, and all links except the root link have saturated by time u for z. Let $\mathcal{E} = \{i \in \mathcal{I} : \sum_{j \neq 1} A_{ij} \geq 1\}$ be the set of the classes that traverse at least one other link besides the root link. The fact that the root link saturates at time u for y but not for z, implies:

$$\sum_{i=1}^I y_i x_i(y; u) = C^0 > \sum_{i=1}^I z_i x_i(z; u), \tag{5.79}$$

and

$$x_i(y) = x_i(y; u) = u = x_i(z; u) \leq x_i(z) \qquad \text{for all } i \notin \mathcal{E}. \tag{5.80}$$

Now consider a network without the root link and the classes that do not belong to the set $\mathcal{E}$. Let $x^{\mathcal{E}}(y)$ and $x^{\mathcal{E}}(z)$ be the optimal rate allocation vectors in the reduced network for $(y_i)_{i \in \mathcal{E}}$ and $(z_i)_{i \in \mathcal{E}}$, respectively. Let $x^{\mathcal{E}}(y; t)$ and $x^{\mathcal{E}}(z; t)$ be the rate allocation vectors produced by the water-filling procedure at time t in the reduced network for $(y_i)_{i \in \mathcal{E}}$ and $(z_i)_{i \in \mathcal{E}}$, respectively.

Now observe that the water-filling procedure will follow identical steps in the original network and in the reduced network up to time u, so that

$$x_i^{\mathcal{E}}(y;u) = x_i(y;u), \qquad x_i^{\mathcal{E}}(z;u) = x_i(z;u) \qquad \text{for all } i \in \mathcal{E}.$$

Noting that $x^{\mathcal{E}}(y) \geq x^{\mathcal{E}}(y;u)$, we have

$$x_i^{\mathcal{E}}(y) \geq x_i(y;u) \qquad \text{for all } i \in \mathcal{E}. \tag{5.81}$$

Furthermore, $x^{\mathcal{E}}(z) = x^{\mathcal{E}}(z;u)$, because all links except the root link have saturated for z by time u, so that the water-filling procedure terminates in the reduced network, and we find

$$x_i^{\mathcal{E}}(z) = x_i(z;u) \qquad \text{for all } i \in \mathcal{E}. \tag{5.82}$$

Since the reduced network is again a tree network, or a collection of tree networks, it is rate-preserving, and hence

$$\sum_{i\in\mathcal{E}} y_i x_i^{\mathcal{E}}(y) \leq \sum_{i\in\mathcal{E}} z_i x_i^{\mathcal{E}}(z). \tag{5.83}$$

Using Equations (5.79)–(5.83), we obtain

$$\begin{aligned}
C^0 &= \sum_{i=1}^{I} y_i x_i(y;u) \\
&= \sum_{i\in\mathcal{E}} y_i x_i(y;u) + \sum_{i\notin\mathcal{E}} y_i x_i(y;u) \\
&\leq \sum_{i\in\mathcal{E}} y_i x_i^{\mathcal{E}}(y) + \sum_{i\notin\mathcal{E}} y_i u \\
&\leq \sum_{i\in\mathcal{E}} z_i x_i^{\mathcal{E}}(z) + \sum_{i\notin\mathcal{E}} z_i x_i(z;u) \\
&= \sum_{i\in\mathcal{E}} z_i x_i(z;u) + \sum_{i\notin\mathcal{E}} z_i x_i(z;u) \\
&= \sum_{i=1}^{I} z_i x_i(z;u) \\
&< C^0,
\end{aligned}$$

which yields a contradiction. Thus the stated monotonicity property also holds in a tree network with $J+1$ links, which completes the proof. □

Proof of Proposition 5.5.2

The proof is by induction on the number of links in a tree network. By definition, the aggregate rate is computed as

$$\sum_{i=1}^{I} R_i(z) = \sum_{i:z_i>0} \Lambda_i(z) + \sum_{i:z_i=0} \rho_i(z).$$

If $J = 1$, the network reduces to a single work-conserving node: $\sum_{i:z_i>0}^{I} \Lambda_i(z) = 1$ if $z \neq 0$. For z in the set $\mathcal{M}(C)$, this leaves two possible cases: $z > 0$ and $\sum_{i=1}^{I} R_i(z) = \sum_{i=1}^{I} \Lambda_i(z) = 1$, or $z = 0$ and $\sum_{i=1}^{I} R_i(z) = \sum_{i=1}^{I} \rho_i(z) \leq 1$. Consequently, for any $z, y \in \mathcal{M}(C)$, $z \geq y$, we obtain $\sum_{i=1}^{I} R_i(z) \geq \sum_{i=1}^{I} R_i(y)$.

Suppose now that tree networks with at most J links are rate-preserving. Consider a tree with $J+1$ links. Since $z \in \mathcal{M}(C)$, $\sum_{i=1}^{I} R_i(z) \leq C^0$. If $\sum_{i=1}^{I} R_i(z) = C^0$, the inequality holds trivially. Suppose now $\sum_{i=1}^{I} R_i(z) < C^0$. This can only occur if there are no classes that traverse the root link only, or these classes have no flows. So we are in one of two cases: (i) the matrix B in Definition 5.5.1 is empty, or equivalently, the set $\mathcal{E}$ defined in the proof of Proposition 5.5.1 includes all classes $i = 1, \ldots, I$; (ii) the matrix B is not empty, but $z_i = 0$ for all classes $i \notin \mathcal{E}$. Since the root link is not saturated, the water-filling procedure as described in the proof of Proposition 5.5.1 produces the same rate allocation, regardless of whether the root link is present or not. In other words, the rate allocation $R(z)$ can be represented as a vector of rate allocations derived independently for each subtree,

$$\begin{aligned} &(i) \quad R(z) = (R^{(1)}(z^{(1)}), \ldots, R^{(m)}(z^{(m)})), \\ \text{or} \\ &(ii) \quad R(z) = (R^{(1)}(z^{(1)}), \ldots, R^{(m)}(z^{(m)}), R^{(B)}(0)), \end{aligned}$$

where $R^{(B)}(0) = (\rho_i)_{i \notin \mathcal{E}}$.

Note that $z, y \in \mathcal{M}(C)$ yields $z^{(k)}, y^{(k)} \in \mathcal{M}^{(k)}(C^{(k)})$, that is $A^{(k)} R^{(k)}(z^{(k)}) \leq C^{(k)}$, while $z \geq y$ implies $z^{(k)} \geq y^{(k)}$ for all $k = 1, \ldots, m$. By the induction assumption, the subtrees are rate-preserving, and hence

$$\sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(z^{(k)}) \geq \sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(y^{(k)}),$$

where $I^{(k)} = \{l \in \{1, 2, \ldots, I\} : \text{ class } l \text{ belongs to } k\text{−th subtree}\}$. Consequently, in case (i),

$$\sum_{i=1}^{I} R_i(z) = \sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(z^{(k)}) \geq \sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(y^{(k)}) = \sum_{i=1}^{I} R_i(y).$$

Observing that $z_i = 0$ implies $y_i = 0$, we obtain in case (ii),

$$\sum_{i=1}^{I} R_i(z) = \sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(z^{(k)}) + \sum_{i \notin \mathcal{E}} \rho_i \geq \sum_{k=1}^{m} \sum_{i \in I^{(k)}} R_i^{(k)}(y^{(k)}) + \sum_{i \notin \mathcal{E}} \rho_i = \sum_{i=1}^{I} R_i(y).$$

□

Chapter 6

Sojourn time asymptotics in a parking lot network

In the present chapter we investigate the asymptotic behavior of the sojourn time for a specific type of bandwidth-sharing network. We focus on a two-link parking lot network as considered in Section 5.6. Although the network topology is seemingly simple, it reveals a few characteristic properties which may be expected to hold in more complex scenarios, and hopefully provides the guidelines for further investigations.

Results for a parking lot network are especially interesting, since such a network operating under a utility-maximizing policy is known [17] to be sensitive in the sense that the stationary distribution of the number of flows depends on the flow size distribution, and not just the mean flow size. For networks with sensitive allocation policies the available results are mostly restricted to approximations for the performance measures of interest rather than an explicit characterization. In [19], Bonald and Proutière studied monotone networks under assumption of Poisson arrivals and derived insensitive lower and upper bounds for the number of flows in the system by means of sample-path comparisons. Massoulié [83] introduced a novel rate allocation policy, which is a modification of the proportional fair policy but is insensitive and has an explicit steady-state distribution. This policy is discussed in further detail later in this chapter.

To the best of our knowledge, sensitive bandwidth-sharing networks have not been studied from a large-deviations perspective. This chapter reports on some first steps in our study of the asymptotic behavior of the number of flows, the workload and the sojourn time in a two-link parking lot network. The derivation of the logarithmic delay asymptotics in the present chapter can be considered as an extension of Chapter 4 to a network scenario. With the network model we face several complications such as the sensitivity of the distribution of the number of flows and the non-work-conserving behavior. The analysis can be split into two main steps:

1. In order to overcome the sensitivity issue, we utilize the results for the modified proportional fair allocation by Massoulié [83] and derive bounds for the

distribution of the number of flows and the workload. Using the monotonicity of the network, we derive upper bounds for the steady-state distribution of the number of flows. The upper bound is completely determined by the traffic loads and the capacity constraints and does not depend on other characteristics of the system. In a similar manner, we show that the workload in the modified system forms a bound for the workload in the original system. Further, we examine finiteness of the MGF of the workload vector.

2. With the above results at hand, we apply large-deviations techniques similar to those in Chapter 4 to derive the logarithmic asymptotics for the sojourn time distribution. The derivation of the large-deviations upper bound is based on the Chernoff bounds, while the derivation of the lower bound requires a change-of measure argument and the fluid-limit results from Chapter 5. We change the interarrival time and flow size distributions in such a way that the root link becomes overloaded. At that stage we invoke the fluid-limit results for the number of flows in an overloaded tree network derived in the previous chapter.

The main result of this chapter concerns the logarithmic asymptotics for the sojourn time distribution of the class which traverses the root link 1 only and competes for bandwidth with flows of class 2. The rate allocation to class-2 flows is bounded by the capacity of link 2. The obtained asymptotics indicate that there are two qualitatively different scenarios for the large-deviations behavior. If the tilted load of class 2 is strictly less than the capacity of link 2, the system asymptotically behaves as a single-link DPS system. In this first case, the decay rate of the class-1 sojourn time coincides with the decay rate in a two-class DPS node as analyzed in Section 4.5. The second scenario corresponds to the situation when the capacity constraint is binding. In this case, class 2 is allocated the full capacity of link 2 while class 1 receives the remaining bandwidth at link 1. The decay rate is then composed of two decay rates in independent PS systems with corresponding flow classes and service rates. The result shows that in both scenarios a large sojourn time is due to a large amount of work generated by both flow classes during the service of the flow under consideration.

This chapter is organized as follows. In Section 6.1 we present a detailed model description and discuss the modified proportional fair allocation. In Section 6.2 we introduce the necessary notation. In Sections 6.3 and 6.4 we derive the bounds for the number of flows in the system and the workload, respectively. The main result is presented in Section 6.5. Finally, Section 6.6 contains suggestions for further research.

6.1 Model description

We consider a parking lot network as described in Section 5.6. The network consists of two links and two classes of flows. The capacities of the links are given

by $c_1 = 1$ and $c_2 = c < 1$. The route of class 1 consists of link 1, and the route of class 2 includes both links. We refer to Figure 5.2 for an illustration.

We assume that class-i flows arrive according to a Poisson process with rate λ_i. The flow sizes have a general distribution with bounded hazard rates (see Assumption 5.5.1 in the previous chapter). The mean flow sizes are given by $1/\mu_i$, $i = 1, 2$. The traffic load of class i is $\rho_i = \lambda_i/\mu_i$. We define the capacity set $\mathcal{C} = \{r \in \mathbb{R}_+^2 : r_1 + r_2 \leq 1, r_2 \leq c\}$, and assume that the vector (ρ_1, ρ_2) lies in the interior of $\mathcal{C}$, i.e.

$$\rho_1 + \rho_2 < 1, \qquad \rho_2 < c. \tag{6.1}$$

Utility-based allocation. We assume that the network operates under an α-fair rate allocation policy as described in Section 1.2.3. For any unweighted α-fair rate allocation (Section 5.6, Appendix 5.E), the rate allocations in case of a two-link parking lot are given by

$$\begin{aligned} \Lambda_1(n) &= \max\left(1 - c, \tfrac{n_1}{n_1+n_2}\right), \quad n_1 > 0, \\ \Lambda_2(n) &= \min\left(c, \tfrac{n_2}{n_1+n_2}\right). \end{aligned} \tag{6.2}$$

By convention, the per-class and per-flow rates $\Lambda_i(n)$ and $\frac{\Lambda_i(n)}{n_i}$ are equal to zero if $n_i = 0$.

