Scheduling in Stochastic Resource-Sharing Systems

Thomas Stieltjes Institute
for Mathematics

Scheduling in Stochastic Resource-Sharing Systems / by Ina Maria Verloop

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-2056-5
NUR: 919
Subject headings: queueing theory, optimal scheduling, performance evaluation, communication networks, bandwidth-sharing networks, parallel-server models
2000 Mathematics Subject Classification: 60K25, 68M20, 90B15, 90B18, 90B22, 90B36

Printed by Ponsen & Looijen b.v.
This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 613.000.436

Scheduling in Stochastic Resource-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 26 november 2009 om 16.00 uur

door

Ina Maria Verloop

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotoren:

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

Copromotor:
dr. R. Núñez Queija

Dankwoord (Acknowledgements)

Dit proefschrift beschrijft het promotieonderzoek dat ik in de periode september 2005 tot en met augustus 2009 heb uitgevoerd op het CWI. Het had niet tot stand kunnen komen zonder de hulp van velen. Ik maak dan ook graag van deze gelegenheid gebruik om een aantal mensen speciaal te bedanken.

Allereerst ben ik Sem Borst en Sindo Núñez Queija veel dank verschuldigd voor de zeer plezierige en stimulerende samenwerking tijdens dit promotieonderzoek. Sem, jouw scherpe en constructieve opmerkingen, welke me altijd de juiste richting opstuurden, heb ik erg gewaardeerd. Sindo, ik ben je ontzettend dankbaar voor onze frequente en altijd zeer leerzame discussies. Kortom, ik had me geen betere begeleiders kunnen wensen. Verder wil ik Onno Boxma bedanken voor het nauwkeurig doornemen van het gehele proefschrift en zijn nuttige suggesties. Het CWI ben ik erkentelijk voor de mij ter beschikking gestelde faciliteiten en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) voor de financiële ondersteuning.

Ik kijk met veel plezier terug op mijn promotietijd, wat niet in de laatste plaats te danken is aan de goede sfeer in de PNA2 groep op het CWI. In het bijzonder wil ik hiervoor mijn mede oio's, Regina, Pascal, Chretien en Wemke, bedanken. Daarnaast ben ik Bala dankbaar voor de gezellige tijd in Anchorage en voor zijn snelle en goede hulp bij LaTeX. Matthieu, ik heb veel geleerd van onze discussies over stochastische orderingen, alsook van je salsalessen waar ik altijd met veel plezier heen ging. Veel dank hiervoor. Verder wil ik graag de oio's in kamer 10.14 op de TU/e bedanken voor hun gastvrijheid en gezelligheid.

My three-month visit at INRIA Paris-Rocquencourt, which was financially supported by EURO-NF, has been a valuable experience abroad. I am grateful to Philippe Robert for offering me this opportunity and making it a pleasant and fruitful stay, as well as to the members of the RAP-group for their generous hospitality.

Tot slot wil ik ook een aantal mensen op persoonlijk vlak bedanken. Als eerste noem ik graag mijn ouders. Ik ben jullie zeer dankbaar voor alle steun in de afgelopen jaren. Annewieke en Lisa, bedankt dat jullie het aandurven om me als paranimfen bij te staan tijdens de verdediging van dit proefschrift. Ook dank ik Renske voor de gezellige uurtjes in Capelle aan den IJssel en Amsterdam. Urtzi, jij hebt zowel op wetenschappelijk als persoonlijk vlak een onmisbare rol in de totstandkoming van dit proefschrift gespeeld. Ik ben je dankbaar voor je onvoorwaardelijke steun en vertrouwen in mij.

Maaike Verloop
September, 2009

Contents

Chapter 1
Introduction

Sharing resources among multiple users is common in daily life. One may think of resources such as lanes on a highway, agents in a call center, the processing capacity of a computer system, the available bandwidth in communication systems, or the transmission power of wireless base stations. In each of these situations, some scheduling mechanism regulates how the resources are shared among competing users. It is not always clear what the "best" way is to do this. Besides efficient use of the available resources in order to meet the demand, issues like fairness and the performance perceived by the users are important as well.

The random nature of arrivals of new users, and of their corresponding service characteristics, motivates the study of queueing-theoretic models. In this thesis we concentrate on three queueing models in particular: single-server systems, bandwidth-sharing networks, and parallel-server models. These models arise in the context of scheduling in communication networks. We are interested in finding scheduling policies that optimize the performance of the system, and evaluating policies that share the resources in a fair manner. Whenever possible, we do this directly for the stochastic queueing model. Otherwise, we resort to asymptotic regimes: we either let the offered work approach the available capacity or consider a related deterministic fluid model.

This first chapter serves as background on the content of the thesis and is organized as follows. In Section 1.1 we describe the essential characteristics of resource-sharing systems and introduce the notions of efficient and fair scheduling. In Section 1.2 we provide several examples of communication networks that motivate our study of resource-sharing systems. The queueing models are introduced, and a literature overview is given in the subsequent sections: in Section 1.3 for single-server queues, in Section 1.4 for bandwidth-sharing networks and in Section 1.5 for parallel-server models. In Section 1.6 we describe the main techniques and concepts used throughout the thesis. Section 1.7 concludes this chapter with an overview of the thesis.

1.1 Scheduling in resource-sharing systems

Deciding how to share the resources among users contending for service is a complicated task. This is in particular due to the following two elements. First of all, it is uncertain at what time new jobs arrive to the system and what amount or what kind of service they require. Second, the capacity of the resources is finite and there may be additional constraints on the way the resources can be shared among the various jobs. For example, some types of jobs might be processed faster by certain specialized resources, some types of jobs might need capacity from several resources simultaneously, etc.

In order to mathematically model the dynamic behavior of a resource-sharing system, we investigate queueing-theoretic models that capture the two elements as mentioned above. A queueing model consists of several servers with finite capacity, which can be allocated to users, possibly subject to additional constraints. The arrivals of new users and the amount and type of service they require, are described by stochastic processes.

The evolution of a queueing model is determined by the employed scheduling policy, which specifies at each moment in time how the capacity of the servers is shared among all users contending for it. An important body of the scheduling literature is devoted to seeking a policy that *optimizes the performance* of the queueing model. The latter may be expressed in terms of performance measures such as throughput, holding cost, user's delay, and the number of users in the system. Besides performance, another important notion is *fairness*. This relates to maintaining some level of "social justice", i.e., fairness in treatment of the users. Fairness is a subjective notion and much research has been devoted to developing quantitative measures [11].

A well-studied queueing model is the work-conserving single-server system, as will be described in Section 1.3. This system works at full speed whenever there is work in the system. Apart from this model, in this thesis we focus on multi-class resource-sharing systems that can be seen as an extension of the single-server queue. More specifically, we study models where the total used capacity might not be constant over time and may depend for instance on the scheduling decision taken or on the types of users presently in the system. The fact that the scheduling decisions affect the total used capacity significantly complicates the task of designing optimal and fair scheduling policies.

In the remainder of this section we introduce in more detail the notions of optimal and fair scheduling. We make a distinction between the static regime and the dynamic regime, which are treated in Sections 1.1.1 and 1.1.2, respectively. In the static regime the population of users is fixed, while the dynamic regime allows for departures and arrivals of new users.

1.1.1 Static setting

In this section we describe the notions of optimal and fair scheduling in a static setting. For a given population of users, indexed by $i = 1, \ldots, I$, we consider different ways to allocate the available capacity among the users. Let x_i be the rate allocated to user i and let $\vec{x} = (x_1, \dots, x_I)$ be the rate allocation vector. The set consisting of all feasible rate allocation vectors is denoted by S. Besides the fact that the capacity of the servers is finite, the shape of S is determined by additional constraints on the way the capacity of the servers can be shared among the users.

In a static setting it is natural to measure the performance in terms of the total throughput $\sum_{i=1}^{I} x_i$. A feasible allocation that maximizes the total throughput may be called optimal in the static setting. However, this optimal allocation does not guarantee that all users are allocated a strictly positive rate. It can be the case that some types of users obtain no capacity at all, which is highly unfair.

A commonly used definition of fairness has its origin in microeconomics. It relies on a social welfare function, which associates with each possible rate allocation the aggregate utility of the users in the system [91]. A feasible allocation is called fair when it maximizes the social welfare function, i.e., an $\vec{x} \in S$ that solves

$$\max_{\vec{x} \in S} \sum_i U_i(x_i), \tag{1.1}$$

with $U_i(x_i)$ the utility of allocating rate x_i to user i. When the functions $U_i(\cdot)$ are strictly concave and the set S is convex and compact, the maximization problem has a unique solution. An important class of utility functions as introduced in [100] is described by

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

with $w_i > 0$ a weight assigned to user i, $i = 1, \dots, I$. The fact that these functions are increasing and strictly concave forces fairness between users: increasing the rate of a user that was allocated a relatively little amount, yields a larger improvement in the aggregate utility. The corresponding allocation that solves the optimization problem (1.1) is referred to as a *weighted α-fair allocation*. The resulting performance of this static fairness notion in a dynamic context is discussed in Section 1.4 for the particular case of bandwidth-sharing networks.

The class of weighted α-fair allocations contains some popular allocation paradigms when $w_i = 1$ for all i. For example, as $\alpha \to 0$ the resulting allocation achieves maximum throughput. Under suitable conditions, the Proportional Fair (PF) and max-min fair allocations (as defined in [24]) arise as special cases when $\alpha = 1$ and $\alpha \to \infty$, respectively, [100]. These notions of fairness have been widely used in the context of various networking areas, see for example [90, 100, 118, 136] for max-min fairness and [71, 100, 111] for PF.

The max-min fair allocation ($\alpha \to \infty$) is commonly seen as the most fair, since it maximizes the minimum rate allocated to any user. On the other extreme, maximizing the throughput ($\alpha \to 0$) can be highly unfair to certain users. The parameter α is therefore often referred to as the fairness parameter measuring the degree of fairness. Typically, realizing fairness and achieving a high throughput are conflicting objectives.

1.1.2 Dynamic setting

In practice, users depart upon service completion and new users arrive into the system over time. As mentioned previously, this can by modeled by queueing-theoretic models. In this section we discuss performance and fairness measures to evaluate different scheduling policies.

A key performance requirement in a dynamic setting is stability. Loosely speaking, stability means that the number of users in the system does not grow unboundedly or, in other words, that the system is able to handle all work requested by users. In this thesis we particularly focus on extensions of the single-server system where the total used capacity may depend on the scheduling decisions taken. Hence, stability conditions strongly depend on the policy employed. We therefore distinguish two types of conditions: (i) stability conditions corresponding to a *particular policy* and (ii) maximum stability conditions. The latter are conditions on the parameters of the *model* under which there exists a policy that makes the system stable.

Besides stability, another important performance measure concerns the number of users present in the system. We note that minimizing the total mean number of users is equivalent to minimizing the mean delay, cf. Little's law. As we will point out in Section 1.3.3, size-based scheduling policies, e.g. the Shortest Remaining Processing Time (SRPT) policy, are popular mechanisms for improving the performance by favoring smaller service requests over larger ones. However, this does not immediately carry over to the models we consider in this thesis. There are two effects to be taken into account. In the short term, it is preferable to favor "small" users that are likely to leave the system soon. In the long term however, a policy that uses the maximum capacity of the system at every moment in time, can empty the work in the system faster. When the total capacity used depends on the way the resources are shared among the classes, the above-described goals can be conflicting.

The objective of optimal scheduling is often contradictory with fair scheduling. For example, giving preference to users based on their size (as is the case with SRPT) may starve users with large service requirements. Similar to the static setting, there is no universally accepted definition of fairness in the dynamic setting. We refer to [11, 155, 156] for an overview on definitions existing in the literature.

In general, it is a difficult task to find fair or efficient policies for the dynamic setting. One may think of a policy as a rule that prescribes a rate allocation for each given population (as the population dynamically changes, the allocation changes as well). It is important to note that the use of fair or efficient allocations from the static setting does not give any guarantee for the behavior of the system in the dynamic setting. For example, maximizing the throughput at every moment in time, might unnecessarily render the system unstable, and hence be certainly suboptimal in the dynamic context (see for example [30, Example 1] and Proposition 3.2.1).

1.2 Motivating examples

In this section we describe several examples of communication networks that motivate the queueing models studied in the thesis. The queueing models are discussed in more detail in Sections 1.3–1.5.

1.2.1 Wired communication networks

The Internet is a packet-switched network, carrying data from source to destination. Each data transfer (flow) is split into several chunks (packets) that are routed individually over a common path from source to destination. Along this path, packets traverse various switches and routers that are connected by links. As a result, data flows contend for bandwidth on these links for the duration of the transfer.

Data flows can be broadly categorized into streaming and elastic traffic. Streaming traffic, corresponding to real-time connections such as audio and video applications, is extremely sensitive to packet delays. It has an intrinsic rate requirement that needs to be met as it traverses the network in order to guarantee satisfactory quality. On the other hand, elastic traffic, corresponding to the transfer of digital documents like Web pages, e-mails, and data files, does not have a stringent rate requirement. Most of the elastic data traffic in the Internet nowadays is regulated by the Transmission Control Protocol (TCP) [65]. This end-to-end control dynamically adapts the transmission rate of packets based on the level of congestion in the network. It ensures a high transmission rate to a user when the load on its path is low, and implies a low rate when links on its path are congested.

Link in isolation

Typically, a given link is transmitting packets generated by several data flows. For example, in Figure 1.1 (left) the white and black packets each correspond to their own data flow. When viewing the system on a somewhat larger time scale (flow level), it can be argued that each data flow is transmitted as a continuous stream through the link, using only a certain fraction of the bandwidth, as depicted in Figure 1.1 (right). In case of homogeneous data flows and routers this implies that the bandwidth is equally shared among the data flows, i.e., the throughput of each data flow is C/n bits per second when there are n flows present on a link in isolation with bandwidth C.

Since the dynamics at the packet level occur at a much faster time scale than the arrivals and departures of data flows, it is reasonable to assume that the bandwidth allocation is adapted instantly after a change in the number of flows. Under this *time-scale separation*, the dynamic bandwidth sharing coincides with the so-called Processor Sharing (PS) queue, where each flow receives a fraction $1/n$ of the total service rate whenever there are n active flows. Hence, PS is a useful paradigm for

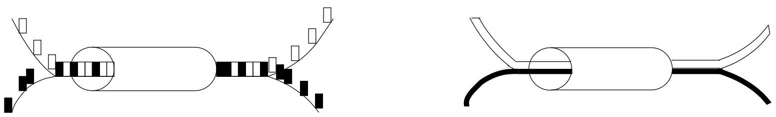


Figure 1.1: Two data flows in a link at packet level (left), and flow level (right).

evaluating the dynamic behavior of elastic data flows competing for bandwidth on a single link [22, 104]. The actual bandwidth shares may in fact significantly differ among competing flows, either due to the heterogeneous end-to-end behavior of data flows or due to differentiation among data flows in routers. An appropriate model for this setting is provided by the Discriminatory Processor Sharing (DPS) queue, where all flows share the bandwidth proportional to certain flow-dependent weights.

Multiple links

Instead of one link in isolation, a more realistic scenario is to consider *several* congested links in the network. Even though individual packets travel across the network on a hop-by-hop basis, when we view the system behavior on a somewhat larger time scale, a data flow claims roughly equal bandwidth on each of the links along its source-destination path *simultaneously*. A mathematical justification for the latter can be found in [153]. The class of weighted α-fair allocations, as described in Section 1.1.1, is commonly accepted to model the flow-level bandwidth allocation as realized by packet-based protocols. For example, the α-fair allocation with $\alpha = 2$ and weights w_k inversely proportional to the source-destination distance, has been proposed as an appropriate model for TCP [108]. In addition, for any α-fair allocation (defined at flow level) there exists a distributed mechanism at packet level that achieves the α-fair allocation [71, 100, 130].

Under the time-scale separation assumption, bandwidth-sharing networks as considered in [94] provide a natural way to describe the *dynamic* flow-level interaction among elastic data flows. See also [70, 153], where bandwidth-sharing networks are obtained as limits of packet-switched networks. In bandwidth-sharing networks, a flow requires simultaneously the same amount of capacity from all links along its source-destination path.

An example of a bandwidth-sharing network is depicted in Figure 1.2. Flows of class 0 request the same amount of bandwidth from all links simultaneously and in each link there is possibly cross traffic present from other routes. This interaction between active flows can cause inefficient use of the available capacity. For example, when there are flows of class 0 present, the capacity of a certain link with no cross traffic may not be fully used when the capacity of another link is already exhausted.

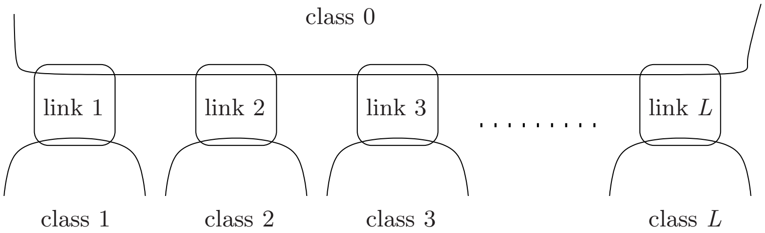


Figure 1.2: Linear bandwidth-sharing network with $L + 1$ classes of data flows.

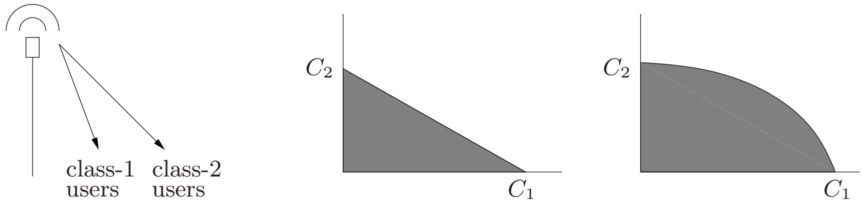


Figure 1.3: A single base station with two classes (left), and the rate region in case of TDMA (middle) and CDMA (right).

1.2.2 Wireless communication networks

In this section we focus on elastic data transfers in a wireless cellular network. Such a network consists of several cells each with their own base station. We concentrate on data transmissions from the base station to the wireless users (laptops, mobiles) in the corresponding cell. The transmission rate at which a user receives data is determined by the control mechanism of the base station. In addition, it is influenced by physical phenomena like signal fading or signal interference with other base stations.

Base station in isolation

We first consider a base station in isolation. There are two basic methods to divide the power of the base station among the users. One method is Time Division Multiple Access (TDMA) in which the base station transmits in each time slot to exactly one user. Another method is Code Division Multiple Access (CDMA) in which the base station transmits simultaneously to several users and the various data streams are jointly coded. Due to power attenuation, users on the edge of the cell will have worse channel conditions compared to users close to the base station. In Figure 1.3 (left) we consider a simple example where a class-1 user (class-2 user) is close to (far from) the base station and its transmission rate equals C_1 (C_2), with $C_1 > C_2$, when being allocated the full power of the base station. The corresponding rate region is depicted in Figures 1.3 (middle) and (right) for TDMA and CDMA, respectively. The northeast boundaries of the capacity regions are obtained when the base station transmits at full power. Note however that the aggregate allocated rate varies depending on the power allocation.

Inter-cell interference

When several neighboring base stations transmit simultaneously, the respective signals may interfere, causing a reduction in the transmission rates. In Figure 1.4 (left) we consider a simple example of two base stations and two classes of users each associated with their own base station. We assume that a base station is either off or is transmitting at full power. When only base station i is on, its transmission rate equals C_i, $i = 1, 2$. However, when both base stations are on, the transmission

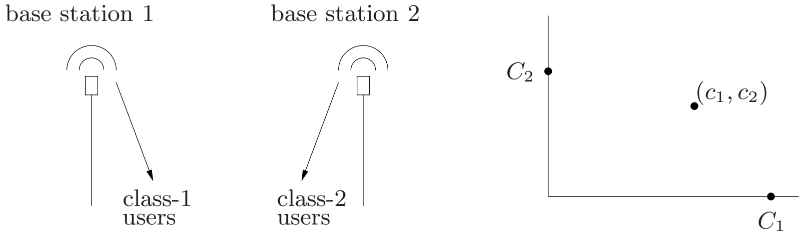


Figure 1.4: Two base stations each with their own class (left), and the rate region (right).

rate of base station i is c_i, $c_i < C_i$, $i = 1, 2$. The corresponding rate region is depicted in Figure 1.4 (right) and we note that the aggregate transmission rate is either C_1, C_2, or $c_1 + c_2$ depending on the activity of the base stations. At present, a base station typically transmits at full power as long as there are users present in its cell. The corresponding flow-level performance is studied in [28] for example. Recently, however, coordination between base stations has been proposed [29, 152], motivating the study of efficient coordinated power control of base stations.

1.3 The single-server system

The classical single-server system consists of a single queue and a single server with fixed capacity. Without loss of generality, the capacity is set equal to one. Users arrive one by one in the system and each user requires a certain amount of service. Let λ denote the arrival rate to the system, so that λ^{-1} is the mean inter-arrival time. The service requirement of a user represents the amount of time that the server needs to serve the user when it would devote its full capacity to this user. This random variable is denoted by B. The capacity of the server may be shared among multiple users at the same time. When a user is not served, it waits in the queue. Preemption of a user in service is allowed. In the case of preemption, a user goes back to the queue awaiting to receive its remaining service requirement. After a user has received its full service, it leaves the system.

A common assumption is that the inter-arrival times are independent and identically distributed (i.i.d.), the service requirements are i.i.d., and the sequences of inter-arrival times and service requirements are independent. This model is referred to as the G/G/1 queue, a notation that was introduced by Kendall [73]. Here the G stands for general. When in addition the inter-arrival times are exponentially distributed, i.e., a Poisson arrival process, the corresponding system is denoted by the M/G/1 queue where the M stands for Markovian or memoryless. When instead the service requirements are exponentially distributed, the queue is referred to as the G/M/1 queue.

In a single-server queue the focus is on work-conserving scheduling policies, that is, policies that always use the full capacity of the server whenever there is work in the system. Obviously, the total unfinished work in the system, the workload, is independent of the work-conserving policy employed. In addition, any work-conserving policy in a G/G/1 queue is stable as long as the traffic load $\rho := \lambda\mathbb{E}(B)$ is strictly less than one [86].

While the workload process and the stability condition are independent of the employed work-conserving policy, this is not the case for the evolution of the queue length process and, hence, for most performance measures. There is a vast body of literature on the analysis of scheduling policies in the single-server queue. In the remainder of this section we mention the results relevant for the thesis. We first give a description of two time-sharing policies: PS and DPS. As explained in Section 1.2.1, these policies provide a natural approach for modeling the flow-level performance of TCP. We conclude this section with an overview of optimal size-based scheduling in the single-server queue.

1.3.1 Processor sharing

Under the Processor Sharing (PS) policy, the capacity is shared equally among all users present in the system. When there are n users in the system, each user receives a fraction $1/n$ of the capacity of the server. Below we present several known results from the literature. For full details and references on the PS queue we refer to [104].

When the arrival process is Poisson and $\rho < 1$, the stationary distribution of the queue length exists and is insensitive to the service requirement distribution apart from its mean. More precisely, the queue length in steady state has a geometric distribution with parameter ρ, i.e., the probability of having n users in the queue is equal to $(1-\rho)\rho^n$, $n = 0, 1, \ldots$, cf. [119]. In particular, this implies that the mean number of users in the system is finite whenever $\rho < 1$. Another appealing property of PS is that a user's slowdown (defined as the user's mean sojourn time divided by its service requirement) equals $1/(1-\rho)$, independent of its service requirement.

For a PS queue with several classes of users, the geometric distribution carries over as well. Consider K classes of users, where class-k users arrive according to a Poisson process with arrival rate λ_k and have service requirements B_k, $k = 1, \ldots, K$. Assuming Poisson arrivals, the probability of having n_k class-k users in the system, $k = 1, \ldots, K$, is equal to

$$(1-\rho) \cdot \frac{(n_1 + \ldots + n_K)!}{n_1! \cdot n_2! \cdot \ldots \cdot n_K!} \cdot \prod_{k=1}^{K} \rho_k^{n_k}, \tag{1.3}$$

with $\rho_k := \lambda_k \mathbb{E}(B_k)$ and $\rho := \sum_{k=1}^{K} \rho_k$, [41, 69]. Another interesting result concerns the remaining service requirements of the users. Given a population of users, the remaining service requirements are i.i.d. and distributed according to the forward recurrence times of their service requirements [41, 69].

1.3.2 Discriminatory processor sharing

The Discriminatory Processor Sharing (DPS) policy, introduced in [77] by Kleinrock, is a multi-class generalization of PS. By assigning different weights to users from different classes, DPS allows class-based differentiation. Let K be the number of classes, and let w_k be the weight associated with class k, $k = 1, \dots, K$. Whenever there are n_k class-k users present, $k = 1, \dots, K$, a class-l user is served at rate

$$\frac{w_l}{\sum_{k=1}^{K} w_k n_k}, \quad l = 1, \dots, K.$$

In case of unit weights, the DPS policy reduces to the PS policy. Despite the similarity, the analysis of DPS is considerably more complicated compared to PS. The geometric queue length distribution for PS does not have any counterpart for DPS. In fact, the queue lengths under DPS are sensitive with respect to higher moments of the service requirements [32]. Despite this fact, in [12] the DPS model was shown to have finite mean queue lengths regardless of the higher-order moments of the service requirements.

The seminal paper [51] provided an analysis of the mean sojourn time conditioned on the service requirement by solving a system of integro-differential equations. As a by-product, it was shown that a user's slowdown behaves like the user's slowdown under PS, as its service requirement grows large, see also [12]. Another asymptotic regime under which the DPS policy has been studied is the so-called heavy-traffic regime, which means that the traffic load approaches the critical value ($\rho \uparrow 1$). For Poisson arrivals and exponentially distributed service requirements, in [113] the authors showed that the scaled joint queue length vector has a proper limiting distribution. Let N_k denote the number of class-k users in steady state, then

$$(1-\rho)(N_1, N_2, \dots, N_K) \xrightarrow{d} X \cdot (\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \dots, \frac{\hat{\rho}_K}{w_K}), \quad \text{as} \quad \rho \uparrow 1,$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\hat{\rho}_k := \lim_{\rho \uparrow 1} \rho_k$, $k = 1, \dots, K$, and X is an exponentially distributed random variable. In Chapter 2 we extend this result for phase-type distributed service requirements. For more results on DPS under several other limiting regimes we refer to the overview paper [5] and to Chapter 2.

For the sake of completeness, we briefly mention a related scheduling policy, Generalized Processor Sharing (GPS) [45, 109]. Under GPS, the capacity is allocated across the non-empty *classes* in proportion to the weights, i.e., class l receives

$$\frac{w_l \mathbf{1}_{(n_l > 0)}}{\sum_{k=1}^{K} w_k \mathbf{1}_{(n_k > 0)}}, \quad l = 1, \dots, K,$$

whenever there are n_k class-k users present, $k = 1, \dots, K$. As opposed to DPS, under GPS each non-empty class is guaranteed a minimum share of the capacity regardless of the number of users present within this class.

1.3.3 Optimal scheduling

There exists a vast amount of literature devoted to optimal scheduling in single-server systems. A well-known optimality result concerns the Shortest Remaining Processing Time (SRPT) policy, which serves at any moment in time the user with the shortest remaining service requirement [120]. In [121, 127] it is proved that SRPT minimizes sample-path wise the number of users present in the single-server system. (Stochastic minimization and other optimality notions used in this section will be introduced in detail in Section 1.6.)

SRPT relies on the knowledge of the (remaining) service requirements of the users. Since this information might be impractical to obtain, a different strand of research has focused on finding optimal policies among the so-called non-anticipating policies. These policies do not use any information based on the (remaining) service requirements, but they do keep track of the attained service of users present in the system. Popular policies like First Come First Served (FCFS), Least Attained Service (LAS), PS and DPS are all non-anticipating. Among all non-anticipating policies, the mean number of users is minimized under the Gittins rule [3, 57]. The latter simplifies to LAS and FCFS for particular cases of the service requirements [3].

The LAS policy [78, Section 4.6], also known as Foreground-Background, which serves at any moment in time the user(s) with the least attained service, has been extensively studied. For an overview we refer to [105]. In case of Poisson arrivals, LAS stochastically minimizes the number of users in the system if and only if the service requirement distribution has a decreasing failure rate (DFR) [3, 114]. This result is based on the fact that under the DFR assumption, as a user obtains more service, it becomes less likely that it will leave the system soon. Therefore, prioritizing the newest users is optimal.

For a service requirement distribution with an increasing failure rate (IFR), any non-preemptive policy, in particular FCFS, stochastically minimizes the number of users in the system [114]. A policy is non-preemptive when at most one user is served at a time and once a user is taken into service this service will not be interrupted. This result can be understood from the fact that under the IFR assumption, as a user obtains more service, it becomes more likely that it will leave the system soon.

We finish this section with an important result for the multi-class single-server system. We associate with each user class a cost c_k and let $\mu_k := 1/\mathbb{E}(B_k)$, where B_k denotes the class-k service requirement. A classical result states that the so-called $c\mu$-rule, the policy that gives strict priority to classes in descending order of $c_k\mu_k$, minimizes the mean holding cost $\mathbb{E}(\sum_k c_k N_k)$. This result holds for the M/G/1 queue among all non-preemptive non-anticipating policies [56] and for the G/M/1 queue among all non-anticipating policies [38, 102]. The optimality of the $c\mu$-rule can be understood from the fact that $1/\mu_k$ coincides in both settings with the expected remaining service requirement of a class-k user *at a scheduling decision epoch*. Hence, at every moment in time, the user with the smallest weighted expected remaining service requirement is served.

1.4 Bandwidth-sharing networks

Bandwidth-sharing networks provide a modeling framework for the dynamic interaction of data flows in communication networks, where a flow claims roughly equal bandwidth on each of the links along its path, as described in Section 1.2.1. Mathematically, a bandwidth-sharing network can be described as follows. It consists of a finite number of nodes, indexed by $l = 1, \dots, L$, which represent the links of the network. Node l has finite capacity C_l. There are K classes of users. Associated with each class is a route that describes which nodes are needed by the users from this class. Let A be the $L \times K$ incidence matrix containing only zeros and ones, such that $A_{lk} = 1$ if node l is used by users of class k and $A_{lk} = 0$ otherwise. Each user requires simultaneously the same capacity from all the nodes on its route. Let s_k denote the aggregate rate allocated to all class-k users. The total capacity used from node l is $\sum_{k=1}^{K} A_{lk} s_k$. Hence, a rate allocation is feasible when $\sum_{k=1}^{K} A_{lk} s_k \leq C_l$, for all $l = 1, \dots, L$.

An example of a bandwidth-sharing network is the so-called linear network as depicted in Figure 1.2. It consists of L nodes and $K = L+1$ classes, for convenience indexed by $j = 0, 1, \dots, L$. Class-0 users require the same amount of capacity from all L nodes simultaneously while class-i users, $i = 1, \dots, L$, require service at node i only. The $L \times (L+1)$ incidence matrix of the linear network is

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

hence the capacity constraints are $s_0 + s_i \leq C_i$, $i = 1, \dots, L$. The corresponding capacity region in the case of a two-node linear network with $C_1 = C_2 = C$ is depicted in Figure 1.5. As this figure indicates, the linear network can be viewed as an extension of the single-server system. More specifically, the system can be interpreted as a single server that handles all classes with the special feature that it can work on classes $1, \dots, L$ simultaneously at full speed.

As explained in Section 1.2.1, the linear network provides a flow-level model for Internet traffic that experiences congestion on each link along its path from other intersecting routes. A linear network also arises in simple models for the mutual interference in wireless networks. Consider the following setting. Users can be either in cell 0, 1 or 2. Users in cells 1 and 2 can be served in parallel by base stations 1 and 2, respectively. Because of interference, a user in cell 0 can only be served when

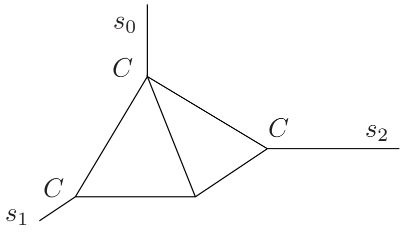


Figure 1.5: Capacity region of a two-node linear network in case $C_1 = C_2 = C$.

either base station 1 or 2 is on and transmits the requested file to the user in cell 0. Hence, class 0 can only be served when both classes 1 and 2 are not served, which can be modeled as a linear network consisting of two nodes. As a further motivating example we could think of write permission in a shared database. Consider L servers that each perform tasks involving read/write operations in some shared database. Read operations can occur in parallel. However, if a server needs to perform a task involving write operations, then the database needs to be locked, and no tasks whatsoever can be performed by any of the other servers. This may be modeled as a linear network with L nodes, where class-0 tasks correspond to the write operations.

An inherent property of bandwidth-sharing networks is that, given a population of users, the total used capacity of the network, $\sum_{k=1}^K \sum_{l=1}^L A_{lk} s_k$, is not necessarily equal to the total available capacity of the network, $\sum_{l=1}^L C_l$. This may even be the case when we restrict ourselves to Pareto-efficient allocations, i.e., allocations where the rate allocated to a class cannot be increased without reducing the rate allocated to another class. For example, one may think of the linear network where at a certain moment in time there are no users of class L present. The Pareto-efficient allocation that serves class 0 makes *full* use of the capacity of the network. However, the Pareto-efficient allocation that serves classes 1 until $L-1$ uses only the capacity of the first $L-1$ nodes, and leaves the capacity of node L unused.