Modified proportional fair allocation. Massoulié [83] introduced a novel rate allocation policy, which in some sense coincides with the proportional fair policy but is insensitive and has an explicit steady-state distribution. Let e_i be a unit vector with 1 in component i and 0 elsewhere, $i = 1, \ldots, I$. The rate allocation under the modified proportional fair policy is defined as

$$\tilde{\Lambda}_i(n) = \begin{cases} \mathrm{e}^{w^*_{\mathcal{C}}(n) - w^*_{\mathcal{C}}(n-e_i)}, & \text{if } n_i > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{6.3}$$

where

$$w^*_{\mathcal{C}}(n) = \sup_{\tilde{\Lambda} \in \mathbb{R}_+^I} (\langle \log \tilde{\Lambda}, n \rangle - w_{\mathcal{C}}(\tilde{\Lambda})).$$

The function $w_{\mathcal{C}}(\tilde{\Lambda})$ is equal to zero if $\tilde{\Lambda} \in \mathcal{C}$ and $+\infty$ if $\tilde{\Lambda} \notin \mathcal{C}$. Note that the function $w^*_{\mathcal{C}}(n)$ coincides with the supremum of the utility function of the proportional fair policy over the capacity set.

Let $\tilde{X}$ denote the number of active flows in the system with the modified proportional fair policy. Since the modified proportional fair policy is insensitive, under the stability conditions (6.1), the queue length process is regenerative and admits a steady-state distribution. The steady-state distribution of $\tilde{X}$ [83] is determined by

$$\tilde{\pi}(n) \equiv \mathbf{P}(\tilde{X}_1 = n_1, \ldots, \tilde{X}_I = n_I) = G\mathrm{e}^{-L(n)}, \tag{6.4}$$

where $G \in (0,\infty)$ denotes the normalizing constant, and

$$L(n) \equiv L(n_1,\ldots,n_I) = w^*_{\mathcal{C}}(n) - \sum_{i=1}^{I} \log(\rho_i) n_i. \tag{6.5}$$

For the two-link parking lot network, applying (6.2), we find

$$L(n) = n_1 \log \frac{\max(1-c, \frac{n_1}{n_1+n_2})}{\rho_1} + n_2 \log \frac{\min(c, \frac{n_2}{n_1+n_2})}{\rho_2}, \tag{6.6}$$

so that

$$\tilde{\pi}(n) = G \left(\frac{\rho_1}{\max\left(1-c, \frac{n_1}{n_1+n_2}\right)} \right)^{n_1} \left(\frac{\rho_2}{\min\left(c, \frac{n_2}{n_1+n_2}\right)} \right)^{n_2}.$$

6.2 Additional notation

The notation used in the present chapter strongly resembles that in Chapter 4. We denote by A_i^n, $n \in \mathbb{N}$, $i = 1,\ldots,I$, the time between the $(n-1)$-st and n-th class-i arrival after time zero. Furthermore, let B_i^n, $n \in \mathbb{Z}$, be the size of the nth class-i flow. We assume that $(A_i^n)_n$ and $(B_i^n)_n$ are mutually independent sequences, each consisting of i.i.d. random variables. We introduce the random walks $S_n^{A_i} = A_i^1 + \ldots + A_i^n$ and $S_n^{B_i} = B_i^1 + \ldots + B_i^n$. We denote the random variable corresponding to a generic interarrival time (flow size) by A_i (B_i, respectively).

Let $N_i(t)$ be the number of class-i arrivals in the time interval $(0,t]$. Recall $N_i(\cdot)$ is a Poisson process with rate λ_i. Denote by $A_i(t)$, $t > 0$, the total amount of class-i work arriving in the time interval $(0,t]$, i.e. ,

$$A_i(t) = \sum_{k=1}^{N_i(t)} B_i^k.$$

Similarly, we define $T_i(t)$ as the total service capacity available for the class-i flows during the time interval $(0,t]$,

$$T_i(t) = \int_0^t \Lambda_i(X(u))\mathrm{d}u,$$

where $X(u)$ denotes the number of flows in the system at time u.

Define the MGFs $\Phi_{B_i}(s) := \mathbf{E}[\mathrm{e}^{sB_i}]$ and $\Phi_{A_i}(s) := \mathbf{E}[\mathrm{e}^{sA_i}] = \frac{\lambda_i}{\lambda_i - s}$. For $s \geq 0$ denote by $\alpha_i(s)$, the asymptotic cumulant function of $A_i(x)$, $x > 0$,

$$\alpha_i(s) = \lim_{x\to\infty} \frac{1}{x} \log \mathbf{E}[\mathrm{e}^{sA_i(x)}].$$

Since the process $A_i(x)$ is a compound Poisson process, the cumulant function is known explicitly,

$$\mathbf{E}[\mathrm{e}^{sA_i(x)}] = \mathrm{e}^{\lambda_i x(\Phi_{B_i}(s)-1)}, \qquad \alpha_i(s) = \lambda_i(\Phi_{B_i}(s) - 1).$$

Further, for any given $u > 0$, we denote by $\delta_i^{(u)}$ the solution of the equation

$$\alpha_i'(s) = u.$$

In other words,

$$\delta_i^{(u)} = (\alpha_i')^{-1}(u) = (\Phi'_{B_i})^{-1}\left(\frac{u}{\lambda_i}\right). \tag{6.7}$$

The value $u = \alpha_i'(\delta_i^{(u)})$ for some $u \geq \rho_i$ can be considered as a new traffic load of class-i flows under exponential tilting with parameter $\delta_i^{(u)}$. We will also use the notation $\rho_i(\delta_i^{(u)}) = \alpha_i'(\delta_i^{(u)}) = u$, when we need to emphasize this interpretation.

The main focus in this chapter is on the sojourn time, say V_1, of a 'tagged' class-1 flow (with flow size B_1^0) arriving at time 0, when the system is assumed to be in steady state. The main goal is to describe the asymptotic behavior of $\mathbf{P}(V_1 > x)$ as $x \to \infty$. However, before proceeding to the sojourn time asymptotics, we derive some results for the queue length and the workload which will be of use in proving the main theorem.

6.3 Queue length bounds

In the present section we are mainly interested in deriving upper bounds for the distribution of the number of flows in the network. The key idea is to use the monotonicity of a parking lot network and the characteristics of the modified proportional fair policy.

Massoulié [83] showed that for all $n \geq 0$ the rate allocation under the proportional fair policy forms an upper bound for the allocation under the modified proportional fair policy. Thus, for all $n \geq 0$,

$$\tilde{\Lambda}(n) \leq \Lambda(n).$$

Let $X(t)$ and $\tilde{X}(t)$ denote the number of active flows at time t in the original and the modified system, respectively. As the parking lot network is a special case of a tree network, Proposition 5.5.1 implies that it is monotone. In view of the above inequality, due to the monotonicity and Theorem 1 in Bonald and Proutière [19], we have for all $t \geq 0$,

$$\tilde{X}(t) \geq X(t). \tag{6.8}$$

Under the stability conditions (6.1) the queue length process admits a steady-state distribution. In steady state,

$$\tilde{X} \geq_{st} X.$$

By $\geq_{st}$ we denote the strong stochastic ordering on $\mathbb{R}_+^2$, that is $\tilde{X} \geq_{st} X$ if and only if $\mathbf{E}[f(\tilde{X})] \geq \mathbf{E}[f(X)]$ for any increasing function $f(\cdot)$.

We need the following auxiliary result.

Lemma 6.3.1.

$$u(1-u)^{\frac{1-c}{c}} \leq c(1-c)^{\frac{1-c}{c}} \tag{6.9}$$

for all $u \in [0,1]$.

Proof. Define the function $f(u)$ as the left-hand side of the above inequality,

$$f(u) = u(1-u)^{\frac{1-c}{c}}.$$

The derivative is determined by

$$f'(u) = (1-u)^{\frac{1-c}{c}} \left(1 - \frac{u}{1-u}\frac{1-c}{c}\right).$$

Noting that the derivative is positive for $u < c$, and negative otherwise, we conclude that for all $u \in [0,1]$, $f(u) \leq f(c)$. □

In the following propositions we derive upper bounds for the marginal distribution of the class-2 queue length and the distribution of the total queue length in the modified system. In view of Inequality (6.8), the upper bound also holds for the original system.

Proposition 6.3.1. *If* $\rho_1 < 1-c$, *then there exists a constant* $A > 0$ *such that for any* $n > 0$ *the stationary queue length distribution satisfies*

$$\mathbf{P}(\tilde{X}_2 = n) \leq A\left(\frac{\rho_2}{c}\right)^n. \tag{6.10}$$

If $\rho_1 > 1-c$, *then for any* $n > 0$ *the stationary queue length distribution satisfies*

$$\mathbf{P}(\tilde{X}_2 = n) \leq K(n)\left(\frac{\rho_2}{c}\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}}\right)^n, \tag{6.11}$$

where

$$K(n) = \frac{\rho_1}{\rho_1 - 1 + c} + \sqrt{2\pi n}\mathrm{e}^{\frac{1}{12c}} - \left(\frac{\rho_1}{1-c}\right)^{-\frac{1-c}{c}n}.$$

If $\rho_1 = 1-c$, *then for any* $n > 0$ *the stationary queue length distribution satisfies*

$$\mathbf{P}(\tilde{X}_2 = n) \leq \left(\frac{1-c}{c}n + \sqrt{2\pi n}\mathrm{e}^{\frac{1}{12c}}\right)\left(\frac{\rho_2}{c}\right)^n. \tag{6.12}$$

Proof. Fix $n_2 > 0$. Define $\eta = \lfloor \frac{1-c}{c} n_2 \rfloor$. Since we are only interested in bounds, we will omit the normalizing constant G in Equation (6.4). From (6.4)–(6.6) we obtain

$$\begin{aligned}
\mathbf{P}(\tilde{X}_2 = n_2) &= \sum_{n_1=0}^{\infty} \mathrm{e}^{-L(n_1,n_2)} \\
&= \sum_{n_1=0}^{\eta} \left(\frac{\rho_2}{c}\right)^{n_2}\left(\frac{\rho_1}{1-c}\right)^{n_1} + \sum_{n_1=\eta+1}^{\infty} \frac{(n_1+n_2)^{(n_1+n_2)}}{(n_2)^{n_2}(n_1)^{n_1}}\rho_2^{n_2}\rho_1^{n_1} \\
&:= J_1 + J_2.
\end{aligned}$$

Consider first the term J_1. Using the summation formula for geometric series, we obtain

$$\begin{aligned} J_1 &= \left(\frac{\rho_2}{c}\right)^{n_2} \frac{1-c}{1-c-\rho_1}\left(1-\left(\frac{\rho_1}{1-c}\right)^{\eta+1}\right) \\ &\leq \left(\frac{\rho_2}{c}\right)^{n_2} \frac{1-c}{1-c-\rho_1}\left(1-\left(\frac{\rho_1}{1-c}\right)\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}n_2}\right). \end{aligned} \tag{6.13}$$

Let us now turn to the term J_2. In order to bound this term, we apply Stirling's formula

$$k^k = \frac{k!\mathrm{e}^k}{\sqrt{2\pi k}}\,\mathrm{e}^{-\epsilon_k}, \tag{6.14}$$

where $\frac{1}{12k+1} < \epsilon_k < \frac{1}{12k}$. Using these approximations, we obtain

$$\begin{aligned} J_2 &\leq \rho_2^{n_2} \sum_{n_1=\eta+1}^{\infty} \binom{n_1+n_2}{n_2} \sqrt{2\pi}\frac{\sqrt{n_1 n_2}}{\sqrt{n_1+n_2}}\rho_1^{n_1}\mathrm{e}^{\frac{1}{12}\left(\frac{1}{n_2}+\frac{1}{n_1}\right)-\frac{1}{12(n_1+n_2)+1}} \\ &\leq \sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\left(\frac{\rho_2}{1-\rho_1}\right)^{n_2}\frac{1}{\rho_1}\sum_{n_1=\eta+1}^{\infty}\binom{n_1+n_2}{n_2}(1-\rho_1)^{n_2}\rho_1^{n_1+1} \\ &\leq \sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\left(\frac{\rho_2}{1-\rho_1}\right)^{n_2}\frac{1}{\rho_1}\sum_{n_1=\eta+1}^{\infty}\binom{n_1+n_2-1}{n_2}(1-\rho_1)^{n_2}\rho_1^{n_1}. \end{aligned}$$

The latter sum is essentially the probability that the number of failures before the n_2-th success is at least $\eta+1$, given that the probability of failure is ρ_1 and the probability of success is $1-\rho_1$. Using the relationship between the negative binomial distribution and the geometric distribution, we can rewrite this probability as

$$\mathbf{P}\left(\sum_{i=1}^{n_2} G_i \geq \eta+1\right),$$

where G_i is a geometrically distributed random variable with $\mathbf{P}(G_i = n) = \rho_1^n(1-\rho_1)$, $n = 0, 1, \ldots$. Consequently,

$$\mathbf{P}\left(\sum_{i=1}^{n_2} G_i \geq \eta+1\right) = \mathbf{P}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} G_i \geq \frac{\eta+1}{n_2}\right) \leq \mathbf{P}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} G_i \geq \frac{1-c}{c}\right). \tag{6.15}$$

Recall that the log moment generating function of the geometric random variable G_i is given by

$$M(t) = \log \mathbf{E}[\mathrm{e}^{tG_i}] = \log(1-\rho_1) - \log(1-\rho_1\mathrm{e}^t), \tag{6.16}$$

with the convex conjugate

$$M^*(y) = \sup_t(ty - M(t)) = \log\left[\left(\frac{y}{(y+1)\rho_1}\right)^y \frac{1}{(1-\rho_1)(y+1)}\right]. \tag{6.17}$$

Applying the Chernoff bound to the latter probability in (6.15), we obtain

$$\mathbf{P}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} G_i \geq \frac{1-c}{c}\right) \leq \mathrm{e}^{-M^*(\frac{1-c}{c})n_2},$$

where

$$M^*\left(\frac{1-c}{c}\right) = \log\left(\left(\frac{1-c}{\rho_1}\right)^{\frac{1-c}{c}} \frac{c}{1-\rho_1}\right). \tag{6.18}$$

Summarizing the above, we derive an upper bound for the term J_2,

$$J_2 \leq \sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\mathrm{e}^{-n_2\log\left(\frac{c}{\rho_2}\left(\frac{1-c}{\rho_1}\right)^{\frac{1-c}{c}}\right)} = \sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\left(\frac{\rho_2}{c}\right)^{n_2}\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}n_2}. \tag{6.19}$$

We now separately consider cases $\rho_1 < 1-c$ and $\rho_1 \geq 1-c$. Let us first assume $\rho_1 < 1-c$. Combining the bounds in (6.13) and (6.19), we derive

$$\mathbf{P}(\tilde{X}_2 = n_2) \leq K_1(n_2)\left(\frac{\rho_2}{c}\right)^{n_2}, \tag{6.20}$$

where

$$K_1(n_2) = \left(\frac{1-c}{1-c-\rho_1} + \left(\sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}} - \frac{\rho_1}{1-c-\rho_1}\right)\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}n_2}\right).$$

Noting that the function $K_1(n_2)$ is bounded from above provides the upper bound.

It remains to consider the case $\rho_1 > 1-c$. Using the bounds (6.13) and (6.19), we obtain

$$\mathbf{P}(\tilde{X}_2 = n_2) \leq K_2(n_2)\left(\frac{\rho_2}{c}\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}}\right)^{n_2}, \tag{6.21}$$

where

$$K_2(n_2) = \frac{\rho_1}{\rho_1 - 1 + c} + \sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}} + \frac{\rho_1}{\rho_1 - 1 + c}\left(\frac{\rho_1}{1-c}\right)^{-\frac{1-c}{c}n_2}.$$

□

Corollary 6.3.1. *If* $\rho_1 \leq 1-c$,

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}(\tilde{X}_2 \geq n) = -\log\frac{c}{\rho_2}. \tag{6.22}$$

If $\rho_1 > 1-c$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{P}(\tilde{X}_2 \geq n) \leq -\log\left(\frac{c}{\rho_2}\left(\frac{1-c}{\rho_1}\right)^{-\frac{1-c}{c}}\right). \tag{6.23}$$

The logarithmic bounds follow in a straightforward manner from Proposition 6.3.1. Since the rate Λ_2 allocated to class-2 flows is bounded by the capacity constraint c, the queue length of class 2 is bounded from below by the queue length in a PS node with capacity c and traffic of class 2 only. Consequently,

$$\mathbf{P}(X_2 \geq n) \geq \mathbf{P}(X_2^{PS} \geq n) = \left(\frac{\rho_2}{c}\right)^n.$$

This implies the asymptotics if $\rho_1 \leq 1 - c$. The derivation in case $\rho_1 \geq 1 - c$ relies on Lemma 6.3.1 and the observation

$$\frac{\rho_2}{c}\left(\frac{\rho_1}{1-c}\right)^{\frac{1-c}{c}} < \frac{\rho_2}{c}\left(\frac{1-\rho_2}{1-c}\right)^{\frac{1-c}{c}}.$$

Remark 6.3.1. The asymptotics for the case $\rho_1 \leq 1 - c$ may be regarded as a positive result. If the load of class 1 is less than $1 - c$, in the large-deviations sense, the class-2 queue in a parking lot behaves as if it always receives the maximum rate c. Thus, the probability of having an extremely large number of class-2 flows is not affected by the preferential rate allocation to class-1 flows. Note that this asymptotic behavior also holds for any utility-based allocation policy.

Lemma 6.3.2. *For any $n > 0$, the stationary total queue length distribution satisfies*

$$\mathbf{P}(\tilde{X}_2 + \tilde{X}_1 = n) \leq (n+1)\left(\max\left(\frac{\rho_2}{c}, \rho_1 + \rho_2\right)\right)^n. \tag{6.24}$$

Proof. Fix $n \geq 1$. Using Equation (6.4), we obtain

$$\begin{aligned}\mathbf{P}(\tilde{X}_2 + \tilde{X}_1 = n) &= \sum_{m=0}^{n} \pi(m, n-m) \leq (n+1) \max_{m\in[0,n]} \pi(m, n-m) \\ &\leq (n+1)\mathrm{e}^{-\inf_{m\in[0,n]} L(m,n-m)}.\end{aligned}$$

Noting that for any constant a and any vector z, $L(az) = aL(z)$ (see (6.6)), we derive

$$\mathbf{P}(\tilde{X}_2 + \tilde{X}_1 = n) \leq (n+1)\mathrm{e}^{-n\inf_{x\in[0,1]} L(x,1-x)} = (n+1)\mathrm{e}^{-nL^*},$$

with $L^* = \inf_{x\in[0,1]} L(x, 1-x)$.

Let us consider the function $L(x, 1-x)$ in more detail,

$$L(x, 1-x) = x\log\frac{\max(1-c, x)}{\rho_1} + (1-x)\log\frac{\min(c, 1-x)}{\rho_2}.$$

Let us first assume $x \in (1-c, 1]$. In this case,

$$L(x, 1-x) = x\log\frac{x}{\rho_1} + (1-x)\log\frac{(1-x)}{\rho_2},$$

$$L'(x,1-x) = \log\frac{x}{\rho_1} + 1 - \log\frac{(1-x)}{\rho_2} - 1 = \log\frac{x}{1-x}\frac{\rho_2}{\rho_1}.$$

Since the derivative is equal to zero at $x^* = \frac{\rho_1}{\rho_2+\rho_1}$, is negative for smaller values and positive for larger values, we conclude that L has a local minimum at x^*. This is however under the assumption $x^* > 1-c$, that is $\frac{\rho_2}{\rho_1} < \frac{c}{1-c}$. If $x^* \leq 1-c$, $L(x,1-x)$ is increasing on the interval $[1-c,1]$.

Consider now the interval $[0,1-c]$. We have

$$L(x,1-x) = x\log\frac{1-c}{\rho_1} + (1-x)\log\frac{c}{\rho_2},$$

$$L'(x,1-x) = \log\frac{1-c}{\rho_1} - \log\frac{c}{\rho_2} = \log\frac{1-c}{c}\frac{\rho_2}{\rho_1}.$$

If $\frac{\rho_2}{\rho_1} > \frac{c}{1-c}$, the function is increasing linearly on $[0,1-c]$, and it is decreasing otherwise. So if $\frac{\rho_2}{\rho_1} > \frac{c}{1-c}$, the minimum on the interval $[0,1-c]$ is attained at $x=0$, and otherwise at $x = 1-c$.

Note also that if $\frac{\rho_2}{\rho_1} > \frac{c}{1-c}$, $L(x,1-x)$ is increasing on $[1-c,1]$. Hence, the global minimum of $L(x,1-x)$ on the entire interval $[0,1]$ is given by

$$x_{min} = \begin{cases} x^*, & \text{if } \frac{\rho_2}{\rho_1} < \frac{c}{1-c}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.25}$$

The optimal values are

$$L^* = \begin{cases} \log\frac{1}{\rho_2+\rho_1}, & \text{if } x_{min} = x^*, \\ \log\frac{c}{\rho_2}, & \text{otherwise.} \end{cases}$$

□

6.4 Workload bounds

In the present section we turn our attention to the workload characteristics. We first compare the workloads in the networks with original and modified proportional fair allocations. Secondly, we prove finiteness of the MGF of the workload for specific arguments. Finally, we state an auxiliary lemma for the MGF of the workload in an ordinary PS link.

Proposition 6.4.1. *The workload W in the system with an α-fair policy and the workload $\tilde{W}$ in the system with the modified proportional fair policy are related as*

$$W \leq_{st} \tilde{W}.$$

Proof. The proof is essentially a compilation of the arguments in [19]. Let X and $\tilde{X}$ denote the number of active flows in the original and the modified systems, respectively. Using the monotonicity of the parking lot network and Inequality (6.8), we obtain that for all t, $i = 1, 2$,

$$\frac{\tilde{\Lambda}_i(\tilde{X}(t))}{\tilde{X}_i(t)} \leq \frac{\Lambda_i(\tilde{X}(t))}{\tilde{X}_i(t)} \leq \frac{\Lambda_i(X(t))}{X_i(t)}.$$

Hence, the residual flow sizes are related in the following manner, for $i = 1, 2$,

$$B^r_{i,j}(t) = \left(B_{i,j} - \int_{a_{i,j}}^{t} \frac{\Lambda_i(X(u))}{X_i(u)} du \right)^+ \leq \left(B_{i,j} - \int_{a_{i,j}}^{t} \frac{\tilde{\Lambda}_{i,j}(\tilde{X}(u))}{\tilde{X}_i(u)} du \right)^+ = \tilde{B}^r_{i,j}(t),$$

where $B_{i,j}$, $B^r_{i,j}(t)$ and $a_{i,j} < t$ denote the initial flow size, the residual flow size and the arrival epoch of the j-th flow of class i, respectively. This implies

$$W_i(t) = \sum_{j=1}^{X_i(t)} B^r_{i,j}(t) \leq \sum_{j=1}^{\tilde{X}_i(t)} \tilde{B}^r_{i,j}(t) = \tilde{W}_i(t).$$

This implies that for any vector $x \geq 0$ and any time t,

$$\mathbf{P}(W_1(t) > x_1, W_2(t) > x_2) \leq \mathbf{P}(\tilde{W}_1(t) > x_1, \tilde{W}_2(t) > x_2).$$

Since under the stability conditions (6.1) the distributions converge as $t \to \infty$, we derive

$$\mathbf{P}(W_1 > x_1, W_2 > x_2) \leq \mathbf{P}(\tilde{W}_1 > x_1, \tilde{W}_2 > x_2),$$

which yields stochastic ordering $W \leq_{st} \tilde{W}$. □

The following proposition states that the MGF with parameter $\delta = (\delta_1, \delta_2)$ of the workload is finite, if the tilted traffic load $\rho(\delta)$ lies in the capacity set $\mathcal{C}$.

Proposition 6.4.2. *Consider a parking lot network operating under the modified proportional fair policy. Suppose $\delta_1, \delta_2 > 0$ satisfy the equation $\alpha_1'(\delta_1) + \alpha_2'(\delta_2) = 1$. Let $u = \alpha_2'(\delta_2)$. Then, if $u \leq c$,*

$$\mathbf{E}[\mathrm{e}^{\delta_1 \tilde{W}_1 + \delta_2 \tilde{W}_2}] < \infty.$$

Proof. The MGF of the total workload can be computed as

$$\mathbf{E}[\mathrm{e}^{\delta_1 \tilde{W}_1 + \delta_2 \tilde{W}_2}] = \mathbf{E}[\beta_1^*(\delta_1)^{\tilde{X}_1} \beta_2^*(\delta_2)^{\tilde{X}_2}] = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \beta_1^*(\delta_1)^{n_1} \beta_2^*(\delta_2)^{n_2} \tilde{\pi}(n_1, n_2).$$

Noting that

$$\beta_i^*(\delta_i) = \frac{\Phi_{B_i}(\delta_i) - 1}{\delta \Phi'_{B_i}(0)} = \frac{\alpha_i(\delta_i)}{\delta_i \rho_i},$$

and substituting Equation (6.4) for $\tilde{\pi}$, we obtain

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{\delta_1\tilde{W}_1+\delta_2\tilde{W}_2}] &= \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\left(\left(\frac{\alpha_1(\delta_1)}{\delta_1\rho_1}\right)^{n_1}\left(\frac{\alpha_2(\delta_2)}{\delta_2\rho_2}\right)^{n_2}\right.\\
&\times \left.\rho_1^{n_1}\rho_2^{n_2}\max\left(1-c,\frac{n_1}{n_1+n_2}\right)^{-n_1}\min\left(c,\frac{n_2}{n_1+n_2}\right)^{-n_2}\right)\\
&= \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\left(\left(\frac{\alpha_1(\delta_1)}{\delta_1}\right)^{n_1}\left(\frac{\alpha_2(\delta_2)}{\delta_2}\right)^{n_2}\right.\\
&\times \left.\max\left(1-c,\frac{n_1}{n_1+n_2}\right)^{-n_1}\min\left(c,\frac{n_2}{n_1+n_2}\right)^{-n_2}\right).
\end{aligned}$$

We now proceed as in the proof of Lemma 6.3.1. Define $\eta = \lfloor\frac{1-c}{c}n_2\rfloor$. We first take the sum with respect to n_1.