The maximum stability conditions of a bandwidth-sharing network are $\sum_{k=1}^K A_{lk}\rho_k < C_l$, for all $l = 1, \ldots, L$, see [59], i.e., the offered load in each node is strictly less than its available capacity. In general, the stability conditions corresponding to a specific policy can be more restrictive than the maximum stability conditions. This becomes for example apparent in the linear network with unit capacities, $C_l = 1$, $l = 1, \ldots, L$. The policy that gives preemptive priority to class-0 users is stable under the maximum stability conditions, $\rho_0 + \rho_i < 1$, for all $i = 1, \ldots, L$. However, the Pareto-efficient policy that gives preemptive priority to classes 1 through L is stable if and only if $\rho_0 < \prod_{i=1}^L (1 - \rho_i)$, which is a more stringent condition. These stability results will be elaborated on in Section 3.2. Note that in [59] it is shown that this instability effect can be avoided. It is proved that any Pareto-efficient policy in a bandwidth-sharing network is stable, provided that it is suitably modified when the number of users in a class becomes too small.

1.4.1 Weighted α-fair sharing

A popular class of policies studied in the context of bandwidth-sharing networks are weighted α-fair bandwidth-sharing policies. In state $\vec{n} = (n_1, \ldots, n_K)$ a weighted α-fair policy allocates $s_k(\vec{n})/n_k$ to each class-k user, with $(s_1(\vec{n}), \ldots, s_K(\vec{n}))$ the solution of the utility optimization problem

$$\begin{aligned} &\text{maximize} \quad \sum_{k=1}^K n_k U_k^{(\alpha)}\left(\frac{s_k}{n_k}\right), \\ &\text{subject to} \quad \sum_{k=1}^K A_{lk} s_k \leq C_l, \quad l = 1, \ldots, L, \end{aligned} \tag{1.4}$$

and $U_k^{(\alpha)}(\cdot)$, $\alpha > 0$, as defined in (1.2). Note that the total rate allocated to class k, s_k, is equally shared among all class-k users, in other words, the intra-class policy is PS.

For a network consisting of one node, the weighted α-fair policy reduces to the DPS policy with weights $w_k^{1/\alpha}$, $k = 1, \ldots, K$. For the linear network with unit capacities, the weighted α-fair rate allocation is given by

$$s_0(\vec{n}) = \frac{(w_0 n_0^\alpha)^{1/\alpha}}{(w_0 n_0^\alpha)^{1/\alpha} + (\sum_{i=1}^K w_i n_i^\alpha)^{1/\alpha}}, \quad s_i(\vec{n}) = \mathbf{1}_{(n_i>0)} \cdot (1 - s_0(\vec{n})), \quad i = 1, \ldots, L,$$

see [30]. For grid and cyclic networks, as described in [30], the weighted α-fair rate allocations can be found in closed form as well.

An important property of weighted α-fair policies in bandwidth-sharing networks concerns stability. In [30] it is proved that when the service requirements and the inter-arrival times are exponentially distributed, weighted α-fair bandwidth-sharing policies ($\alpha > 0$) achieve stability under the maximum stability conditions, $\sum_{k=1}^K A_{lk}\rho_k < C_l$, for all $l = 1, \ldots, L$, see also [139, 159]. For phase-type distributed service requirements, maximum stability is proved for the Proportional Fair (PF) policy ($\alpha = 1$ and unit weights) [93]. In [31, 34, 82] stability is investigated when the set of feasible allocations is not given by (1.4). The authors of [31] prove that for any convex set of feasible allocations, PF and the max-min fair policy ($\alpha \to \infty$ and unit weights) provide stability under the maximum stability conditions. In [34, 82] stability is investigated when the set of feasible allocations is non-convex or time-varying. It is shown that the stability condition depends on the parameter α, and that for some special cases the stability condition becomes tighter as α increases.

1.4.2 Flow-level performance

Very little is known about the way α-fair sharing affects the performance perceived by users. Closed-form analysis of weighted α-fair policies has mostly remained elusive, except for so-called hypercube networks (a special case is the linear network) with unit capacities. For those networks, the steady-state distribution of the numbers of users of the various classes under PF is of product form and insensitive to the service requirement distributions [30, 32]. For all other situations, the distributions of the numbers of users under weighted α-fair policies are sensitive with respect to higher moments of the service requirement distributions [32]. In [33], insensitive stochastic bounds on the number of users in any class are derived for the special case of tree networks. A related result can be found in [134] where the authors focus on exponentially distributed service requirements and obtain an upper bound on the total mean number of users under PF.

A powerful approach to study the complex dynamics under weighted α-fair policies is to investigate asymptotic regimes. For example, in [49] the authors study the max-min fair policy under a large-network scaling and give a mean-field approximation. Another asymptotic regime is the heavy-traffic setting where the load on at least one node is close to its capacity. In this regime, the authors of [68, 72, 160] study weighted α-fair policies under fluid and diffusion scalings and investigate diffusion approximations for the numbers of users of the various classes. In addition, when the load on *exactly* one node tends to its capacity, the authors of [160] identify a cost function that is minimized in the diffusion scaling by the weighted α-fair policy. For the linear network, heavy-traffic approximations for the scaled mean numbers of users are derived in [81]. Bandwidth-sharing networks in an overloaded regime, that is when the load on one or several of the nodes exceeds the capacity, are considered in [46]. The growth rates of the numbers of users of the various classes under weighted α-fair policies are characterized by a fixed-point equation.

Motivated by the optimality results in the single-server system, research has focused on improving weighted α-fair policies using performance benefits from size-based scheduling. In [1] the authors propose to deploy SRPT as intra-class policy, instead of PS, in order to reduce the number of users in each class. Another approach is taken in [157, 158], where weighted α-fair policies are studied with dynamic per-user weights that depend on the remaining service requirements. Simulations show that the performance can improve considerably over the standard α-fair policies.

1.5 The parallel-server model

The parallel-server model consists of L multi-skilled servers that can work in parallel on K classes of users. A class might be served more efficiently on one server than on another. We denote by $\mu_{kl} := 1/\mathbb{E}(B_{kl})$ the mean service rate of a class-k user at server l, where B_{kl} denotes the service requirement of a class-k user when server l works at full speed on this user. Figure 1.6 (left) shows a parallel-server model with two classes of users and two servers.

The parallel-server system may be viewed as a simple model for a parallel computer system where processors have overlapping capabilities and the capacity of the processors needs to be allocated among several tasks. Other applications are service facilities like call centers. An agent can be specialized in a certain type of calls, but can also handle other types at a relatively low speed. In the thesis we will specif-

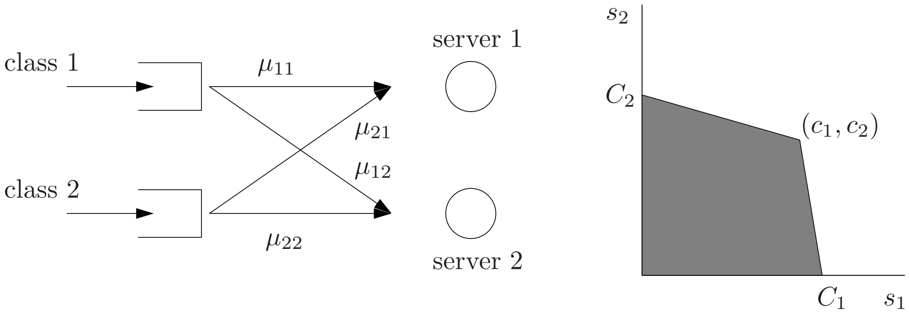


Figure 1.6: Parallel two-server model with two classes (left), and the capacity region when $c_1 + c_2 > \max(C_1, C_2)$ (right).

ically focus on a parallel two-server model with two classes of users, where both servers can work simultaneously on the same user. This model may represent the interference of two base stations in a cellular wireless network, as described in the next paragraph.

Consider a parallel two-server model with two classes where both servers can work simultaneously on the same user. We define c_1, c_2, μ_1 and μ_2 such that they satisfy $\mu_{11} = c_1\mu_1$, $\mu_{12} = (C_1 - c_1)\mu_1$, $\mu_{21} = (C_2 - c_2)\mu_2$ and $\mu_{22} = c_2\mu_2$, with $C_1, C_2 > 0$. In case of exponentially distributed service requirements, we can now give an equivalent representation of the parallel two-server model with two classes. In this equivalent model description, class-k users have a mean service requirement of $1/\mu_k$, $k = 1, 2$. When each class is served by its own server, class k receives capacity c_k (since then its departure rate is $\mu_{kk} = c_k\mu_k$). However, when both servers work together on class k, this class receives capacity C_k (since then its departure rate is $\mu_{k1}+\mu_{k2} = c_k\mu_k+(C_k-c_k)\mu_k$). The corresponding capacity region is depicted in Figure 1.6 (right) in case $c_1 + c_2 > \max(C_1, C_2)$, where s_k denotes the capacity allocated to class k. The application to interference in wireless networks becomes now apparent: the capacity region coincides with that in Figure 1.4 (right) and is a simplification for the region of Figure 1.3 (right). Interestingly, the shape of the capacity region, when setting $C_1 = C_2 = 1$ (without loss of generality), indicates that the parallel two-server model with two classes may be viewed as an extension of the single-server system. There is one main server with capacity one that handles both classes of users. This server has the special feature that when the server works on both classes in parallel, its capacity becomes $c_1 + c_2$.

The above-described parallel two-server model with two classes has been well studied under the simple priority rule that server k gives preemptive priority to class k, $k = 1, 2$, and helps the other server when there is no queue of class k. Under this policy, the model is also referred to as the coupled-processors model for which the joint queue length distribution has been analyzed in [50] for exponential service requirements. In [42] the joint workload distribution is characterized in the case of general service requirements. Both results in [42, 50] require the solution of a Riemann-Hilbert boundary value problem. A diffusion approximation for the queue lengths has been obtained in [25, 26] for a heavily-loaded system with general service requirements.

The maximum stability conditions of a parallel-server model can be explicitly described: There exists a policy that makes the parallel-server model stable if and only if there exist $x_{kl} \geq 0$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, such that $\sum_k x_{kl} \leq 1$, and $\lambda_k < \sum_l x_{kl}\mu_{kl}$, with λ_k the arrival rate of class k [132, 135]. Due to the specialized servers, Pareto-efficient policies in parallel-server models are not necessarily stable under the maximum stability conditions. In [137] policy-dependent stability conditions are characterized for the parallel two-server model with $\mu_{21} = 0$.

Obtaining closed-form expressions for performance measures and finding efficient scheduling policies in parallel-server models is a notoriously difficult task. For results obtained in this area we refer to the overview in [128]. In the remainder of this section we describe those relevant for the thesis.

1.5.1 Threshold-based policies

Popular policies studied in the context of parallel-server models rely on thresholds. Decisions are taken based on whether or not queue lengths exceed class-dependent thresholds. For example, in the case of the parallel two-server model with two classes a threshold-based policy could be that both servers serve their own class. However, when the number of class-1 users exceeds a threshold, server 2 helps server 1 to reduce the work in class 1. In the case of phase-type distributed service requirements, the exact stability conditions have been obtained for this policy [107, 137]. In particular, it is shown that the threshold should be sufficiently large in order for the system to be stable.

A general class of threshold-based policies for parallel-server models is proposed in [129]. An important observation made there is that finding reasonable values for the thresholds is not trivial since performance can be quite sensitive to the threshold values. The authors of [129] derive approximate formulas for the queue lengths based on vacation models and illustrate how these can be used to obtain suitable threshold values. In [107] the authors consider the parallel two-server model with two classes of users and propose another class of threshold-based policies. Besides determining the stability conditions, they evaluate the robustness against misestimation of load. Approximations for mean response times are given in [106], also incorporating switching times when a server switches between queues. Threshold-based policies that achieve optimality in a heavy-traffic setting are described in [19, 20].

1.5.2 Max-Weight policies

Max-Weight policies were first introduced in [135] and have been extensively studied ever since, see for example [89, 125, 132]. The generalized $c\mu$-rule [99], including the Max-Weight policy as a special case, is analyzed in [89] for a parallel-server model. This rule myopically maximizes the rate of decrease of certain instantaneous holding cost. More precisely, when server l is free, it starts serving a user from class k' such that $k' = \arg\max_k \mu_{kl} \frac{\mathrm{d}C_k(n_k)}{\mathrm{d}n_k}$, whenever there are n_k class-k users present, $k = 1, \dots, K$, and serves this user until it leaves the system. The function $C_k(n_k)$ can be interpreted as the cost of having n_k class-k users present in the system. The class of Max-Weight policies corresponds to functions of the type $C_k(n_k) = \gamma_k n_k^{\beta+1}$, with $\beta, \gamma_k > 0$. In that case, the policy can be described by cones in $\mathbb{R}_+^K$ such that the decision taken by the Max-Weight policy is based on which cone the queue length vector currently belongs to. Related projective-cone schedulers have been studied in [8, 116] where the decision is based on which cone the *workload vector* currently belongs to.

Under fairly mild conditions, Max-Weight policies achieve maximum stability for a large class of queueing networks [125, 132, 135]. However, the framework does not allow for linear holding cost, i.e., $\beta = 0$. In fact, a myopic policy based on a linear cost function can render the system unnecessarily unstable. Besides stability, another important characteristic is that these policies are *robust* in the sense that they do not rely on any information of the inter-arrival processes.

Heavy-traffic results for the parallel-server model have been obtained in [89, 132], where it is in particular shown that the generalized $c\mu$-rule minimizes the holding cost, $\sum_{k=1}^{K} C_k(N_k(t))$, sample-path wise in the diffusion limit. Here $N_k(t)$ denotes the number of class-k users at time t under the generalized $c\mu$-rule. More details on the heavy-traffic results can be found in Section 8.6.2.

1.5.3 Optimal scheduling in heavy traffic

Determining the optimal policy in a parallel-server model has so far proved analytically infeasible. Most research in this area has focused on heavily-loaded systems under a (complete) resource pooling condition. The latter means that as the system approaches its capacity, the individual servers can be effectively combined to act as a single pooled resource. As mentioned in Section 1.5.2, the generalized $c\mu$-rule minimizes the scaled cost sample-path wise in heavy traffic. A complementary result is obtained in [19, 20], where the authors prove that certain threshold-based policies minimize the scaled average discounted number of users in a heavy-traffic setting, see Section 8.6.1 for more details. In [10, 61, 62] several discrete-review policies are proposed (the system is reviewed at discrete points in time, and decisions are based on the queue lengths at the review moment) for which heavy-traffic optimality results hold as well. It is important to note that Max-Weight policies are robust, while efficient threshold-based and discrete-review policies may depend on the inter-arrival characteristics.

1.6 Methodology

When seeking efficient policies, our goal is to minimize the number of users present in the system, or more generally, the so-called holding cost. Because of Little's law, minimizing the total mean number of users is equivalent to minimizing the mean sojourn time, and thus equivalent to maximizing the user's throughput defined as the ratio between the mean service requirement and the mean sojourn time.

We first discuss several notions of optimality. The strongest notion we consider relates to stochastic ordering. Two random variables X and Y are stochastically ordered, $X \leq_{st} Y$, when $\mathbb{P}(X > s) \leq \mathbb{P}(Y > s)$ for all $s \in \mathbb{R}$. Equivalently, $X \leq_{st} Y$ if and only if there exist two random variables X' and Y' defined on a common probability space, such that $X \overset{d}{=} X'$, $Y \overset{d}{=} Y'$, and $X' \leq Y'$ with probability one [101, 117]. We call a policy $\tilde{\pi}$ stochastically optimal when it stochastically minimizes the holding cost at any moment in time, i.e.,

$$\sum_{k=1}^{K} c_k N_k^{\tilde{\pi}}(t) \leq_{st} \sum_{k=1}^{K} c_k N_k^{\pi}(t), \quad \text{for all } t \geq 0, \quad \text{and for all } \pi \in \Pi,$$

where c_k is a positive cost associated with class k, Π is a predetermined set of policies to which the search is restricted, and $N_k^{\pi}(t)$ denotes the number of class-k users at time t under policy π, $k = 1, \ldots, K$. A weaker notion of optimality is obtained when taking the expectation on both sides, i.e., a policy is called optimal when it minimizes the mean holding cost, $\mathbb{E}(\sum_{k=1}^{K} c_k N_k(t))$, at any moment in time. When optimal policies in the transient regime do not exist, we further weaken the notion of optimality. We then focus on policies that stochastically minimize the long-run holding cost, $\lim_{m\to\infty} \frac{1}{m}\int_0^m \sum_{k=1}^K c_k N_k(t)\mathrm{d}t$, or that minimize the average long-run holding cost,

$$\lim_{m\to\infty} \frac{1}{m}\mathbb{E}\Big(\int_0^m \sum_{k=1}^K c_k N_k(t)\mathrm{d}t\Big).$$

The latter notion is referred to as average-cost optimal. Unfortunately, it is not always within reach to explicitly determine optimal policies. In such cases, we resort to asymptotic regimes such as a fluid scaling and a heavy-traffic regime. Optimality definitions in these regimes will be described in more detail in Sections 1.6.3 and 1.6.4.

In the remainder of this section we sketch the four main techniques used in the thesis: sample-path comparison, stochastic dynamic programming, fluid scaling, and the heavy-traffic regime. As such, this section serves as a reference framework throughout the thesis. In Chapters 4, 7 and 8 we apply a sample-path comparison technique to characterize policies that minimize the mean holding cost at any moment in time. Similar techniques are used in Chapter 3 to obtain stability conditions. Another technique used in Chapters 4 and 8 is dynamic programming in order to find either stochastically-optimal policies or to determine characterizations of average-cost optimal policies. Fluid-scaled processes and asymptotically fluid-optimal policies are investigated in Chapters 5 and 8. Chapters 2, 6 and 8 contain results for systems in a heavy-traffic regime.

1.6.1 Sample-path comparison

Sample-path comparison is a useful tool in the control of queueing networks. A sample path corresponds to one particular realization of the stochastic process. As the name suggests, sample-path comparison techniques aim to compare, sample path by sample path, stochastic processes defined on a common probability space.

When for each sample path the same ordering on two processes holds, these processes are ordered sample-path wise. This is closely related to stochastic ordering of processes. Processes $\{X(t)\}_t$ and $\{Y(t)\}_t$ are stochastically ordered, $\{X(t)\}_t \leq_{st} \{Y(t)\}_t$, if and only if $(X(t_1),\ldots,X(t_m)) \leq_{st} (Y(t_1),\ldots,Y(t_m))$ for any m and all $0 \leq t_1 < t_2 < \ldots < t_m < \infty$, [101]. Hence, if there exist two processes $\{X'(t)\}_t$ and $\{Y'(t)\}_t$ defined on a common probability space (i.e., these two processes are coupled) that are ordered sample-path wise and satisfy $\{X'(t)\}_t \overset{d}{=} \{X(t)\}_t$ and $\{Y'(t)\}_t \overset{d}{=} \{Y(t)\}_t$, then the processes $\{X(t)\}_t$ and $\{Y(t)\}_t$ are stochastically ordered.

In queueing networks, a rather intuitive way of coupling processes corresponding to different policies is to consider the same realizations of the arrival processes and service requirements. However, often more ingenious couplings are needed in order to obtain the desired comparison. We refer to [47, 84] for an overview on sample-path comparison methods and applications to queueing networks. In [92] (see also [85]) necessary and sufficient conditions on the transition rates are given in order for a stochastic order-preserving coupling to exist between two Markov processes.

The optimality of the $c\mu$-rule (denoted by $\pi^{c\mu}$) in the G/M/1 queue can be proved using sample-path arguments [84]. Here we describe the proof in the case of two classes, since it illustrates the basic steps taken in most of the sample-path proofs in the thesis. Assume $c_1\mu_1 \geq c_2\mu_2$ so that the $c\mu$-rule amounts to giving preemptive priority to class 1, see Section 1.3.3. When the system is initially empty and the same realizations of arrivals and service requirements are considered under all policies, the following inequalities hold sample-path wise:

$$W_1^{\pi^{c\mu}}(t) \leq W_1^{\pi}(t) \tag{1.5}$$

and

$$W_1^{\pi^{c\mu}}(t) + W_2^{\pi^{c\mu}}(t) \leq W_1^{\pi}(t) + W_2^{\pi}(t), \tag{1.6}$$

for all $t \geq 0$ and for all policies π, where $W_k^{\pi}(t)$ denotes the workload in class k under policy π at time t. Multiplying (1.5) by $c_1\mu_1 - c_2\mu_2 \geq 0$ and (1.6) by $c_2\mu_2$, and using that $\mathbb{E}(W_k^{\pi}(t)) = \mathbb{E}(N_k^{\pi}(t))/\mu_k$ for non-anticipating policies (results from the memoryless property of the exponentially distributed service requirements), it follows that $c_1\mathbb{E}(N_1^{\pi^{c\mu}}(t)) + c_2\mathbb{E}(N_2^{\pi^{c\mu}}(t)) \leq c_1\mathbb{E}(N_1^{\pi}(t)) + c_2\mathbb{E}(N_2^{\pi}(t))$, for all $t \geq 0$ and for all non-anticipating policies π.

1.6.2 Stochastic dynamic programming

Markov decision theory is a useful framework for modeling decision making in Markovian queueing systems. So-called stochastic dynamic programming techniques, based on Bellman's principle of optimality [21], allow to study a wide range of optimization problems. Although these techniques are well developed, only a few special queueing networks allow for an explicit solution of the optimal policy, see the survey on Markov decision problems (MDP's) in the control of queues [131]. Even when not explicitly solvable, characterizations of the optimal policies can often still be obtained. We refer to the textbooks [110, 117] for a full overview on MDP's.

In the simplest setting, an MDP is described as follows. At equidistant points in time, $t = 0, 1, \ldots$, a decision maker observes the state of the system, denoted by x, and chooses an action a from the action space $A(x)$. The state at the next decision epoch, denoted by y, is described by the transition probabilities $p(x, a, y)$ depending on the current state and the action chosen. There is a direct cost $C(x)$ each time state x is visited. The corresponding Markov decision chain can be described by $\{X_t, A_t\}_t$, where X_t and A_t represent the state and action at time t, respectively.

Markov decision theory allows optimization under finite-horizon, infinite-horizon discounted, and average-cost criteria. Here we focus on the latter, that is, we search for a policy that minimizes

$$\limsup_{m\to\infty} \frac{1}{m}\mathbb{E}(\sum_{t=0}^{m-1} C(X_t)).$$

An average-cost optimal policy does not necessarily need to exist when the state space is infinite. There exist, however, sufficient conditions under which existence is guaranteed, see for example [123]. In that case, if $(g, V(\cdot))$ is a solution of the average-cost optimality equations

$$g + V(x) = C(x) + \min_{a \in A(x)} \sum_y p(x,a,y)V(y), \text{ for all states } \ x, \tag{1.7}$$

then g equals the minimum average cost and a stationary policy that realizes the minimum in (1.7) is average-cost optimal [117, Chapter V.2]. The function $V(\cdot)$ is referred to as the value function.

There are two main dynamic programming techniques: the policy iteration algorithm and the value iteration algorithm. The latter is used throughout the thesis. Value iteration consists in analyzing the functions $V_m(\cdot)$, $m = 0, 1, \ldots$, defined as

$$\begin{aligned} V_0(x) &= 0 \\ V_{m+1}(x) &= C(x) + \min_{a \in A(x)} \{\sum_y p(x,a,y)V_m(y)\}, \ m = 0, 1, \ldots. \end{aligned} \tag{1.8}$$

The functions $V_{m+1}(x)$ are interesting by themselves. They represent the minimum achievable expected cost over a horizon $m+1$ when starting in state x, i.e., the term $\mathbb{E}(\sum_{t=0}^{m} C(X_t)|X_0 = x)$ is minimized. Under certain conditions it holds that $V_m(\cdot) - mg \to V(\cdot)$ and $V_{m+1}(\cdot) - V_m(\cdot) \to g$ as $m \to \infty$ [64]. In addition, the minimizing actions in (1.8) converge to actions that constitute an average-cost optimal policy [64, 124]. As a consequence, if properties such as monotonicity, convexity, and submodularity [79] are satisfied for $V_m(\cdot)$, for all $m = 0, 1, \ldots$, then the same is true for the value function $V(\cdot)$. Together with (1.7) this helps in the characterization of an optimal policy.

For a finite state space, the value iteration algorithm is useful to numerically determine an approximation of the average-cost optimal policy. This consists in recursively computing the functions $V_{m+1}(\cdot)$ until the difference between $\max_x(V_{m+1}(x) - V_m(x))$ and $\min_x(V_{m+1}(x) - V_m(x))$ is sufficiently small. Since the state spaces considered in the thesis are infinite, in all our numerical experiments we apply the value iteration algorithm after appropriate truncation of the state space.

In a Markovian queueing system, without loss of generality, one can focus on policies that make decisions at transition epochs. The times between consecutive decision epochs are state-dependent and exponentially distributed. We can however equivalently consider the uniformized Markov process [110]: After uniformization, the transition epochs (including "dummy" transitions that do not alter the system state) are generated by a Poisson process of uniform rate. As such, the model can be reformulated as a discrete-time MDP, obtained by embedding at transition epochs.

Throughout the thesis we use value iteration to find either (characterizations of) average-cost optimal policies (as described above), or stochastically optimal policies. The latter is done by setting the direct cost equal to zero, $C(\cdot) = 0$, and allowing a terminal cost at the end of the horizon, $V_0(\cdot) = \tilde{C}(\cdot)$. In that case, the term $V_{m+1}(x)$ represents the minimum achievable expected terminal cost when the system starts in state x at $m+1$ time units away from the horizon, i.e., the term $\mathbb{E}(\tilde{C}(X_{m+1})|X_0 = x)$ is minimized. Setting $\tilde{C}(\cdot) = \mathbf{1}_{(\tilde{c}(\cdot) > s)}$, with $\tilde{c}(\cdot)$ some cost function, this corresponds to minimizing $\mathbb{P}(\tilde{c}(X_m) > s|X_0 = x)$. The minimizing action in (1.8) is an optimal action at $m + 1$ time units from the horizon. Hence, if the optimal actions do not depend on the time horizon m and on the value for s, then the corresponding stationary policy stochastically minimizes the cost $\tilde{c}(X_t)$ for all t.

1.6.3 Fluid scaling

The analysis of fluid-scaled processes has proved to be a powerful approach to investigate stability and optimal scheduling in queueing networks. A well-known result is [44], where stability of a multi-class queueing network is linked to that of the corresponding fluid-scaled model. For more details on fluid analysis, we refer to [40, 97, 115] and references therein. In this section we describe the fluid scaling of interest and focus on its application to optimal scheduling.

Consider a sequence of processes, indexed by $r \in \mathbb{N}$, such that $N_k^r(t)$ denotes the number of class-k users at time t in a queueing network with K classes of users when the initial queue lengths equal $N_k^r(0) = rn_k$, $n_k \geq 0$, $k = 1, \ldots, K$. The fluid-scaled number of users is obtained when both time and space are scaled linearly, i.e.,

$$\overline{N}_k^r(t) := \frac{N_k^r(rt)}{r},\ k = 1, \ldots, K.$$

Whenever fluid scaling is applied in this thesis, we assume exponential inter-arrival times and service requirements, and consider non-anticipating policies. More general service requirements are allowed when posing additional conditions on the intra-class policies. Due to the functional strong law of large numbers [40], loosely speaking, each converging subsequence of $\overline{N}^r(t)$ converges to some process $\overline{N}(t)$, which has continuous characteristics and deterministic fluid input [44]. This limit is referred to as a *fluid limit.*

When it does not seem possible to derive optimal policies for the stochastic queueing network, fluid-scaling techniques can help to obtain approximations instead. In order to do so, a deterministic *fluid control model* is considered, which is a first-order approximation of the stochastic network by only taking into account the mean drifts. For example, in a multi-class single-server queue, on average λ_k class-k users arrive per time unit, and on average $\mu_k := 1/\mathbb{E}(B_k)$ class-k users depart when class k is given full priority. Hence, in this case the fluid control model is described by the process $(n_1(t), \ldots, n_K(t))$ that satisfies

$$n_k(t) = n_k + \lambda_k t - \mu_k U_k(t), \text{ and } n_k(t) \geq 0,\ t \geq 0,\ k = 1, \ldots, K,$$

with $U_k(t) = \int_0^t u_k(v)\mathrm{d}v$ and where $u_k(\cdot)$ are feasible control functions, i.e.,

$$\sum_{k=1}^{K} u_k(v) \leq 1, \quad \text{and} \quad u_k(v) \geq 0,\ k = 1, \ldots, K, \text{ for all } v \geq 0.$$

In this thesis we call a fluid control optimal when it minimizes $\int_0^\infty \sum_{k=1}^K c_k n_k(t)\mathrm{d}t$. The optimal trajectories in the fluid control model are denoted by $n_1^*(t), \ldots, n_K^*(t)$.

In the literature, optimal fluid controls have been obtained by using Pontryagin's maximum principle, see for example [14] or by solving a separated continuous linear program, see for example [154].

Motivated by the close relation between stability of the stochastic queueing network and its associated fluid model [44], researchers became interested in connections between optimal scheduling in the stochastic network and the far simpler fluid control problem [13, 95, 97]. A crucial question is how to make a translation from the optimal control in the fluid model to a stable and efficient policy in the actual stochastic network. The optimal fluid control provides intuition on what a good policy in the stochastic network should try to achieve, however, difficulties can arise around the boundaries of the state space where a straightforward translation is not always adequate, see for example [53] and Chapters 5 and 8.

Once a translation to the stochastic network has been made, one needs to show that this policy is close to optimal. We use the following concept. Given that the system is stable, a policy π is called *asymptotically fluid-optimal* when

$$\lim_{r\to\infty} \mathbb{E}\Big(\int_0^D \sum_{k=1}^K c_k \overline{N}_k^{\pi,r}(t)\mathrm{d}t\Big) = \int_0^D \sum_{k=1}^K c_k n_k^*(t)\mathrm{d}t,$$

for all D sufficiently large. The main step to prove that a policy is asymptotically fluid-optimal consists in showing that the fluid limit of the stochastic network under this policy coincides with the optimal trajectories in the fluid control model, $n_1^*(t),\ldots,n_K^*(t)$. We refer to [15, 53, 88, 90, 96] and Chapters 5 and 8 for several examples of multi-class queueing networks for which asymptotically fluid-optimal policies have been derived.

Under suitable conditions, an average-cost optimal policy is asymptotically fluid-optimal [16], [53], [97, Theorem 10.0.5]. Unfortunately, no guarantee exists for the average cost of an asymptotically fluid-optimal policy. In fact, the asymptotically fluid-optimality definition aims at emptying the system efficiently starting from large initial state conditions, while average-cost optimality is concerned with the steady-state behavior of the system. In numerical experiments it has been observed that the average cost under asymptotically fluid-optimal policies is close to optimal. A first step towards a formal connection has been made in [96]. There, asymptotically fluid-optimal policies are proposed for which bounds on the average cost exist. In heavy traffic, these bounds (scaled with $1-\rho$) are tight and coincide with the optimal (scaled) average cost.

1.6.4 Heavy-traffic regime

Under a heavy-traffic regime the system is investigated as the traffic load approaches the capacity limit of the system. Analyzing the system in this regime can provide useful intuition as to how the system behaves when it is close to saturation. Typical heavy-traffic results relate to optimal control, queue length approximations, and state-space collapse (reduction in dimension of a multi-dimensional stochastic process).

The earliest heavy-traffic result is due to Kingman [76] who considered the steady-state behavior of a single-server queue under a non-preemptive policy (for service requirements with finite second moments). He proved that the steady-state queue length, scaled with $1-\rho$, converges in distribution to an exponential random variable as $\rho \to 1$. For PS or DPS the queue length is of the order $(1-\rho)^{-1}$ as well, see for example Chapter 2, but this is not true in general. For example, under LAS it can be either smaller or larger than $(1-\rho)^{-1}$, depending on the service requirement distribution [105].

So-called diffusion-scaled processes are commonly studied in a heavy-traffic setting to describe the *transient* behavior. A sequence of traffic parameters, indexed by r, is considered that converges at an appropriate rate to a heavily-loaded system. Let $N_k^r(t)$ denote the number of class-k users in the r-th system and define the diffusion-scaled number of users by

$$\hat{N}_k^r(t) := N_k^r(rt)/\sqrt{r}.$$

Due to the functional central limit theorem, the limit of such a diffusion-scaled process typically involves a reflected Brownian motion [40, 80]. We refer to [25, 68, 89] for several examples of queueing networks where diffusion-scaled processes have been analyzed.

For the single-server queue the diffusion scaling consists in letting

$$\lim_{r\to\infty} \rho^r = 1 \quad \text{such that} \quad \lim_{r\to\infty} \sqrt{r}\mu^r(\rho^r - 1) = \theta \in \mathbb{R}.$$

It is known that the diffusion-scaled number of users in a non-preemptive single-server system converges to a reflected Brownian motion with negative drift [40, 80]. Note that the stationary distribution of the latter process is exponential [40, Theorem 6.2], which coincides with the exponential distribution as mentioned earlier for the scaled steady-state process. For general networks, it is not obvious whether this interchange of the heavy-traffic limit and steady-state limit is allowed, and it has only been proved for some special cases, see for example [55] and Remark 2.6.2.