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{\delta_1\tilde{W}_1+\delta_2\tilde{W}_2}] &= \sum_{n_2=0}^{\infty}\left(\frac{\alpha_2(\delta_2)}{c\delta_2}\right)^{n_2}\sum_{n_1=0}^{\eta}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{n_1}\\
&+ \sum_{n_2=0}^{\infty}\left(\frac{\alpha_2(\delta_2)}{\delta_2}\right)^{n_2}\sum_{n_1=\eta+1}^{\infty}\left(\frac{\alpha_1(\delta_1)}{\delta_1}\right)^{n_1}\frac{(n_1+n_2)^{(n_1+n_2)}}{(n_2)^{n_2}(n_1)^{n_1}}\\
&:= J_1 + J_2.
\end{aligned}$$

Consider first the term J_1. Using the summation formula for geometric series, we obtain

$$\sum_{n_1=0}^{\eta}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{n_1} = \frac{1}{1-\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}}\left(1-\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}n_2+1}\right).$$

Hence,

$$J_1 = \sum_{n_2=0}^{\infty}\left(\frac{\alpha_2(\delta_2)}{c\delta_2}\right)^{n_2}\frac{1}{1-\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}}\left(1-\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}n_2+1}\right).$$

Let us now turn to the term J_2. Applying Stirling's formula (6.14), we obtain

$$\begin{aligned}
J_2 &\leq \sum_{n_2=0}^{\infty}\left(\frac{\alpha_2(\delta_2)}{\delta_2}\right)^{n_2}\sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\left(\frac{1}{1-\frac{\alpha_1(\delta_1)}{\delta_1}}\right)^{n_2}\frac{\delta_1}{\alpha_1(\delta_1)}\\
&\times \sum_{n_1=\eta+1}^{\infty}\binom{n_1+n_2-1}{n_2}\left(1-\frac{\alpha_1(\delta_1)}{\delta_1}\right)^{n_2}\left(\frac{\alpha_1(\delta_1)}{\delta_1}\right)^{n_1}.
\end{aligned}$$

The latter sum is essentially the probability that the number of failures before the n_2-th success is at least $\eta+1$, given that the probability of failure is $\frac{\alpha_1(\delta_1)}{\delta_1}$

($\frac{\alpha_1(\delta_1)}{\delta_1} < 1-u < 1$) and the probability of success is $1-\frac{\alpha_1(\delta_1)}{\delta_1}$. Using the relationship between the negative binomial distribution and the geometric distribution, we obtain that the latter sum is bounded by

$$\mathbf{P}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} G_i \geq \frac{1-c}{c}\right) \leq \mathrm{e}^{-M^*(\frac{1-c}{c})n_2},$$

where G_i is a geometrically distributed random variable with parameter $\frac{\alpha_1(\delta_1)}{\delta_1}$, and the convex conjugate (cf. (6.17)) is given by

$$M^*\left(\frac{1-c}{c}\right) = \log\left(\left(\frac{1-c}{\frac{\alpha_1(\delta_1)}{\delta_1}}\right)^{\frac{1-c}{c}} \frac{c}{1-\frac{\alpha_1(\delta_1)}{\delta_1}}\right).$$

Summarizing the above, we derived an upper bound for the term J_2,

$$J_2 \leq \sum_{n_2=0}^{\infty}\left(\frac{\alpha_2(\delta_2)}{c\delta_2}\right)^{n_2}\sqrt{2\pi n_2}\mathrm{e}^{\frac{1}{12c}}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}n_2}.$$

Thus, we obtain for some constants A_1, A_2, A_3

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{\delta_1\tilde{W}_1+\delta_2\tilde{W}_2}] \quad &\leq \quad \sum_{n_2=0}^{\infty}\left[A_1\left(\frac{\alpha_2(\delta_2)}{c\delta_2}\right)^{n_2}\right. \\ &+ \quad \left.(A_2+A_3\sqrt{n_2})\left(\frac{\alpha_2(\delta_2)}{c\delta_2}\right)^{n_2}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}n_2}\right].\end{aligned}$$

In order for the MGF to be finite we need

$$\frac{\alpha_2(\delta_2)}{c\delta_2} < 1, \quad \text{and} \quad \frac{\alpha_2(\delta_2)}{c\delta_2}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}} < 1.$$

If $u \leq c$, the first inequality holds naturally,

$$\frac{\alpha_2(\delta_2)}{c\delta_2} < \frac{\alpha_2'(\delta_2)}{c} = \frac{u}{c} \leq 1.$$

For the second inequality we have

$$\frac{\alpha_2(\delta_2)}{c\delta_2}\left(\frac{\alpha_1(\delta_1)}{(1-c)\delta_1}\right)^{\frac{1-c}{c}} < \frac{u}{c}\left(\frac{1-u}{1-c}\right)^{\frac{1-c}{c}},$$

which does not exceed 1 for any $u \in [0,1]$ by Lemma 6.3.1. □

In a single-class single-link PS system the proof of finiteness of the MGF is significantly simpler. The following lemma states that the MGF with argument δ of the workload is finite, if $\rho(\delta)$ does not exceed the capacity r of the link. The proof is based on the argument in [82] for the case $r = 1$.

Lemma 6.4.1. *Consider an egalitarian PS system with capacity r. Suppose the traffic intensity $\rho < r$. Then*

$$\mathbf{E}[\mathrm{e}^{\delta W}] < \infty,$$

where $\delta = (\alpha')^{-1}(r)$.

Proof. Since $\rho < r$, the PS system with capacity r reaches steady state and the workload W can be identified with the waiting time under FCFS. Hence, $W = \sup_{n\geq 0}(S_n^B - rS_n^A)$, and

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{\delta W}] &= \int_0^\infty \mathbf{P}\left(\sup_{n\geq 0} \mathrm{e}^{\delta(S_n^B - rS_n^A)} > x\right) dx \leq \sum_{n=0}^\infty \int_0^\infty \mathbf{P}\left(\mathrm{e}^{\delta(S_n^B - rS_n^A)} > x\right) dx \\
&= \sum_{n=0}^\infty \mathbf{E}[\mathrm{e}^{\delta S_n^B}]\mathbf{E}[\mathrm{e}^{-\delta r S_n^A}] = \sum_{n=0}^\infty (\Phi_A(-r\delta)\Phi_B(\delta))^n \qquad (6.26) \\
&= \sum_{n=0}^\infty \left(\frac{\lambda}{\lambda + r\delta}\Phi_B(\delta)\right)^n. \qquad (6.27)
\end{aligned}$$

Note that due to strict convexity of the cumulant function $\alpha(\cdot)$, $\alpha(\delta) < \alpha'(\delta)\delta = r\delta$. Consequently, $\frac{\lambda}{\lambda+r\delta}\Phi_B(\delta) < \frac{\lambda}{\lambda+\alpha(\delta)}\Phi_B(\delta) = 1$, which implies finiteness of the MGF. □

6.5 Class-1 delay asymptotics

In the present section we investigate the asymptotic behavior of the sojourn time of class-1 flows. We consider a tagged class-1 flow that arrives into the system at time 0 and has size B_1^0. We derive the large-deviations asymptotics for the sojourn time V_1 of the tagged customer.

We need to make two technical assumptions. We assume the flow size distributions have bounded hazard rates (see Assumption 5.5.1). We also assume the following.

Assumption 6.5.1. *There exists a solution $\delta_1^*, \delta_2^* > 0$ to*

$$\begin{cases} \alpha_1'(\delta_1^*) + \alpha_2'(\delta_2^*) = 1, \\ \alpha_2'(\delta_2^*) \leq c, \end{cases}$$

such that $\Phi_i(\delta_i) < \infty$ for all δ_i in a neighborhood of δ_i^*, $i = 1, 2$.

In preparation for the main theorem we state an auxiliary lemma.

Lemma 6.5.1. *Consider the function*

$$H(u) = (\alpha_1')^{-1}(1-u) - (\alpha_2')^{-1}(u), \qquad u \in [0,1].$$

The equation $H(u) = 0$ has a unique solution u^*, and $u^* \in (\rho_2, 1-\rho_1)$.

Proof. The derivative of $H(u)$ is given by

$$H'(u) = -\frac{1}{\alpha_1''(\delta_1^{(1-u)})} - \frac{1}{\alpha_2''(\delta_1^{(u)})},$$

which is strictly negative due to the strict convexity of cumulant functions. Since $\rho_0 + \rho_1 < 1$, we obtain $H(\rho_2) = \delta_1^{(1-\rho_2)} > \delta_1^{(\rho_1)} = 0$ and $H(1-\rho_1) = -\delta_2^{(1-\rho_1)} < -\delta_2^{(\rho_2)} = 0$. Since the function $H(\cdot)$ is continuous and strictly decreasing, we conclude that there exists $u^* \in (\rho_2, 1-\rho_1)$ such that $\delta_1^{(1-u^*)} = \delta_2^{(u^*)}$. □

The main result of the chapter is the following theorem.

Theorem 6.5.1. *Let u^* be the solution of the equation $(\alpha_2')^{-1}(u^*) = (\alpha_1')^{-1}(1-u^*)$, that is*

$$\left(\Phi'_{B_2}\right)^{-1}\left(\frac{u^*}{\lambda_2}\right) = \left(\Phi'_{B_1}\right)^{-1}\left(\frac{1-u^*}{\lambda_1}\right),$$

and let $\delta^* = \delta_2^{(u^*)} = \delta_1^{(1-u^*)}$. Then,

$$\lim_{x\to\infty}\frac{1}{x}\log \mathbf{P}(V_1 > x) = \begin{cases} \alpha_1(\delta^*) + \alpha_2(\delta^*) - \delta^*, & \text{if } u^* \le c, \\ \alpha_1(\delta_1^{(1-c)}) + \alpha_2(\delta_2^{(c)}) - (1-c)\delta_1^{(1-c)} - c\delta_2^{(c)}, & \text{if } u^* > c. \end{cases} \tag{6.28}$$

This result provides insight in the manner a large sojourn time occurs. Note first that the solution u^* corresponds to the traffic load of class 2 in an exponentially tilted system; $1-u^*$ is the traffic load of class 1. Hence, in the scenario $u^* < c$, the capacity constraint of the second link is not binding. This implies that the system dynamics asymptotically coincide with the dynamics in a single-link DPS system. See also Section 4.5 for the large-deviations results in a DPS system. The second scenario, $u^* \ge c$, corresponds to the situation when the second link is saturated, that is the class-2 rate allocation achieves its maximum c. In either scenario, the asymptotics indicate that a large sojourn time is predominantly caused by a large amount of work generated by both classes during the service of the tagged flow.

The proof of the theorem consists of two parts: derivation of the large-deviations upper bound based on the Chernoff bound and derivation of the lower bound based on a change-of-measure approach. We first present the derivation of the upper bound.

6.5.1 Proof of the upper bound

In order to obtain an upper bound, we distinguish between two cases: (a) $u^* \le c$, and (b) $u^* > c$.

Case (a) Let us first assume $u^* \le c$.

The event $\{V_1 > x\}$ implies that the total workload of class-1 and class-2 flows at time epoch x, $W_1(x) + W_2(x) = B_1^0 + W_1 + W_2 + A_1(x) + A_2(x) - (C_1(x) +$

$C_2(x)$), is positive. Since during the time interval $[0, x]$ there is always class-1 work, $C_1(x) + C_2(x) = x$. Hence, we can write

$$\mathbf{P}(V_1 > x) \leq \mathbf{P}\left(B_1^0 + W_1 + W_2 + A_1(x) + A_2(x) > x\right). \tag{6.29}$$

Applying the Chernoff bound with parameter $\delta = \delta^*$ to the above probability, we obtain

$$\mathbf{P}(V_1 > x) \leq \Phi_{B_1}(\delta^*)\mathbf{E}[\mathrm{e}^{\delta^*(W_1+W_2)}]\mathrm{e}^{(\alpha_1(\delta^*)+\alpha_2(\delta^*)-\delta^*)x}.$$

The MGF of B_1 is finite by definition of δ^*. Since, under the assumption $u^* \leq c$, $\alpha_2'(\delta^*) \leq \alpha_2'(\delta_2^c) = c$, finiteness of the term $\mathbf{E}[\mathrm{e}^{\delta^*(W_1+W_2)}]$ follows from Propositions 6.4.1 and 6.4.2. Taking logarithms, dividing by x, and letting $x \to \infty$, we obtain

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) \leq \alpha_1(\delta^*) + \alpha_2(\delta^*) - \delta^*.$$

Case (b). Suppose now $u^* > c$.

The event $\{V_1 > x\}$ implies that the workload of class-1 flows at time epoch x, $W_1(x) = W_1 + B_1^0 + A_1(x) - C_1(x)$, is positive. Hence, we can write

$$\begin{aligned}
\mathbf{P}(V_1 > x) &\leq \mathbf{P}\left(W_1 + B_1^0 + A_1(x) > C_1(x)\right) \\
&= \mathbf{P}\left(W_1 + B_1^0 + A_1(x) > C_1(x), A_2(x) + W_2 \geq cx\right) \\
&+ \mathbf{P}\left(W_1 + B_1^0 + A_1(x) > C_1(x), A_2(x) + W_2 < cx\right) := I + II.
\end{aligned}$$

Consider first the term I. Due to (6.2), the rate allocated to class-1 flows is at least $1 - c$, implying $C_1(x) \geq (1 - c)x$. Consequently,

$$I \leq \mathbf{P}\left(W_1 + B_1^0 + A_1(x) > (1 - c)x, A_2(x) + W_2 \geq cx\right).$$

Due to Proposition 6.4.1, we have $W \leq_{st} \tilde{W}$, where $\tilde{W}$ denotes the workload in the system operating under the modified proportional fair policy. Hence,

$$I \leq \mathbf{P}\left(\tilde{W}_1 + B_1^0 + A_1(x) > (1 - c)x, A_2(x) + \tilde{W}_2 \geq cx\right).$$

Since the queue length distribution under the modified proportional fair policy is known (see (6.4)), the latter can be written as

$$\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} \tilde{\pi}(n_1, n_2)\mathbf{P}\left(\sum_{i=1}^{n_1} B_{1,i}^r + B_1^0 + A_1(x) > (1 - c)x, A_2(x) + \sum_{i=1}^{n_2} B_{2,i}^r \geq cx\right)$$

$$= \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} \tilde{\pi}(n_1, n_2)\mathbf{P}\left(\sum_{i=1}^{n_1} B_{1,i}^r + B_1^0 + A_1(x) > (1 - c)x\right)\mathbf{P}\left(A_2(x) + \sum_{i=1}^{n_2} B_{2,i}^r \geq cx\right).$$

Applying the Chernoff bound independently to the above probabilities with parameters $\delta_1 = \delta_1^{(1-c)}$ and $\delta_2 = \delta_2^{(c)}$, we obtain

$$\begin{aligned} I &\leq \mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})+\alpha_2(\delta_2^{(c)})-(1-c)\delta_1^{(1-c)}-c\delta_2^{(c)})x}\Phi_{B_1}(\delta_1^{(1-c)})\mathbf{E}[(\beta_1^*(\delta_1^{(1-c)}))^{\tilde{X}_1}(\beta_2^*(\delta_2^{(c)}))^{\tilde{X}_2}] \\ &= \mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})+\alpha_2(\delta_2^{(c)})-(1-c)\delta_1^{(1-c)}-c\delta_2^{(c)})x}\Phi_{B_1}(\delta_1^{(1-c)})\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}\tilde{W}_1+\delta_2^{(c)}\tilde{W}_2}]. \end{aligned}$$

The MGF of B_1 is finite by definition of $\delta_1^{(1-c)}$. Since, $\alpha_2'(\delta_2^{(c)}) = c$, finiteness of the MGF $\mathbf{E}[\mathrm{e}^{\delta_1\tilde{W}_1+\delta_2\tilde{W}_2}]$ follows from Proposition 6.4.2. Summarizing the above discussion, we obtain that for some constant M_1,

$$I \leq M_1\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})+\alpha_2(\delta_2^{(c)})-(1-c)\delta_1^{(1-c)}-c\delta_2^{(c)})x}.$$

Let us now turn to term II. Due to the capacity constraints, the rate allocation $\Lambda_2 \leq c$. Hence, we can bound the class-2 workload from below by the workload in a single-node PS queue with capacity c and traffic generated by class-2 flows only, $W_2 \geq W_2^{PS(c)}$. Since class 1 is non-empty during the time interval $[0, x]$, by the bandwidth allocation policy, it follows that link 1 is fully utilized, i.e. $x - C_1(x) = C_2(x)$. Using the fact that $C_2(x) \leq W_2 + A_2(x)$, we derive

$$\begin{aligned} II &\leq \mathbf{P}(W_1 + B_1^0 + A_1(x) > (1-c)x, A_2(x) + W_2 < cx) \\ &\leq -\int_0^c \mathbf{P}(W_1^{PS(1-c)} + B_1^0 + A_1(x) > (1-u)x)d\mathbf{P}\left(A_2(x) + W_2^{PS(c)} \geq ux\right). \end{aligned}$$

Applying the Chernoff bound with parameter $\delta_1 = \delta_1^{(1-c)}$ to the integrand and using integration by parts, we obtain that the latter expression can be bounded from above by

$$\begin{aligned} &-\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}B_1}]\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}W_1^{PS(1-c)}}]\int_0^c \mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})-\delta_1^{(1-c)}(1-u))x}d\mathbf{P}\left(A_2(x) + W_2^{PS(c)} \geq ux\right) \\ &\leq \mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}B_1}]\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}W_1^{PS(1-c)}}]\Big(\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})-\delta_1^{(1-c)})x} \\ &\qquad + \int_0^c \mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})-\delta_1^{(1-c)}(1-u))x}\mathbf{P}\left(A_2(x) + W_2^{PS(c)} \geq ux\right)\delta_1^{(1-c)}xdu\Big). \end{aligned}$$

Applying the Chernoff bound with parameter $\delta_2 = \delta_2^{(c)}$ to the probability under the integral sign we obtain

$$\begin{aligned} II &\leq \mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}B_1}]\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)}W_1^{PS(1-c)}}]\Big(\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})-\delta_1^{(1-c)})x} \\ &\quad + \mathbf{E}[\mathrm{e}^{\delta_2^{(c)}W_2^{PS(c)}}]\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)})+\alpha_2(\delta_2^{(c)})-\delta_1^{(1-c)})x}\delta_1^{(1-c)}x\int_0^c \mathrm{e}^{u(\delta_1^{(1-c)}-\delta_2^{(c)})x}du\Big). \end{aligned}$$

Notice that $\rho_1 < 1-c$, since by Lemma 6.5.1, $\rho_1 \leq 1-u^*$ and we assume $u^* > c$. Since $\rho_2 < c$, the MGFs $\mathbf{E}[\mathrm{e}^{\delta_2^{(c)} W_2^{PS(c)}}]$ and $\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)} W_1^{PS(1-c)}}]$ are finite by Lemma 6.4.1. Under the assumption $u^* > c$, due to Lemma 6.5.1, $\delta_1^{(1-c)} - \delta_2^{(c)} > 0$ and the integral can be bounded from above by $c\mathrm{e}^{c(\delta_1^{(1-c)} - \delta_2^{(c)})x}$. Consequently,

$$\begin{aligned} II &\leq \mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)} B_1}]\mathbf{E}[\mathrm{e}^{\delta_1^{(1-c)} W_1^{PS(1-c)}}] \\ &\times \left(\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)}) - \delta_1^{(1-c)})x} + \mathbf{E}[\mathrm{e}^{\delta_2^{(c)} W_2^{PS(c)}}]\mathrm{e}^{(\alpha_1(\delta_1^{(1-c)}) + \alpha_2(\delta_2^{(c)}) - (1-c)\delta_1^{(1-c)} - c\delta_2^{(c)})x}\right). \end{aligned}$$

Notice that

$$\begin{aligned} &\alpha_1(\delta_1^{(1-c)}) + \alpha_2(\delta_2^{(c)}) - (1-c)\delta_1^{(1-c)} - c\delta_2^{(c)} \\ &= \alpha_1(\delta_1^{(1-c)}) - \delta_1^{(1-c)} + \alpha_2(\delta_2^{(c)}) + c(\delta_1^{(1-c)} - \delta_2^{(c)}) \\ &> \alpha_1(\delta_1^{(1-c)}) - \delta_1^{(1-c)}. \end{aligned}$$

Applying the principle of the largest term [41], taking logarithms, dividing by x, and letting $x \to \infty$ we derive

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) \leq \alpha_1(\delta_1^{(1-c)}) + \alpha_2(\delta_2^{(c)}) - (1-c)\delta_1^{(1-c)} - c\delta_2^{(c)}.$$

6.5.2 Proof of the lower bound

We now proceed by deriving the lower bound. Introduce a probability measure $\mathbf{P}_{\delta_i(\varepsilon)}$ for $\delta_i(\varepsilon) \geq 0$ in such a way that:

$$\mathbf{P}_{\delta_i(\varepsilon)}(A_i \in dx) = \mathrm{e}^{-\alpha_i(\delta_i(\varepsilon))x}\mathbf{P}(A_i \in dx)/\Phi_{A_i}(-\alpha_i(\delta_i(\varepsilon))), \tag{6.30}$$

$$\mathbf{P}_{\delta_i(\varepsilon)}(B_i \in dx) = \mathrm{e}^{\delta_i(\varepsilon)x}\mathbf{P}(B_i \in dx)/\Phi_{B_i}(\delta_i(\varepsilon)), \tag{6.31}$$

for $i = 1, 2$. The parameters $\delta_i(\varepsilon) > 0$ are chosen to satisfy the following properties:

$$\begin{aligned} &\rho_1(\delta_1(\varepsilon)) + \rho_2(\delta_2(\varepsilon)) = 1 + \varepsilon, \\ &\rho_2(\delta_2(\varepsilon)) = \rho_2(\delta_2(0)) - \varepsilon, \end{aligned} \tag{6.32}$$

for some sufficiently small $\varepsilon > 0$. Here $\delta_2(0) = \min(\delta^*, \delta_2^{(c)})$.

Under the new measure the work arrival process $A_i(x)$ is a compound Poisson process with arrival rate $\lambda_i(\varepsilon) = \lambda_i/\Phi_{A_i}(-\alpha_i(\delta_i(\varepsilon))) = \lambda_i\Phi_{B_i}(\delta_i(\varepsilon))$ and flow sizes with the MGFs $\Phi^{\varepsilon}_{B_i}(s) = \Phi_{B_i}(s + \delta_i(\varepsilon))/\Phi_{B_i}(\delta_i(\varepsilon))$. Hence, we can use the Wald martingale [7] w.r.t. probability $\mathbf{P}_{\delta_1(\varepsilon)} \times \mathbf{P}_{\delta_2(\varepsilon)}$ associated with the processes $A_i(x)$,

$$M_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(x) = \mathrm{e}^{\alpha_1(\delta_1(\varepsilon))x - \delta_1(\varepsilon)A_1(x)}\mathrm{e}^{\alpha_2(\delta_2(\varepsilon))x - \delta_2(\varepsilon)A_2(x)}. \tag{6.33}$$

By Theorem XIII.3.2 in [7], we have the following identity

$$\mathbf{P}(V_1 > x) = \mathbf{E}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left[M_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(x)\mathbf{1}(V_1 > x)\right].$$

Let us introduce the event $S_\varepsilon(x) = \{A_i(x) \le (\rho_i(\delta_i(\varepsilon)) + \varepsilon/2)x,\ \forall u \in [0,x],\ i = 1,2\}$. Then,

$$\mathbf{P}(V_1 > x) \ge \mathbf{E}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left[M_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(x)\mathbf{1}(V_1 > x)\mathbf{1}(S_\varepsilon^c(x))\right]. \tag{6.34}$$

Taking logarithms in (6.34), dividing by x and letting $x \to \infty$, we obtain

$$\begin{aligned}&\liminf_{x\to\infty} \tfrac{1}{x}\log \mathbf{P}(V_1 > x)\\ &\ge \alpha_2(\delta_2(\varepsilon)) - \delta_2(\varepsilon)(\rho_2(\delta_2(\varepsilon)) + \varepsilon/2) + \alpha_1(\delta_1(\varepsilon)) - \delta_1(\varepsilon)(\rho_1(\delta_1(\varepsilon)) + \varepsilon/2) \qquad (6.35)\\ &\quad + \liminf_{x\to\infty} \tfrac{1}{x}\log \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x, S_\varepsilon^c(x)).\end{aligned}$$

Consider the last term in (6.35). We now show that $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x, S_\varepsilon^c(x))$ decays subexponentially, that is $\log \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x, S_\varepsilon^c(x)) = o(x)$. We start by bounding it from below,

$$\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x, S_\varepsilon^c(x)) \ge \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x) - \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(S_\varepsilon^c(x)).$$

Consider the second probability:

$$\begin{aligned}\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(S_\varepsilon^c(x)) &\le \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(\frac{A_1(x)}{x} > \rho_1(\delta_1(\varepsilon)) + \varepsilon/2\right)\\ &+ \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(\frac{A_2(x)}{x} > \rho_2(\delta_2(\varepsilon)) + \varepsilon/2\right). \qquad (6.36)\end{aligned}$$

It is easy to see that

$$\lim_{x\to\infty}\frac{1}{x}\mathbf{E}_{\delta_i(\varepsilon)}[A_i(x)] = \lambda_i \mathbf{E}_{\delta_i(\varepsilon)}B_i = \rho_i(\delta_i(\varepsilon)) < \rho_i(\delta_i(\varepsilon)) + \varepsilon/2, \tag{6.37}$$

and

$$\lim_{x\to\infty}\frac{1}{x}\log \mathbf{E}_{\delta_i(\varepsilon)}\left[\mathrm{e}^{sA_i(x)}\right] = \alpha_i(\delta_i(\varepsilon) + s) - \alpha_i(\delta_i(\varepsilon)), \tag{6.38}$$

which is finite around $s = 0$ by Assumption 6.5.1. Applying the Chernoff bound, we obtain

$$\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(\frac{A_1(x)}{x} > \rho_1(\delta_1(\varepsilon)) + \varepsilon/2\right) \le \mathrm{e}^{\gamma_i x},$$

where $\gamma_i = \inf_{s\ge 0}(\alpha_i(\delta_i(\varepsilon)+s) - \alpha_i(\delta_i(\varepsilon)) - (\rho_i(\delta_i(\varepsilon)) + \varepsilon/2)\delta_i(\varepsilon))$, which is negative in view of (6.37)-(6.38) and Assumption 6.5.1. Thus, we conclude that $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(S_\varepsilon^c(x))$ decays exponentially fast in x.