Optimal scheduling in heavy traffic is a well-studied field, typically focusing on policies with non-preemptive intra-class policies. For example, in [98] it is proved that a generalized Max-Weight policy is approximately optimal in the sense that its average cost is at most $|\log(1-\rho)|$ worse than that of the optimal average-cost policy, implying optimality in heavy traffic. Other optimality results relate to diffusion-scaled networks, where the goal is to find a policy that minimizes some diffusion-scaled cost (either sample-path wise or on average) as $r \to \infty$ [19, 61, 89, 99, 160]. Asymptotically optimal policies in heavy traffic can serve as useful approximations for the optimal policy in the original system when the load is high.

1.7 Overview of the thesis

In this chapter we presented several concepts related to resource-sharing systems, with special attention for the single-server system and two extensions of this model:

the linear network and the parallel two-server model with two classes of users. In the remainder of the thesis we concentrate on these three systems for which we investigate efficient scheduling policies and evaluate policies that share the resources in a fair manner.

In Chapter 2 we focus on the single-server system and analyze a generalization of the DPS policy. More specifically, we consider phase-type distributed service requirements and allow customers to have different weights in various phases of their service. In our main result we establish a state-space collapse for the steady-state queue length vector in heavy traffic. This result has several interesting consequences. We derive that in heavy traffic the remaining service requirement of any customer is distributed according to the forward recurrence time of its service requirement. In addition, we obtain that the scaled holding cost stochastically reduces as customers with lower variability in their service requirement obtain larger weights. Chapter 2 presents the results that appeared in [143, 144].

In Chapter 3 we turn to the linear bandwidth-sharing network. We investigate fundamental stability properties of size-based scheduling mechanisms, such as SRPT and LAS, applied in a linear network. The results indicate that instability effects may occur when users with long routes have relatively large service requirements compared to the ones with short routes. For networks with sufficiently many nodes, instability phenomena may in fact arise at arbitrarily low traffic loads. When instead the long routes have relatively small service requirements, size-based scheduling policies are stable under the maximum stability conditions. This chapter is based upon [146].

Chapter 4 focuses on optimal scheduling within the class of non-anticipating policies for the linear bandwidth-sharing network with exponentially distributed service requirements. We observe that policies that minimize the mean holding cost strongly depend on the mean service requirements of the various classes. For certain settings, simple priority rules are optimal. In the case of a two-node linear network, an optimal policy can be characterized in the remaining cases by "switching curves", i.e., the policy dynamically switches between several priority rules. Knowledge of optimal policies allows to evaluate the performance of the class of α-fair bandwidth-sharing policies. Through numerical experiments we observe that the gap between α-fair policies and optimal policies is not that large provided the system load is moderate. In addition, the performance under α-fair policies is quite insensitive to α, as long as this value is not too small. Chapter 4 presents the results published in [147, 151].

Chapter 5 is a continuation of Chapter 4. In Chapter 4 it was shown that in a two-node linear network an optimal policy is characterized by switching curves, however, an exact characterization of these curves was in general not possible. In this chapter we set out to study these switching curves in asymptotic regimes. We find that linear switching curves are optimal for the related fluid control problem. Using this fact, we derive that, in most cases, policies characterized by these linear switching curves are asymptotically fluid-optimal in the original stochastic model as well. For some scenarios however, fluid-based switching curves may result in a policy that not only is far from optimal, but may in fact be unstable. In that case, the diffusion scaling is appropriate and efficient switching-curve policies have a square-root shape. Through numerical experiments we assess the potential gain that switching curve policies can achieve over weighted α-fair policies in a moderately-loaded regime, and find that the latter can approach the optimal performance when choosing the weights appropriately. Chapter 5 builds upon the analysis of [148, 150].

While in Chapters 4 and 5 we concentrated on exponentially distributed service requirements, in Chapter 6 we turn to generally distributed service requirements. Since deriving a strictly optimal policy for the linear network does not seem possible, we instead consider a heavy-traffic regime. Motivated by the size-based scheduling results for single-server systems, we focus on (anticipating) policies that separate within a class the large requests from the small ones. Such policies turn out to be asymptotically optimal in heavy traffic for service requirements with bounded support. In addition, we show that these size-based policies may outperform α-fair policies, which are non-anticipating, by an arbitrarily large factor when the load is sufficiently high. This chapter presents the results published in [145].

In Chapter 7 we consider a multi-class queueing system with general inter-arrival times and service requirements, and give sufficient conditions in order to compare sample-path wise the workload and the number of users under different policies. This allows us to evaluate the performance of the system under various policies in terms of stability and the mean holding cost. In particular, for the linear network under weighted α-fair policies we obtain stability results and, in the case of exponentially distributed service requirements, establish monotonicity of the mean holding cost with respect to the fairness parameter α and the relative weights. In order to broaden the comparison results, we investigate a heavy-traffic regime and perform numerical experiments. In addition, we study a single-server system with two user classes, and show that under DPS and Generalized Processor Sharing the mean holding cost is monotone with respect to the relative weights. This result is in line with the monotonicity result obtained for DPS under a heavy-traffic scaling in Chapter 2. Chapter 7 is based upon [141, 142].

In Chapter 8 we turn our attention to a parallel two-server model with two classes of users and set out to study optimal non-anticipating scheduling policies for exponentially distributed service requirements. For some settings we can determine the optimal policy exactly, but in general this is analytically infeasible. We therefore seek asymptotically fluid-optimal policies, using similar techniques as in Chapter 5. We investigate the fluid control model for which we show that the optimal control is described by a switching curve. Using this fact, we derive that policies characterized by either linear or exponential switching curves are asymptotically fluid-optimal in the original stochastic model. For a moderately-loaded system, we numerically compare these fluid-based policies with Max-Weight and threshold-based policies, which are known to be optimal in a heavy-traffic setting. We observe that the fluid-based and the threshold-based policies perform well, while significant performance gains can be achieved over Max-Weight policies. Chapter 8 is based upon [149].

Chapter 2
Heavy-traffic analysis of discriminatory processor sharing

Efficient scheduling in a single-server system is a well-studied field, as described in Section 1.3.3. In this chapter we focus on Discriminatory Processor Sharing (DPS) policies and are interested in how the choice of the weight parameters affects the performance of the system in steady state. In fact, we analyze a generalization of the DPS queue with phase-type distributed service requirements, and allow customers to have different weights in various phases of their service. Since the steady-state analysis will not be tractable in general, we study the system in heavy-traffic conditions.

In the main result of this chapter we establish a state-space collapse for the steady-state queue length vector in heavy traffic. The result shows that in the limit, the queue length vector is the product of an exponentially distributed random variable and a deterministic vector. The reduction of dimensionality of a multi-dimensional stochastic process under heavy-traffic scaling has been demonstrated previously in other queueing models, see for example [19, 68, 132]. In addition, our main result allows to derive several interesting results concerning the residual service requirements and monotonicity properties of the holding cost.

Our work is inspired by the heavy-traffic analysis in [113] for the traditional DPS model with exponentially distributed service requirements. After developing a procedure to determine all moments of the queue length distributions from systems of linear equations, the authors show that the variability of the queue length vector is of a lower order than the mean queue lengths, which directly leads to state-space collapse of the queue length vector. Here we follow a different and more direct approach by investigating the joint probability generating function of the queue lengths. This function is shown to satisfy a partial differential equation, which takes a convenient form after passing to the heavy-traffic limit, allowing a closed-form solution in that case.

Generalized DPS models similar to the one studied in this chapter were previously considered in [23, 58, 63]. The analysis in [58] is particularly relevant for the present study. Through appropriate choices for a quite general functional of the queue length process, [58] determined the heavy-traffic distributions of the marginal queue lengths and sojourn times, when the service requirements have finite second moments. Our results are complementary to those: On one hand we restrict the focus to the queue lengths, and on the other hand we study the *joint* queue length distribution. That way, we establish a state-space collapse for the queue length vector.

Several papers have analyzed (discriminatory) processor sharing mechanisms assuming overload conditions and general service requirement distributions. For example, the authors of [7] characterize the queue length growth rates of the standard DPS model by a fixed-point equation, generalizing the analogous result for the PS model [66]. More recently, further extensions to bandwidth-sharing networks [46] and a network setting similar to ours [23] have been obtained. In all these references the *transient* behavior of the queue lengths is studied under overload conditions, while we investigate the convergence of the (scaled) *steady-state* distribution as the critical load is approached.

As phase-type distributions lie dense in the class of all probability distributions, in practice the restriction to this class is not seen as being essential. In the present chapter, an important caveat must be accounted for, though. Our analysis relies on heavy-traffic scaling techniques which typically require finite second moments of the service requirements. Since all phase-type distributions (with a finite number of phases) have a finite second moment, this restriction is implicit in our modeling approach. Indeed, our results show that the second moments appear in a natural fashion in the heavy-traffic limit. We believe that our results do extend to all distributions with a finite second moment (not necessarily phase-type), but we do not investigate this here.

Allowing the relative service weights of customers to change over time as they acquire service, effectively opens up a way to implement size-based scheduling by assigning relatively high weights in service phases that are more likely to lead to a quick service completion. Using the heavy-traffic result, we investigate how the choice of the weights influences the asymptotic performance of the system. In particular, we prove that the scaled holding cost reduces as more preference is given to customers in service phases with a small expected remaining service requirement.

The standard DPS queue with phase-type service requirement distributions is a special case of our model. The state-space collapse allows to show that in a heavy-traffic setting, conditioned on the number of customers, the remaining service requirements of the various customers are independent and distributed according to the forward recurrence times. In addition, we derive that the scaled holding cost in the standard DPS queue reduces as more preference is given to classes according to the forward recurrence times of the service requirements. The applicability of this result for a moderately loaded system is investigated by numerical experiments.

The present chapter is organized as follows. In Section 2.1 we introduce the general Markovian framework and state the main result, which establishes a state-space collapse of the joint queue length vector. As a preparation for the proof of the main result, the functional equation for the generating function of the joint queue length process is studied in Section 2.2 and, under the heavy-traffic scaling, in Section 2.3. The proof of the main result is given in Section 2.4. Section 2.5 discusses size-based scheduling. Section 2.6 applies the state-space collapse result to the standard DPS queue with phase-type distributed service requirements. In addition, it presents the implications for residual service requirements and monotonicity properties of the holding cost. Concluding remarks can be found in Section 2.7.

2.1 General framework and main result

We consider a general Markovian system with J customer types. Customers arrive according to a Poisson process with rate λ, and an arriving customer is of type i with probability p_{0i}, $i = 1, \dots, J$. Type-i customers have an exponentially distributed service requirement with mean $1/\mu_i$. After service completion, they become of type j with probability p_{ij}, $j = 1, \dots, J$, and leave the system with probability $p_{i0} := 1 - \sum_{j=1}^{J} p_{ij}$. Let P be a $J \times J$ matrix with $P = (p_{ij})$, $i, j = 1, \dots, J$. We assume that customers require a finite service amount with probability one, so that all customers eventually leave the system. This implies $\lim_{n\to\infty} P^n = 0$, and hence, $(I - P)^{-1}$ is well defined. In addition, we assume that none of the J types are redundant (i.e., eventually all types are observed); this assumption is formalized following equation (2.1) below.

The J customer types share a common resource of capacity one. There are strictly positive weights $g_1, \dots, g_J$ associated with each of the types. Whenever there are q_i type-i customers, $i = 1, \dots, J$, present in the system, each type-j customer is served at rate

$$\frac{g_j}{\sum_{i=1}^{J} g_i q_i}, \quad j = 1, \dots, J.$$

We denote the number of type-i customers in steady-state by Q_i.

The above-described framework is a generalization of the DPS queue with phase-type distributed service requirements: It represents an M/PH/1 DPS queue where customers may have different weights in various phases of their service.

We let R_i denote the remaining service requirement until departure for a customer that is now of type i. Note that this includes service in all subsequent stages as the customer changes from one type to another. Since the service time of each type is exponentially distributed, the expected remaining service requirements can be interpreted as absorption times in an appropriate Markov chain and therefore satisfy the following system of linear equations: $\mathbb{E}(R_i) = \frac{1}{\mu_i} + \sum_{j=1}^{J} p_{ij}\mathbb{E}(R_j)$, $i = 1, \dots, J$. Let $\mathbb{E}(\vec{R}) = (\mathbb{E}(R_1), \dots, \mathbb{E}(R_J))^T$ and $\vec{m} = (1/\mu_1, \dots, 1/\mu_J)^T$, so that we can write

$$\mathbb{E}(\vec{R}) = (I - P)^{-1}\vec{m}.$$

Denote the total traffic load by

$$\rho := \lambda \sum_{j=1}^{J} p_{0j}\mathbb{E}(R_j).$$

Let γ_i represent the expected number of times a customer is of type i during its visit in the network. Hence, $\gamma_1, \ldots, \gamma_J$ satisfy the following equations

$$\gamma_i = p_{0i} + \sum_{j=1}^{J} \gamma_j p_{ji}, \quad i = 1, \ldots, J, \tag{2.1}$$

i.e., $\vec{\gamma}^T = \vec{p}_0^T (I - P)^{-1}$, with $\vec{\gamma} = (\gamma_1, \ldots, \gamma_J)^T$ and $\vec{p}_0 = (p_{01}, \ldots, p_{0J})^T$. Note that $\frac{\gamma_i}{\mu_i}$ represents the expected cumulative amount of service a customer requires while being of type i during its visit in the network. Our assumption that none of the J types is redundant, entails that $\vec{\gamma}$ is a vector with strictly positive elements. We denote the load corresponding to customers while they are of type i by

$$\rho_i := \lambda \frac{\gamma_i}{\mu_i}.$$

Hence, for the total traffic load ρ we may equivalently write

$$\rho = \lambda \sum_{j=1}^{J} p_{0j} \mathbb{E}(R_j) = \lambda \vec{p}_0^T \mathbb{E}(\vec{R}) = \lambda \vec{p}_0^T (I - P)^{-1} \vec{m} = \lambda \vec{\gamma}^T \vec{m} = \lambda \sum_{j=1}^{J} \frac{\gamma_j}{\mu_j} = \sum_{j=1}^{J} \rho_j. \tag{2.2}$$

Our main result shows that the steady-state distribution of the multi-dimensional queue length process takes a rather simple form when the system is near saturation, i.e., $\rho \uparrow 1$, which is commonly referred to as the heavy-traffic regime. This regime can be obtained by fixing the $\vec{p}_0, P$, and $\vec{m}$, and letting

$$\lambda \uparrow \hat{\lambda} := \frac{1}{\vec{p}_0^T (I - P)^{-1} \vec{m}}, \tag{2.3}$$

since then $\rho = \lambda \vec{p}_0^T (I - P)^{-1} \vec{m} \uparrow 1$. Although approaching heavy traffic in this way is natural, the results remain valid for any other sequence of parameters (belonging to stable systems) that reaches heavy traffic in the limit. In heavy traffic, we denote by

$$\hat{\rho}_i := \hat{\lambda} \frac{\gamma_i}{\mu_i}$$

the load corresponding to customers while they are of type i ($\sum_{j=1}^{J} \hat{\rho}_j = 1$).

We can now state our main result, which establishes a state-space collapse for the queue length vector in the heavy-traffic regime.

Proposition 2.1.1. *Consider the general Markovian framework. When scaled with $1 - \rho$, the queue length vector has a proper limiting distribution as $(\rho_1, \ldots, \rho_J) \to (\hat{\rho}_1, \ldots, \hat{\rho}_J)$, such that $\rho \uparrow 1$,*

$$(1 - \rho)(Q_1, Q_2, \ldots, Q_J) \xrightarrow{d} (\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) \overset{d}{=} X \cdot \left(\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J} \right), \tag{2.4}$$

where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean

$$\mathbb{E}(X) = \frac{\sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j)}{\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \mathbb{E}(R_j)}. \tag{2.5}$$

The proof will be given in Section 2.4. Here we give some intuition for the result. Proposition 2.1.1 shows that in heavy traffic, the multi-dimensional queue length process essentially reduces to a one-dimensional random process: it can be expressed as a random variable X times a deterministic vector. Given this reduced variability of the process, the value of the deterministic vector can be understood as follows. Note that, in general

$$\rho_j = \mathbb{E}\left(\frac{g_j Q_j}{\sum_{i=1}^{J} g_i Q_i} \cdot \mathbf{1}_{(\sum_{i=1}^{J} Q_i > 0)} \right), \tag{2.6}$$

since the expression within the expectation operator reflects the capacity allocated to type j. Here the function $\mathbf{1}_A$ denotes the indicator function, i.e., $\mathbf{1}_A = 1$ if A is true, and 0 otherwise. Using that the process reduces to one dimension in heavy traffic, in the limit we may replace Q_j/Q_i by a ratio of constants a_j/a_i. Together with (2.6) and under the assumption that the scaled queue length will be strictly positive in heavy traffic, this implies that $a_j = (\sum_{i=1}^{J} g_i a_i) \frac{\hat{\rho}_j}{g_j}$. The pre-factor $\sum_i g_i a_i$ is common to all a_j, which explains the appearance of the vector $(\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J})$ in Proposition 2.1.1.

Numerical illustration of Proposition 2.1.1: We consider two types of customers and choose $g_1 = 2, g_2 = 1, \mu_1 = 2, \mu_2 = 5, p_{01} = 0.6, p_{02} = 0.4, p_{12} = 0.3, p_{21} = 0.1$. In Figures 2.1 and 2.2 we plot the joint queue length probabilities (obtained by simulation) for loads $\rho = 0.8$ ($\rho_1 \approx 0.59, \rho_2 \approx 0.21$), $\rho = 0.90$ ($\rho_1 \approx 0.66, \rho_2 \approx 0.24$) and $\rho = 0.99$ ($\rho_1 \approx 0.73, \rho_2 \approx 0.26$), respectively. The horizontal and vertical axes correspond to Q_1 and Q_2 respectively. As a consequence of the state-space collapse stated in Proposition 2.1.1, in heavy traffic the probabilities will lie on a straight line with slope $\frac{g_1}{\hat{\rho}_1} \frac{\hat{\rho}_2}{g_2} \approx 0.72$, starting from the origin. In Figures 2.1 and 2.2 we see that as the load increases, the probable states indeed tend to concentrate more around this line. For load $\rho = 0.99$, this effect is clearly visible; the probable queue length states are strongly concentrated around the line with slope 0.72.

2.2 Functional equation

Before focusing on the heavy-traffic regime, we derive a functional equation for the generating function of the joint queue length process. Denote by $\vec{Q}$ and $\vec{q}$ the vectors $(Q_1, Q_2, \ldots, Q_J) \geq \vec{0}$ and $(q_1, q_2, \ldots, q_J) \geq \vec{0}$, respectively. The equilibrium

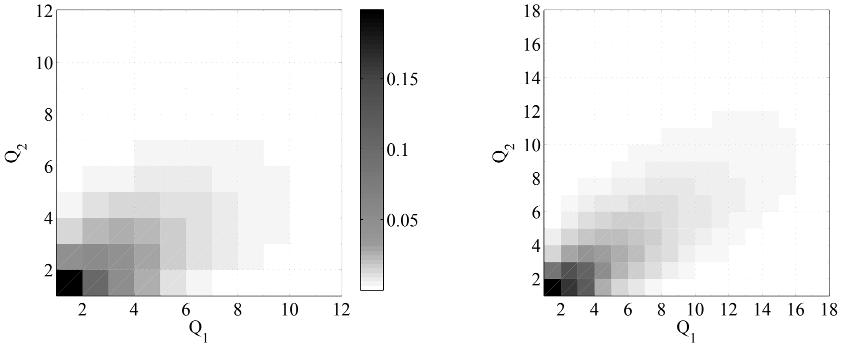


Figure 2.1: Joint queue length probabilities for load $\rho = 0.8$ (left) and $\rho = 0.90$ (right), respectively.

distribution $\pi(\vec{q}) := \mathbb{P}(\vec{Q} = \vec{q})$ satisfies

$$\lambda\pi(\vec{0}) = \sum_{i=1}^{J} \mu_i p_{i0} \pi(\vec{e}_i), \tag{2.7}$$

and for $\vec{q} \neq \vec{0}$,

$$\left(\lambda + \frac{\sum\limits_{i=1}^{J} g_i q_i \mu_i}{\sum\limits_{i=1}^{J} g_i q_i}\right) \pi(\vec{q}) = \sum_{i=1}^{J} \lambda p_{0i} \delta_{q_i} \pi(\vec{q} - \vec{e}_i) + \sum_{i=1}^{J} \frac{g_i(q_i+1)}{\sum\limits_{j=1}^{J} g_j q_j + g_i} \cdot \mu_i p_{i0} \pi(\vec{q} + \vec{e}_i)$$
$$+ \sum_{i=1}^{J}\sum_{j=1}^{J} \delta_{q_j} \cdot \frac{g_i(q_i+1)}{\sum\limits_{m=1}^{J} g_m q_m + g_i - g_j} \cdot \mu_i p_{ij} \pi(\vec{q} + \vec{e}_i - \vec{e}_j), \tag{2.8}$$

where $\delta_q = 1$ if $q > 0$, and $\delta_q = 0$ otherwise, and with $\vec{e}_i$ the i-th unit vector. It will be notationally convenient to use the following transformation:

$$R(\vec{0}) = 0 \quad \text{and} \quad R(\vec{q}) = \frac{\pi(\vec{q})}{\sum\limits_{j=1}^{J} g_j q_j}, \quad \text{for } \vec{q} \neq \vec{0}.$$

Also, let $p(\vec{z})$ and $r(\vec{z})$ denote the generating functions of $\pi(\vec{q})$ and $R(\vec{q})$, respectively, where $\vec{z} = (z_1, \ldots, z_J)$ and $|z_i| < 1$ for $i = 1, \ldots, J$:

$$\begin{aligned} p(\vec{z}) &= \mathbb{E}(z_1^{Q_1} \cdot \ldots \cdot z_J^{Q_J}) = \sum_{q_1=0}^{\infty} \cdots \sum_{q_J=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_J^{q_J} \pi(\vec{q}), \\ r(\vec{z}) &= \mathbb{E}\left(\frac{z_1^{Q_1} \cdot \ldots \cdot z_J^{Q_J}}{\sum_{j=1}^{J} Q_j g_j}\right) = \sum_{q_1=0}^{\infty} \cdots \sum_{q_J=0}^{\infty} z_1^{q_1} \cdot \ldots \cdot z_J^{q_J} R(\vec{q}), \end{aligned}$$

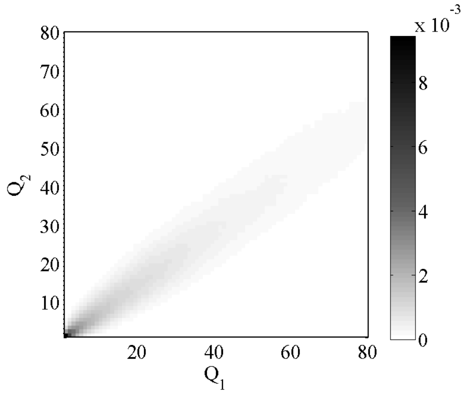


Figure 2.2: Joint queue length probabilities for load $\rho = 0.99$.

where we use the convention that $1/\sum_{j=1}^{J} Q_j g_j = 0$ when $\vec{Q} = \vec{0}$. Note that

$$g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} = \sum_{q_1,\ldots,q_J:\sum_{j=1}^{J} q_j>0} \frac{g_i q_i}{\sum_{j=1}^{J} g_j q_j} z_1^{q_1} \cdot \ldots \cdot z_J^{q_J} \pi(\vec{q}). \tag{2.9}$$

Multiplying (2.8) by $z_1^{q_1} \ldots z_J^{q_J}$, summing both sides over $q_1, q_2, \ldots, q_J$, and adding equation (2.7), we obtain from (2.9) that

$$\lambda p(\vec{z}) + \sum_{i=1}^{J} \mu_i g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} = \sum_{i=1}^{J} \lambda p_{0i} z_i p(\vec{z}) + \sum_{i=1}^{J} \mu_i g_i p_{i0} \frac{\partial r(\vec{z})}{\partial z_i} + \sum_{i=1}^{J}\sum_{j=1}^{J} \mu_i g_i p_{ij} z_j \frac{\partial r(\vec{z})}{\partial z_i}. \tag{2.10}$$

Since $\pi(\vec{0}) = 1 - \rho$, it follows from (2.9) that

$$\sum_{i=1}^{J} g_i z_i \frac{\partial r(\vec{z})}{\partial z_i} + 1 - \rho = p(\vec{z}). \tag{2.11}$$

Together with (2.10) this gives the following partial differential equation for $r(\vec{z})$:

$$\begin{aligned} &\lambda(1-\rho)(1 - \sum_{i=1}^{J} p_{0i} z_i) \\ &= \sum_{i=1}^{J} \left(\mu_i g_i (p_{i0} + \sum_{j=1}^{J} p_{ij} z_j - z_i) - \lambda g_i z_i (1 - \sum_{j=1}^{J} p_{0j} z_j) \right) \frac{\partial r}{\partial z_i}. \end{aligned} \tag{2.12}$$

This equation turns out to be very useful to analyze the joint queue length distribution in heavy traffic, as it allows for an explicit solution in that asymptotic regime. That is the topic of the next two sections. Note that equation (2.12) was derived in [113] for the case of exponentially distributed service requirements.

2.3 Heavy-traffic scaling

It will be convenient to use the change of variables $z_i = \mathrm{e}^{-s_i}$ with $s_i > 0$, $i = 1, \ldots, J$. Denote $\vec{s} = (s_1, \ldots, s_J)$ and $\mathrm{e}^{-(1-\rho)\vec{s}} = (\mathrm{e}^{-(1-\rho)s_1}, \ldots, \mathrm{e}^{-(1-\rho)s_J})$. If

$$\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \lim_{\rho\uparrow 1} \mathbb{E}(\mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \ldots \cdot \mathrm{e}^{-(1-\rho)s_J Q_J}) \tag{2.13}$$

exists, then there is a (possibly defective) random vector $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ such that $(1-\rho)(Q_1, Q_2, \ldots, Q_J)$ converges in distribution to $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$, and the distribution of $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ is uniquely determined by the limit in (2.13) (cf. the Continuity theorem [52]). For now, we assume that the limit exists; we come back to this assumption in Section 2.4. In this section we give two lemmas that describe properties of $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$. In particular, in Lemma 2.3.2 we obtain a partial differential equation, which will be the key element in the proof of the main result stated in Proposition 2.1.1.

In order to describe the behavior of the generating function, we define

$$\hat{r}(\vec{s}) := \mathbb{E}\left(\frac{1 - \mathrm{e}^{-s_1 \hat{Q}_1} \cdot \ldots \cdot \mathrm{e}^{-s_J \hat{Q}_J}}{\sum_{j=1}^J \hat{Q}_j g_j} \right),$$

where we use the convention that $1/\sum_{j=1}^J \hat{Q}_j g_j = 0$ when $\sum_{j=1}^J \hat{Q}_j = 0$. The “1” in the numerator is to ensure that the expression between brackets remains bounded when the $\hat{Q}_j$'s are all near zero. We can now state the following lemma.

Lemma 2.3.1. *If* $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ *exists, then it satisfies:*

$$\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{i=1}^J g_i \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \tag{2.14}$$

Proof: From (2.11) we have

$$\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \lim_{\rho\uparrow 1} \sum_{i=1}^J g_i \frac{\partial r(\vec{z})}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}}. \tag{2.15}$$

By definition of $r(\vec{z})$ we can write

$$\begin{aligned}
\lim_{\rho\uparrow 1} \frac{\partial r(\vec{z})}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} &= \lim_{\rho\uparrow 1} \frac{\partial \mathbb{E}\left(\frac{z_1^{Q_1} \cdot \ldots \cdot z_J^{Q_J}}{\sum_{j=1}^J Q_j g_j} \right)}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} \\
&= \lim_{\rho\uparrow 1} \mathbb{E}\left(\frac{Q_i}{\sum_{j=1}^J Q_j g_j} \cdot \frac{\mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \ldots \cdot \mathrm{e}^{-(1-\rho)s_J Q_J}}{\mathrm{e}^{-(1-\rho)s_i}} \right) \\
&= \mathbb{E}\left(\frac{\hat{Q}_i}{\sum_{j=1}^J \hat{Q}_j g_j} \cdot \mathrm{e}^{-s_1 \hat{Q}_1} \cdot \ldots \cdot \mathrm{e}^{-s_J \hat{Q}_J} \right) \\
&= \frac{\partial \hat{r}(\vec{s})}{\partial s_i}.
\end{aligned} \tag{2.16}$$

In the third step we used that $\frac{Q_i}{\sum_{j=1}^{J} Q_j g_j} \cdot \mathrm{e}^{-(1-\rho)s_1 Q_1} \cdot \ldots \cdot \mathrm{e}^{-(1-\rho)s_J Q_J}$ is upper bounded by $\frac{1}{\min_j(g_j)}$, and, cf. the continuous mapping theorem [27], converges in distribution to $\frac{\hat{Q}_i}{\sum_{j=1}^{J} \hat{Q}_j g_j} \cdot \mathrm{e}^{-s_1 \hat{Q}_1} \cdot \ldots \cdot \mathrm{e}^{-s_J \hat{Q}_J}$. From (2.15) and (2.16) we obtain (2.14). □

In the following lemma we show that the partial differential equation as given in (2.12) simplifies considerably in the heavy-traffic regime.

Lemma 2.3.2. *If* $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ *exists, then the function* $\hat{r}(\vec{s})$ *satisfies the following partial differential equation:*

$$0 = \sum_{i=1}^{J} F_i(\vec{s}) \frac{\partial \hat{r}(\vec{s})}{\partial s_i} = \vec{F}(\vec{s}) \cdot \bigtriangledown \hat{r}(\vec{s}), \quad \forall\ \vec{s} \geq \vec{0},$$

where $\vec{F}(\vec{s}) = (F_1(\vec{s}), \ldots, F_J(\vec{s}))$, *and*

$$F_i(\vec{s}) = g_i\Big(\mu_i(-s_i + \sum_{j=1}^{J} p_{ij} s_j) + \hat{\lambda} \sum_{j=1}^{J} p_{0j} s_j\Big), \tag{2.17}$$

with $\hat{\lambda}$ *as defined in (2.3).*

Proof: Taking $\vec{z}$ equal to $\mathrm{e}^{-(1-\rho)\vec{s}}$ in (2.12), dividing both sides by $1-\rho$ and taking the limit of $\rho \uparrow 1$, this gives

$$\begin{aligned} 0 = \lim_{\rho\uparrow 1} \sum_{i=1}^{J} \Bigg(\mu_i g_i & \frac{1 - \mathrm{e}^{-(1-\rho)s_i} + \sum\limits_{j=1}^{J} p_{ij}(\mathrm{e}^{-(1-\rho)s_j} - 1)}{1-\rho} \\ & - \lambda g_i \mathrm{e}^{-(1-\rho)s_i} \sum_{j=1}^{J} p_{0j} \frac{1 - \mathrm{e}^{-(1-\rho)s_j}}{1-\rho} \Bigg) \cdot \frac{\partial r(\vec{z})}{\partial z_i}\Big|_{\vec{z}=\mathrm{e}^{-(1-\rho)\vec{s}}} \\ = \sum_{i=1}^{J} g_i \cdot \Bigg(\mu_i(s_i - \sum_{j=1}^{J} p_{ij} s_j) - \hat{\lambda} \sum_{j=1}^{J} p_{0j} s_j \Bigg) \cdot \frac{\partial \hat{r}(\vec{s})}{\partial s_i}. \end{aligned} \tag{2.18}$$

In the second step we used (2.16) and the fact that $\lim_{\rho\uparrow 1} \frac{x^{1-\rho}-1}{1-\rho} = \ln(x)$. □

2.4 Proof of the main result

This section contains the proof of the main result stated in Proposition 2.1.1. It consists of two steps. First we show in Section 2.4.1 that

$$(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) \overset{d}{=} (\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J}) \cdot X, \tag{2.19}$$

for some random variable X. Second, we demonstrate in Section 2.4.2 that X is exponentially distributed with mean as given in (2.5). With these two partial results, the proof can be completed as follows: In Section 2.3 we assumed that $\lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}})$ exists, thereby showing in Sections 2.4.1 and 2.4.2 that there is a unique limit. For any converging subsequence this analysis can be performed, in particular for the lim sup and lim inf, which implies that the limit itself exists. This establishes the state-space collapse $(1-\rho)(Q_1, Q_2, \ldots, Q_J) \xrightarrow{d} (\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ with $(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J)$ taking only values on the line described in (2.19).

2.4.1 State-space collapse

In this section we give the proof of (2.19). The proof is based on the fact that the probability generating function satisfies the partial differential equation as described in Lemma 2.3.2. From this partial differential equation it can be derived that the function $\hat{r}(\vec{s})$ is constant on the $(J-1)$-dimensional hyperplane

$$H_c := \{\vec{s} \geq \vec{0} : \sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j = c\}, \quad c > 0,$$

as will be shown in Lemma 2.4.2. From this it follows that the function $\hat{r}(\vec{s})$ depends on $\vec{s}$ only through $\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j$, so there is a function $\hat{r}^* : \mathbb{R} \to \mathbb{R}$ such that $\hat{r}(\vec{s}) = \hat{r}^*(\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j)$. From Lemma 2.3.1 and $\frac{\partial \hat{r}(\vec{s})}{\partial s_i} = \frac{\hat{\rho}_i}{g_i} \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j}$, we then obtain

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{-\sum_{i=1}^{J} s_i \hat{Q}_i}) &= \lim_{\rho\uparrow 1} p(\mathrm{e}^{-(1-\rho)\vec{s}}) = \sum_{i=1}^{J} g_i \frac{\partial \hat{r}(\vec{s})}{\partial s_i} = \sum_{i=1}^{J} \hat{\rho}_i \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j} \\ &= \frac{\mathrm{d}\hat{r}^*(v)}{\mathrm{d}v}\Big|_{v=\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j},\end{aligned}$$

which again depends on $\vec{s}$ only through $\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j$. Equivalently, we can write

$$\mathbb{E}(\mathrm{e}^{-\sum_{i=1}^{J} s_i \hat{Q}_i}) = \mathbb{E}(\mathrm{e}^{-\frac{g_1}{\hat{\rho}_1}\hat{Q}_1 \sum_{i=1}^{J} \frac{\hat{\rho}_i}{g_i} s_i} \cdot \mathrm{e}^{-s_2 \frac{\hat{\rho}_2}{g_2}(\frac{g_2}{\hat{\rho}_2}\hat{Q}_2 - \frac{g_1}{\hat{\rho}_1}\hat{Q}_1)} \cdot \ldots \cdot \mathrm{e}^{-s_J \frac{\hat{\rho}_J}{g_J}(\frac{g_J}{\hat{\rho}_J}\hat{Q}_J - \frac{g_1}{\hat{\rho}_1}\hat{Q}_1)}).$$

Since this only depends on $\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} s_j$, it implies that $\frac{g_i}{\hat{\rho}_i}\hat{Q}_i = \frac{g_j}{\hat{\rho}_j}\hat{Q}_j$ almost surely for all i, j, and we obtain:

$$(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) = \left(\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J}\right) \cdot \frac{g_1}{\hat{\rho}_1}\hat{Q}_1, \quad \text{almost surely,}$$

or equivalently

$$(\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_J) \overset{d}{=} \left(\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \ldots, \frac{\hat{\rho}_J}{g_J}\right) \cdot X,$$

with X distributed as $\frac{g_1}{\hat{\rho}_1}\hat{Q}_1$.

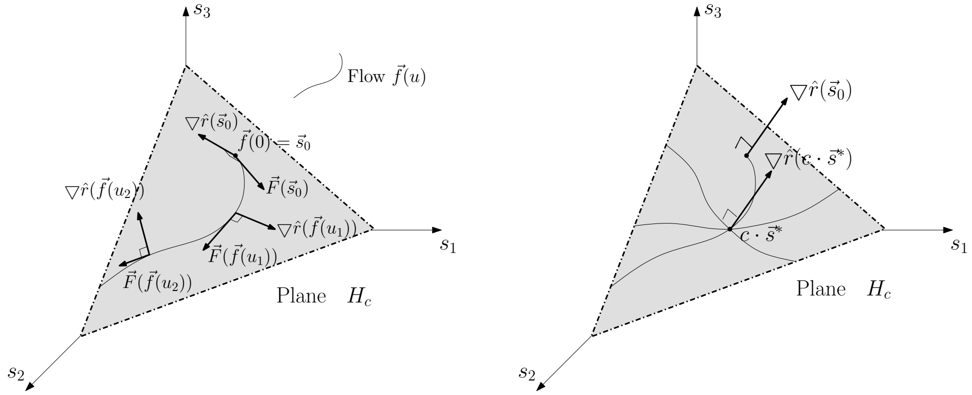

s_3
Flow $\vec{f}(u)$
$\bigtriangledown\hat{r}(\vec{s}_0)$
$\vec{f}(0) = \vec{s}_0$
$\vec{F}(\vec{s}_0)$
$\bigtriangledown\hat{r}(\vec{f}(u_2))$
$\bigtriangledown\hat{r}(\vec{f}(u_1))$
$\vec{F}(\vec{f}(u_1))$
$\vec{F}(\vec{f}(u_2))$
s_1
Plane H_c
s_2
s_3
$\bigtriangledown\hat{r}(\vec{s}_0)$
$\bigtriangledown\hat{r}(c \cdot \vec{s}^*)$
$c \cdot \vec{s}^*$
s_1
Plane H_c
s_2

Figure 2.3: Geometrical interpretation of the proof of Lemma 2.4.2 when $J = 3$.

Before we proceed to prove that the generating function $\hat{r}(\vec{s})$ is constant on the hyperplane H_c we first give a geometric interpretation for this fact in the particular case of $J = 3$. In Figure 2.3 (left) we depict the hyperplane H_c for $J = 3$. For a given $\vec{s}_0 \in H_c$, we draw a flow curve $\vec{f}(u)$, $u \geq 0$, defined such that the tangent at every point is precisely $\vec{f}'(u) := \vec{F}(\vec{f}(u))$ and $\vec{f}(0) = \vec{s}_0 \in H_c$. We will see in the proof of Lemma 2.4.2 that the vector $\vec{F}(\vec{s})$ is parallel to the hyperplane H_c, for all $\vec{s} \in H_c$, thus the flow $\vec{f}(u)$ stays in the hyperplane H_c for all $u \geq 0$. By Lemma 2.3.2, the vector $\vec{F}(\vec{s})$ and the gradient $\bigtriangledown\hat{r}(\vec{s})$ are perpendicular, for all $\vec{s}$, so $\vec{f}'(u) = \vec{F}(\vec{f}(u)) \perp \bigtriangledown\hat{r}(\vec{f}(u))$. Thus the function $\hat{r}$ has the same value in every point on a given flow $\vec{f}(u)$. In Figure 2.3 (right) we draw several flows in the hyperplane H_c. In the proof of Lemma 2.4.2 we will see that all flows starting in the hyperplane H_c converge to one common point $c \cdot \vec{s}^*$. Since the function $\hat{r}(\cdot)$ is continuous and constant on each flow trajectory, it follows that $\hat{r}(\vec{s})$ is constant on the whole hyperplane H_c, or equivalently, $\bigtriangledown\hat{r}(\vec{s}) \perp H_c$.

The following technical lemma is used in the proof of Lemma 2.4.2.

Lemma 2.4.1. *Consider the matrix A defined as*

$$\begin{pmatrix} g_1\Big(-\mu_1 + \mu_1 p_{11} + \hat{\lambda} p_{01}\Big) & g_1\Big(\mu_1 p_{12} + \hat{\lambda} p_{02}\Big) & \dots & g_1\Big(\mu_1 p_{1J} + \hat{\lambda} p_{0J}\Big) \\ g_2\Big(\mu_2 p_{21} + \hat{\lambda} p_{01}\Big) & g_2\Big(-\mu_2 + \mu_2 p_{22} + \hat{\lambda} p_{02}\Big) & \dots & g_2\Big(\mu_2 p_{2J} + \hat{\lambda} p_{0J}\Big) \\ \vdots & \vdots & \ddots & \vdots \\ g_J\Big(\mu_J p_{J1} + \hat{\lambda} p_{01}\Big) & g_J\Big(\mu_J p_{J2} + \hat{\lambda} p_{02}\Big) & \dots & g_J\Big(-\mu_J + \mu_J p_{JJ} + \hat{\lambda} p_{0J}\Big) \end{pmatrix},$$

where $\hat{\lambda}$ is as defined in (2.3). One eigenvalue of A is 0 (with multiplicity 1), and all the other eigenvalues have a strictly negative real part. In addition, there exists a vector $\vec{\eta} \geq \vec{0}$ with $\sum_{j=1}^{J} \eta_j = 1$ such that $\vec{s}^*$ with $s_j^* := \frac{g_j}{\hat{\rho}_j}\eta_j$ is an eigenvector of A corresponding to the eigenvalue 0, and $\vec{s}^* \in H_1$.

Proof: Define D as the diagonal matrix $\mathrm{diag}[d_1, d_2, \ldots, d_J]$ with $d_i := \frac{\hat{\rho}_i}{g_i}$, and define $S := DAD^{-1}$, i.e., S equals

$$\begin{pmatrix} g_1\Big(-\mu_1 + \mu_1 p_{11} + \hat{\lambda} p_{01}\Big) & \hat{\rho}_1 \frac{g_2}{\hat{\rho}_2}\Big(\mu_1 p_{12} + \hat{\lambda} p_{02}\Big) & \ldots & \hat{\rho}_1 \frac{g_J}{\hat{\rho}_J}\Big(\mu_1 p_{1J} + \hat{\lambda} p_{0J}\Big) \\ \hat{\rho}_2 \frac{g_1}{\hat{\rho}_1}\Big(\mu_2 p_{21} + \hat{\lambda} p_{01}\Big) & g_2\Big(-\mu_2 + \mu_2 p_{22} + \hat{\lambda} p_{02}\Big) & \ldots & \hat{\rho}_2 \frac{g_J}{\hat{\rho}_J}\Big(\mu_2 p_{2J} + \hat{\lambda} p_{0J}\Big) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_J \frac{g_1}{\hat{\rho}_1}\Big(\mu_J p_{J1} + \hat{\lambda} p_{01}\Big) & \hat{\rho}_J \frac{g_2}{\hat{\rho}_2}\Big(\mu_J p_{J2} + \hat{\lambda} p_{02}\Big) & \ldots & g_J\Big(-\mu_J + \mu_J p_{JJ} + \hat{\lambda} p_{0J}\Big) \end{pmatrix}.$$

Hence, the matrix S is similar to A and therefore A, S and S^T have the same eigenvalues. Using (2.1), it follows that

$$-\mu_i \hat{\rho}_i + \sum_{j=1}^{J} \hat{\rho}_j (\mu_j p_{ji} + \hat{\lambda} p_{0i}) = \hat{\lambda}(-\gamma_i + p_{0i} + \sum_{j=1}^{J} \gamma_j p_{ji}) = 0, \quad i = 1, \ldots, J. \quad (2.20)$$

Hence, the sum of each row in S^T (sum of each column in S) is equal to 0, and the off-diagonal elements in S^T are all positive. This implies that the matrix S^T is a generator corresponding to a finite-state continuous-time Markov chain. This Markov chain is irreducible, as will be shown at the end of the proof, and hence it has a unique equilibrium distribution $\vec{\eta}$, i.e., $\vec{\eta} S^T = \vec{0}$ and $\sum_{j=1}^{J} \eta_j = 1$. This implies that 0 is an eigenvalue with multiplicity 1 of the matrix S^T, and, cf. [9, Proposition 6.2], the real parts of all other eigenvalues are strictly negative. Since the eigenvalues of A and S^T coincide, the same holds for the matrix A. The eigenvector of A corresponding to the eigenvalue 0 is given by $\vec{s}^* = D^{-1}\vec{\eta}$, since $A\vec{s}^* = D^{-1}DAD^{-1}\vec{\eta} = D^{-1}S\vec{\eta} = \vec{0}$.

It remains to be shown that the Markov chain corresponding to the generator S^T is irreducible. Note that, since $\gamma_i > 0$, also $\hat{\rho}_i > 0$ for all $i = 1, \ldots, J$. Let

$$\mathcal{J}_0 := \{j = 1, \ldots, J : p_{0j} > 0\},$$

denote the non-empty set of types that receive external arrivals. Let

$$\mathcal{J}_n := \{j = 1, \ldots, J : \text{ there exist } j_0, \ldots, j_{n-1} \text{ with } p_{0j_0} \cdot p_{j_0 j_1} \cdot \ldots \cdot p_{j_{n-1} j} > 0\},$$

$n = 1, \ldots, J-1$, denote the set of types such that there is a strictly positive probability that a customer becomes of this type after n steps. Since $J < \infty$ and eventually all types are observed, we have that $\cup_{n=0}^{J-1} \mathcal{J}_n = \{1, \ldots, J\}$.

Now consider the $J \times J$ matrix S^T. If $j \in \mathcal{J}_0$, then the (j,i)-th element of S^T, $\hat{\rho}_i \frac{g_j}{\hat{\rho}_j}(\mu_i p_{ij} + \hat{\lambda} p_{0j})$, is strictly positive for all $i \neq j$. Thus, in the Markov chain corresponding to the generator S^T, from any state in $\mathcal{J}_0$ one can reach all other states. In order to prove irreducibility it is now sufficient to show that from any state in $\{1, \ldots, J\} \backslash \mathcal{J}_0$, some state in $\mathcal{J}_0$ can be reached.

Assume $j \in \mathcal{J}_1$. By definition, there exists an $i \in \mathcal{J}_0$ such that $p_{ij} > 0$. Hence, the (j,i)-th element of the matrix S^T, $\hat{\rho}_i \frac{g_j}{\hat{\rho}_j}(\mu_i p_{ij} + \hat{\lambda} p_{0j})$, is strictly positive. This implies that from every state in $\mathcal{J}_1$, a state in $\mathcal{J}_0$ can be reached. Now consider a state $j \in \mathcal{J}_2$. By definition, there exists a state $i \in \mathcal{J}_1$ such that $p_{ij} > 0$. Similarly to the previous case, this implies that the (j,i)-th element of S^T is strictly positive, and we can conclude that from every state in $\mathcal{J}_2$, a state in $\mathcal{J}_1$ can be reached. Proceeding along these lines, it can be shown that from every state in $\mathcal{J}_n$, a state in $\mathcal{J}_{n-1}$ can be reached, $n = 1, \ldots, J-1$. Since $\cup_{n=0}^{J-1}\mathcal{J}_n = \{1, \ldots, J\}$, we obtain that every state outside $\mathcal{J}_0$ can reach a state in the set $\mathcal{J}_0$, which concludes the proof. □

The following lemma shows that the generating function $\hat{r}(\vec{s})$ is constant on H_c.

Lemma 2.4.2. *For any $c > 0$, the function $\hat{r}(\vec{s})$ is constant on H_c.*

Proof: From (2.20) it follows that

$$\begin{aligned}\sum_{i=1}^{J} \frac{\hat{\rho}_i}{g_i} \cdot F_i(\vec{s}) &= \sum_{i=1}^{J} \hat{\rho}_i \cdot \left(\mu_i(-s_i + \sum_{j=1}^{J} p_{ij} s_j) + \hat{\lambda} \sum_{j=1}^{J} p_{0j} s_j \right) \\ &= \sum_{i=1}^{J} (-\mu_i \hat{\rho}_i + \sum_{j=1}^{J} \hat{\rho}_j (\mu_j p_{ji} + \hat{\lambda} p_{0i})) s_i = 0.\end{aligned}$$

This implies that for all $\vec{s} \in H_c$, the vector $\vec{F}(\vec{s})$ is parallel to the hyperplane H_c. Since $\vec{F}$ is C^1, for each state $\vec{s} \geq \vec{0}$ there exists a unique flow $\vec{f}(u) = (f_1(u), \ldots, f_J(u))$, parametrized by $u \geq 0$, such that

$$\vec{f}(0) = \vec{s} \quad \text{and} \quad \frac{\mathrm{d} f_i(u)}{\mathrm{d} u} = F_i(\vec{f}(u)), \quad \text{for all } i \text{ and } u \geq 0. \tag{2.21}$$

Since $\vec{F}(\vec{s})$ is parallel to H_c for all $\vec{s} \in H_c$, when started in H_c, the flow $\vec{f}(u)$ will stay in H_c. Another important property of this flow $\vec{f}(u)$ is that

$$\frac{\mathrm{d}\hat{r}(\vec{f}(u))}{\mathrm{d}u} = \sum_{i=1}^{J} \frac{\mathrm{d} f_i(u)}{\mathrm{d}u} \cdot \left. \frac{\partial \hat{r}(\vec{s})}{\partial s_i} \right|_{\vec{s}=\vec{f}(u)} = 0,$$

which follows from the chain rule, Lemma 2.3.2, and equation (2.21). Hence, along each flow $\vec{f}(u)$, which lies in H_c, the function $\hat{r}(\vec{f}(u))$ is constant. We will now show that each flow in H_c converges to a certain point $c \cdot \vec{s}^* \geq \vec{0}$ as $u \to \infty$.

Relation (2.21) can be written as $\vec{f}(0) = \vec{s}$ and $\vec{f}'(u) = A\vec{f}(u)$, with A as defined in Lemma 2.4.1, see (2.17). In Lemma 2.4.1 it is proved that one eigenvalue of A is 0 with eigenvector $\vec{s}^* \geq \vec{0}$, $\vec{s}^* \in H_1$, and all the other eigenvalues have a strictly negative real part. Hence, the solution of $\vec{f}'(u) = A\vec{f}(u)$ with $\vec{f}(0) \in H_c$ can be written as $\vec{f}(u) = c \cdot \vec{s}^* + \vec{g}(u)$, where $\lim_{u\to\infty} \vec{g}(u) = \vec{0}$ and $\vec{s}^* \geq \vec{0}$. This implies that all the flows in the hyperplane H_c converge to one common point $c \cdot \vec{s}^* \geq \vec{0}$.

Since the continuous function $\hat{r}(\vec{s})$ is constant along one flow, and all flows in the hyperplane H_c converge to $c \cdot \vec{s}^* \in H_c$, we obtain that the function $\hat{r}(\vec{s})$ is constant on H_c. □

2.4.2 Determining the common factor

In the previous section we showed that $(\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_J) \overset{d}{=} (\frac{\hat{\rho}_1}{g_1}, \frac{\hat{\rho}_2}{g_2}, \dots, \frac{\hat{\rho}_J}{g_J}) \cdot X$, with X some random variable. In this section we determine the distribution of X. In order to do so, we consider the total workload in the network, denoted by W. When scaled with $(1-\rho)$ the total workload has a proper distribution as $\rho \uparrow 1$, see [76]:

$$(1-\rho)W \xrightarrow{d} \hat{W},$$

where $\hat{W}$ is exponentially distributed with mean

$$\mathbb{E}(\hat{W}) = \sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j). \tag{2.22}$$

The total workload can be represented as

$$W = \sum_{j=1}^{J} \sum_{h=1}^{Q_j} R_{j,h},$$

with $R_{j,h}$ the residual service requirement of the h-th type-j customer. Note that the remaining service requirements of all customers in phase j are i.i.d. and have the same phase-type distribution independent of $\vec{Q}$, more precisely, $R_{j,h} \overset{d}{=} R_j$ for all h. Hence,

$$\begin{aligned} \mathbb{E}(\mathrm{e}^{-sW}) &= \mathbb{E}(\mathrm{e}^{-s\sum_{j=1}^{J}\sum_{h=1}^{Q_j} R_{j,h}}) = \mathbb{E}(\prod_{j=1}^{J} \mathbb{E}(\mathrm{e}^{-s\sum_{h=1}^{Q_j} R_{j,h}} | \vec{Q})) \\ &= \mathbb{E}(\prod_{j=1}^{J} (\mathbb{E}(\mathrm{e}^{-sR_j}))^{Q_j}) = \mathbb{E}(\mathrm{e}^{\sum_{j=1}^{J} Q_j \ln(\mathbb{E}(\mathrm{e}^{-sR_j}))}), \end{aligned}$$

for $s > 0$. For the scaled workload we can therefore write

$$\begin{aligned} \mathbb{E}(\mathrm{e}^{-s\hat{W}}) &= \lim_{\rho\uparrow 1} \mathbb{E}(\mathrm{e}^{-(1-\rho)sW}) = \lim_{\rho\uparrow 1} \mathbb{E}(\mathrm{e}^{\sum_{j=1}^{J} \frac{\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))}{(1-\rho)s}(1-\rho)sQ_j}) \\ &= \mathbb{E}(\mathrm{e}^{-s\sum_{j=1}^{J} \mathbb{E}(R_j)\hat{Q}_j}), \end{aligned} \tag{2.23}$$

where in the last step we used that $\mathrm{e}^{\sum_{j=1}^{J} \frac{\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))}{(1-\rho)s}(1-\rho)sQ_j}$ is bounded by 1 and converges in distribution to $\mathrm{e}^{-s\sum_{j=1}^{J} \mathbb{E}(R_j)\hat{Q}_j}$. The latter follows from $\frac{\ln(\mathbb{E}(\mathrm{e}^{-(1-\rho)sR_j}))}{(1-\rho)s}$ $\to -\mathbb{E}(R_j)$, as $\rho \uparrow 1$. From (2.23) we obtain that

$$\hat{W} \overset{d}{=} \sum_{j=1}^{J} \mathbb{E}(R_j)\hat{Q}_j,$$

and together with (2.19) this gives

$$\hat{W} \overset{d}{=} X \cdot \sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \mathbb{E}(R_j). \tag{2.24}$$

Since $\hat{W}$ is exponentially distributed, the same is true for X. Taking expectations in (2.24), from (2.22) we obtain

$$\mathbb{E}(X) = \frac{\sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j)}{\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \mathbb{E}(R_j)},$$

which concludes the proof of Proposition 2.1.1.

2.5 Size-based scheduling

Allowing the relative service weights of customers to change over time as they acquire service, opens up a way to implement size-based scheduling by assigning relatively high weights in service phases that are more likely to lead to a quick service completion. In this section we investigate how the choice of the weights influences the performance of the system. With each type of customers we associate a cost $c_j \geq 0$, $j = 1, \ldots, J$. As performance measure we take the holding cost $\sum_{j=1}^{J} c_j Q_j$.

Recall that we consider the general Markovian framework where type-j customers have weight g_j. In this section we will write $Q_j^{(g)}$ ($\hat{Q}_j^{(g)}$) instead of Q_j ($\hat{Q}_j$) to emphasize the dependence on the weights $g_1, \ldots, g_J$. From Proposition 2.1.1 we obtain that the scaled holding cost, $(1-\rho)\sum_{j=1}^{J} c_j Q_j^{(g)}$, converges in distribution to an exponentially distributed random variable with mean

$$\sum_{j=1}^{J} c_j \mathbb{E}(\hat{Q}_j^{(g)}) = \frac{\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \cdot c_j}{\sum_{j=1}^{J} \frac{\hat{\rho}_j}{g_j} \cdot \mathbb{E}(R_j)} \cdot \sum_{j=1}^{J} \hat{\rho}_j \mathbb{E}(R_j), \tag{2.25}$$

as $\rho \uparrow 1$. Using this expression, we obtain the following monotonicity result in the heavy-traffic regime: The holding cost decreases “stochastically” as more preference is given to customers of types with a large value of $\frac{c_i}{\mathbb{E}(R_i)}$.

Proposition 2.5.1. *Consider the general Markovian framework and consider two policies with weights $(g_1, \ldots, g_J)$ and $(\tilde{g}_1, \ldots, \tilde{g}_J)$, respectively. Let $c_j \geq 0$, $j = 1, \ldots, J$. Without loss of generality we assume that the types are ordered such that $\frac{c_1}{\mathbb{E}(R_1)} \geq \frac{c_2}{\mathbb{E}(R_2)} \geq \ldots \geq \frac{c_J}{\mathbb{E}(R_J)}$.*

If $\frac{g_j}{g_{j+1}} \leq \frac{\tilde{g}_j}{\tilde{g}_{j+1}}$, for all $j = 1, \ldots, J-1$, then

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^{J} c_j Q_j^{(g)} \geq_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{j=1}^{J} c_j Q_j^{(\tilde{g})}.$$

Proof: We have that $(1-\rho)\sum_{j=1}^{J} c_j Q_j^{(g)}$ converges in distribution to an exponentially distributed random variable with mean as stated in (2.25). Hence, it only remains to check that

$$\frac{\sum_{j=1}^{J}\frac{c_j\hat{\rho}_j}{g_j}}{\sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}\mathbb{E}(R_j)} \geq \frac{\sum_{j=1}^{J}\frac{c_j\hat{\rho}_j}{\tilde{g}_j}}{\sum_{j=1}^{J}\frac{\hat{\rho}_j}{\tilde{g}_j}\mathbb{E}(R_j)}.$$

This holds since

$$\begin{aligned}
&\left(\sum_{j=1}^{J}\frac{c_j\hat{\rho}_j}{g_j}\right)\cdot\left(\sum_{j=1}^{J}\frac{\hat{\rho}_j}{\tilde{g}_j}\mathbb{E}(R_j)\right)\\
&= \sum_{j,i:j\neq i}\hat{\rho}_j\hat{\rho}_i\Big(\frac{1}{g_j\tilde{g}_i}c_j\mathbb{E}(R_i)+\frac{1}{g_i\tilde{g}_j}c_i\mathbb{E}(R_j)\Big)+\sum_{j=1}^{J}\hat{\rho}_j^2\frac{1}{g_j\tilde{g}_j}c_j\mathbb{E}(R_j)\\
&\geq \sum_{j,i:j\neq i}\hat{\rho}_j\hat{\rho}_i\Big(\frac{1}{g_i\tilde{g}_j}c_j\mathbb{E}(R_i)+\frac{1}{g_j\tilde{g}_i}c_i\mathbb{E}(R_j)\Big)+\sum_{j=1}^{J}\hat{\rho}_j^2\frac{1}{g_j\tilde{g}_j}c_j\mathbb{E}(R_j)\\
&= \left(\sum_{j=1}^{J}\frac{c_j\hat{\rho}_j}{\tilde{g}_j}\right)\cdot\left(\sum_{j=1}^{J}\frac{\hat{\rho}_j}{g_j}\mathbb{E}(R_j)\right).
\end{aligned}$$

Here we used that $c_i\mathbb{E}(R_j)(\frac{1}{g_i\tilde{g}_j}-\frac{1}{g_j\tilde{g}_i}) \geq c_j\mathbb{E}(R_i)(\frac{1}{g_i\tilde{g}_j}-\frac{1}{g_j\tilde{g}_i})$, which follows from the fact that $\frac{g_i}{g_j}\leq\frac{\tilde{g}_i}{\tilde{g}_j}$ and $\frac{c_i}{\mathbb{E}(R_i)}\geq\frac{c_j}{\mathbb{E}(R_j)}$, for $i\leq j$. □

As described in Section 1.3.3, the $c\mu$-rule minimizes the mean holding cost in an (i) M/G/1-queue among all non-preemptive policies as well as in an (ii) G/M/1-queue among all preemptive non-anticipating policies. In both systems the *expected remaining service requirement* of a class-k customer at a scheduling decision epoch is $1/\mu_k$. Hence, the $c\mu$-rule gives priority according to the cost c_k divided by the expected remaining service requirement of a class-k customer. Proposition 2.5.1 can be seen as an extension of the $c\mu$-rule for DPS-based policies in the heavy-traffic regime: the performance improves as larger weights are assigned according to the values of $\frac{c_j}{\mathbb{E}(R_j)}$, $j=1,\ldots,J$. In particular, we obtain that a policy that gives preemptive priority to type $i=\arg\max_{j=1,\ldots,J}\frac{c_j}{\mathbb{E}(R_j)}$ minimizes the scaled holding cost in heavy traffic among all DPS-based policies.

2.6 The standard DPS queue in heavy traffic

In this section we specialize the results obtained so far to the standard DPS queue with phase-type distributed service requirements. In order to show how this queue fits into the Markovian framework of Section 2.1, let us give a brief description of the standard DPS queue.

We consider a single-server system with capacity one and Poisson arrivals with rate λ. With probability p_k an arrival is a class-k customer. Class-k customers have phase-type distributed service requirements, B_k, with a finite number of phases. In particular, this implies that the second moment of B_k is finite. Let

$$\varrho_k := \lambda p_k \mathbb{E}(B_k)$$

be the load associated with class-k customers. The capacity is shared among the customers of the various classes in accordance with the DPS policy. When there are n_k class-k customers present in the system, $k = 1, \dots, K$, each class-k customer is served at rate

$$\frac{w_k}{\sum_{l=1}^K w_l n_l},$$

where w_k is the weight associated with class k. It is important to note that the weight for a class-k customer is *independent* of the current phase of its service requirement. Denote by N_k the number of class-k customers in the DPS queue in steady-state.

The DPS queue with phase-type distributed service requirements fits as follows into the Markovian framework as described in Section 2.1. Within each customer *class* of the DPS queue, we distinguish between customers residing in different service phases, and represent them in the general framework as different customer *types*. Denoting the number of phases of the class-k phase-type distribution with J_k, the total number of types is $J := \sum\limits_{k=1}^{K} J_k$. With slight abuse of terminology, we also refer to a class-k customer in the j^{th} service phase as being of type $\sum_{l=1}^{k-1} J_l + j$. We use $k(j)$ to denote the customer class to which type-j customers belong. If types i and j belong to the same customer class, then they are associated the same weight, i.e., $g_i = g_j = w_{k(j)}$ when $k(i) = k(j)$. The p_{0j} in the general framework is taken such that for $l = k(j)$, p_{0j}/p_l is the probability that a class-l customer starts with service phase j. In the DPS queue, no transitions are possible between types belonging to different customer classes, hence for the general framework this implies that $p_{ij} = 0$ if $k(i) \neq k(j)$. If a class-$k(i)$ customer finishes phase i, then p_{ij} is the probability that it continues in phase j (with $k(i) = k(j)$). The number of class-l customers in the DPS model can be written as $N_l = \sum_{j:k(j)=l} Q_j$.

The mean service requirement of a class-l customer may be written as $\mathbb{E}(B_l) = \sum_{j:k(j)=l} \frac{p_{0j}}{p_l} \mathbb{E}(R_j)$. Hence, the load in class l can be expressed by

$$\varrho_l = \lambda p_l \mathbb{E}(B_l) = \lambda \sum_{j:k(j)=l} p_{0j} \mathbb{E}(R_j). \tag{2.26}$$

For the DPS queue, the set of equations as given in (2.1) simplifies: per class there is a set of equations that can be solved independently. For class l, the corresponding γ_i's can be found from the following set of equations:

$$\gamma_i = p_{0i} + \sum_{j:k(j)=l} \gamma_j p_{ji}, \quad \text{for all } \ i \ \text{ s.t. } \ k(i) = l.$$

Applying the same reasoning as we followed to obtain equation (2.2), it follows that an equivalent representation of ϱ_l is

$$\varrho_l = \lambda \sum_{j:k(j)=l} \frac{\gamma_j}{\mu_j} = \sum_{j:k(j)=l} \rho_j. \tag{2.27}$$

Note that the total load in the DPS queue equals $\sum_{l=1}^K \varrho_l = \sum_{l=1}^K \sum_{j:k(j)=l} \rho_j =: \rho$, i.e., it coincides indeed with the total load in the general framework.

Before proceeding with the main result of this section, we first give expressions for the forward recurrence time of the service requirements. For class l, we denote this random variable by B_l^{fwd}. From renewal theory we know that the associated distribution is

$$\mathbb{P}(B_l^{fwd} \leq x) := \frac{1}{\mathbb{E}B_l} \int_0^x \mathbb{P}(B_l > y) dy, \tag{2.28}$$

and hence $\mathbb{E}(B_l^{fwd}) = \frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)}$. Alternatively we can write

$$\mathbb{P}(B_l^{fwd} \leq x) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{P}(R_j \leq x), \tag{2.29}$$

see [9, Chapter III, Corollary 5.3]. Intuitively relation (2.29) can be explained as follows: Note that $\frac{\gamma_j}{p_l}$ represents the expected number of visits to phase j during the lifetime of the random variable B_l, with $k(j) = l$. As a consequence, $\gamma_j/(p_l \mu_j)$ is the expected time spent in phase j. Thus, with probability

$$\frac{\frac{\gamma_j}{p_l \mu_j}}{\sum_{i:k(i)=l} \frac{\gamma_i}{p_l \mu_i}} = \frac{\rho_j}{\sum_{i:k(i)=l} \rho_i} = \frac{\rho_j}{\varrho_l},$$

the residual life time equals the residual service requirement starting in phase j, and this gives relation (2.29). Combining (2.28) and (2.29), we obtain that the mean forward recurrence time of B_l satisfies

$$\frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \mathbb{E}(B_l^{fwd}) = \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \cdot \mathbb{E}(R_j). \tag{2.30}$$

We now show the state-space collapse for the standard DPS queue with phase-type distributed service requirements. When passing $\rho \to 1$ as described in Section 2.1, we actually fix the service requirement distributions and the class probabilities p_k, while increasing the arrival rate. In particular, the heavy-traffic scaling as considered in Section 2.1, $\lambda \uparrow \hat{\lambda} = \left(\vec{p}_0^T (I-P)^{-1} \vec{m}\right)^{-1}$, is equivalent with $\lambda \uparrow (\sum_l p_l \mathbb{E}(B_l))^{-1}$, since $\sum_{l=1}^K p_l \mathbb{E}(B_l) = \sum_{j=1}^J p_{0j} \mathbb{E}(R_j) = \vec{p}_0^T (I-P)^{-1} \vec{m}$. We denote the limiting loads of all classes by $\hat{\varrho}_l = \hat{\lambda} p_l \mathbb{E}(B_l)$, $l = 1, \ldots, K$ (or equivalently, $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$).