Let us now consider the remaining term $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x)$. For the tagged class-1 flow we can write

$$\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x) = \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(B_1^0 > \int_0^x \frac{\Lambda_1(e_1 + Q^{\varepsilon,p}(u))}{1 + Q_1^{\varepsilon,p}(u)}du\right),$$

where $Q^{\varepsilon,p}$ denotes the number of flows in the system with a single permanent (tagged) class-1 flow, $e_1 = (1,0)$. Due to the monotonicity of tree networks, the per-flow rate allocations in the system with the permanent flow are bounded from above by the per-flow rate allocations in the original queue. Thus, we can conclude that the sample paths of Q_1^{ε} and $Q_1^{\varepsilon,p}$ are related as $Q_1^{\varepsilon} \leq Q_1^{\varepsilon,p}$. For $\delta_1(\varepsilon)$, $\delta_2(\varepsilon)$ chosen as in (6.32), Theorem 5.5.1 and Proposition 5.6.1 imply that there exist constants $\beta_i > 0$, such that as $x \to \infty$, $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(Q_i^{\varepsilon}(x)/x \geq \beta_i\right)$ converges to one. Consequently, for x large enough

$$\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(Q_1^{\varepsilon,p}(u) \geq \beta_1 u,\ u \in [0,x]\right) > 0.$$

Since the rate allocated to class 1 does not exceed one, the probability $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x)$ is greater than or equal to

$$\begin{aligned}
&\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(B_1^0 > \int_0^x \frac{1}{1+Q_1^{\varepsilon,p}(u)}du, Q_1^{\varepsilon,p}(u) \geq \beta_1 u, u \in [0,x]\right)\\
&\quad \geq \quad \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(B_1^0 > \int_0^x \frac{1}{1+\beta_1 u}du\right)\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(Q_1^{\varepsilon,p}(u) \geq \beta_1 u,\ \in [0,x]\right)\\
&\quad = \quad \mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(B_1^0 > \frac{1}{\beta_1}\log\left(1+\beta_1 x\right)\right)\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}\left(Q_1^{\varepsilon,p}(u) \geq \beta_1 u,\ u \in [0,x]\right).
\end{aligned}$$

Since the distribution of B_1 is assumed to have bounded hazard rate, the distribution of B_1 under the change of measure also has bounded hazard rate,

$$\frac{f_{\delta_1(\varepsilon),B_1}(x)}{\mathbf{P}_{\delta_1(\varepsilon)}(B_1 > x)} = \frac{\mathrm{e}^{\delta_1(\varepsilon)x} f_{B_1}(x)}{\int_x^{\infty} \mathrm{e}^{\delta_1(\varepsilon)y} f_{B_1}(y)dy} \leq \frac{f_{B_1}(x)}{\mathbf{P}(B_1 > x)} \leq M.$$

Hence,

$$-\log \mathbf{P}_{\delta_1(\varepsilon)}(B_1 > a\log x) = \int_0^{a\log x} \frac{f_{\delta_1(\varepsilon),B_1}(u)}{\mathbf{P}_{\delta_1(\varepsilon)}(B_1 > u)}du \leq Ma\log x,$$

for any constant $a > 0$, and for some constant $M \in (0,\infty)$. This implies that the probability $\mathbf{P}_{\delta_1(\varepsilon),\delta_2(\varepsilon)}(V_1 > x)$ behaves like $\mathrm{e}^{o(x)}$.

Summarizing the above discussion, we obtain

$$\begin{aligned}
\liminf_{x\to\infty} \frac{1}{x}\log \mathbf{P}(V_1 > x) \quad &\geq \quad \alpha_2(\delta_2(\varepsilon)) - \delta_2(\varepsilon)(\rho_2(\delta_2(\varepsilon)) + \varepsilon/2)\\
&+ \quad \alpha_1(\delta_1(\varepsilon)) - \delta_1(\varepsilon)(\rho_1(\delta_1(\varepsilon)) + \varepsilon/2).
\end{aligned}$$

The remaining step is to let $\varepsilon \to 0$. Due to continuity and monotonicity of $\alpha_i(\cdot)$ and $\rho_i(\cdot)$, we obtain in the first case $\delta_2(\varepsilon) \to \delta_2 = \min(\delta^*, \delta_2^{(c)})$, $\delta_1(\varepsilon) \to \delta_1 = \max(\delta^*, \delta_1^{(1-c)})$. This completes the proof of the lower bound. □

6.5.3 Example: exponential flow sizes

In this subsection we determine the decay rate of the sojourn time distribution in the system with exponentially distributed flow sizes. In this case, the cumulant function can be computed explicitly,

$$\alpha_i(s) = \lambda_i(\Phi_{B_i}(s) - 1) = \lambda_i \left(\frac{\mu_i}{\mu_i - s} - 1 \right), \qquad i = 1, 2.$$

The solution $\delta_i^{(u)}$ of equation

$$\alpha_i'(\delta_i^{(u)}) = \lambda_i \frac{\mu_i}{(\mu_i - \delta_i^{(u)})^2} = u, \qquad u > 0,\ i = 1, 2.$$

is given by

$$\delta_i^{(u)} = \mu_i \left(1 - \sqrt{\frac{\rho_i}{u}} \right), \qquad i = 1, 2,$$

yielding,

$$\alpha_i(\delta_i^{(u)}) = \lambda_i \left(\frac{\mu_i}{\mu_i - \delta_i^{(u)}} - 1 \right) = \lambda_i \left(\sqrt{\frac{u}{\rho_i}} - 1 \right), \qquad i = 1, 2.$$

Substituting the above expressions into Equation (6.28), we obtain

$$\begin{aligned} &\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) \\ &= \begin{cases} -\left(\sqrt{\lambda_1} - \sqrt{(1-u^*)\mu_1} \right)^2 - \left(\sqrt{\lambda_2} - \sqrt{u^*\mu_2} \right)^2, & \text{if } u^* < c, \\ -\left(\sqrt{\lambda_1} - \sqrt{(1-c)\mu_1} \right)^2 - \left(\sqrt{\lambda_2} - \sqrt{c\mu_2} \right)^2, & \text{if } u^* \geq c. \end{cases} \end{aligned} \tag{6.39}$$

Note that the decay rate of the sojourn time is determined as the sum of the decay rate of the busy period in the M/M/1 system with capacity $1 - u^*$ $(1 - c)$ and the class-1 flows and the decay rate of the busy period the M/M/1 system with capacity u^* (c) and the class-2 flows.

Assuming in addition $\mu \equiv \mu_1 \equiv \mu_2$, we can obtain an explicit expression for the value u^*. Solving equation $\delta_1^{(1-u^*)} = \delta_2^{(u^*)}$, we derive

$$u^* = \frac{\rho_2}{\rho_1 + \rho_2}, \qquad \delta^{(u^*)} = \mu(1 - \sqrt{\rho_1 + \rho_2}).$$

Substitution of these expressions into Equation (6.28) gives

$$\begin{aligned} &\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_1 > x) \\ &= \begin{cases} -\left(\sqrt{\lambda_1 + \lambda_2} - \sqrt{\mu} \right)^2, & \text{if } \rho_2 < \frac{c}{1-c}\rho_1, \\ -\left(\sqrt{\lambda_1} - \sqrt{(1-c)\mu} \right)^2 - \left(\sqrt{\lambda_2} - \sqrt{c\mu} \right)^2, & \text{if } \rho_2 \geq \frac{c}{1-c}\rho_1. \end{cases} \end{aligned} \tag{6.40}$$

The above equation implies that if $\mu_1 = \mu_2$ and $u^* < c$, the decay rate coincides with the decay rate in the M/M/1 system with a single traffic class with the arrival rate $\lambda_1 + \lambda_2$ and the mean flow size $1/\mu$.

6.6 Open questions

We conclude this chapter by discussing directions for further research.

6.6.1 Class-2 asymptotics

In the present chapter we investigated the behavior of the sojourn times of class-1 flows in a two-link parking lot. An important question that has not been addressed, concerns the large-deviations delay characteristics of class 2. At this stage we may only conjecture the delay asymptotics.

Conjecture 6.6.1. *Let u^* be the solution of the equation $(\alpha_2')^{-1}(u^*) = (\alpha_1')^{-1}(1-u^*)$, that is*

$$\left(\Phi'_{B_2}\right)^{-1}\left(\frac{u^*}{\lambda_2}\right) = \left(\Phi'_{B_1}\right)^{-1}\left(\frac{1-u^*}{\lambda_1}\right),$$

and let $\delta^* = \delta_2^{(u^*)} = \delta_1^{(1-u^*)}$. Then,

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbf{P}(V_2 > x) = \begin{cases} \alpha_1(\delta^*) + \alpha_2(\delta^*) - \delta^*, & \text{if } u^* < c, \\ \\ \alpha_2(\delta_2^{(c)}c) - c\delta_2^{(c)}, & \text{if } u^* \geq c. \end{cases} \tag{6.41}$$

The conjecture can be motivated in the following manner. As discussed earlier in Section 6.5, if the workload u^* of class 2 in a tilted system is strictly less than the capacity of link 2, the asymptotic behavior of the system coincides with the behavior of a DPS link. Thus, the decay rates of both classes 1 and 2 are given by the decay rate of the residual busy period when link 1 is critically loaded, cf. Theorem 4.5.1. If the capacity constraint is binding, i.e. $u^* = c$, then asymptotically class 2 behaves as in a single-server PS system with capacity c unaffected by class-1 flows.

The crucial difference in the analysis of class 1 and class 2 is due to the difference in the number of links they traverse. Since class-1 flows use link 1 only, it can be analyzed as in a two-class single-link system. We note that the large-deviations analysis in the previous section bears a strong resemblance to the derivations in single-server PS and DPS systems (see [82] and Chapter 4). Class 2 in its turn utilizes two links simultaneously, which in a large-deviations sense leads to two different regimes: critical load on link 1 and on link 2. See also the conjecture for networks with general topology in the next subsection.

The derivation of the upper bound for class 2 is significantly more involved in comparison to class 1. While the rate allocation policy guarantees a certain minimum rate to class-1 flows, it does not provide such protection to class 2. Furthermore, the presence of class-2 flows does not necessarily imply saturation of the root link and can not guarantee work-conserving behavior of the network. For instance,

the bound as in Equation (6.29) is not valid for the sojourn time of class 2. The lower bound seems to follow naturally from the rate allocation policy. Due to the capacity constraint, the departure rate of the class-2 flows is lower than in a PS system with capacity c. Thus, the sojourn time in a PS system with traffic of class 2 provides one natural lower bound. The other lower bound may be derived by removing the capacity constraint c at the second link. However, albeit the argument is very intuitive, it seems difficult to make it rigorous. At this point we leave this topic for future investigations.

6.6.2 More general networks

Although in the present chapter we focused on a two-link network, the heuristics developed can be extended to networks with a more general topology. The large-deviations analysis in the previous sections indicates that there are two main scenarios that (in combination) can cause a large sojourn time in a system with light-tailed flow sizes. These are (i) a large amount of work upon arrival of the tagged flow and (ii) a large amount of work brought by other flows during the sojourn time of the tagged flow. In a single-node PS system scenario (ii) dominates the sojourn time large-deviations [82]. Based on the similar results derived in the present chapter, we expect the same effect in more general networks.