Proposition 2.6.1. *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights $w_1, \ldots, w_K$. When scaled with $1-\rho$, the queue length vector has a proper distribution as $\rho \to 1$,*

$$(1-\rho)(N_1, N_2, \ldots, N_K) \xrightarrow{d} (\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_K) \overset{d}{=} X \cdot (\frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \ldots, \frac{\hat{\varrho}_K}{w_K}), \qquad (2.31)$$

where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean

$$\mathbb{E}(X) = \frac{\sum_k p_k \mathbb{E}(B_k^2)}{\sum_k p_k \mathbb{E}(B_k^2)/w_k}, \qquad (2.32)$$

which is equal to 1 when $w_k = 1$ for all k, i.e., in the case of a standard PS queue.

Remark 2.6.2. In the case of exponentially distributed service requirements, in [68] a related result is proved. The authors consider a sequence of systems indexed by r such that $\varrho_k^r \to \hat{\varrho}_k$, $\rho^r = \sum_{k=1}^K \varrho_k^r \uparrow 1$, and $\sqrt{r}(1-\rho^r) \to 1$, as $r \to \infty$, and obtain that $(1-\rho^r)N^r(rt)$ converges in distribution to

$$\frac{\hat{W}(t)}{\sum_{k=1}^K \frac{\hat{\varrho}_k}{w_k \mu_k}} \cdot (\frac{\hat{\varrho}_1}{w_1}, \ldots, \frac{\hat{\varrho}_K}{w_K}), \qquad (2.33)$$

with $\hat{W}(t)$ the diffusion-scaled workload process, being equal to a reflected Brownian motion with negative drift. The stationary distribution of the latter process is exponential with mean $\sum_{k=1}^K \frac{\hat{\varrho}_k}{\mu_k}$, see also Section 1.6.4. Hence, for exponentially distributed service requirements, the stationary limit of (2.33) coincides with the heavy-traffic limit of the steady-state queue lengths (2.31) as derived in Proposition 2.6.1. Interestingly, this shows that the heavy-traffic limit and the steady-state limit commute in the case of exponentially distributed service requirements.

Proof of Proposition 2.6.1: Recall that the DPS queue with phase-type distributed service requirements is a special case of the general framework of Section 2.1 when the parameters are chosen as described in the beginning of this section. In particular, recall that $g_i = g_j = w_l$ when $k(i) = k(j) = l$. Since $N_l = \sum_{j:k(j)=l} Q_j$, $\hat{\varrho}_l = \sum_{j:k(j)=l} \hat{\rho}_j$ (see (2.27)), and since for the general framework relation (2.4) holds, relation (2.31) follows directly where X is an exponentially distributed random variable with mean as given in (2.5). We are left with showing that (2.5) reduces to (2.32).

From (2.26) and (2.30), and since type-j customers belong to class $k(j)$ and have weight $g_j = w_{k(j)}$, we obtain that

$$\sum_{j=1}^J \frac{\rho_j}{g_j} \mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l} \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \mathbb{E}(R_j) = \sum_{l=1}^K \frac{\varrho_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^K \frac{\lambda p_l}{w_l} \frac{\mathbb{E}(B_l^2)}{2}. \qquad (2.34)$$

Similarly, we have that

$$\sum_{j=1}^{J} \rho_j \mathbb{E}(R_j) = \sum_{l=1}^{K} \varrho_l \sum_{j:k(j)=l} \frac{\rho_j}{\varrho_l} \mathbb{E}(R_j) = \sum_{l=1}^{K} \varrho_l \frac{\mathbb{E}(B_l^2)}{2\mathbb{E}(B_l)} = \sum_{l=1}^{K} \lambda p_l \frac{\mathbb{E}(B_l^2)}{2}. \qquad (2.35)$$

Obviously, equations (2.34) and (2.35) remain valid in heavy traffic. Equation (2.32) follows after substituting (2.34) and (2.35) into (2.5). □

Note that, although the limiting distribution depends on the second moments of the service requirement distributions through $\mathbb{E}(X)$, the impact of the second moment on $\mathbb{E}(X)$ is uniformly bounded, and in particular

$$\min_k w_k \leq \mathbb{E}(X) \leq \max_k w_k,$$

cf. [2]. Similar partial insensitivity results have also been proved for the mean sojourn time conditioned on the service requirement, [12], and the tail index of the sojourn time distribution, [36].

The state-space collapse as demonstrated above, allows us to show further interesting properties for the DPS queue. In Section 2.6.1 we obtain heavy-traffic results for the residual service requirements of the customers in the various classes. In Section 2.6.2, monotonicity in the weights of the standard DPS queue is investigated.

2.6.1 Residual service requirements

The distribution of the residual service requirement of a customer, without having knowledge on the current phase of its service requirement, depends on the used scheduling policy. For example, in a FCFS queue the residual service requirement for customers waiting to be served is given by their original service requirement. In case of a standard PS queue, the residual service requirements are independent random variables distributed according to the forward recurrence times of the service requirements. Given that there are n_k class-k customers in the system, let $B_{k,h}^r$ denote the remaining service requirement of the h-th class-k customer, $k = 1, \ldots, K$, $h = 1, \ldots, n_k$. The following result is known for PS:

$$\begin{aligned} &\mathbb{P}(B_{k,h}^r \leq x_{k,h},\ N_k = n_k, k = 1, \ldots, K, h = 1, \ldots, n_k) \\ &= \mathbb{P}(N_k = n_k, k = 1, \ldots, K) \prod_{k=1}^{K} \prod_{h=1}^{n_k} \mathbb{P}(B_k^{fwd} \leq x_{k,h}), \end{aligned}$$

with $x_{k,h} \geq 0$, and $\mathbb{P}(N_k = n_k, k = 1, \ldots, K)$ as given in (1.3). In this section we show that in a heavy-traffic setting a similar result holds for the DPS queue.

Obviously, in the heavy-traffic limit, there will be an infinite number of customers present in the system. Therefore, we concentrate on the first $y_k < \infty$ class-k customers, $k = 1, \ldots, K$. In the following proposition we show that the scaled number of customers in the various classes and the remaining service requirements of any finite subset of customers are independent in a heavy-traffic setting. In particular, the remaining service requirement of a class-k customer is distributed according to the forward recurrence time of its service requirement B_k. It will be convenient to set $B^r_{k,h} = 0$ whenever $h > N_k$, $k = 1, \ldots, K$.

Proposition 2.6.3. *Assume phase-type distributed service requirements, and consider a standard DPS queue with weights $w_1, \ldots, w_K$. Then,*

$$\lim_{\rho \uparrow 1} \mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K s_l(1-\rho)N_l - \sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}}\right)$$
$$= \mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K s_l \hat{N}_l}\right) \cdot \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} B_l^{fwd}}\right),$$

for $y_l \in \{0, 1, \ldots\}$ and $s_{l,h}, s_l > 0$, $l = 1, \ldots, K$, $h = 1, \ldots, y_l$.

Recall that $(\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_K) \stackrel{d}{=} X \cdot \left(\frac{\hat{\varrho}_1}{w_1}, \frac{\hat{\varrho}_2}{w_2}, \ldots, \frac{\hat{\varrho}_K}{w_K}\right)$, where X is exponentially distributed with mean $\mathbb{E}(X) = \frac{\sum_{l=1}^K p_l \mathbb{E}(B_l^2)}{\sum_{l=1}^K p_l \mathbb{E}(B_l^2)/w_l}$, cf. Proposition 2.6.1.

Proof of Proposition 2.6.3: It will be convenient to first analyze the conditional expectation $\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{Q}\right)$. In order to do so, we condition on the type of the h-th class-l customer, which we denote by $I_{l,h}$ and takes values in $\{i : k(i) = l\}$. For convenience, if $h > \sum_{j:k(j)=l} Q_j$, then $I_{l,h}$ has no significance. Let $\vec{I} = (I_{1,1}, \ldots, I_{1,y_1}, \ldots, I_{K,1}, \ldots, I_{K,y_K})$, which takes values in the set

$$\mathcal{I} := \{\vec{i} : k(i_{1,1}) = 1, \ldots, k(i_{1,y_1}) = 1, \ \ldots, \ k(i_{K,1}) = K, \ldots, k(i_{K,y_K}) = K\}.$$

Conditioning on the types of the customers, we can write

$$\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{Q}\right) = \sum_{\vec{i} \in \mathcal{I}} \mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{I} = \vec{i}, \ \vec{Q}\right) \cdot \mathbb{P}(\vec{I} = \vec{i} | \vec{Q}). \tag{2.36}$$

Define the random variable Y_l as

$$Y_l := \min(y_l, \sum_{j:k(j)=l} Q_j) = \min(y_l, N_l), \quad l = 1, \ldots, K,$$

and note that $\mathbb{P}(Y_l = y_l) = \mathbb{P}(\sum_{j:k(j)=l} Q_j > y_l) \to 1$, as $\rho \uparrow 1$, cf. Proposition 2.1.1. Since y_l is a deterministic value, this implies convergence of Y_l to y_l in probability. By definition, if the h-th class-l customer is of type $i_{l,h}$, then the corresponding residual service requirement has the same distribution as $R_{i_{l,h}}$, $h = 1, \ldots, y_l$. Hence,

$$\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^K \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{I} = \vec{i}, \ \vec{Q}\right) = \prod_{l=1}^K \prod_{h=1}^{Y_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} R_{i_{l,h}}}\right),$$
$$\to \prod_{l=1}^K \prod_{h=1}^{y_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} R_{i_{l,h}}}\right), \quad \text{as } \rho \uparrow 1, \tag{2.37}$$

where the convergence holds in probability (since the conditional expectation on the left-hand side converges to a deterministic value).

Given the population vector $\vec{Q}$, the first chosen class-l customer is of type i, $k(i) = l$, with probability $\frac{Q_i}{\sum_{j:k(j)=l} Q_j}$. The next chosen class-$l$ customer is of type j, $k(j) = l$, with probability $\frac{Q_j - 1_{(i=j)}}{\sum_{j:k(j)=l} Q_j - 1}$, etc. So we obtain

$$\begin{aligned}
&\mathbb{P}(\vec{I} = \vec{i}|\vec{Q}) \\
&= \frac{Q_{i_{1,1}}}{\sum_{j:k(j)=1} Q_j} \cdot \frac{Q_{i_{1,2}} - \mathbf{1}_{(i_{1,1}=i_{1,2})}}{\sum_{j:k(j)=1} Q_j - 1} \cdot \ldots \cdot \frac{Q_{i_{1,Y_1}} - \sum_{h=1}^{Y_1-1} \mathbf{1}_{(i_{1,h}=i_{1,Y_1})}}{\sum_{j:k(j)=1} Q_j - (Y_1 - 1)} \cdot \ldots \cdot \\
&\quad \frac{Q_{i_{K,1}}}{\sum_{j:k(j)=K} Q_j} \cdot \frac{Q_{i_{K,2}} - \mathbf{1}_{(i_{K,1}=i_{K,2})}}{\sum_{j:k(j)=K} Q_j - 1} \cdot \ldots \cdot \frac{Q_{i_{K,Y_K}} - \sum_{h=1}^{Y_K-1} \mathbf{1}_{(i_{K,h}=i_{K,Y_K})}}{\sum_{j:k(j)=K} Q_j - (Y_K - 1)}.
\end{aligned}$$

The latter converges in probability to

$$\prod_{l=1}^{K} \prod_{h=1}^{y_l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{\varrho}_l}, \quad \text{as} \quad \rho \uparrow 1,$$

where we used that $(1-\rho)(Q_1, \ldots, Q_J) \xrightarrow{d} X \cdot (\hat{\rho}_1/g_1, \ldots, \hat{\rho}_J/g_J)$ (see Proposition 2.1.1), the fact that Y_l converges in probability to y_l, the continuous mapping theorem [27], $g_{i_{l,h}} = w_{k(i_{l,h})} = w_l$, and (2.27). Together with (2.29), (2.36) and (2.37) we now obtain

$$\begin{aligned}
&\mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K} \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{Q}\right) \to \sum_{\vec{i} \in \mathcal{I}} \prod_{l=1}^{K} \prod_{h=1}^{y_l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{\varrho}_l} \cdot \mathbb{E}\left(\mathrm{e}^{-s_{l,h} R_{i_{l,h}}}\right) \\
&= \prod_{l=1}^{K} \prod_{h=1}^{y_l} \sum_{i_{l,h}:k(i_{l,h})=l} \frac{\hat{\rho}_{i_{l,h}}}{\hat{\varrho}_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} R_{i_{l,h}}}\right) = \prod_{l=1}^{K} \prod_{h=1}^{y_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} B_l^{fwd}}\right),
\end{aligned}$$

in probability as $\rho \uparrow 1$. By the law of total expectation we therefore have

$$\begin{aligned}
&\lim_{\rho \uparrow 1} \mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J} s_j (1-\rho) Q_j - \sum_{l=1}^{K} \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}}\right) \\
&= \lim_{\rho \uparrow 1} \mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J} s_j (1-\rho) Q_j} \mathbb{E}\left(\mathrm{e}^{-\sum_{l=1}^{K} \sum_{h=1}^{y_l} s_{l,h} B^r_{l,h}} \middle| \vec{Q}\right)\right) \\
&= \mathbb{E}\left(\mathrm{e}^{-\sum_{j=1}^{J} s_j \hat{Q}_j}\right) \cdot \prod_{l=1}^{K} \prod_{h=1}^{y_l} \mathbb{E}\left(\mathrm{e}^{-s_{l,h} B_l^{fwd}}\right).
\end{aligned} \tag{2.38}$$

The result now follows by setting $s_j = \tilde{s}_{k(j)}$, $j = 1, \ldots, J$, in equation (2.38) and noting that $\sum_{j:k(j)=l} \hat{Q}_j = \hat{N}_l$. □

2.6.2 Monotonicity in the weights

In this section, we investigate how the choice of the weights influences the holding cost for the standard DPS queue. We denote by $d_k \geq 0$ the cost associated with a class-k customer. Note that this is a different setting compared to Section 2.5, where a cost was assigned per *type*. As we will see in the proposition below, the scaled holding cost stochastically decreases when relatively larger weights are assigned to classes according to the values of $d_k/\mathbb{E}(B_k^{fwd})$, $k = 1, \dots, K$. Note that $\mathbb{E}(B_k^{fwd}) = \frac{\mathbb{E}(B_k^2)}{2\mathbb{E}(B_k)} = \frac{1}{2}\left(\frac{Var(B_k)}{\mathbb{E}(B_k)} + \mathbb{E}(B_k)\right)$, hence customers belonging to classes with highly variable service requirement distributions should be given lower priority. In addition, from Proposition 2.6.3 it follows that $\mathbb{E}(B_k^{fwd}) = \mathbb{E}(B_k^r)$ in heavy traffic. This shows a connection with the $c\mu$-rule where priority is given according to the cost d_k divided by the expected residual service requirement of a class-k customer, see also Section 2.5.

Proposition 2.6.4. *Assume phase-type distributed service requirements and consider two standard DPS queues with weights $(w_1, \dots, w_K)$ and $(\tilde{w}_1, \dots, \tilde{w}_K)$. Let $d_k \geq 0$, $k = 1, \dots, K$. Without loss of generality we assume that the classes are ordered such that $d_1/\mathbb{E}(B_1^{fwd}) \geq \dots \geq d_K/\mathbb{E}(B_K^{fwd})$.*

If $\frac{w_k}{w_{k+1}} \leq \frac{\tilde{w}_k}{\tilde{w}_{k+1}}$, for all $k = 1, \dots, K-1$, then

$$\lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^{K} d_k N_k^{DPS(w)} \geq_{st} \lim_{\rho\uparrow 1}(1-\rho)\sum_{k=1}^{K} d_k N_k^{DPS(\tilde{w})}, \tag{2.39}$$

where $N_k^{DPS(w)}$ denotes the number of class-k customers in the DPS queue with weights $w_1, \dots, w_K$.

Proof: From Proposition 2.6.1 we obtain that $(1-\rho)\sum_{k=1}^{K} d_k N_k^{DPS(w)}$ converges in distribution to an exponentially distributed random variable with mean

$$\frac{\sum_{k=1}^{K}\frac{d_k\hat{\varrho}_k}{w_k}}{\sum_k p_k\frac{\mathbb{E}(B_k^2)}{w_k}} \cdot \sum_k p_k\mathbb{E}(B_k^2),$$

hence we need to check that

$$\frac{\sum_{k=1}^{K}\frac{d_k\hat{\varrho}_k}{w_k}}{\sum_{k=1}^{K}\frac{\hat{\varrho}_k}{w_k}\frac{\mathbb{E}(B_k^2)}{\mathbb{E}(B_k)}} \geq \frac{\sum_{k=1}^{K}\frac{d_k\hat{\varrho}_k}{\tilde{w}_k}}{\sum_{k=1}^{K}\frac{\hat{\varrho}_k}{\tilde{w}_k}\frac{\mathbb{E}(B_k^2)}{\mathbb{E}(B_k)}}.$$

This follows using similar arguments as in the proof of Proposition 2.5.1 and noting that $\frac{\mathbb{E}(B_k^2)}{2\mathbb{E}(B_k)} = \mathbb{E}(B_k^{fwd})$. □

Note that the $c\mu$-rule can be obtained in the limit from a DPS policy by letting the ratios w_k/w_{k+1}, $k = 1, \dots, K-1$, all go to ∞. When the service requirements are exponentially distributed, it holds that $d_k/\mathbb{E}(B_k^{fwd}) = d_k\mu_k$, so that the optimality of the $c\mu$-rule in heavy traffic is obtained as a special case of Proposition 2.6.4.

In Section 7.5 we will study monotonicity properties similar to (2.39) for the DPS queue with exponential service requirements *outside* the heavy-traffic setting. For general service requirements, however, monotonicity will not necessarily hold in

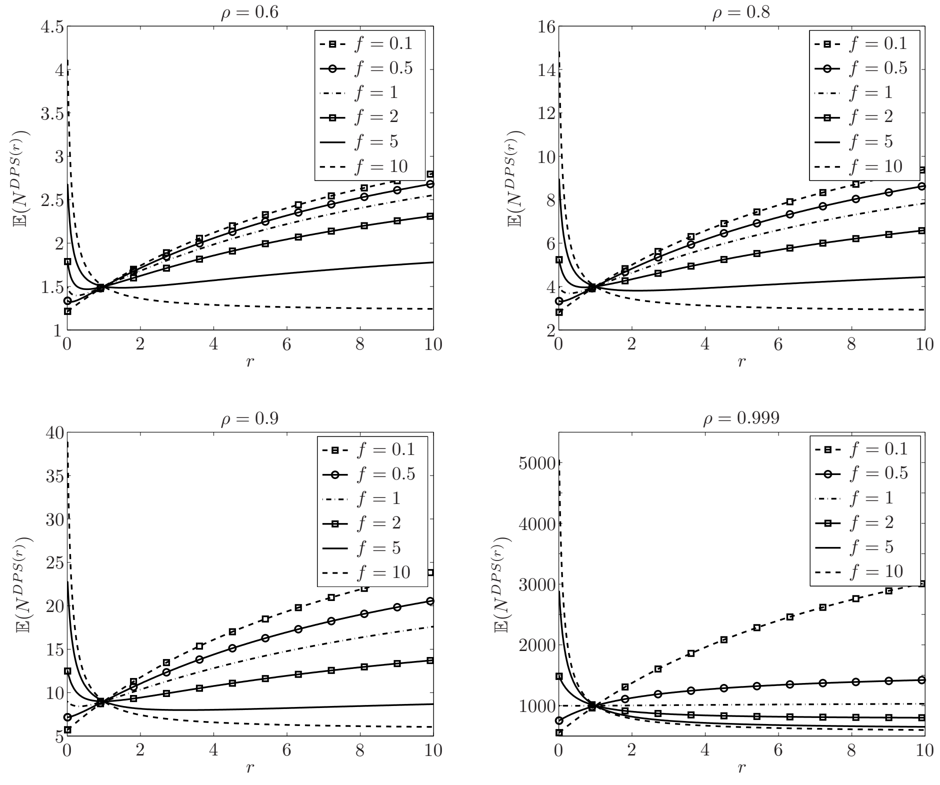


Figure 2.4: Total mean number of customers under a DPS policy with weights $w_1 = 1$ and $w_2 = r$. Class-1 service requirements are hyper-exponentially distributed and class-2 service requirements are exponentially distributed. The load $\rho = \varrho_1 + \varrho_2$ equals 0.6, 0.8, 0.9 and 0.999, respectively.

a moderately-loaded queue. This is further explained in the example below where the behavior of the DPS queue with hyper-exponential service requirements is numerically investigated for several values of the load.

Numerical illustration of Proposition 2.6.4: We consider a DPS queue with two classes. Class-1 customers have hyper-exponentially distributed service requirements, i.e., with a certain probability p a class-1 customer has an exponentially distributed service requirement with mean $1/\mu_{11}$ and with probability $1-p$ it has an exponentially distributed service requirement with mean $1/\mu_{12}$. Class-2 customers have exponentially distributed service requirements with mean $1/\mu_2$. Furthermore, we assume the load is equally distributed between classes 1 and 2, i.e., $\varrho_1 = \varrho_2$. We will be interested in the total number of customers in the system, hence we set

$d_1 = d_2 = 1$. Note that

$$\mathbb{E}(B_1^{fwd}) = \frac{p/\mu_{11}^2 + (1-p)/\mu_{12}^2}{p/\mu_{11} + (1-p)/\mu_{12}} \text{ and } \mathbb{E}(B_2^{fwd}) = 1/\mu_2.$$

Without loss of generality we set $w_1 = 1$ and $w_2 = r$, with $r > 0$. Proposition 2.6.4 states that in a heavily-loaded system the steady-state total number of customers is stochastically increasing in r when $\mathbb{E}(B_1^{fwd}) < \mathbb{E}(B_2^{fwd})$, is constant in r when $\mathbb{E}(B_1^{fwd}) = \mathbb{E}(B_2^{fwd})$, and is stochastically decreasing in r when $\mathbb{E}(B_1^{fwd}) > \mathbb{E}(B_2^{fwd})$. Note that when $r = 1$, the policy reduces to standard PS, and in that case the total mean number of customers is given by $\frac{\rho}{1-\rho}$.

In Figure 2.4 we plot the total mean number of customers as a function of the weight parameter r (denoted by $\mathbb{E}(N^{DPS(r)})$). We consider the case $\mu_{11} = 0.1$, $\mu_{12} = 10$, and $\mu_2 = 1$, while choosing several values for $f := \mathbb{E}(B_1^{fwd})/\mathbb{E}(B_2^{fwd})$. The total mean number of customers is obtained by solving a system of linear equations as described in [51]. For $\rho = \varrho_1 + \varrho_2$ we chose the following values: 0.6, 0.8, 0.9 and 0.999. We see that in the latter case, a heavily-loaded system, the total mean number of customers is increasing when $f < 1$, constant when $f = 1$, and decreasing when $f > 1$. As the total load decreases, the monotonicity no longer necessarily holds. This can be explained as follows. Since $\mu_{11} < \mu_2 < \mu_{12}$, the $c\mu$-rule suggests to prioritize class-1 customers in phase 2, while the class-1 customers in phase 1 should receive lowest priority. In the DPS queue no differentiation can be made between customers residing in different phases. Therefore, the way the weight r affects the mean total number of customers depends on the typical mix of numbers of class-1 customers residing in the two phases. In heavy traffic, this mix is characterized by the loads corresponding to the work of class 1 residing in phases 1 and 2, cf. Proposition 2.1.1, and is hence independent of r. However, away from heavy traffic, this mix may itself be influenced by r, leading to the observed non-monotonic behavior in the figures.

2.7 Concluding remarks

We have studied a multiple-phase network of which the DPS queue with phase-type distributed service requirements is a special case. In our main result we have shown that, in heavy-traffic conditions, the queue length process exhibits a so-called state-space collapse. Based on this result, we found that the DPS model in heavy traffic inherits several well known properties of PS (not necessarily in heavy traffic). For example, in the limit, the (scaled) number of customers present in a DPS model is exponentially distributed, which is the continuous analogue of the geometric queue length distribution of the PS queue. In addition, in a heavy-traffic regime the residual service requirements are independent and distributed according to the forward recurrence times, which is true for PS as well.

We have investigated the performance of a DPS queue in heavy traffic as a function of the weights and showed that the scaled holding cost reduces as customers with smaller weighted residual service requirements get larger weights. In Chapter 7 we will investigate monotonicity properties of the DPS queue outside the heavy-traffic setting.

This chapter can serve as a first step towards analyzing the steady-state queue lengths for the class of weighted α-fair policies, of which DPS is a special case. It would be interesting to investigate whether a heavy-traffic analysis similar to the one performed in this chapter can be carried out for the linear bandwidth-sharing network. This will not be a trivial extension, since the work-conserving property, which was used to derive the exponentially distributed random variable as described in Section 2.4.2, does not carry over to the linear network.

Chapter 3
Stability and size-based scheduling in a linear network

Size-based scheduling policies, such as SRPT and LAS, provide popular mechanisms in single-server systems for improving the overall performance by favoring smaller service requests over larger ones, see Section 1.3.3. In this chapter we examine the merits of size-based scheduling in the linear bandwidth-sharing network. More precisely, the capacity among the various classes is allocated based on the sizes of the service requirements of the users. We explore fundamental stability properties of such size-based scheduling policies.

Due to concurrent resource possession in a linear network, size-based scheduling policies may use the capacity of the nodes inefficiently and persistently leave critical resources underutilized, even when congestion builds up. As a result, SRPT and LAS may unnecessarily cause instability, and will then certainly not yield good performance. Rather than aiming at a general characterization of the stability conditions, in this chapter we focus on various (limiting) regimes of the service requirements. This appears already sufficiently rich to exhibit the instability effects. In particular, we prove that this occurs when the users with long routes have larger service requirements than the ones with shorter routes. For networks with sufficiently many nodes, instability phenomena can in fact arise at arbitrarily low traffic loads. In the opposite regime, where the users with long routes have smaller service requirements than the ones with shorter routes, size-based scheduling strategies are less prone to instability effects.

It is worth drawing a distinction with the situation in queueing networks with feedback where the usual necessary stability conditions are not sufficient either, as first exemplified in Lu & Kumar [87] for priority scheduling and later studied in Bramson [37] for FCFS. In these networks, users visit the various nodes along their route through the network in succession, whereas users in bandwidth-sharing networks require service at all nodes along their route simultaneously. The way in which the queues build up in those feedback networks is also qualitatively different, and typically involves oscillatory behavior.

This chapter is organized as follows. In Section 3.1 we present a model description and discuss some preliminary results. The three subsequent sections examine fundamental stability properties under several size-based scheduling policies. In Section 3.2 this is done for SEPT (Shortest Expected Processing Time) policies, which can be described by simple priority rules. We turn the attention to SRPT and LAS in Sections 3.3 and 3.4, respectively. In Section 3.5 we make some concluding remarks.

3.1 Model and preliminaries

We consider a linear network with L nodes and $L+1$ classes, where class i requires service at node i only, $i = 1, \ldots, L$, while class 0 requires service at all L nodes simultaneously, see Figure 1.2. For convenience, we assume each of the nodes to have a unit service rate. Class-j users arrive according to independent Poisson processes of rate λ_j, and have generally distributed service requirements B_j with distribution function $B_j(x) = \mathbb{P}(B_j < x)$, $j = 0, 1, \ldots, L$. Define the traffic load of class j by $\rho_j := \lambda_j \mathbb{E}(B_j)$. Let $A_j(0,t)$ denote the amount of work from class j that arrives in the interval $(0,t]$, and note that $\lim_{t\to\infty} \frac{A_j(0,t)}{t} = \rho_j$, almost surely (a.s.).

The queue of class-j users is referred to as Q_j, $j = 0, \ldots, L$. In bandwidth-sharing networks, the queue is a purely virtual entity in the sense that the users do not actually reside in physical queues, but rather keep the bulk of the backlogged work stored in their own buffers. Denote by $N_j(t)$ the length of Q_j at time t, i.e., the number of class-j users in the system at time t.

In this chapter we focus on scheduling policies that can be described by sized-based priority ranking. For example, priority is given based on some class parameter (SEPT), remaining service requirement (SRPT), or amount of attained service (LAS). Since class-0 users require simultaneous service at all nodes, without any further arbitration mechanism, capacity can be left unused. Therefore the priority ranking needs to be augmented with a further arbitration mechanism to arrive at the rate allocation to the various classes. We will distinguish between two options: (i) *weak priority*, which means that the capacity in node i that is left unused, is re-allocated to class i; (ii) *strict priority*, which implies that this capacity is left unused.

In this chapter we use the following definitions for stability of a single queue, a node and the complete system.

Definition 3.1.1 (Stability). For a given policy, Q_j, $j = 0, \ldots, L$, is stable when

$$\liminf_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(N_j(t)=0)} \mathrm{d}t > 0, \quad \text{a.s.}$$

Node i, $i = 1, \ldots, L$, is stable when both Q_0 and Q_i are stable. The system is stable when all nodes are stable.

By the Poisson assumption, the process $\vec{N}(t) = (N_0(t), \ldots, N_L(t))$ is a regenerative process with regeneration state $\vec{0}$. A common definition for stability used in the literature, as well as in the chapters of this thesis (with the exception of this chapter), is that the process is stable when it has a finite mean recurrence time to state $\vec{0}$. (In the case of a Markov process this is equivalent to state $\vec{0}$ being positive recurrent.) Note that the process $\vec{N}(t)$ has a finite mean recurrence time to state $\vec{0}$ if and only if

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(\vec{N}(t)=\vec{0})}\mathrm{d}t > 0, \quad \text{a.s,} \tag{3.1}$$

see [138]. A necessary condition for the linear network to have a finite mean recurrence time is $\rho_0+\rho_i \leq 1$ for all $i = 1,\dots,L$, see for example [59]. Note that when the system is stable according to Definition 3.1.1, i.e., $\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(N_j(t)=0)}\mathrm{d}t > 0$, a.s., for all $j = 0,\dots,L$, this does not imply that (3.1) is satisfied or, equivalently, that the process $\vec{N}(t)$ has a finite mean recurrence time. Hence, the stability notion used in this chapter is slightly weaker than the definition that is commonly used. The advantage of using the weaker notion of stability of the system is that stability of the individual queues implies that the system is stable as well. This implication would not go through when using the definition of finite mean recurrence time instead.

For a given policy, let $s_j(t)$ denote the service rate allocated to class j at time t. We define by

$$\underline{\sigma}_j := \liminf_{T\to\infty} \frac{1}{T}\int_0^T s_j(t)\mathrm{d}t \quad\text{and}\quad \overline{\sigma}_j := \limsup_{T\to\infty} \frac{1}{T}\int_0^T s_j(t)\mathrm{d}t,$$

the random variables denoting respectively the minimum and maximum long-term average service rate of class j, $j = 0,\dots,L$. We have the following lemma.

Lemma 3.1.2. *It holds that $\underline{\sigma}_j \leq \overline{\sigma}_j \leq \rho_j$, a.s.*
If Q_j is stable, $j = 1,\dots,L$, then $\overline{\sigma}_j = \rho_j$, a.s.

Proof: The statement $\underline{\sigma}_j \leq \overline{\sigma}_j \leq \rho_j$ follows immediately from $\int_0^T s_j(t)\mathrm{d}t \leq A_j(0,T) + W_j(0)$ and the fact that $\lim_{T\to\infty} A_j(0,T)/T = \rho_j$, a.s. Here $W_j(t)$ denotes the workload in class j at time t. The second statement deserves more elaboration. Note that

$$\liminf_{T\to\infty} \frac{W_j(T)}{T} = \liminf_{T\to\infty} \left(\frac{A_j(0,T)}{T} - \frac{1}{T}\int_0^T s_j(t)\mathrm{d}t \right) = \rho_j - \overline{\sigma}_j.$$

Hence, it remains to be shown that if Q_j is stable, then $\liminf_{T\to\infty} W_j(T)/T = 0$ a.s., or equivalently, if $\liminf_{T\to\infty} W_j(T)/T > 0$ with strictly positive probability, then Q_j is unstable.

Assume $\liminf_{T\to\infty} W_j(T)/T > 0$ with strictly positive probability. Hence, with strictly positive probability we have $\lim_{T\to\infty} W_j(T) = \infty$, and there exists a $\bar{T} > 0$ such that $\mathbf{1}_{(W_j(t)>0)} = 1$ for all $t > \bar{T}$. We can conclude that $\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(W_j(t)>0)}\mathrm{d}t = 1$ with strictly positive probability, i.e., Q_j is unstable.
□

For a given policy, it will be convenient to define the function $c_j(t)$ as follows: $c_j(t) := s_j(t)$ when $N_j(t) > 0$, and otherwise the term $c_j(t)$ is defined as the maximal capacity that could have been allocated to class j (if it would have been present) at time t without reducing the rates allocated to other users. We define by

$$\underline{c}_j := \liminf_{T\to\infty} \frac{1}{T}\int_0^T c_j(t)\mathrm{d}t \quad \text{and} \quad \overline{c}_j := \limsup_{T\to\infty} \frac{1}{T}\int_0^T c_j(t)\mathrm{d}t,$$

the random variables denoting respectively the minimum and maximum long-term average of $c_j(t)$. In the next lemma we describe the stability conditions in terms of $\underline{c}_j$ and $\overline{c}_j$. The terms $\underline{c}_j$ and $\overline{c}_j$ depend on the employed scheduling policy. In general, they are difficult to obtain, since they are highly influenced by the interaction with the other classes $i \neq j$.