Conjecture 6.6.2. *Introduce the set* $\mathcal{C}_j(\alpha) = \{\delta \in \mathbb{R}_+^I : \sum_{i=1}^I A_{ji}\alpha_i'(\delta_i) \leq C_j\}$, $j = 1, \ldots, J$. *Then, for all* $k = 1, \ldots, I$,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_k > x) = \max_{j:A_{jk}=1} \sup_{\delta\in\mathcal{C}_j(\alpha)} \sum_{i=1}^{I} A_{ji} \left(\alpha_i(\delta_i) - \alpha_i'(\delta_i)\delta_i\right). \tag{6.42}$$

In plain words, we conjecture that the most probable way for a large sojourn time to occur is due to a critical traffic load at a link on its route. The sojourn time decay rate can be viewed as the minimum of the per-link decay rates when each link is viewed in isolation. The per-link decay rate is a sum of the decay rates of the flows of all the classes which share the link in such a way that their total load or equivalently, the total rate allocation does not exceed the link capacity. The decay rate is fully determined by the distributions of the arrival process and the flow sizes.

In the case of a two-link parking lot, $I = 2$, $J = 2$, the conjecture takes the following form. Take $k = 1$. Since class 1 uses link 1 only, we have

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_k > x) = \sup_{\delta\in\mathcal{C}_1(\alpha)} (\alpha_1(\delta_1) - \alpha_1'(\delta_1)\delta_1 + \alpha_2(\delta_2) - \alpha_2'(\delta_2)\delta_2), \tag{6.43}$$

with $\mathcal{C}_1(\alpha) = \{\delta_1, \delta_2 > 0 : \alpha_1'(\delta_1) + \alpha_2'(\delta_2) \leq 1,\ \alpha_2'(\delta_2) \leq c\}$, which after the optimization procedure gives exactly (6.28). For $k = 2$, we have

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbf{P}(V_k > x) = \max\{\gamma_1, \gamma_2\}$$

where

$$\gamma_1 = \sup_{\delta \in \mathcal{C}_1(\alpha)} (\alpha_1(\delta_1) - \alpha_1'(\delta_1)\delta_1 + \alpha_2(\delta_2) - \alpha_2'(\delta_2)\delta_2),$$

$$\gamma_2 = \sup_{\delta_2 \in \mathcal{C}_2(\alpha)} (\alpha_2(\delta_2) - \alpha_2'(\delta_2)\delta_2), \qquad \mathcal{C}_2(\alpha) = \{\delta_2 > 0 : \alpha_2'(\delta_2) \leq c\}.$$

This result is consistent with Conjecture 6.6.1

A rigorous proof for a general topology and general rate allocation policy is very challenging. The first crucial step is to verify that scenario (i) has little effect on the delay asymptotics. In principle, this task requires the knowledge of the steady-state workload distribution. In our proof for the two-link network, we succeeded to eliminate the problem by deriving sufficiently sharp upper bounds for the workload which allow for the large-deviations derivations. This approach may be extended in a straightforward manner to different network topologies with the proportional fair rate allocation which in consequence may lead to the large-deviations asymptotics.

Another important issue is the non-work-conserving nature of the network. It significantly complicates the analysis even in networks with a simple topology, see the discussion in the previous subsection.

An important role in the proof is played by the insight that for bandwidth-sharing networks under overload the queue length increases roughly at a linear rate. For the two-link parking lot network, we have proved in Section 5.6 using a fluid-limit approach that the number of flows grows at least at a linear rate if the stability conditions are not satisfied. An appropriate change of measure and the fluid-limit result enabled us to derive the asymptotic lower bound. Unfortunately such results are only available for a certain class of networks and overload conditions, see Chapter 5.

Bibliography

[1] Abate, J., Choudhury, G.L., Whitt, W. (1994). Waiting-time tail probabilities in queues with long tail service-time distributions. *Queueing Systems* 16:311–338.

[2] Abate, J., Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10:5–88.

[3] Abate, J., Whitt, W. (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems* 25:173–233.

[4] Altman, E., Avrachenkov, K.E., Ayesta, U. (2006). A survey on discriminatory processor sharing. *Queueing Systems* 53:53–63.

[5] Altman, E., Jimenez, T., Kofman, D. (2004). DPS queues with stationary ergodic service times and the performance of TCP in overload. In: *Proceedings of IEEE Infocom*, Hong-Kong.

[6] Anantharam, V. (1989). How large delays build up in a GI/G/1 queue. *Queueing Systems* 5:345–367.

[7] Asmussen, S. (2003). *Applied Probability and Queues*. Springer-Verlag, New York.

[8] Avrachenkov, K.E., Ayesta, U., Brown, P., Núñez-Queija, R. (2005). Discriminatory processor sharing revisited. In: *Proceedings of IEEE Infocom*, Miami, 784–795.

[9] Bazaraa, M.S., Sherali, H.D., Shetty, C.M. (1993). *Nonlinear Programming: Theory and Algorithms.* John Wiley, New Jersey.

[10] Bekker, R., Borst, S.C., Núñez-Queija, R. (2004). Performance of TCP-friendly streaming sessions in the presence of heavy-tailed elastic flows. *Performance Evaluation* 61:143–162.

[11] Ben Fredj, S., Bonald, T., Régnié, G., Roberts, J.W. (2001). Statistical bandwidth sharing: a study of congestion at flow level. In: *Proceedings of ACM SIGCOMM*, San Diego, 111–122.

[12] Ben-Tal, A., Nemirovski, A. (2001). *Lectures on Modern Convex Optimization.* MPS-SIAM Series on Optimization, SIAM, Philadelphia.

[13] Van den Berg, J.L. (1990). Sojourn Times in Feedback and Processor-Sharing Queues. *PhD Thesis*, Utrecht University.

[14] Bonald, T., Massoulié, L. (2001). Impact of fairness on Internet performance. In: *Proceedings of ACM Sigmetrics & Performance Conference*, Boston, 82–91.

[15] Bonald, T., Massoulié, L., Proutière, A., Virtamo, J. (2006). A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems* 53:65–84.

[16] Bonald, T., Proutière, A. (2002). Insensitivity in processor-sharing networks. *Performance Evaluation* 49:193–209.

[17] Bonald, T., Proutière, A. (2003). Insensitive bandwidth sharing in data networks. *Queueing Systems* 44:69–100.

[18] Bonald, T., Proutière, A. (2004). On stochastic bounds for balanced fairness. *Performance Evaluation* 55:25–50.

[19] Bonald, T., Proutière, A. (2004). On stochastic bounds for monotonic processor-sharing networks. *Queueing Systems* 47:81–106.

[20] Bonald, T., Roberts, J.W. (2001). Performance modeling of elastic traffic in overload. In: *Proceedings of ACM Sigmetrics & Performance Conference*, Boston, 342–343.

[21] Bonnans, J.F., Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems.* Springer, New York.

[22] Borst, S.C., Boxma, O.J., Morrison, J.A., Núñez-Queija, R. (2003). The equivalence between processor sharing and service in random order. *Operations Research Letters* 31:254–262.

[23] Borst, S.C., Egorova, R., Zwart, A.P. (2009). Fluid limits for bandwidth-sharing networks in overload. In preparation.

[24] Borst, S.C., Núñez-Queija, R., van Uitert, M.J.G. (2002). User-level performance of elastic traffic in a differentiated-services environment. *Performance Evaluation* 49:507–519.

[25] Borst, S.C., Núñez-Queija, R., Zwart, A.P. (2006). Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* 53:31–51.

[26] Borst, S.C., Van Ooteghem, D., Zwart, A.P. (2005). Tail asymptotics for discriminatory processor sharing queues with heavy-tailed service requirements. *Performance Evaluation* 61:281–298.

[27] Boyd, S., Vanderberghe, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge.

[28] Bramson, M. (1994). Instability of FIFO queueing networks. *Annals of Applied Probability* 4:414–431.

[29] Bramson, M. (2005). Stability of networks for max-min fair routing. Presentation at the 13th INFORMS Applied Probability Conference, Ottawa.

[30] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Theory of Probability and its Applications* 10:323–331.

[31] Chang, C.-S. (1995). Sample path large deviations and intree networks. *Queueing Systems* 20:7–36.

[32] Chen, H., Yao, D.D. (2001). *Fundamentals of Queueing Networks.* Springer, New York.

[33] Cheung, S.-K. (2007). Processor-Sharing Queues and Resource Sharing in Wireless LANs. *PhD Thesis*, University of Twente.

[34] Chiang, M., Shah, D., Tang, A. (2006). Stochastic stability of network utility maximization: General file size distribution. In: *Proceedings of 44th Allerton Conference on Computation, Communication and Control*, Urbana.

[35] Coffman, E.G., Muntz, R., Trotter, H. (1970). Waiting time distributions for processor-sharing systems. *Journal of the ACM* 17:123–130.

[36] Cohen, J.W. (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* 12:245–284.

[37] Cohen, J.W. (1982). *The Single Server Queue.* North Holland, Amsterdam.

[38] Cox, D.R., Smith, W.L. (1961). *Queues.* Methuen, London.

[39] Dai, J.G. (1995). On positive Harris recurrence of multi-class queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability* 5:49–77.

[40] Delcoigne, F., Proutière, A., Régnié, G. (2004). Modeling integration of streaming and data traffic. *Performance Evaluation* 55:185–209.

[41] Dembo, A., Zeitouni, O. (1998). *Large Deviation Techniques and Applications.* Springer, New York.

[42] Den Iseger, P. (2006). Numerical inversion of Laplace transforms using a Gaussian quadrature. *Probability in the Engineering and Informational Sciences* 20:1–44.

[43] Denisov, D., Zwart, A.P. (2007). On a theorem of Breiman and a class of random difference equations. *Journal of Applied Probability* 44:1031–1046.

[44] Egorova, R., Borst, S.C., Zwart, A.P. (2007). Bandwidth-sharing networks in overload. *Performance Evaluation* 64:978–993.

[45] Egorova, R., Borst, S.C., Zwart, A.P. (2008). Bandwidth sharing in overloaded networks. In: *Proceedings of IEEE CISS*, Princeton, 36–41.

[46] Egorova, R., Mandjes, M.R.H., Zwart, A.P. (2007). Sojourn time asymptotics in processor sharing queues with varying service rate. *Queueing Systems* 56:169–181.

[47] Egorova, R., Zwart, A.P. (2007). Tail behavior of conditional sojourn times in processor-sharing queues. *Queueing Systems* 55:107–121.

[48] Egorova, R., Zwart, A.P. (2009). Sojourn time asymptotics in a parking lot network. In preparation.

[49] Egorova, R., Zwart, A.P., Boxma, O.J. (2006). Sojourn time tails in the M/D/1 processor sharing queue. *Probability in the Engineering and Informational Sciences* 20:429–446.

[50] Erlang, A.K. (1909). Sandsynlighedsregning og telefonsamtaler. (in Danish) *Nyt Tidsskrift for Matematic B* 20:33–39.

[51] Fayolle, G., De la Fortelle, A., Lasgouttes, J.-M., Massoulié, L., Roberts, J.W. (2001). Best-effort networks: modeling and performance analysis via large network asymptotics. In: *Proceedings of IEEE Infocom*, Anchorage, 709–716.

[52] Fayolle, G., Mitrani, I., Iasnogorodski, R. (1980). Sharing a processor among many job classes. *Journal of the ACM* 27:519–532.

[53] Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Volume II.* Wiley, New York.

[54] Flatto, L. (1997). The waiting time distribution for the random order service M/M/1 queue. *Annals of Applied Probability* 7:382–409.

[55] Glynn, P.W., Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* 31A:131–156.

[56] Grishechkin, S. (1992). On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability* 24:653–698.

[57] Gromoll, H.C., Puha, A.L., Williams, R.J. (2002). Fluid limit of a heavily loaded processor sharing queue. *Annals of Applied Probability* 12:1–63.

[58] Gromoll, H.C., Robert, Ph., Zwart, A.P. (2008). Fluid limits for processor sharing queues with impatience. *Mathematics of Operations Research* 33:375–402.

[59] Gromoll, H.C., Robert, Ph., Zwart, A.P., Bakker, R. (2006). The impact of reneging in processor sharing queues. *Performance Evaluation Review*, 34:87–96.

[60] Gromoll, H.C., Williams, R.J. (2009). Fluid limit of a network with fair bandwidth sharing and general document size distribution. *Annals of Applied Probability*, to appear.

[61] Gromoll, H.C., Williams, R.J. (2007). Fluid model for a data network with alpha-fair bandwidth sharing and general document size distributions: two examples of stability. *Markov Processes and Related Topics*, 4:253–265.

[62] Guillemin, F., Robert, Ph., Zwart, A.P. (2004). Tail asymptotics for processor-sharing queues. *Advances in Applied Probability* 36:525–543.

[63] Jean-Marie, A. (2005). On overloaded queues. Presentation at 30th Conference on the Mathematics of Operations Research, Lunteren, The Netherlands. http://www.lirmm.fr/∼ajm/Talks/18Jan2005/transparents.pdf.