Lemma 3.1.3. *If $\rho_j < \underline{c}_j$, a.s., then Q_j, $j = 0, \dots, L$, is stable. If Q_j is stable, then $\rho_j \leq \overline{c}_j$, a.s.*

Proof: Note that $c_j(t) \leq 1$, so that $\mathbf{1}_{(N_j(t)=0)} \geq \mathbf{1}_{(N_j(t)=0)}c_j(t) = c_j(t) - s_j(t)$, for all t. Hence, we obtain that

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(N_j(t)=0)}\mathrm{d}t \geq \underline{c}_j - \overline{\sigma}_j, \quad a.s. \tag{3.2}$$

From this it follows that if $\underline{c}_j > \overline{\sigma}_j$, a.s., then Q_j is stable. Since $\overline{\sigma}_j \leq \rho_j$, the first statement in the lemma is proved.

The second statement in the lemma follows from the fact that $s_j(t) \leq c_j(t)$, and hence, $\overline{\sigma}_j \leq \overline{c}_j$, a.s., together with the fact that $\overline{\sigma}_j = \rho_j$ if Q_j is stable. □

Obviously, $\rho_j < 1$ is a necessary condition for Q_j to be stable, $j = 0, \dots, L$. In the following lemma, useful *sufficient* stability conditions are presented under certain conditions on the policies.

Lemma 3.1.4. *(i) A sufficient condition for stability of Q_i, $i = 1, \dots, L$, is $\rho_0+\rho_i < 1$, provided that under the employed policy node i operates at the full service rate whenever Q_i is non-empty.*
(ii) A sufficient condition for stability of node i is $\rho_0 + \rho_i < 1$, provided that under the employed policy node i operates at the full service rate whenever Q_0 or Q_i are non-empty.
(iii) A sufficient condition for stability of the system is $\sum_{i=0}^L \rho_i < 1$, provided that under the employed policy at least one of the nodes operates at the full service rate whenever the system is non-empty.

Proof: Statement (ii) follows from the fact that node i behaves as a work-conserving single-server queue with load $\rho_0 + \rho_i$. Statement (iii) follows from the fact that the total workload of classes $0, 1, \dots, L$, is stochastically dominated by that in a work-conserving system where classes $1, \dots, L$, are never served at the same time. Statement (i) deserves more elaboration. When node i operates at the full service rate whenever Q_i is non-empty, we have $\mathbf{1}_{(N_i(t)>0)} \leq s_0(t) + s_i(t)$. Together with Lemma 3.1.2 this implies

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(N_i(t)>0)} \mathrm{d}t \leq \rho_0 + \rho_i, \quad \text{a.s.}$$

Hence, Q_i is stable if $\rho_0 + \rho_i < 1$. □

For the scheduling policies considered in this chapter (SEPT, SRPT and LAS), the third property in Lemma 3.1.4 is always satisfied, while the first property only holds for the variants with weak priority. In the remainder of this chapter, stability conditions for size-based scheduling policies are further investigated. We do not aim at deriving stability conditions in the full setting of general service requirements. Instead, we focus on two particular regimes of the service requirements. Either the class-0 users have, in some sense, relatively large service requirements compared to the class-i users, $i = 1, \ldots, L$, or the class-0 users have relatively small service requirements.

3.2 SEPT scheduling

In preparation for the analysis of SRPT and LAS, we first consider the Shortest Expected Processing Time first (SEPT) policy with preemption. SEPT simply gives preemptive priority to class-i users over class-0 users when $\mathbb{E}(B_i) < \mathbb{E}(B_0)$ and vice versa when $\mathbb{E}(B_i) > \mathbb{E}(B_0)$, $i = 1, \ldots, L$.

3.2.1 Large class-0 users

When $\mathbb{E}(B_0) > \mathbb{E}(B_1), \ldots, \mathbb{E}(B_L)$, i.e., large class-0 users, SEPT scheduling in a linear network corresponds to the policy that gives preemptive priority to classes $1, \ldots, L$ over class 0.

Proposition 3.2.1. *Under the priority rule that gives preemptive priority to classes $1, \ldots, L$ over class 0, Q_i is stable if and only if $\rho_i < 1$, $i = 1, \ldots, L$. In addition, Q_0 is stable if $\rho_0 < \Pi_{i=1}^L (1 - \rho_i)$, and unstable if $\rho_0 > \Pi_{i=1}^L (1 - \rho_i)$.*

Proof: When classes $1 \ldots, L$ are given preemptive priority, class i behaves as in an isolated M/G/1 queue with class i only. Therefore, Q_i is stable if and only if $\rho_i < 1$, $i = 1, \ldots, L$. Let N_i, $i = 1, \ldots, L$, be the random variable with the time-average distribution of $N_i(t)$. Since $N_1, \ldots, N_L$ are independent, we have $\mathbb{P}(N_1 = 0, \ldots, N_L = 0) = \Pi_{i=1}^L (1 - \rho_i)$. Class 0 is served when there are no class-i users present, $i = 1, \ldots, L$. By Lemma 3.1.3, we obtain that if

$$\rho_0 < \mathbb{P}(N_1 = 0, \ldots, N_L = 0) = \Pi_{i=1}^L (1 - \rho_i), \tag{3.3}$$

then Q_0 is stable. In addition, if $\rho_0 > \mathbb{P}(N_1 = 0, \ldots, N_L = 0) = \Pi_{i=1}^L (1 - \rho_i)$, then Q_0 is unstable. □

Note that the above condition is more stringent than the maximum stability condition. In fact, the system can be unstable for arbitrarily low values of ρ_0 if the number of traversed nodes is large. The instability can arise here since this priority policy can leave a substantial portion of the capacity unused, regardless of how large the number of class-0 users is. In Sections 3.3 and 3.4 we show that SRPT and LAS inherit these difficulties.

3.2.2 Small class-0 users

When $\mathbb{E}(B_0) < \mathbb{E}(B_1), \ldots, \mathbb{E}(B_L)$, i.e., small class-0 users, SEPT scheduling in a linear network corresponds to the policy that gives preemptive priority to class 0 over all class-i users, $i = 1, \ldots, L$. Under this priority rule, the system is stable under the maximum stability conditions, see the next proposition.

Proposition 3.2.2. *Under the priority rule that gives preemptive priority to class-0 users, Q_0 is stable if and only if $\rho_0 < 1$. In addition, Q_i is stable if $\rho_0 + \rho_i < 1$, and unstable if $\rho_0 + \rho_i > 1$, $i = 1, \ldots, L$.*

Proof: When class 0 is given preemptive priority, class 0 behaves as in an isolated M/G/1 queue with class 0 only. Therefore, Q_0 is stable if and only if $\rho_0 < 1$. Since $N_0(t)$ is a regenerative process, we have that $\lim_{T\to\infty} N_0(T)/T \to 0$, a.s., when Q_0 is stable. Together with

$$\lim_{T\to\infty} \frac{W_0(T)}{T} = \rho_0 - \lim_{T\to\infty} \frac{1}{T} \int_0^T s_0(t)\mathrm{d}t,$$

this implies $\underline{\sigma}_0 = \overline{\sigma}_0 = \rho_0$. Class-$i$ users, $i = 1, \ldots, L$, behave as in an isolated priority queue with classes 0 and i, hence we have $c_i(t) = 1 - s_0(t)$, so that $\underline{c}_i = \overline{c}_i = 1 - \rho_0$. By Lemma 3.1.3, we then obtain the result. □

3.2.3 Intermediate-size class-0 users

In order to cover the full range of service requirements, we extend the model with class-i' users, $i = 1, \ldots, L$, that require service from node i only, arrive according to a Poisson process of rate $\lambda_{i'}$, and have exponentially distributed service requirements with mean $\mathbb{E}(B_{i'})$, such that $\mathbb{E}(B_i) < \mathbb{E}(B_0) < \mathbb{E}(B_{i'})$ for all $i = 1, \ldots, L$. Denote the traffic load of class i' by $\rho_{i'} := \lambda_{i'}\mathbb{E}(B_{i'})$, $i = 1, \ldots, L$.

Under the SEPT policy, class-j users, $j = 0, 1, \ldots, L$, are not affected by the presence of class-i' users, $i = 1, \ldots, L$. It thus follows from the results in Subsection 3.2.1 that Q_i, $i = 1, \ldots, L$, is stable if $\rho_i < 1$ and that Q_0 is stable if $\rho_0 < \Pi_{i=1}^{L}(1 - \rho_i)$.

In order to establish sufficient stability conditions, it is important to know whether we have weak or strict SEPT. Strict SEPT only allows a class-i' user to be served when there are no class-0 and class-i users in the system. In contrast, weak SEPT also allows a class-i' user to be served when there are class-0 users in the system which are however blocked from service by class-j users, $j \neq i$, and there are no class-i users present.

Weak SEPT

For weak SEPT, class-i' users can be served during the time that Q_i is empty. However, class-0 users may be served during this time as well. Note that $c_{i'}(t) \geq 1 - s_0(t) - s_i(t)$. Hence, it follows from Lemma 3.1.2 that $\underline{c}_{i'} \geq 1 - \overline{\sigma}_0 - \rho_i$. Together with Lemma 3.1.3 we obtain that $\rho_{i'} < 1 - \overline{\sigma}_0 - \rho_i$, $i = 1, \ldots, L$, is a sufficient stability condition for $Q_{i'}$. In order to determine the value of $\overline{\sigma}_0$, we need to distinguish whether Q_0 is stable or not, i.e., whether $\rho_0 < \Pi_{i=1}^L (1 - \rho_i)$ or not. If Q_0 is stable, then $\overline{\sigma}_0 = \rho_0$, and thus the stability condition for $Q_{i'}$ becomes $\rho_0 + \rho_i + \rho_{i'} < 1$. If Q_0 is unstable, then it follows from (3.2) that with strictly positive probability $\underline{c}_0 \leq \overline{\sigma}_0$. Since $s_0(t) \leq c_0(t)$, we also have that $\overline{\sigma}_0 \leq \overline{c}_0$. Class 0 is served when there are no users from classes $1, \ldots, L$ present, hence $\underline{c}_0$ and $\overline{c}_0$ are both equal to $\mathbb{P}(N_1 = 0, \ldots, N_L = 0) = \Pi_{i=1}^L (1 - \rho_i)$ a.s. (see Proposition 3.2.1). Hence, a sufficient stability condition for $Q_{i'}$ takes the form

$$\rho_{i'} < 1 - \Pi_{j=1}^L (1 - \rho_j) - \rho_i = (1 - \rho_i)(1 - \prod_{j \neq i} (1 - \rho_j)).$$

Strict SEPT

Under strict SEPT, class i' is only served when no users of class 0 and class i are present. Hence, by Lemma 3.1.3, a sufficient stability condition for $Q_{i'}$ may be expressed by $\rho_{i'} < \liminf_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_{(N_0(t)=0, N_i(t)=0)} \mathrm{d}t$, a.s., $i = 1, \ldots, L$. In general no tractable expression appears to exist for the latter.

3.3 SRPT scheduling

We turn the attention to SRPT scheduling. A class-0 user receives the total capacity of all nodes whenever it has the smallest remaining service requirement among all users. Otherwise, in case of weak SRPT, the total capacity in node i is given to the class-i user with the smallest remaining service requirement. However, in case of strict SRPT the total capacity in node i is only given to a class-i user when this user has indeed the smallest remaining service requirement among all class-0 and class-i users, and otherwise the capacity is lost. Possible ties (which occur with non-zero probability in case of discrete service requirement distributions) are assumed to be broken at random.

It will be convenient to introduce

$$\rho_j(x) := \lambda_j \mathbb{E}(B_j \mathbf{1}_{(B_j < x)}) = \lambda_j \int_0^{x^-} y \mathrm{d}B_j(y), \tag{3.4}$$

the traffic load of class j when all class-j users of size x or larger are rejected. It is important to note that users of size exactly x are excluded in this definition.

In order to establish exact stability conditions, we need to impose some additional assumptions on the service requirement distributions, as will be done in the next subsections.

3.3.1 Large class-0 users

In this subsection we consider class-0 users with large service requirements, compared to all other classes.

Stability of Q_i, for $i = 1, \ldots, L$

Define $m_j := \inf\{x : B_j(x) > 0\}$ and $M_j := \sup\{x : B_j(x) < 1\}$ as the minimum and maximum values of the class-j service requirements, $j = 0, \ldots, L$. We focus on the case where class 0 has larger service requirements than all classes i, i.e., $m_0 > M_i$, for $i \neq 0$. Thus, a class-0 user can only enter service when there are no class-i users in the system. When a class-0 user is in service and a class-i user arrives, the service is preempted when the remaining service requirement of the class-0 user is larger than that of the arriving class-i user.

Evidently, $\rho_i < 1$ is a necessary condition for stability of Q_i, $i = 1, \ldots, L$, because otherwise Q_i would be unstable even in the absence of any class-0 users. The next proposition shows that for weak SRPT with $m_0 > M_i$ this condition is sufficient as well.

Proposition 3.3.1. *If the policy is weak SRPT and $m_0 > M_i$ for all $i = 1, \ldots, L$, then the condition for stability of Q_i is $\rho_i < 1$.*

Proof: As observed above, the fact that $m_0 > M_i$ implies that class i receives preemptive priority over class 0, unless a class-0 user has a smaller remaining service requirement than all class-i users (so at most M_i) and is being served. In the presence of this class-0 user, it depends on the other classes whether class 0 or class i is being served. When this class-0 user leaves the system, no new class-0 users are taken into service under SRPT as long as class i is present ($m_0 > M_i$). Hence, as long as Q_i remains non-empty, it will be prevented from service for at most a duration M_i, since weak SRPT does not leave any capacity in node i unused when class i is present. It follows that class i behaves as in an isolated queue with class i only and random service interruptions whose total duration during each busy period is bounded by M_i. Lemma 3.3.2 implies that a queue with service interruptions of bounded size in each busy period, is stable for any $\rho_i < 1$. □

Lemma 3.3.2. *Consider an M/G/1 queue with traffic load ρ and with service interruptions. Assume that the total duration of the service interruptions in any contiguous period during which the queue is continuously backlogged is stochastically bounded by a random variable D with $\mathbb{E}(D) < \infty$. If $\rho < 1$, then for any work-conserving policy the queue is stable.*

Proof: Let the random variable BP denote the length of a busy period in an ordinary M/G/1 queue without service interruptions. Obviously, $\mathbb{E}(BP) < \infty$ when $\rho < 1$.

Now consider the queue with service interruptions as described in the lemma. Let the random variable C denote the length of a contiguous period during which the queue is continuously backlogged. With each user we can associate a sub-busy period during which that user is served, as well as users that arrived during that service time (not counting those that arrive when that service time is interrupted), those that arrived during the service of those users and so on. The period C can now be split into the following three components: the service interruptions, the sub-busy period of the user that arrived at an empty system when there is no service interruption (this user may not be present) and the sub-busy periods of the users that arrived during a service interruption. The expected number of users that arrive while the service is interrupted is bounded by $\lambda\mathbb{E}(D)$. We can therefore write

$$\mathbb{E}(C) \leq \mathbb{E}(D) + (1 + \lambda\mathbb{E}(D))\mathbb{E}(BP) < \infty.$$

This implies $\mathbb{P}(N = 0) = \frac{1/\lambda}{1/\lambda+\mathbb{E}(C)} > 0$, which establishes the stability of the queue. □

The next proposition indicates that for strict SRPT the condition $\rho_i < 1$ is in general not sufficient for Q_i to be stable.

Proposition 3.3.3. *If the policy is strict SRPT and $m_0 > M_j$, for all $j = 1, \ldots, L$, then the condition for stability of Q_i is*

$$\rho_i < 1 \text{ and } \rho_j(M_i) < 1 \text{ for all } j \neq 0, i. \tag{3.5}$$

Proof: We first prove that the above condition is sufficient. Since $m_0 > M_i$, class i receives preemptive priority over class 0, and will be entitled to service, unless a class-0 user is present with a smaller remaining service requirement than all class-i users. Although the service of such a class-0 user may repeatedly be interrupted by arriving class-j users, $j = 1, \ldots, L$, the latter users all have service requirements of at most M_i.

Consider a period that starts with an arrival of class i in an empty Q_i, and finishes when Q_i is empty again. Let D_i denote the total amount of time that class i is prevented from service during this period. By Lemma 3.3.2, Q_i is stable for any $\rho_i < 1$ if $\mathbb{E}(D_i) < \infty$. It remains to be shown that $\mathbb{E}(D_i) < \infty$ when $\rho_j(M_i) < 1$ for all $j \neq 0, i$.

Define T_i as the time it takes for a class-0 user with a remaining service requirement of $r_0 = M_i$, to receive the last M_i part of its service. D_i can be bounded from above by T_i, since class i only notices the class-0 user, when $r_0 \leq M_i$. Note that at the moment that the class-0 user is being served and r_0 reaches the level M_i, because of SRPT, it is necessary that there are no other users present with a remaining service requirement smaller than M_i.

Denote by $r_0(t)$ the smallest remaining service requirement of all the class-0 users present at time t. A class-0 user with a remaining service requirement smaller than M_i is being served until a user of size smaller than $r_0(t_1)$ arrives at time t_1 (so it necessarily is of class i, $i = 1, \ldots, L$). This user preempts the class-0 user. The class-0 user can resume its service when all newly arrived users with size not larger than $r_0(t_1)$ have left the system. This period is called a busy period of classes $1, \ldots, L$.

After this busy period the class-0 user can enter service again, until at a certain time t_2 a user arrives with size smaller than $r_0(t_2)$ (such a user is necessarily of class i, $i = 1, \dots, L$). A new busy period starts of class-i users, $i = 1, \dots, L$, with sizes smaller than or equal to $r_0(t_2)$. This pattern repeats itself until the class-0 user has received its complete service and leaves the system.

Note that an upper bound for such a busy period of classes $1, \dots, L$, is obtained when instead we look at the busy period of users from classes $1, \dots, L$ with size smaller than or equal to M_i. Hence, if $\rho_j(M_i) < 1$ for all $j = 1, \dots, L$, then the class-0 user can be served during at least a fraction of time $\Pi_{j=1}^{L}(1 - \rho_j(M_i))$ in the period T_i. Since the class-0 user needs a total of M_i service, we can conclude that $\Pi_{j=1}^{L}(1 - \rho_j(M_i))\mathbb{E}(T_i) \leq M_i$, and hence $\mathbb{E}(T_i) < \infty$. The fact that $\mathbb{E}(D_i) \leq \mathbb{E}(T_i) < \infty$ concludes the proof that (3.5) is sufficient.

It remains to be shown that (3.5) is necessary as well. Clearly, $\rho_i < 1$ is a necessary condition. To show that the second condition is necessary too, suppose it is not satisfied, i.e., $\rho_j(M_i) > 1$ for some $j \neq 0, i$. The following can now happen with positive probability. Suppose at time $t = 0$ the system is empty. A first user arrives at time T_1 and is of class 0 with a service requirement B_0. Define $d_j := \sup\{x : \rho_j(x) \leq 1\}$, hence $d_j < M_i$. There is a d, $d_j < d \leq M_i$, such that there arrive class-j users with sizes in the interval $(d_j, d]$. Now assume at time $T_1 + B_0 - r$, $r \in (d, M_i)$, a second user arrives, which is of class i and has a service requirement smaller than r. Hence, the class-0 user is interrupted from service by this class-i user. Since $\rho_j(d) > 1$ and since we consider SRPT, we can define T^* as the last moment that class j had no users with sizes in the interval $(d_j, d]$. Assume in the interval $(T_1 + B_0 - r, T^*)$ there are always users of class i present with (remaining) service requirement less than or equal to r. Hence, at time T^*, the class-0 user has still size r. After time T^*, there are always class-j users present with service requirements in the interval $(d_j, d]$. Since $d < r$, the class-0 user, which has size r, and therefore also class-i users of size larger than r, will never be served again (strict SRPT). Hence Q_i will grow indefinitely from time T^* onward. The above-described trajectory occurs with positive probability. □

Stability of Q_0

We now turn to the stability of Q_0. To determine sufficient conditions for stability of Q_0, we will consider the network in a limiting regime, obtained by scaling the dynamics of some classes with a common parameter ϵ and passing $\epsilon \downarrow 0$. This technique is usually referred to as analytic perturbation, and has successfully been applied to study steady-state performance as a function of ϵ, as $\epsilon \downarrow 0$, see for instance [6, 43].

We consider a sequence of systems, indexed by ϵ, where the class-i arrival rate in the ϵ-system is $\lambda_i^{(\epsilon)} := \lambda_i/\epsilon$ and the class-i service requirements are distributed as ϵB_i, for $i = 1, \dots, L$. Note that the traffic load of class i in the ϵ-system is $\rho_i^{(\epsilon)} = \frac{\lambda_i}{\epsilon}\epsilon\mathbb{E}(B_i) = \rho_i$, independent of ϵ. Let $\rho_i^{(\epsilon)}(x)$ be the equivalent of (3.4) for the ϵ-system. Furthermore, as $\epsilon \downarrow 0$, the class-i service requirements become extremely small compared to class 0, so we are in the situation of large class-0 users.

In the ϵ-system we make a distinction between class-i users with original size smaller or larger than $\sqrt{\epsilon}$. Denote by $N_i^{\sqrt{\epsilon}}(t)$ the number of class-i users in the ϵ-system with original size smaller than $\sqrt{\epsilon}$ present at time t. We define $A_i^{\sqrt{\epsilon}}$, for $i = 1, \dots, L$, as a period where class-i users with original size smaller than $\sqrt{\epsilon}$ are served in the ϵ-system. Note that in this period the total capacity of node i is allocated to a class-i user with original size smaller than $\sqrt{\epsilon}$. We define $I_i^{\sqrt{\epsilon}}$, for $i = 1, \dots, L$, as a period where class-i users with original size smaller than $\sqrt{\epsilon}$ are not served in the ϵ-system.

It is possible that $N_i^{\sqrt{\epsilon}}(t) > 0$ during a period $I_i^{\sqrt{\epsilon}}$, but that these class-i users are blocked from service by a class-i user with an original size larger than $\sqrt{\epsilon}$ or a class-0 user, both with a remaining service requirement smaller than $\sqrt{\epsilon}$. The latter will occur at most a fraction of order $\sqrt{\epsilon}$ of the time, so class-i users with original size smaller than $\sqrt{\epsilon}$ will receive priority over the other users virtually all the time as $\epsilon \downarrow 0$. Thus, class i restricted to $\sqrt{\epsilon}$ will approximately behave as in an isolated queue with class i restricted to $\sqrt{\epsilon}$ only as $\epsilon \downarrow 0$. Moreover, since

$$\rho_i^{(\epsilon)}(\sqrt{\epsilon}) = \frac{\lambda_i}{\epsilon}\mathbb{E}(\epsilon B_i \mathbf{1}_{(\epsilon B_i < \sqrt{\epsilon})}) = \lambda_i \mathbb{E}(B_i \mathbf{1}_{(B_i < \frac{1}{\sqrt{\epsilon}})}) \to \rho_i$$

and classes i for $i = 1, \dots, L$ will behave roughly independently, this suggests that $\lim_{T\to\infty} \frac{1}{T} E_{[0,T]}(I_1^{\sqrt{\epsilon}}, \dots, I_L^{\sqrt{\epsilon}}) \to \Pi_{i=1}^L (1-\rho_i)$, as $\epsilon \downarrow 0$, as is confirmed by the next proposition. Here $E_{[0,T]}(A)$ denotes the amount of time that the event A occurs during the interval $[0, T]$.

Proposition 3.3.4. *Consider the ϵ-systems under the weak SRPT policy. If $\rho_0 + \rho_i < 1$ for $i = 1, \dots, L$, then*

$$\lim_{\epsilon\downarrow 0} \lim_{T\to\infty} \frac{1}{T} E_{[0,T]}(I_1^{\sqrt{\epsilon}}, \dots, I_L^{\sqrt{\epsilon}}) = \Pi_{i=1}^L (1-\rho_i), \quad a.s.$$

Proof: Let us introduce a reference system with class i only and with the same arrival process and service requirements as in the original system, but where class-i users with sizes larger than $\sqrt{\epsilon}$ are rejected. The number of class-i users in the reference system is denoted by $\hat{N}_i^{\sqrt{\epsilon}}(t)$. Define $\hat{A}_i^{\sqrt{\epsilon}}$ and $\hat{I}_i^{\sqrt{\epsilon}}$ as the active and idle periods of the reference system, respectively.

If $\rho_0 + \rho_i < 1$, $i = 1, \dots, L$, then $Q_1, \dots, Q_L$ are stable in the original system and in the reference system (Lemma 3.1.4 (i)). For the reference system (which is an isolated work-conserving queue) this implies that

$$\lim_{T\to\infty} \frac{1}{T} E_{[0,T]}(\hat{A}_i^{\sqrt{\epsilon}}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(\hat{N}_i^{\sqrt{\epsilon}}(t) > 0)} \mathrm{d}t = \rho_i^{(\epsilon)}(\sqrt{\epsilon}). \tag{3.6}$$

In addition, for the original system we obtain from Lemma 3.1.2 that $\rho_i^{(\epsilon)}(\sqrt{\epsilon}) =$

$\limsup_{T\to\infty} \frac{1}{T} E_{[0,T]}(A_i^{\sqrt{\epsilon}})$. For $i = 1, \ldots, L$, we now have

$$\begin{aligned}\limsup_{T\to\infty} \frac{1}{T} E_{[0,T]}(\hat{I}_i^{\sqrt{\epsilon}}, A_i^{\sqrt{\epsilon}}) &= \limsup_{T\to\infty} \frac{1}{T}(E_{[0,T]}(A_i^{\sqrt{\epsilon}}) - E_{[0,T]}(A_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}})) \\ &\leq \limsup_{T\to\infty} \frac{1}{T} E_{[0,T]}(A_i^{\sqrt{\epsilon}}) - \liminf_{T\to\infty} \frac{1}{T} E_{[0,T]}(A_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}}) \\ &= \lim_{T\to\infty} \frac{1}{T} E_{[0,T]}(\hat{A}_i^{\sqrt{\epsilon}}) - \liminf_{T\to\infty} \frac{1}{T} E_{[0,T]}(A_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}}) \\ &\leq \limsup_{T\to\infty} \frac{1}{T}(E_{[0,T]}(\hat{A}_i^{\sqrt{\epsilon}}) - E_{[0,T]}(A_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}})) \\ &= \limsup_{T\to\infty} \frac{1}{T} E_{[0,T]}(I_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}}). \qquad (3.7)\end{aligned}$$

We proceed to derive an upper bound for the latter expression. Observe that when the reference system is active at time t, i.e., $\hat{N}_i^{\sqrt{\epsilon}}(t) > 0$, it holds that $N_i^{\sqrt{\epsilon}}(t) > 0$, because $N_i^{\sqrt{\epsilon}}(t) \geq \hat{N}_i^{\sqrt{\epsilon}}(t)$. Thus, in order for the event $(I_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}})$ to occur, it must be that $N_i^{\sqrt{\epsilon}}(t) > 0$, and hence there is a class-i user with original size smaller than $\sqrt{\epsilon}$, but this user is not served. As noted earlier, this can only arise when a class-i user with original size larger than $\sqrt{\epsilon}$ or a class-0 user is present, both with a remaining service requirement smaller than $\sqrt{\epsilon}$.

We have the bound $E_{[0,t]}(I_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}}) \leq \sum_{n=1}^{N_{[0,t]}} D_{i,n}^{\sqrt{\epsilon}}$, with $N_{[0,t]}$ denoting the number of class-i users with original size larger than $\sqrt{\epsilon}$ and class-0 users, that are served during the interval $[0, t]$; the index n is used to denote the n-th such user, and $D_{i,n}^{\sqrt{\epsilon}}$ is the amount of time that class-i users with original size smaller than $\sqrt{\epsilon}$ are prevented from service because of user n. For weak SRPT, we have $D_{i,n}^{\sqrt{\epsilon}} \leq \sqrt{\epsilon}$, since no capacity is left unused in the presence of class-i users and user n needs only its last $\sqrt{\epsilon}$ amount of service. Using the strong law of large numbers, we can conclude that

$$\limsup_{T\to\infty} \frac{E_{[0,T]}(I_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}})}{T} \leq \lim_{T\to\infty} \frac{N_{[0,T]}}{T}\sqrt{\epsilon} \leq (\lambda_0 + \lambda_i^{(\epsilon)}\mathbb{P}(B_i^{(\epsilon)} > \sqrt{\epsilon}))\sqrt{\epsilon}, \quad \text{a.s.} \qquad (3.8)$$

Furthermore we have $\lim_{\epsilon\downarrow 0} \lambda_i^{(\epsilon)}\mathbb{P}(B_i^{(\epsilon)} > \sqrt{\epsilon})\sqrt{\epsilon} = \lim_{\epsilon\downarrow 0} \frac{\lambda_i}{\sqrt{\epsilon}}\mathbb{P}(B_i > \frac{1}{\sqrt{\epsilon}})$. It can be shown that when $\mathbb{E}(B_i) < \infty$, this limit equals 0. Hence, by (3.8) we obtain

$$\lim_{\epsilon\downarrow 0} \lim_{T\to\infty} \frac{E_{[0,T]}(I_i^{\sqrt{\epsilon}}, \hat{A}_i^{\sqrt{\epsilon}})}{T} = 0 \quad \text{for all } i = 1, \ldots, L, \quad \text{a.s.} \qquad (3.9)$$

For any converging subsequence T_k of $\frac{E_{[0,T]}(I_1^{\sqrt{\epsilon}}, \ldots, I_L^{\sqrt{\epsilon}})}{T}$ it follows from (3.7) and (3.9) that

$$\lim_{\epsilon\downarrow 0}\lim_{k\to\infty}\frac{E_{[0,T_k]}(I_1^{\sqrt{\epsilon}},\ldots,I_L^{\sqrt{\epsilon}})}{T_k}=\lim_{\epsilon\downarrow 0}\lim_{k\to\infty}\frac{E_{[0,T_k]}(\hat{I}_1^{\sqrt{\epsilon}},\ldots,\hat{I}_L^{\sqrt{\epsilon}})}{T_k}$$
$$-\frac{E_{[0,T_k]}(\hat{I}_1^{\sqrt{\epsilon}},\ldots,\hat{I}_L^{\sqrt{\epsilon}},A_i^{\sqrt{\epsilon}},\text{ for an } i\neq 0)}{T_k}+\frac{E_{[0,T_k]}(I_1^{\sqrt{\epsilon}},\ldots,I_L^{\sqrt{\epsilon}},\hat{A}_i^{\sqrt{\epsilon}},\text{ for an } i\neq 0)}{T_k}$$
$$=\lim_{\epsilon\downarrow 0}\lim_{k\to\infty}\frac{E_{[0,T_k]}(\hat{I}_1^{\sqrt{\epsilon}},\ldots,\hat{I}_L^{\sqrt{\epsilon}})}{T_k}$$
$$=\Pi_{i=1}^{L}(1-\rho_i),\quad\text{a.s.},$$

where in the last step we use that $\hat{I}_1^{\sqrt{\epsilon}},\ldots,\hat{I}_L^{\sqrt{\epsilon}}$ are independent, together with relation (3.6) and $\rho_i^{(\epsilon)}(\sqrt{\epsilon})\to\rho_i$, for $i=1,\ldots,L$. □

From Proposition 3.3.4 we can now derive the stability condition for ϵ small enough in an ϵ-system.

Corollary 3.3.5. *Consider the ϵ-system under the weak SRPT policy. If $\rho_0+\rho_i<1$, for an $i=1,\ldots,L$, then Q_i is stable.*
(i) If in addition $\rho_0<\Pi_{i=1}^{L}(1-\rho_i)$, then there exists an $\bar{\epsilon}$ such that Q_0 is stable in the ϵ-system for every $\epsilon<\bar{\epsilon}$.
(ii) Conversely, if $\rho_0>\Pi_{i=1}^{L}(1-\rho_i)$, then there exists an $\bar{\epsilon}$ such that Q_0 is unstable in the ϵ-system for every $\epsilon<\bar{\epsilon}$.

Proof: Stability of Q_i when $\rho_0+\rho_i<1$ follows from Lemma 3.1.4 (i). The remainder of the proof is concerned with stability of Q_0.

There is at least one node i that can work at full rate on class-i users with service requirement larger than $\sqrt{\epsilon}$ or on class 0, whenever no users of classes $1,\ldots,L$ with service requirement smaller than $\sqrt{\epsilon}$ are being served. Hence, it follows from Lemma 3.1.3 that a sufficient condition for Q_0 to be stable is $\rho_0+\sum_{i=1}^{L}(\rho_i-\rho_i^{(\epsilon)}(\sqrt{\epsilon}))<\liminf_{T\to\infty}\frac{1}{T}E_{[0,T]}(I_1^{\sqrt{\epsilon}},\ldots,I_L^{\sqrt{\epsilon}})$, a.s. Since $\rho_i^{(\epsilon)}(\sqrt{\epsilon})\to\rho_i$, for $i=1,\ldots,L$, Proposition 3.3.4 implies that if $\rho_0<\Pi_{i=1}^{L}(1-\rho_i)$, then there exists an $\bar{\epsilon}$ such that Q_0 is stable in the ϵ-system for every $\epsilon<\bar{\epsilon}$. This proves statement (i) of the corollary.

Conversely, we have that if $\rho_0>\Pi_{i=1}^{L}(1-\rho_i)$, then there exists an $\bar{\epsilon}$ such that $\rho_0-\rho_0(\sqrt{\epsilon})>\limsup_{T\to\infty}\frac{1}{T}E_{[0,T]}(I_1^{\sqrt{\epsilon}},\ldots,I_L^{\sqrt{\epsilon}})$, for all $\epsilon\leq\bar{\epsilon}$. This follows from Proposition 3.3.4 and from the fact that $\lim_{\epsilon\downarrow 0}\rho_0(\sqrt{\epsilon})=\lim_{\epsilon\downarrow 0}\lambda_0\mathbb{E}(B_0\mathbf{1}_{(B_0<\sqrt{\epsilon})})=0$. Since class-0 users with sizes larger than $\sqrt{\epsilon}$ receive at most capacity when class-i users with service requirement strictly less than $\sqrt{\epsilon}$ are not served, Lemma 3.1.3 implies that Q_0 is not stable, which proves statement (ii) of the corollary. □

3.3.2 Small class-0 users

We now turn the attention to small class-0 users. We consider the setting where $M_0<m_i$, for $i=1,\ldots,L$, as defined in Section 3.3.1.

Stability of Q_0

In contrast to Corollary 3.3.5, the minimal condition $\rho_0 < 1$ can in this case already be *sufficient* for stability of Q_0. Moreover, for strict SRPT it is the exact stability condition for Q_0.

Observation 3.3.6. *If the policy is strict SRPT and $M_0 < m_i$ for all $i = 1, \dots, L$, then the condition for stability of Q_0 is $\rho_0 < 1$.*

This may be deduced as follows. The fact that $M_0 < m_i$ implies that class 0 receives preemptive priority over class i, and will be entitled to service, unless a class-i user, for some $i = 1, \dots, L$, has a smaller remaining service requirement than all class-0 users (so at most M_0). Class 0 has to wait until those users with a remaining service requirement smaller than all class-0 users have left the network. Since it is strict SRPT, no new class-$1, \dots, L$ users are taken into service. Thus, as long as Q_0 remains non-empty after the arrival of a new class-0 user, it will be prevented from service for at most a period M_0. By Lemma 3.3.2, Q_0 is stable for any $\rho_0 < 1$.

Under weak SRPT, $\rho_0 < 1$ does not always give a stable Q_0. However, first we illustrate a situation in which it is a sufficient condition. For that purpose we consider deterministic service requirements and $L = 2$ nodes.

Proposition 3.3.7. *Assume class j has a deterministic service requirement d_j, $j = 0, 1, 2$, with $d_1 \neq d_2$ and $d_0 < d_1, d_2$, or $d_1 = d_2 > 2d_0$. For the network under consideration with the weak SRPT policy, Q_0 is stable if and only if $\rho_0 < 1$.*

Proof: The fact that $d_0 < d_i$ implies that class 0 receives preemptive priority over class i, and will be entitled to service, unless a class-i user, for some $i = 1, \dots, L$, has a smaller remaining service requirement than d_0. Although class-$1, \dots, L$ users may continue to be served for a while, the delay incurred by a newly arrived class-0 user is bounded as will be shown below. Thus, as long as Q_0 remains non-empty after the arrival of a new class-0 user, it will be prevented from service for at most a bounded period. By Lemma 3.3.2, such a queue is stable for any $\rho_0 < 1$.

It remains to be shown that the delay incurred by a newly arrived class-0 user is bounded. Suppose that class 0 could be prevented from entering service indefinitely. Then at a certain point in time we have for example a class-1 user with a remaining service requirement $r_1 < d_0$ as well as class-0 and class-2 users of which none have received any service. Because of weak SRPT, the class-1 and class-2 users are served. When the class-1 user leaves the system, the class-2 user has a remaining service requirement of $r_2 = d_2 - r_1$. When r_2 is smaller than d_0, this class-2 user is served and because of weak SRPT, a class-1 user also receives service. In order for this to repeat indefinitely, it is necessary that

$$r_1 < d_0, \quad 0 < r_2 = d_2 - r_1 < d_0, \quad 0 < d_1 - d_2 + r_1 < d_0,$$

$$0 < 2d_2 - d_1 - r_1 < d_0, \quad 0 < 2d_1 - 2d_2 + r_1 < d_0, \; \dots,$$

or equivalently,

$$k(d_1 - d_2) + r_1 < d_0 \quad \text{and} \quad k(d_2 - d_1) + d_2 - r_1 < d_0, \quad \forall\, k \geq 0. \tag{3.10}$$

When $d_1 \neq d_2$, we can choose a K, such that for all $r_1 < d_0$ there exists a $k = k(r_1) < K$ for which (3.10) is not satisfied. When $d_1 = d_2 > 2d_0$, we may choose $K = 1$. We can conclude that at some point in time class-1 and class-2 users have remaining service requirements larger than d_0, so a class-0 user can enter service. In fact, the delay for class 0 is bounded by $(K+1)d_0$, independent of r_1. □

In general, $\rho_0 < 1$ is not a sufficient condition for stability of Q_0 under weak SRPT, as may be illustrated again with deterministic service requirements and $L = 2$ nodes. Take $d_1 = d_2 = d$ with $d_0 < d < 2d_0$ and assume that Q_1 and Q_2 are both unstable. In that case, the staggered service pattern of class-1 and class-2 users described in the proof of the above proposition may in fact replicate itself ad infinitum and class 0 can never return to service. Hence, Q_0 may also become unstable with non-zero probability. If Q_1 or Q_2 is stable, which is the case if $\rho_0 + \rho_1 < 1$ or $\rho_0 + \rho_2 < 1$, then with probability 1 the above cycle cannot repeat indefinitely, and it may in fact be checked that Q_0 is stable.

Stability of Q_i, for $i = 1, \ldots, L$

Finally, we investigate the conditions for stability of Q_i. Under weak SRPT, it follows from Lemma 3.1.4 that $\rho_0 + \rho_i < 1$ is a sufficient condition for stability of Q_i.

Under strict SRPT, $\rho_0 + \rho_i < 1$ will in general not be sufficient for stability of Q_i, $i = 1, \ldots, L$. We will show this by considering again $L = 2$ nodes and deterministic service requirements d_j, $j = 0, 1, 2$, $d_0 < d_1, d_2$. By Lemma 3.1.3, if $\rho_0 + \rho_i < \liminf_{T\to\infty} \frac{1}{T}\int_0^T (c_0(t) + c_i(t))\mathrm{d}t$, a.s., then node i, and hence Q_i, is stable. It holds that $c_0(t) + c_i(t) = 0$ when at time t there are no class-i users with a remaining service requirement smaller than d_0 present and there are new class-0 users in the system which cannot be served because of the presence of a class-j user with a remaining service requirement smaller than d_0, $j \neq 0, i$. Otherwise $c_0(t) + c_i(t) = 1$. An explicit expression for the long-run average of $c_0(t) + c_i(t)$, or for $\underline{c}_0$ and $\underline{c}_i$, appears hard to find.

3.4 LAS scheduling

In this section we consider LAS scheduling. In each node, the users with the least attained service are granted the right to an equal share of the capacity at that node. Class-0 users only receive the minimum of the granted shares at the nodes. This may leave capacity unused at some of the nodes. As with SRPT, we again distinguish two variants of LAS. With weak LAS, the unused capacity is re-allocated to the other class at that node (if there are users of that class). In case of strict LAS, the unused capacity is simply lost.

The subsequent analysis is facilitated by a particular property of LAS: the users with a total service requirement x are not influenced by users that have received more than x in service. It will be convenient to define the following quantities, which we refer to as truncated loads:

$$\tilde{\rho}_j(x) := \lambda_j \int_0^{x^-} y \mathrm{d}B_j(y) + \lambda_j x \mathbb{P}(B_j \geq x) = \rho_j(x) + \lambda_j x \mathbb{P}(B_j \geq x),$$

where $\rho_j(x)$ was previously defined in (3.4). Thus, $\tilde{\rho}_j(x)$ represents the load due to class-j users *truncated* at size x (users larger than or equal to x contribute an amount x, rather than zero as in $\rho_j(x)$). We call the system obtained by truncating the sizes of class-j users at x_j, $j = 0, \ldots, L$, the $(x_0, \ldots, x_L)$-truncated system. If $x_0 = \ldots = x_L = x$ we simply refer to the "x-truncated" system. The ∞-truncated system corresponds to the original one.

Property 3.4.1. *From the perspective of users of size x, the system dynamics are identical to those of the x-truncated system. In addition, if there is an $\bar{x}_0$ such that $\mathbb{P}(B_0 \leq \bar{x}_0) = 1$, then from the perspective of class-i users of size $x_i > \bar{x}_0$ for an $i = 1, \ldots, L$, the system behaves identically to the $(\infty, \bar{x}_0, \ldots, \bar{x}_0, x_i, \bar{x}_0, \ldots, \bar{x}_0)$-truncated system, with x_i in the $i + 1$-th component.*

While the first claim is immediate from the arguments above, the second statement deserves some elaboration. The influence of class j, with $j \neq 0, i$, on class i is through class-0 users. If no class-0 user is larger than $\bar{x}_0$, then class-j users larger than $\bar{x}_0$ have no effect on the class-0 users, and therefore no influence on the class-i users either.

By choosing x small enough, we can ensure that $\sum_{j=0}^{L} \tilde{\rho}_j(x) < 1$. Hence, by Lemma 3.1.4 (iii) there exists a stable x-truncated system, for some $x > 0$. It follows from Property 3.4.1 that class-j users of size at most x experience a stable system if and only if Q_j is stable in the x-truncated system, for $j = 0, \ldots, L$. In addition, stability is monotone with respect to truncation: if $(x_0, \ldots, x_L) \geq (y_0, \ldots, y_L)$ component-wise and Q_j is stable in the $(x_0, \ldots, x_L)$-truncated system, then so is Q_j in the $(y_0, \ldots, y_L)$-truncated system.

In the remainder of this section, we impose additional assumptions on the service requirements in order to obtain stability conditions.

3.4.1 Large class-0 users

In this section we consider large class-0 users. We consider an ϵ-system in which class-0 users arrive according to a Poisson process of rate $\lambda_0^{(\epsilon)} := \epsilon\lambda_0$ and sizes are distributed as B_0/ϵ, $\epsilon > 0$.

In Proposition 3.4.2 we derive that the maximum stability conditions $\rho_0 + \rho_i < 1$, $i = 1, \ldots, L$, are in general not sufficient for stability under LAS scheduling. In particular, we prove that if class 0 has extremely large service requirements compared to all other classes, then $\rho_0 \leq \prod_{i=1}^{L}(1 - \rho_i)$ is a necessary stability condition. The proof of this proposition uses that, in the limit as $\epsilon \downarrow 0$, all classes $i \neq 0$ behave as if there is no class-0 traffic.

Proposition 3.4.2. *Assume the policy is either weak or strict LAS. If there exists an $\bar{\epsilon}$ such that Q_0 is stable in the ϵ-system, for all $0 < \epsilon < \bar{\epsilon}$, then it must be that $\rho_0 \leq \prod_{i=1}^{L}(1-\rho_i)$.*

Proof: Let us focus on the ϵ-system. When Q_0 is stable, this must be true in particular for traffic due to class-0 users with service requirements larger than h. Once these users have received an amount of service equal to h, they can at most be served when no users are present with attained service less than h. By Lemma 3.1.3, stability of Q_0 implies that

$$\begin{aligned}\rho_0 - \tilde{\rho}_0^{(\epsilon)}(h) &\leq \limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(N_j^{(\epsilon,h)}(t)=0,\ \forall j=0,\ldots,L)}\mathrm{d}t \\ &\leq \limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(N_i^{(\epsilon,h)}(t)=0,\ \forall i=1,\ldots,L)}\mathrm{d}t, \qquad (3.11)\end{aligned}$$

a.s., where $N_i^{(\epsilon,h)}(t)$ denotes the number of class-i users with attained service less than h present at time t in the ϵ-system. For now, assume that

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}_{(N_i^{(\epsilon,h)}(t)=0,\ \forall i=1,\ldots,L)}\mathrm{d}t \leq \prod_{i=1}^{L}(1-\tilde{\rho}_i(h)) \qquad (3.12)$$

for all $\epsilon > 0$ and $h > 0$. Setting $h = h(\epsilon) = 1/\sqrt{\epsilon}$, we obtain from (3.11) and (3.12), together with the fact that

$$\tilde{\rho}_0^{(\epsilon)}(h(\epsilon)) = \lambda_0 \mathbb{E}(B_0 \mathbf{1}_{(B_0<\sqrt{\epsilon})}) + \lambda_0\sqrt{\epsilon}\mathbb{P}(B_0 \geq \sqrt{\epsilon}) \to 0, \text{ as } \epsilon \downarrow 0, \qquad (3.13)$$

and

$$\tilde{\rho}_i(h(\epsilon)) = \lambda_i \mathbb{E}(B_i \mathbf{1}_{(B_i<1/\sqrt{\epsilon})}) + \lambda_i\frac{1}{\sqrt{\epsilon}}\mathbb{P}(B_i \geq \frac{1}{\sqrt{\epsilon}}) \to \rho_i, \text{ as } \epsilon \downarrow 0, \qquad (3.14)$$

that $\rho_0 < \Pi_{i=1}^{L}(1-\rho_i)$, and hence the proposition is proved.

We show (3.12) by comparing the workloads in classes $i = 1,\ldots,L$, with those in a reference system where class 0 is omitted. Since ϵ will remain fixed in the remainder of the proof, we suppress the dependence on ϵ for notational convenience. Let us denote the workload of class i at time t in the h-truncated system by $W_i^h(t)$, and that in the h-truncated reference system by $\hat{W}_i^h(t)$. We further represent, both for the original and the reference system, the amount of traffic of class i truncated at h that arrives in the time interval $(s,t]$ by $A_i^h(s,t)$. In the original system we also define the amount of service given to class-0 users with attained service less than h in the time interval $(s,t]$ by $B_0^h(s,t)$, and the capacity wasted in $(s,t]$ at node i while there is at least one class-i user that has received at most h in service by $U_i^h(s,t)$.

Both systems are in the same state at time 0. Without loss of generality, assume they both start empty. Then for $i = 1, \ldots, L$,

$$\begin{aligned} W_i^h(t) &= \sup_{s\in[0,t]} \{A_i^h(s,t) + B_0^h(s,t) + U_i^h(s,t) - (t-s)\} \geq \sup_{s\in[0,t]} \{A_i^h(s,t) - (t-s)\} \\ &= \hat{W}_i^h(t), \end{aligned}$$

so that

$$\begin{aligned} &\limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(N_i^{(h)}(t)=0,\ \forall i=1,\ldots,L)} \mathrm{d}t = \limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(W_i^h(t)=0,\ \forall i=1,\ldots,L)} \mathrm{d}t \\ &\leq \limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(\hat{W}_i^h(t)=0,\ \forall i=1,\ldots,L)} \mathrm{d}t = \prod_{i=1}^L (1 - \tilde{\rho}_i(h)), \end{aligned}$$

where the last equality follows from the independence of the various classes in the reference system. □

For weak LAS, we prove that the necessary condition $\rho_0 < \Pi_{i=1}^L(1-\rho_i)$ is also a sufficient condition for stability of the ϵ-system in the limiting regime.

Proposition 3.4.3. *Assume the policy is weak LAS. If $\rho_0 < \Pi_{i=1}^L(1-\rho_i)$, then for ϵ small enough the ϵ-system is stable.*

Proof: Under weak LAS, Q_i is stable when $\rho_0 + \rho_i < 1$. Hence, it remains to be shown that Q_0 is stable. We will use the same notation as in the proof of Proposition 3.4.2. In particular, we distinguish between users with attained service smaller and larger than h. When $N_i^{(\epsilon,h)}(t) = 0$, for all $i = 1, \ldots, L$, at least one of the nodes works at full speed on class-i users with attained service larger than h, $i = 1, \ldots, L$, and on class-0 users, whenever present. By Lemma 3.1.3 and Lemma 3.1.4 (iii) it is therefore sufficient for stability of Q_0 to have

$$\rho_0 + \textstyle\sum_{i=1}^L (\rho_i - \tilde{\rho}_i(h)) < \liminf_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{1}_{(N_i^{(\epsilon,h)}(t)=0,\ \forall i=1,\ldots,L)} \mathrm{d}t. \tag{3.15}$$

If we can show that, whenever $\rho_0 < \Pi_{i=1}^L(1-\rho_i)$, there exists an h (may depend on ϵ) such that the above holds for ϵ small enough, then the proposition is proved.

Let $s_j^{(\epsilon,h)}(t)$ denote the service rate allocated at time t to class-j users with attained service less than h. If $N_i^{(\epsilon,h)}(t) > 0$, then $1 - s_0^{(\epsilon,h)}(t) - s_i^{(\epsilon,h)}(t) = 0$, because the unused capacity in node i is re-allocated to class-i users with attained service less than h, when the service discipline is weak LAS. Hence, the wasted capacity in node i while there are class-i users present that have received at most h in service, $U_i^{(\epsilon,h)}(0,t)$, equals 0. We can conclude that

$$\begin{aligned} W_i^{(\epsilon,h)}(t) &= \sup_{s\in[0,t]} \{A_i^h(s,t) + B_0^{(\epsilon,h)}(s,t) - (t-s)\} \qquad (3.16) \\ &\leq \sup_{s\in[0,t]} \{A_i^h(s,t) - (1-g(\epsilon))(t-s)\} + \sup_{s\in[0,t]} \{B_0^{(\epsilon,h)}(s,t) - g(\epsilon)(t-s)\}. \end{aligned}$$

Let $h = h(\epsilon)$ and choose the function $g(\epsilon)$ such that $\tilde{\rho}_0^{(\epsilon)}(h(\epsilon)) < g(\epsilon) < 1$. Then, $V_i^{(\epsilon)}(t) := \sup_{s\in[0,t]}\{A_i^{h(\epsilon)}(s,t) - (1-g(\epsilon))(t-s)\}$ represents the workload in a queue with capacity $1 - g(\epsilon)$ that serves only class-i users truncated at size $h(\epsilon)$. Using independence, we obtain

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{(V_i^{(\epsilon)}(t)=0,\ \forall i=1,\ldots,L)}\mathrm{d}t = \Pi_{i=1}^L(1 - \frac{\tilde{\rho}_i(h(\epsilon))}{1-g(\epsilon)}), \quad \text{a.s.} \tag{3.17}$$

Note that $V_0^{(\epsilon)}(t) := \sup_{s\in[0,t]}\{B_0^{(\epsilon,h(\epsilon))}(s,t) - g(\epsilon)(t-s)\} \le \sup_{s\in[0,t]}\{A_0^{h(\epsilon)}(s,t) - g(\epsilon)(t-s)\}$, where the right-hand side can be interpreted as the workload in a queue with capacity $g(\epsilon)$ that serves only class-0 users truncated at size $h(\epsilon)$ (and hence its time-average limit exists). Hence,

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{(V_0^{(\epsilon)}(t)>0)}\mathrm{d}t &\le \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{(\sup_{s\in[0,t]}\{A_0^{h(\epsilon)}(s,t)-g(\epsilon)(t-s)\}>0)}\mathrm{d}t \\ &= \tilde{\rho}_0^{(\epsilon)}(h(\epsilon))/g(\epsilon), \quad \text{a.s.} \end{aligned} \tag{3.18}$$

By (3.16) we have

$$\begin{aligned}\mathbf{1}_{(W_i^{(\epsilon,h(\epsilon))}(t)=0,\ \forall i=1,\ldots,L)} &\ge \mathbf{1}_{(V_i^{(\epsilon)}(t)=0,\ \forall i=1,\ldots,L, V_0^{(\epsilon)}(t)=0)} \\ &\ge \mathbf{1}_{(V_i^{(\epsilon)}(t)=0,\ \forall i=1,\ldots,L)} - \mathbf{1}_{(V_0^{(\epsilon)}(t)>0)}.\end{aligned}$$

Together with (3.17) and (3.18), we obtain that

$$\liminf_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}_{(N_i^{(\epsilon,h(\epsilon))}(t)=0,\ \forall i=1,\ldots,L)}\mathrm{d}t \ge \Pi_{i=1}^L(1 - \frac{\tilde{\rho}_i(h(\epsilon))}{1-g(\epsilon)}) - \frac{\tilde{\rho}_0^{(\epsilon)}(h(\epsilon))}{g(\epsilon)}, \tag{3.19}$$

a.s. Setting $h(\epsilon) = 1/\sqrt{\epsilon}$, we have $\tilde{\rho}_0^{(\epsilon)}(h(\epsilon)) \to 0$ and $\tilde{\rho}_i(h(\epsilon)) \to \rho_i$, for $i = 1,\ldots,L$ (see (3.13) and (3.14)). Choosing $g(\cdot)$ such that $\lim_{\epsilon\downarrow 0}\tilde{\rho}_0^{(\epsilon)}(h(\epsilon))/g(\epsilon) = 0$ and $\lim_{\epsilon\downarrow 0} g(\epsilon) = 0$, we obtain that the right-hand side in (3.19) converges to $\Pi_{i=1}^L(1-\rho_i)$ as $\epsilon \downarrow 0$. This proves that (3.15) is satisfied when $\rho_0 < \Pi_{i=1}^L(1-\rho_i)$. □

3.4.2 Small class-0 users

In this section we consider class-0 users with small service requirements, compared to the service requirements of class-i users, $i = 1,\ldots,L$. As before, we study a sequence of systems indexed by ϵ and let $\epsilon \downarrow 0$. In the ϵ-system, class-0 users arrive according to a Poisson process of rate $\lambda_0^{(\epsilon)} := \lambda_0/\epsilon$ and the sizes are distributed as ϵB_0.

The next proposition shows that in the ϵ-system, the maximum stability conditions can be arbitrarily close to the sufficient stability conditions when we consider $B_0^{(\epsilon)}$ truncated at $h(\epsilon)$, with $\lim_{\epsilon\downarrow 0} h(\epsilon) = 0$.

Proposition 3.4.4. *Assume the policy is weak or strict LAS and consider a function $h(\epsilon)$ such that $\lim_{\epsilon\downarrow 0} h(\epsilon) = 0$. If $\rho_0 + \rho_i < 1$ for an $i \neq 0$, then there exists an $\bar{\epsilon}$ such that for all $0 < \epsilon < \bar{\epsilon}$, node i is stable in the ϵ-system with $B_0^{(\epsilon)}$ truncated at $h(\epsilon)$.*

Proof: Assume $\rho_0 + \rho_i < 1$, for an $i = 1, \dots, L$. We have $\lim_{\epsilon \downarrow 0} h(\epsilon) = 0$, so there is an $\bar{\epsilon}$ such that

$$\rho_0 + \rho_i + \sum_{j=1, j \neq i}^{L} \tilde{\rho}_j(h(\epsilon)) < 1, \text{ for all } \epsilon < \bar{\epsilon}. \tag{3.20}$$

From this it follows that $\tilde{\rho}_0^{(\epsilon)}(h(\epsilon)) + \sum_{j=1}^{L} \tilde{\rho}_j(h(\epsilon)) < 1$, for all $\epsilon < \bar{\epsilon}$, which, by Lemma 3.1.4 (iii), is a sufficient condition for stability of the $h(\epsilon)$-truncated system. According to Property 3.4.1, for $\epsilon < \bar{\epsilon}$, Q_0 is stable when $B_0^{(\epsilon)}$ is truncated at $h(\epsilon)$.

For class i, Property 3.4.1 implies that Q_i is stable in the $(h(\epsilon), \infty, \dots, \infty)$-truncated system if and only if Q_i is stable in the $(h(\epsilon), \dots, h(\epsilon), \infty, h(\epsilon), \dots, h(\epsilon))$-truncated system, with ∞ in the $i+1$-th component. Because of Lemma 3.1.4 (iii), for the latter it is sufficient to have (3.20), which holds for all $\epsilon < \bar{\epsilon}$. □

Remark 3.4.5. The fact that we can choose $h(\epsilon)$ such that $\lim_{\epsilon \downarrow 0} h(\epsilon)/\epsilon = \infty$, and thus $\mathbb{P}(B_0^{(\epsilon)} \leq h(\epsilon)) \to 1$, as $\epsilon \downarrow 0$, suggests that the non-truncated ϵ-system can be arbitrarily closely approximated by the truncated one. However, the proof of Proposition 3.4.4 relies on the truncation of $B_0^{(\epsilon)}$. In the particular case that B_0 is bounded from above by a constant M, Proposition 3.4.4 does imply that the condition $\rho_0 + \rho_i < 1$ is sufficient for stability of node i in the ϵ-system for ϵ small enough (take $h(\epsilon) = \epsilon M$).

3.5 Concluding remarks

We have explored the fundamental stability properties of size-based scheduling policies in linear bandwidth-sharing networks. In particular, we established the exact stability conditions for SRPT and LAS scheduling in various limiting regimes. Despite its simplicity, the linear network appears already sufficiently rich to exhibit instability phenomena that may occur for general network topologies and route structures. The results indicate that, due to the simultaneous resource possession, size-based scheduling among classes may fail to use the available resources efficiently, and cause instability effects, even at arbitrarily low traffic loads.

In particular, the results in this chapter imply that the prototypical size-based scheduling policies will certainly not yield optimal performance in bandwidth-sharing networks when applied among classes. Instead, proper tuning of the parameters of weighted α-fair bandwidth-sharing policies, which are stable under the maximum stability conditions, or applying size-based scheduling within classes, might provide more promising approaches for improving the performance of the network. This will be further explored in Chapters 4, 5, and 6. It is noteworthy that in single-link scenarios, weighted α-fair policies essentially reduce to DPS policies, which are known to cover the entire achievable mean-holding cost region (for non-anticipating policies) in the case of exponentially distributed service requirements [56, Lemma 6.3].

Chapter 4
Optimal scheduling in a linear network

The focus of the present and the next two chapters is on optimal scheduling in a linear bandwidth-sharing network. As observed in Chapter 3, the size-based scheduling policies SRPT and LAS, which have optimality properties in single-server models, may cause instability effects in a network scenario and, hence, do certainly not yield optimal performance. This indicates that for a general setting, optimal policies are hard to obtain. In this chapter we therefore mainly focus on exponentially distributed service requirements and restrict the attention to non-anticipating policies, i.e., policies that do not have knowledge of the remaining service requirements. We seek policies that minimize in some sense the holding cost. Our main result is that an optimal policy can be characterized by so-called switching curves, i.e., the policy dynamically switches between several priority rules. In particular, for special choices of the mean service requirements, it reduces to simple priority rules.

A popular class of policies studied in the context of bandwidth-sharing networks are the α-fair bandwidth-sharing policies. These policies achieve stability under the maximum stability conditions, provided $\alpha > 0$. However, it is not well understood to what extent their performance leaves potential room for improvement. Armed with the knowledge of an optimal policy in the linear network, we compare numerically its performance with various α-fair bandwidth-sharing policies. Our results indicate that, for a moderately-loaded system, the optimal policy achieves only modest improvements over an optimized α-fair policy. In their turn, the performance of α-fair policies is fairly insensitive to the value of α, as long as this value is not too small.

This chapter is organized as follows. In Section 4.1 we provide a model description and discuss some preliminaries. In Section 4.2 we derive sample-path comparisons for the workload processes under various scheduling policies. These results are used in Section 4.3 to show that for certain settings simple priority rules minimize the mean holding cost. In the case of two nodes, we use dynamic programming techniques to show that such policies are in fact stochastically optimal. In addition, we show that an optimal policy can be characterized by a switching curve. Numerical experiments can be found in Section 4.4. We summarize our results in Section 4.5.

4.1 Model and preliminaries

We consider a linear network with L nodes and $L+1$ classes, where class i requires service at node i only, $i = 1, \dots, L$, while class 0 requires service at all L nodes simultaneously, see Figure 1.2. For convenience, we assume each of the nodes to have a unit service rate. Hence, a rate allocation is feasible when it belongs to the capacity region $S := \{(s_0, \dots, s_L) \in \mathbb{R}_+^{L+1} : s_0 + s_i \leq 1, \text{ for all } i = 1, \dots, L\}$, which is depicted in Figure 1.5 for the case $L = 2$. Class-j users arrive according to independent Poisson processes of rate λ_j, and have generally distributed service requirements B_j, $j = 0, \dots, L$. Define the traffic load of class j by $\rho_j := \lambda_j \mathbb{E}(B_j)$. For a given policy π, denote by $N_j^\pi(t)$ the number of class-j users and by $W_j^\pi(t)$ the workload in class j, at time t. We further define N_j^π and W_j^π as random variables with the corresponding steady-state distributions (when they exist).

For any point in time, a policy decides how the capacity is divided between the various classes. We assume that the numbers of users in the various classes are observable to a policy. The class containing all policies is denoted by Π and the class containing all (possibly preemptive) non-anticipating policies is denoted by $\bar{\Pi} \subset \Pi$. We also define two classes of priority rules, which play a central role in this chapter:

- Π^*: $\pi^* \in \Pi^*$ when π^* gives preemptive priority to class 0 whenever it is backlogged. Otherwise, all other classes with a backlog are served simultaneously.
- Π^{**}: $\pi^{**} \in \Pi^{**}$ when π^{**} simultaneously serves all classes $i = 1, \dots, L$ whenever at least one user of each class is present. Otherwise class 0 is served. When class 0 is empty, any other class with at least one user present is served.

Policies in Π^* and Π^{**} ensure that each node operates at full rate whenever it is nonempty. Hence, those policies achieve a stable system under the maximum stability conditions $\rho_0 + \rho_i < 1$, $i = 1, \dots, L$, see Lemma 3.1.4 (ii).

Besides the stability conditions, it is possible to derive closed-form expressions for other performance measures for a policy $\pi^* \in \Pi^*$ as well. Note that class 0 does not notice the presence of other classes under a policy π^*. The mean amount of class-0 work is therefore given by the Pollaczek-Khintchine formula:

$$\mathbb{E}(W_0^{\pi^*}) = \frac{\lambda_0 \mathbb{E}({B_0}^2)}{2(1-\rho_0)}.$$

With a policy π^*, any class $i \neq 0$ sees its service being interrupted by busy periods of class 0, so that [133]:

$$\mathbb{E}(W_i^{\pi^*}) = \frac{\lambda_0 \mathbb{E}({B_0}^2) + \lambda_i \mathbb{E}({B_i}^2)}{2(1-\rho_0-\rho_i)} - \frac{\lambda_0 \mathbb{E}({B_0}^2)}{2(1-\rho_0)}.$$

These formulas hold for any service requirement distribution and intra-class policy. In the special case of exponentially distributed service requirements and a non-anticipating intra-class policy, the mean number of users can simply be obtained from $\mathbb{E}(N_i^\pi) = \mu_i \mathbb{E}(W_i^\pi)$, with $\mu_i = 1/\mathbb{E}(B_i)$ (and thus $\mathbb{E}({B_i}^2) = 2/\mu_i^2$). In particular, if $\pi^* \in \Pi^* \cap \bar{\Pi}$, then

$$\mathbb{E}(N_0^{\pi^*}) = \frac{\rho_0}{1-\rho_0}, \tag{4.1}$$

$$\mathbb{E}(N_i^{\pi^*}) = \frac{\rho_i}{1-\rho_0-\rho_i} + \frac{\mu_i}{\mu_0}\left(\frac{\rho_0}{1-\rho_0-\rho_i} - \frac{\rho_0}{1-\rho_0}\right), \quad i = 1,\dots,L. \tag{4.2}$$

For a policy $\pi^{**} \in \Pi^{**}$ there is no closed-form expression available for the mean workloads. For $L = 2$, determining these is equivalent to solving a boundary-value problem [42]: the service rate allocated to any class i depends on the workloads of both other classes.

4.2 Workload

In this section we compare (sample-path wise) the workloads of the various classes under different policies. In fact, the results presented in this section are valid for generally distributed inter-arrival times.

Let π^i be a policy that is work-conserving in node i, for an $i = 1,\dots,L$, i.e., the capacity of node i is fully used whenever that node is backlogged. Obviously, such a policy stochastically minimizes the total workload process in node i. More specifically, if $W_0^{\pi^i}(0) + W_i^{\pi^i}(0) \leq_{st} W_0^{\pi}(0) + W_i^{\pi}(0)$ for a policy $\pi \in \Pi$, then

$$\{W_0^{\pi^i}(t) + W_i^{\pi^i}(t)\}_{t\geq 0} \leq_{st} \{W_0^{\pi}(t) + W_i^{\pi}(t)\}_{t\geq 0}. \tag{4.3}$$

This is obtained by considering the same realizations of the arrival processes and service requirements under both policies. Note that policies in Π^* and Π^{**} are work-conserving in each node, so (4.3) holds for all $i = 1,\dots,L$, when $\pi^i \in \Pi^* \cup \Pi^{**}$. We call $W_{0,j,k}^{\pi}(t) := W_0^{\pi}(t) + W_j^{\pi}(t) + W_k^{\pi}(t)$ the aggregate workload in nodes j and k. Besides minimizing the workload in each node at any point in time, any policy $\pi^{**} \in \Pi^{**}$ also minimizes the aggregate workload in at least one pair of nodes (these need not always be the same) as is formalized in the following lemma. This result will be useful for the analysis in the next section.