[64] Jean-Marie, A., Robert, Ph. (1994). On the transient behavior of the processor sharing queue. *Queueing Systems* 17:129–136.

[65] Jelenković, P.R., Momčilović, P. (2003). Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operations Research* 28:587–608.

[66] Kalashnikov, V., Tsitsiashvili, G. (1999). Tails of waiting times and their bounds. *Queueing Systems* 32:257–283.

[67] Kella, O., Zwart, A.P., Boxma, O.J. (2005). Some transient properties of symmetric M/G/1 queues. *Journal of Applied Probability* 42:223–234.

[68] Kelly, F.P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.

[69] Kelly, F.P. (2008). The mathematics of traffic in networks. In: Gowers, T., Barrow-Green, J., Leader, I. (editors). *The Princeton Companion to Mathematics.* Princeton University Press.

[70] Kelly, F.P., Maulloo, A., Tan, D. (1998). Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society* 49:237–252.

[71] Kelly, F.P., Williams, R.J. (2004). Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability* 14:1055–1083.

[72] Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chains. *Annals of Mathematical Statistics* 24:338–354.

[73] Kesidis, G., Walrand, J., Chang, C.-S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking* 1:424–428.

[74] Kim, J., Kim, B. (2004). Sojourn time distribution in the M/M/1 queue with discriminatory processor-sharing. *Performance Evaluation* 58:341–365.

[75] Kirk, W.A., Sims, B. (2001). *Handbook of Metric Fixed Point Theory*. Kluwer Academic Publishers, Dordrecht.

[76] Kleinrock, L. (1964). Analysis of a time-shared processor. *Naval Research Logistics Quarterly* 11:59–73.

[77] Kleinrock, L. (1976). *Queueing Systems*. John Wiley, New York.

[78] Kleinrock, L. (1976). Time-shared systems: A theoretical treatment. *Journal of the ACM* 14:242–261.

[79] Van Leeuwaarden, J.S.H., Löpker, A.H., Janssen, A.J.E.M. (2008). Connecting renewal age processes and M/D/1 processor sharing queues through stick breaking. Eurandom report 2008-017. www.eurandom.tue.nl/reports/2008/017-AL-abstract.pdf.

[80] Lu, S.H., Kumar, P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control* 36:1406–1416.

[81] Mandjes, M., Nuyens, M. (2004). Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the Engineering and Informational Sciences* 19:351–361.

[82] Mandjes, M., Zwart, A.P. (2006). Large deviations for sojourn times in processor sharing queues. *Queueing Systems* 52:237–250.

[83] Massoulié, L. (2007). Structural properties of proportional fairness: stability and insensitivity. *Annals of Applied Probability* 17:809–839.

[84] Massoulié, L., Roberts, J.W. (1999). Bandwidth sharing: objectives & algorithms. In: *Proceedings of IEEE Infocom*, New York, 1395–1403.

[85] Mo, J., Walrand, J. (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* 8:556–567.

[86] Morrison, J.A. (1985). Response-time distribution for a processor-sharing system. *SIAM Journal on Applied Mathematics* 45:152–167.

[87] Núñez-Queija, R. (2000). Processor-sharing models for integrated-services networks. *PhD thesis,* Eindhoven University of Technology.

[88] Nuyens, M., Zwart, A.P. (2005). A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems* 54:85–97.

[89] Ott, T.J. (1984). The sojourn-time distribution in the M/G/1 queue with processor sharing. *Journal of Applied Probability* 21:360–378.

[90] Padhye, J., Firoiu, V., Towsley, D., Kurose, J. (2000). Modeling TCP Reno performance: A simple model and its empirical validation. *IEEE/ACM Transactions on Networking* 8:133–145.

[91] Palmowski, Z., Rolski, T. (2006). On the exact asymptotics of the busy period in GI/G/1 queues. *Advances in Applied Probability* 38:792–803.

[92] Puha, A.L., Stolyar, A.L., Williams, R.J. (2006). The fluid limit of an overloaded processor sharing queue. *Mathematics of Operations Research* 31:316–350.

[93] Stolyar, A.L., Ramanan, K. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Annals of Applied Probability* 11:1–48.

[94] Ramaswami, V. (1984). The sojourn time in the GI/M/1 queue with processor sharing. *Journal of Applied Probability* 21:437–442.

[95] Rege, K.M., Sengupta, B. (1994). A decomposition theorem and related results for the discriminatory processor sharing queue. *Queueing Systems* 18:333–351.

[96] Rege, K.M., Sengupta, B. (1996). Queue length distribution for the discriminatory processor sharing queue. *Operations Research* 44:653–657.

[97] Roberts, J.W. (2000). Engineering for quality of service. In: Park, K., Willinger, W. (editors). *Self-Similar Network Traffic and Performance Evaluation.* Wiley, New York, 401–420.

[98] Roberts, J.W. (2001). Traffic theory and the Internet. *IEEE Communications Magazine* 39:94–99.

[99] Roberts, J.W., Massoulié, L. (1998). Bandwidth sharing and admission control for elastic traffic. In: *Proceedings of ITC Specialist Seminar*, Yokohama.

[100] Ross, S.M. (1996). *Stochastic Processes*. John Wiley, New York.

[101] Rybko, A.N., Stolyar, A.L. (1992). Ergodicity of stochastic processes describing the operations of open queueing networks. *Problemy Peredachi Informatsii* 28:3–26.

[102] Sakata, M., Noguchi, S., Oizumi, J. (1969). Analysis of a processor shared queueing model for time sharing systems. *Proceedings of the 2nd Hawaii International Conference on System Sciences*, 625–628.

[103] Sarangapani, J. (2007). *Wireless Ad Hoc and Sensor Networks: Protocols, Performance, and Control*. CRC Press.

[104] Schassberger, R. (1984). A new approach to the M/G/1 processor sharing queue. *Advances in Applied Probability* 16:802–813.

[105] Sengupta, B. (1992). An approximation for the sojourn-time distribution for the GI/G/1 processor-sharing queue. *Stochastic Models* 8:35–57.

[106] Sengupta, B., Jagerman, D.L. (1985). A conditional response time of the M/M/1 processor sharing queue. *AT&T Technical Journal* 64:409–421.

[107] Tijms, H.C. (2003). *A First Course in Stochastic Models*. John Wiley, Chichester.

[108] Van Kessel, G., Núñez-Queija, R., Borst, S.C. (2004). Asymptotic regimes and approximations for discriminatory processor sharing. *Performance Evaluation Review* 32:44–46.

[109] Van Uitert, M.J.G. (2003). Generalized Processor Sharing Queues. *PhD thesis,* Eindhoven University of Technology.

[110] De Veciana, G., Lee, T.-L., Konstantopoulos, T. (1999). Stability and performance analysis of networks supporting services with rate control – could the Internet be unstable? In: *Proceedings of IEEE Infocom*, New York, 802–810.

[111] De Veciana, G., Lee, T.-L., Konstantopoulos, T. (2001). Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking* 9:2–14.

[112] Virtamo, J. (2003). Insensitivity of a network of symmetric queues with balanced service rates. Internal Report.
http://netlab.hut.fi/tutkimus/fit/publ/insens_symm_qn.pdf.

[113] Widder, D.V. (1946). *The Laplace Transform*. Princeton University Press, Princeton.

[114] Whitt, W. (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems* 2:71–107.

[115] Whitt, W. (2002). *Stochastic-Process Limits*. Springer, New York.

[116] Yashkov, S.F. (1983). A derivation of response time distribution for a M/G/1 processor-sharing queue. *Problems of Control and Information Theory* 12:133–148.

[117] Yashkov, S.F. (1987). Processor-sharing queues: some progress in analysis. *Queueing Systems* 2:1–17.

[118] Yashkov, S.F. (1993). On a heavy-traffic limit theorem for the M/G/1 processor-sharing queue. *Stochastic Models* 9:467–471.

[119] Yi, Y., Chiang, M. (2007). Stochastic network utility maximization. *European Transactions on Telecommunications* 19:421–442.

[120] Zhang, Z.-L. (1997). Large deviations and the generalized processor sharing scheduling for a two-queue system. *Queueing Systems* 26:229–254.

[121] Zwart, A.P. (2001). Queueing Systems with Heavy Tails. *PhD Thesis*, Eindhoven University of Technology.

[122] Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems* 35:141–166.

Summary

The processor-sharing discipline was originally introduced as a modeling abstraction for the design and performance analysis of the processing unit of a computer system. Under the processor-sharing discipline, all active tasks are assumed to be processed simultaneously, receiving an equal share of the server capacity. Various extensions of the basic egalitarian discipline have been developed in order to capture scenarios with heterogeneous service shares and network settings. Over the past several years, the processor-sharing discipline has received renewed attention as a powerful tool in modeling and analyzing dynamic bandwidth sharing among elastic transfers in communication networks like the Internet.

The sojourn time of a customer, i.e. the amount of time a customer spends in the system from his arrival until his service completion, is the most important performance measure for processor-sharing systems. In this monograph we study various asymptotic properties of the sojourn time distribution. The advantage of considering the asymptotic behavior is that the analysis often provides insight into the typical scenario for a long sojourn time to occur. Moreover, the resulting asymptotic formulas can be used for approximate analysis, providing accurate estimates in situations where numerical procedures become unreliable. In order to analyze the sojourn time asymptotics, we apply several probabilistic and analytic techniques, such as Laplace transforms, branching arguments, large-deviations methods and fluid limits.

The main focus in this thesis is on the PS queue where the service time has a light-tailed distribution. This case has received relatively little attention compared to the case of heavy-tailed distributions. Exact asymptotics (of highly uncommon and interesting form) were only available for the M/M/1 queue and were obtained by analytical methods that did not provide insight into the nature of the underlying rare event.

In Chapter 2 we analyze the asymptotic behavior of the sojourn time distribution in the classical single-node PS queue. We derive exact tail asymptotics for the sojourn time distribution in the queue with Poisson arrivals and deterministic service times. The proof involves a geometric random sum representation of the sojourn time, and a connection with Yule processes. Numerical experiments demonstrate a remarkable accuracy of the asymptotic approximation.

In Chapter 3 we consider the M/G/1 queue, and investigate the tail behavior of the sojourn time distribution for a request of a given length. An exponential

asymptote is proved for general service times in two special cases: when the traffic load is sufficiently high and when the request length is sufficiently small. Using the branching process technique, we derive exact asymptotics of exponential type for the sojourn time in the M/M/1 queue. We study the accuracy of the exponential asymptote using numerical methods.

In Chapter 4 we study the GI/GI/1 queue operating under a PS discipline with stochastically varying service rate. The focus is on logarithmic estimates of the tail of the sojourn time distribution, under the assumption that the service time distribution has a light tail. The analysis in this chapter relies predominantly on large-deviations techniques. Furthermore, we extend our results to a similar system operating under the discriminatory processor-sharing discipline.

In Chapters 5 and 6 we analyze the behavior of alpha-fair bandwidth-sharing networks which can be regarded as generalizations of a processor-sharing discipline from a single node to a network with several shared links. In Chapter 5 we focus on an overload scenario where the traffic load on one or several of the links exceeds the capacity. In order to characterize the overload behavior, we examine the fluid limit, which emerges from a suitably scaled version of the number of flows of the various classes. We derive a functional equation characterizing the fluid limit. We show that any strictly positive solution must be unique, which in particular implies the convergence of the scaled number of flows to the fluid limit for nonzero initial states when the traffic load is sufficiently high. In addition, we establish the uniqueness of the fluid limit for networks with a tree topology. For the case of a zero initial state and zero-degree homogeneous rate allocation functions, we show that there exists a uniquely determined linear solution to the fluid-limit equation, and obtain a fixed-point equation for the corresponding asymptotic growth rates. The results are illustrated for parking lot, linear and star networks as important special cases. We briefly discuss extensions to models with user impatience.

In Chapter 6 we derive the asymptotics for the sojourn time distribution in a specific type of bandwidth-sharing network: a parking lot network. Such networks can be practically useful in modeling access networks consisting of several multiplexing stages. Using large-deviations techniques and the fluid-limit results from Chapter 5, we obtain the logarithmic asymptote under the assumption that flow sizes have a light-tailed distribution. In addition, we derive stochastic bounds for the number of flows and the workload in the system.

About the author

Regina Robertovna Egorova was born in Batumi, Georgia, on March 18, 1980. She studied Applied Mathematics at the Ufa State Aviation Technical University, Russia, where she graduated with honors from the Faculty of General Sciences in July 2002. In July 2004 she received her M.Sc. degree in Risk Analysis and Environmental Modeling from Delft University of Technology, The Netherlands. Subsequently, she became a Ph.D. student at CWI (Center for Mathematics and Computer Science, Amsterdam) and Eindhoven University of Technology under the supervision of Sem Borst, Onno Boxma and Bert Zwart. Regina defends her thesis on February 5, 2009. As of September 2008, she works as a financial engineer at Cardano in Rotterdam.