Lemma 4.2.1. *Consider the same realizations of the arrival processes and service requirements for a policy $\pi^{**} \in \Pi^{**}$ and a policy $\pi \in \Pi$. If for $t = 0$ there exist nodes j and k with $j \neq k$, such that*

$$W_{0,j,k}^{\pi^{**}}(t) \leq W_{0,j,k}^{\pi}(t), \tag{4.4}$$

then, for any $t > 0$, there exist j and k (not necessarily the same as at time $t = 0$) with $j \neq k$ such that (4.4) holds.

In particular, when $L = 2$, the lemma states that any policy in Π^{**} *stochastically* minimizes the total workload in the system. We note that there is no policy that achieves the same for $L > 2$. In the short term it is favorable to *not* serve class 0 whenever there are users present of at least two other classes (classes $1,\dots,K$).

However, in the long run, serving class 0 when present may allow all other classes to be served at an even higher rate later and possibly empty the entire system sooner.

Proof of Lemma 4.2.1: We show by contradiction that u, defined as

$$u := \inf\{t > 0 : W^{\pi^{**}}_{0,j,k}(t) > W^{\pi}_{0,j,k}(t), \quad \text{for all } j, k = 1, \dots, L,\ j \neq k\},$$

cannot be finite. Let us suppose $u < \infty$. Inequality (4.4) can only be violated for all pairs j and k immediately after time u, when it holds with equality at time u for some j and k, which we fix for the remainder of the proof. In addition, for the equality to cease to be valid, policy π^{**} should not be serving both classes j and k at full rate, hence we have for example $W^{\pi^{**}}_j(u) = 0$. Policy π^{**} is work-conserving in all nodes. Hence, from (4.3) we obtain $W^{\pi^{**}}_{0jk}(t) = W^{\pi^{**}}_0(t) + W^{\pi^{**}}_k(t) \leq W^{\pi}_0(t) + W^{\pi}_k(t) \leq W^{\pi}_{0jk}(t)$, for all $u \leq t < T_j$, with T_j the moment of the first class-j arrival after time u. This contradicts the definition of u. □

4.3 Optimality results

In the remainder of this chapter we focus on exponentially distributed service requirements and write $\mu_j := 1/\mathbb{E}(B_j)$, $j = 0, 1, \dots, L$. We are interested in non-anticipating policies that minimize in some sense the holding cost $\sum_{j=0}^{L} c_j N_j(t)$, where c_j is an arbitrary nonnegative cost associated with class j, $j = 0, \dots, L$.

To put our results in context, we recall that the $c\mu$-rule is known to minimize the mean holding cost in a single-server multi-class queue with exponentially distributed service requirements among all non-anticipating policies [38, 102]. The rationale behind this rule is that it maximizes the weighted departure rate at all times. The problem of how to allocate the capacity of the nodes among the various users in a linear network is more complex. Besides trying to maximize the total weighted departure rate of the system, we must take into account that giving more preference to class 0 may make better use of the available capacity. For example, when $c_i\mu_i > c_0\mu_0$ for all $i = 1, \dots, L$, giving preemptive priority to classes $1, \dots, L$, myopically maximizes the total weighted departure rate of the system. However, such a policy unnecessarily causes instability when $\Pi_{i=1}^{L}(1 - \rho_i) < \rho_0$, see Proposition 3.2.1. In general, there can be a trade-off between maximizing the total weighted departure rate and using the full capacity in each backlogged node. For relatively 'large' values of μ_0 these two objectives are compatible and priority rules are optimal. More precisely, this is so when $c_0\mu_0 \geq \sum_{i \geq 1, i \neq j} c_i\mu_i$ for all $j \neq 0$, which will be the setting of Section 4.3.1. In Section 4.3.2 we treat the other case, i.e., $c_0\mu_0 < \sum_{i \geq 1, i \neq j} c_i\mu_i$ for an $j \neq 0$, and describe the general structure of an optimal policy.

4.3.1 Priority rules and optimality

In this section, we prove that priority rules minimize the *mean* holding cost when $c_0\mu_0 > \sum_{i \geq 1, i \neq j} c_i\mu_i$ for all $j \neq 0$. In the case of two nodes ($L = 2$), we establish *stochastic* optimality.

Mean holding cost

Because of the memoryless property of the exponential distribution and since we will only consider non-anticipating policies, the workload, $W_j(t)$, is distributed as $\sum_{k=1}^{N_j(t)} E_k^j$, where E_k^j are i.i.d. exponential random variables with mean $1/\mu_j$. After taking expectations, this gives $\mathbb{E}(W_j(t)) = \frac{\mathbb{E}(N_j(t))}{\mu_j}$. This relation allows to use the results of Section 4.2 to readily derive that, in certain cases, priority rules minimize the mean holding cost. These optimality results are in fact valid for generally distributed inter-arrival times.

Proposition 4.3.1. *(This proposition holds for generally distributed inter-arrival times.) Assume exponentially distributed service requirements. Let $\pi \in \bar{\Pi}$, $\pi^* \in \Pi^* \cap \bar{\Pi}$, and assume $W_j^{\pi^*}(0) \leq_{st} W_j^{\pi}(0)$, for all $j = 0, \ldots, L$. If $\sum_{i=1}^L c_i\mu_i \leq c_0\mu_0$, then $\mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi^*}(t)) \leq \mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi}(t))$, for all time $t \geq 0$.*

Proof: Giving preemptive priority to class 0 stochastically minimizes the workload of class 0, i.e., $\{W_0^{\pi^*}(t)\}_{t\geq 0} \leq_{st} \{W_0^{\pi}(t)\}_{t\geq 0}$. Hence, we have

$$\mathbb{E}(N_0^{\pi^*}(t)) \leq \mathbb{E}(N_0^{\pi}(t)). \tag{4.5}$$

By (4.3), policy π^* stochastically minimizes the workload in each node, which implies by the relation $\mathbb{E}(W_i(t)) = \frac{\mathbb{E}(N_i(t))}{\mu_i}$ that

$$\frac{1}{\mu_0}\mathbb{E}(N_0^{\pi^*}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi^*}(t)) \ \leq \frac{1}{\mu_0}\mathbb{E}(N_0^{\pi}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t)), \tag{4.6}$$

for all $i = 1, \ldots, L$. Multiplying (4.5) by $\frac{c_0\mu_0 - \sum_{i=1}^L c_i\mu_i}{\mu_0} \geq 0$, multiplying (4.6) by $c_i\mu_i \geq 0$, for all $i = 1, \ldots, L$, and summing these $L+1$ inequalities gives $\mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi^*}(t)) \leq \mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi}(t))$. □

Proposition 4.3.2. *(This proposition holds for generally distributed inter-arrival times.) Assume exponentially distributed service requirements. Let $\pi \in \bar{\Pi}$, $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$, and assume $W_j^{\pi^{**}}(0) \leq_{st} W_j^{\pi}(0)$, for all $j = 0, \ldots, L$. If $\sum_{i=1}^L c_i\mu_i \geq c_0\mu_0 \geq \sum_{i=1, i\neq j}^L c_i\mu_i$ for all $j \neq 0$, then $\mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi^{**}}(t)) \leq \mathbb{E}(\sum_{j=0}^L c_j N_j^{\pi}(t))$, for all time $t \geq 0$.*

Proof: Note that both the stochastic inequalities in (4.3) and in (4.4) are obtained by coupling the arrival and service processes. Consider one such realization. At time t we then have

$$W_0^{\pi^{**}}(t) + W_i^{\pi^{**}}(t) \leq W_0^{\pi}(t) + W_i^{\pi}(t), \quad \text{for all} \ \ i \neq 0, \tag{4.7}$$

and there are classes $j, k \neq 0$, $j \neq k$, such that

$$W_0^{\pi^{**}}(t) + W_j^{\pi^{**}}(t) + W_k^{\pi^{**}}(t) \leq W_0^{\pi}(t) + W_j^{\pi}(t) + W_k^{\pi}(t). \tag{4.8}$$

Now multiply (4.7) by $c_0\mu_0 - \sum_{l=1, l\neq i}^L c_l\mu_l \geq 0$, for $i = j, k$ and by $c_i\mu_i$ for all $i \neq 0, j, k$, multiply inequality (4.8) by $\sum_{i=1}^L c_i\mu_i - c_0\mu_0 \geq 0$, and sum these $L+1$

inequalities to obtain $\sum_{j=0}^{L} c_j\mu_j W_j^{\pi^{**}}(t) \leq \sum_{j=0}^{L} c_j\mu_j W_j^{\pi}(t)$. Taking expectations on both sides gives the result. □

The above-obtained results extend to the case of hyperexponentially distributed service requirements with parameters p_{ik}, $\sum_{k=1}^{K_i} p_{ik} = 1$, and μ_{ik}, $k = 1, \dots, K_i$. Optimality in expectation of a policy $\pi^* \in \Pi^* \cap \bar{\Pi}$ or of a policy $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$ can then be established when either $\sum_{i=1}^{L} \mu_i^{max} \leq \mu_0^{min}$ or $\sum_{i=1}^{L} \mu_i^{min} \geq \mu_0^{max}$ and $\mu_0^{min} \geq \sum_{i=1, i\neq j}^{L} \mu_i^{max}$, for all $j = 1, \dots, L$, respectively, with $\mu_j^{min} = \min_{k=1,\dots,K_j} \mu_{jk}$ and $\mu_j^{max} = \max_{k=1,\dots,K_j} \mu_{jk}$.

Stochastic optimality

It is worth noting that despite the stochastic workload relations, the above arguments cannot be strengthened to prove that priority rules in fact stochastically minimize the holding cost for the given parameter values. This can, however, be accomplished using a dynamic programming approach in the case of two nodes and unit costs, $c_j = 1$, $j = 0, \dots, L$. (This agrees with the stochastic optimality of the $c\mu$-rule, which has only been proved in the case of unit costs [114].) The obtained stochastic optimality results concern exponentially distributed service requirements. For an extension to phase-type distributed service requirements we refer to [140].

We consider the uniformized Markov chain, that is, transition epochs are generated by a Poisson process of uniform rate $\nu = \sum_{j=0}^{L} \lambda_j + \sum_{j=0}^{L} \mu_j$. Since ν is finite, we may assume $\nu = 1$ without loss of generality. We then focus on the discrete-time Markov chain embedded at transition epochs, and, for transparency of notation, denote the number of class-j users after t steps by $N_j(t)$, $j = 0, \dots, L$. We define the value functions $V_m(\cdot) : \mathbb{Z}_+^{L+1} \to \mathbb{R}$, $m = 0, 1, \dots$, as follows. Let $\vec{x} = (x_0, \dots, x_L) \in \mathbb{Z}^{L+1}$. Then, $V_0(\vec{x}) := \tilde{C}(\vec{x})$, with $\tilde{C}(\cdot) : \mathbb{Z}_+^{L+1} \to \mathbb{R}$ a terminal cost, and for $m = 1, 2, \dots,$

$$\begin{aligned} V_{m+1}(\vec{x}) &:= C(\vec{x}) + \sum_{i=0}^{L} \lambda_i V_m(\vec{x} + \vec{e}_i) \\ &\quad + \min_{\vec{s}\in S} \left\{ \sum_{j=0}^{L} \mathbf{1}_{(x_j>0)} \mu_j s_j V_m(\vec{x} - \vec{e}_j) + (1 - \sum_{j=0}^{L} \lambda_j - \sum_{j=0}^{L} \mathbf{1}_{(x_j>0)} \mu_j s_j) V_m(\vec{x}) \right\} \\ &= C(\vec{x}) + \sum_{i=0}^{L} \lambda_i V_m(\vec{x} + \vec{e}_i) + \sum_{i=0}^{L} \mu_i V_m(\vec{x}) \\ &\quad + \min_{\vec{s}\in S} \left\{ \sum_{j=0}^{L} \mathbf{1}_{(x_j>0)} \mu_j s_j \left(V_m(\vec{x} - \vec{e}_j) - V_m(\vec{x}) \right) \right\}, \end{aligned} \tag{4.9}$$

with $C(\cdot) : \mathbb{Z}_+^{L+1} \to \mathbb{R}$ the direct cost, S the capacity region, and $\vec{e}_j \in \mathbb{Z}^{L+1}$ the $(j+1)$-th unit vector, $j = 0, \dots, L$. Setting $C(\vec{x}) = 0$ and $\tilde{C}(\vec{x}) = \mathbf{1}_{(\sum_{j=0}^{L} x_j > y)}$, we obtain $V_{m+1}(\vec{x}) = \min_{\pi\in\bar{\Pi}} \mathbb{P}(\sum_{j=0}^{L} N_j^{\pi}(m+1) > y | \vec{N}(0) = \vec{x})$, with $\vec{N}(t) =$

$(N_0(t), \ldots, N_L(t))$. Hence, if we can show that for all $y \geq 0$ and all $m \in \{0, 1, \ldots\}$ we can choose the same minimizing action in (4.9) (the optimal action may depend on the state $\vec{x}$), then the corresponding stationary policy stochastically minimizes the total number of users at every instant in time.

The following lemma states that the value functions are non-decreasing.

Lemma 4.3.3. *If $C(\cdot)$ and $\tilde{C}(\cdot)$ are non-decreasing in x_j, for all j, then $V_m(\cdot)$ is non-decreasing in x_j, for all j and $m = 0, 1, \ldots$.*

Proof The statement follows directly from the definition of $V_m(\cdot)$. □

The set S is convex, hence the minimizing action in (4.9) will be one of the extreme points of S. From the lemma above, it can be concluded that the minimizer will not be $\vec{0} \in S$, since $\sum_{j=0}^{L} \mathbf{1}_{(x_j>0)} \mu_j s_j \left(V_m(\vec{x} - \vec{e}_j) - V_m(\vec{x})\right) \leq 0$ for all $\vec{s} \in S$. Hence we can rewrite the function $V_{m+1}(\cdot)$ as follows:

$$\begin{aligned} V_{m+1}(\vec{x}) &= C(\vec{x}) + \sum_{i=0}^{L} \lambda_i V_m(\vec{x} + \vec{e}_i) \\ &\quad + \min(\mu_0 V_m((x_0 - 1)^+, \ldots, x_L) + \sum_{i=1}^{L} \mu_i V_m(\vec{x}), \\ &\qquad \mu_0 V_m(\vec{x}) + \sum_{i=1}^{L} \mu_i V_m(x_0, \ldots, x_{i-1}, (x_i - 1)^+, x_{i+1} \ldots, x_L)). \end{aligned} \tag{4.10}$$

From now on we focus on the case $L = 2$. In order to obtain stochastic optimality results, we prove three lemmas that establish convenient properties of the functions $V_m(\cdot)$, $m = 0, 1, \ldots$, without specifying the functions $C(\cdot)$ and $\tilde{C}(\cdot)$. The proofs may be found in Appendix 4.A.

The first lemma shows that under certain conditions on the cost functions, the minimizing action in (4.10) will be to always serve class 0 rather than classes 1 or 2 alone, independent of the remaining time horizon.

Lemma 4.3.4. *Assume $C(\cdot)$ and $\tilde{C}(\cdot)$ are non-decreasing in x_0, x_1 and x_2. If both $Z = C$ and $Z = \tilde{C}$ satisfy for $i, j = 1, 2$, with $i \neq j$,*

$$\mu_0 Z(\vec{x} - \vec{e}_0) + \mu_i Z(\vec{x}) \leq \mu_0 Z(\vec{x}) + \mu_i Z(\vec{x} - \vec{e}_i), \tag{4.11}$$

for all $x_0, x_i > 0, x_j \geq 0$, then the same is true for $Z = V_m$ for all m.

The next lemma shows that under certain conditions on the cost functions it is better to serve class 0 rather than classes 1 and 2 simultaneously, independent of the remaining time horizon.

Lemma 4.3.5. *Assume $C(\cdot)$ and $\tilde{C}(\cdot)$ are non-decreasing in x_0, x_1 and x_2. If both $Z = C$ and $Z = \tilde{C}$ satisfy (4.11) for $i = 1$ and $i = 2$, and, in addition, satisfy*

$$\mu_0 Z(\vec{x} - \vec{e}_0) + (\mu_1 + \mu_2) Z(\vec{x}) \leq \mu_0 Z(\vec{x}) + \mu_1 Z(\vec{x} - \vec{e}_1) + \mu_2 Z(\vec{x} - \vec{e}_2), \tag{4.12}$$

for all $x_0, x_1, x_2 > 0$, then the same is true for $Z = V_m$ for all m.

Combining Lemmas 4.3.4 and 4.3.5 we obtain that any policy $\pi^* \in \Pi^* \cap \bar{\Pi}$ is stochastically optimal when $\mu_1 + \mu_2 \leq \mu_0$.

Proposition 4.3.6. *Assume exponentially distributed service requirements. Let $\pi^* \in \Pi^* \cap \bar{\Pi}$, $\pi \in \bar{\Pi}$, and assume $\vec{N}^{\pi^*}(0) = \vec{N}^{\pi}(0)$. If $\mu_1 + \mu_2 \leq \mu_0$, then $\sum_{j=0}^{2} N_j^{\pi^*}(t) \leq_{st} \sum_{j=0}^{2} N_j^{\pi}(t)$, for all time $t \geq 0$.*

Proof: We set $C(\vec{x}) = 0$. If $\mu_1 + \mu_2 \leq \mu_0$, then the non-decreasing function $\tilde{C}(\vec{x}) = \mathbf{1}_{(\sum_{j=0}^{2} x_j > y)}$ satisfies (4.11) for $i = 1$ and $i = 2$, and (4.12). Lemmas 4.3.4 and 4.3.5 imply that serving class 0 (whenever possible) is always the minimizing action in (4.10). □

In the following lemma we show that under certain conditions on the cost functions, it can also be better to serve classes 1 and 2 whenever both are present, rather than class 0. Again this is independent of the remaining time horizon.

Lemma 4.3.7. *Assume $C(\cdot)$ and $\tilde{C}(\cdot)$ are non-decreasing in x_0, x_1 and x_2. If both $Z = C$ and $Z = \tilde{C}$ satisfy*

$$\mu_0 Z(\vec{x}) + \mu_1 Z(\vec{x} - \vec{e}_1) + \mu_2 Z(\vec{x} - \vec{e}_2) \leq \mu_0 Z(\vec{x} - \vec{e}_0) + (\mu_1 + \mu_2) Z(\vec{x}), \qquad (4.13)$$

for all $x_0, x_1, x_2 > 0$, then the same is true for $Z = V_m$ for all m.

Combining Lemmas 4.3.4 and 4.3.7 we obtain that any policy $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$ is stochastically optimal when $\mu_1, \mu_2 \leq \mu_0$ and $\mu_1 + \mu_2 \geq \mu_0$.

Proposition 4.3.8. *Assume exponentially distributed service requirements. Let $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$, $\pi \in \bar{\Pi}$, and assume $\vec{N}^{\pi^{**}}(0) = \vec{N}^{\pi}(0)$. If $\mu_1, \mu_2 \leq \mu_0$ and $\mu_1 + \mu_2 \geq \mu_0$, then $\sum_{j=0}^{2} N_j^{\pi^{**}}(t) \leq_{st} \sum_{j=0}^{2} N_j^{\pi}(t)$, for all time $t \geq 0$.*

Proof: We set $C(\vec{x}) = 0$. If $\mu_1, \mu_2 \leq \mu_0$ and $\mu_1 + \mu_2 \geq \mu_0$, then the non-decreasing function $\tilde{C}(\vec{x}) = \mathbf{1}_{(\sum_{j=0}^{2} x_j > y)}$ satisfies (4.11) for $i = 1$ and $i = 2$, and (4.13). This implies that the minimizing action in (4.10) is to serve classes 1 and 2 whenever both are present (Lemma 4.3.7) and to serve class 0 otherwise (Lemma 4.3.4). □

4.3.2 General structure of an average-cost optimal policy

Again assuming exponential service requirements, we now explore the remaining case when there exists a $j = 1, \ldots, L$, such that $\sum_{i=1, i \neq j}^{L} c_i \mu_i \geq c_0 \mu_0$. It may still be better to sometimes serve class 0, even if that does not maximize the total weighted departure rate in the short run. Doing so creates the potential to serve classes $1, \ldots, L$ simultaneously in the future, and therefore achieve a higher degree of parallelism. Hence, as the number of users varies, the system will dynamically switch between several priority rules. We focus on an average-cost optimal policy, i.e., a policy that minimizes $\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}(\int_0^T \sum_{j=0}^{L} c_j N_j^{\pi}(t)) \mathrm{d}t$ over all policies $\pi \in \bar{\Pi}$. According to [123, Corollary 20] such an average-cost optimal policy exists.

We focus on the case of two nodes and hence consider service rates such that $\mu_0 < \mu_i$ for at least one $i = 1, 2$. Intuitively it is clear that when there are users of both classes 1 and 2 present, serving them will be optimal. When there are only users of classes 0 and 1 present (no class-2 users) and $c_1\mu_1 < c_0\mu_0$, serving class 0 seems appropriate, since it uses the full capacity in both nodes, and it maximizes the weighted departure rate. However, when $c_0\mu_0 < c_1\mu_1$, there is no obvious choice: serving class 1 will maximize the weighted departure rate, but leaves node 2 unused. In contrast, serving class 0 uses the full capacity of the network, but the weighted departure rate is not maximized. The next proposition states that in such situations, the average-optimal policy can be characterized by a switching curve that determines which class should be served. Note that no closed-form expression for this curve can be obtained using dynamic programming techniques. Finding approximations for the switching curve will be the subject of Chapter 5.

Proposition 4.3.9. *Assume exponentially distributed service requirements. Consider a non-anticipating policy that is described by switching curves $h_i(\cdot), i = 1, 2$:*

- *When $N_1(t), N_2(t) > 0$, classes 1 and 2 are served simultaneously.*
- *When $N_{3-i}(t) = 0$ and $c_0\mu_0 \geq c_i\mu_i$, for an $i = 1, 2$, class 0 is served.*
- *When $N_{3-i}(t) = 0$ and $c_0\mu_0 < c_i\mu_i$, for an $i = 1, 2$, class 0 is served if $N_i(t) < h_i(N_0(t))$, and otherwise class i is served.*

If $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, then there exist switching curves such that the policy described above is an average-cost optimal policy (among all non-anticipating policies).

In the remainder of this section we give a proof of Proposition 4.3.9. We focus again on the discrete-time Markov chain and value functions $V_m(\cdot)$ as defined in Section 4.3.1. Choosing $C(\vec{x}) = \sum_{j=0}^{2} c_j x_j$ and $\tilde{C}(\vec{x}) = 0$ implies that the objective is to find an average-cost optimal policy, i.e., a policy that minimizes the average holding cost $\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}(\int_0^T \sum_{j=0}^{2} c_j N_j(t)\mathrm{d}t)$. For now, we do not consider any particular choice of the cost functions and derive, under certain conditions on the cost functions, inequalities for the value functions.

In case there are no class-2 users present, the optimal action is described by a switching curve if only if the value function $V(\cdot) = \lim_{m\to\infty} V_m(\cdot) - mg$, with g the minimum average cost, satisfies Properties A and B below. By symmetry, similar properties are required for the existence of a switching curve when there are no class-1 users.

Property A: If it is optimal to serve class 1 in state $(x_0, x_1, 0)$, then this is optimal in state $(x_0, x_1 + 1, 0)$ as well. Or equivalently, if

$$\begin{aligned}&(\mu_0 + \mu_2)V(x_0, x_1, 0) + \mu_1 V(x_0, x_1 - 1, 0)\\ &\leq \mu_0 V(x_0 - 1, x_1, 0) + (\mu_1 + \mu_2)V(x_0, x_1, 0),\end{aligned}$$

then

$$\begin{aligned}&(\mu_0 + \mu_2)V(x_0, x_1 + 1, 0) + \mu_1 V(x_0, x_1, 0)\\ &\leq \mu_0 V(x_0 - 1, x_1 + 1, 0) + (\mu_1 + \mu_2)V(x_0, x_1 + 1, 0).\end{aligned}$$

Note that this property is implied by the following inequality:

$$\begin{aligned}&\mu_0 V(x_0,x_1+1,0)+\mu_0 V(x_0-1,x_1,0)+2\mu_1 V(x_0,x_1,0)\\ &\le \mu_0 V(x_0,x_1,0)+\mu_0 V(x_0-1,x_1+1,0)+\mu_1 V(x_0,x_1-1,0)+\mu_1 V(x_0,x_1+1,0).\end{aligned} \tag{4.14}$$

Property B: If it is optimal to serve class 0 in state $(x_0,x_1,0)$, then this is optimal in state $(x_0+1,x_1,0)$ as well. Or equivalently, if

$$\begin{aligned}&\mu_0 V(x_0-1,x_1,0)+(\mu_1+\mu_2)V(x_0,x_1,0)\\ &\le (\mu_0+\mu_2)V(x_0,x_1,0)+\mu_1 V(x_0,x_1-1,0),\end{aligned}$$

then

$$\begin{aligned}&\mu_0 V(x_0,x_1,0)+(\mu_1+\mu_2)V(x_0+1,x_1,0)\\ &\le (\mu_0+\mu_2)V(x_0+1,x_1,0)+\mu_1 V(x_0+1,x_1-1,0).\end{aligned}$$

This property is implied by the following inequality:

$$\begin{aligned}&2\mu_0 V(x_0,x_1,0)+\mu_1 V(x_0+1,x_1,0)+\mu_1 V(x_0,x_1-1,0)\\ &\le \mu_0 V(x_0+1,x_1,0)+\mu_0 V(x_0-1,x_1,0)+\mu_1 V(x_0+1,x_1-1,0)+\mu_1 V(x_0,x_1,0).\end{aligned} \tag{4.15}$$

To derive the main result of a switching curve structure, we show that inequalities (4.14) and (4.15) (and hence Properties A and B), as well as the two analogous versions of them (when class 1 is empty), hold for $V_m(\cdot)$, $m=0,1,\ldots$. The proof, which may be found in Appendix 4.B, follows by induction on m. In each step, three auxiliary inequalities are proved as well, representing submodularity and supermodularity of the value functions [79].

Lemma 4.3.10. *Assume $C(\cdot)$ and $\tilde{C}(\cdot)$ are non-decreasing in x_0, x_1 and x_2. If $Z=C$ and $Z=\tilde{C}$ satisfy equation (4.13), as well as the following four inequalities, for all $x_0>0, x_1, x_2 \ge 0$,:*

$$\begin{aligned}&\mu_0 Z(\vec{x}+\vec{e}_1)+\mu_0 Z(\vec{x}-\vec{e}_0)+2\mu_1 Z(\vec{x})\\ &\le \mu_0 Z(\vec{x})+\mu_0 Z(\vec{x}-\vec{e}_0+\vec{e}_1)+\mu_1 Z(x_0,(x_1-1)^+,x_2)+\mu_1 Z(\vec{x}+\vec{e}_1),\end{aligned} \tag{4.16}$$

$$\begin{aligned}&\mu_0 Z(\vec{x}+\vec{e}_2)+\mu_0 Z(\vec{x}-\vec{e}_0)+2\mu_2 Z(\vec{x})\\ &\le \mu_0 Z(\vec{x})+\mu_0 Z(\vec{x}-\vec{e}_0+\vec{e}_2)+\mu_2 Z(x_0,x_1,(x_2-1)^+)+\mu_2 Z(\vec{x}+\vec{e}_2),\end{aligned} \tag{4.17}$$

$$\begin{aligned}&2\mu_0 Z(\vec{x})+\mu_1 Z(\vec{x}+\vec{e}_0)+\mu_1 Z(x_0,(x_1-1)^+,x_2)\\ &\le \mu_0 Z(\vec{x}+\vec{e}_0)+\mu_0 Z(\vec{x}-\vec{e}_0)+\mu_1 Z(x_0+1,(x_1-1)^+,x_2)+\mu_1 Z(\vec{x}),\end{aligned} \tag{4.18}$$

$$\begin{aligned}&2\mu_0 Z(\vec{x})+\mu_2 Z(\vec{x}+\vec{e}_0)+\mu_2 Z(x_0,x_1,(x_2-1)^+)\\ &\le \mu_0 Z(\vec{x}+\vec{e}_0)+\mu_0 Z(\vec{x}-\vec{e}_0)+\mu_2 Z(x_0+1,x_1,(x_2-1)^+)+\mu_2 Z(\vec{x}),\end{aligned} \tag{4.19}$$

and the following three properties, for all $x_0, x_1, x_2 \geq 0$,:

$$\textit{Supermodularity in the components } x_0 \textit{ and } x_2 : \tag{4.20}$$
$$Z(x_0, x_1, (x_2-1)^+) + Z((x_0-1)^+, x_1, x_2) \leq Z(\vec{x}) + Z((x_0-1)^+, x_1, (x_2-1)^+),$$

$$\textit{Supermodularity in the components } x_0 \textit{ and } x_1 : \tag{4.21}$$
$$Z(x_0, (x_1-1)^+, x_2) + Z((x_0-1)^+, x_1, x_2) \leq Z(\vec{x}) + Z((x_0-1)^+, (x_1-1)^+, x_2),$$

$$\textit{Submodularity in the components } x_1 \textit{ and } x_2 : \tag{4.22}$$
$$Z(x_0, (x_1-1)^+, (x_2-1)^+) + Z(\vec{x}) \leq Z(x_0, (x_1-1)^+, x_2) + Z(x_0, x_1, (x_2-1)^+),$$

then the same is true for $Z = V_m$ for all m.

We are now able to prove Proposition 4.3.9.

Proof of Proposition 4.3.9: We set $\tilde{C}(\vec{x}) = 0$. If $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, then the non-decreasing cost function $C(\vec{x}) = \sum_{j=0}^{2} c_j x_j$ satisfies all conditions in Lemmas 4.3.7 and 4.3.10. Hence, the minimizing action in (4.10) is to serve classes 1 and 2 whenever both are present (Lemma 4.3.7), independent of the value m. When choosing $x_2 = 0$ in (4.16) and (4.18) and $x_1 = 0$ in (4.17) and (4.19), we obtain the desired inequalities for the existence of a switching curve. In addition, if $c_0\mu_0 \geq c_i\mu_i$, $i \neq 0$, then $C(\vec{x}) = \sum_{j=0}^{2} c_j x_j$ satisfies the relation (4.11). Hence, in that case the minimizing action in (4.10) is to serve class 0 rather than class i alone. □

4.4 Numerical evaluation of α-fair policies

In this section we compare the performance of the optimal policy with that of α-fair bandwidth-sharing policies with unit weights $w_j = 1$, $j = 0, \ldots, L$, as defined in Section 1.4.1. Recall that when $\alpha = 1$, the α-fair policy is also referred to as the Proportional Fair (PF) policy. In that case, the steady-state distribution of the number of users is of product form and insensitive to the service requirements. In particular, the mean numbers of users equal [94]

$$\mathbb{E}(N_0^{PF}) = \frac{\rho_0}{1-\rho_0}\left(1 + \sum_{i=1}^{L} \frac{\rho_i}{1-\rho_0-\rho_i}\right), \tag{4.23}$$
$$\mathbb{E}(N_i^{PF}) = \frac{\rho_i}{1-\rho_0-\rho_i}, \quad i = 1, \ldots, L. \tag{4.24}$$

Comparing the total mean number of users under PF and a policy $\pi^* \in \Pi^* \cap \bar{\Pi}$ already provides important insight. Assuming exponential service requirements, from (4.1), (4.2), (4.23), and (4.24) we obtain that

$$\mathbb{E}(\sum_{j=0}^{L} c_j N_j^{\pi^*}) - \mathbb{E}(\sum_{j=0}^{L} c_j N_j^{PF}) = \frac{\rho_0}{1-\rho_0} \sum_{i=1}^{L} \frac{\rho_i}{1-\rho_0-\rho_i}\left(c_i \frac{\mu_i}{\mu_0} - c_0\right). \tag{4.25}$$

Note that for $\mu_0 < \bar{\mu}_0 := \frac{\sum_{i=1}^{L} \lambda_i c_i/(1-\rho_0-\rho_i)}{c_0 \sum_{i=1}^{L} \rho_i/(1-\rho_0-\rho_i)}$ (relatively large class-0 users), PF does better than π^*, and the difference in (4.25) is unbounded as $\mu_0 \downarrow 0$. For $\mu_0 > \bar{\mu}_0$ (relatively small class-0 users), it is better to prioritize class 0. In fact, π^* achieves the minimum mean holding cost among all strategies in $\bar{\Pi}$, when $c_0\mu_0 \geq \sum_{i=1}^{L} c_i\mu_i$. Still, the difference is bounded by $-\frac{\rho_0}{1-\rho_0} \sum_{i=1}^{L} \frac{c_0\rho_i}{1-\rho_0-\rho_i}$. Thus, PF performs reasonably well compared to π^* over a wide range of parameter values.

We now proceed to numerically investigate whether the latter finding holds in greater generality. We consider a linear network with two nodes ($L = 2$). For general α-fair bandwidth-sharing policies ($\alpha \neq 1$) we conduct simulations in order to estimate the mean numbers of users. The optimal policy is computed using value iteration after truncating the state space. In cases where the optimal policy is known explicitly, we verified that the results obtained by value iteration are accurate. We examined a wide range of scenarios. Since the results are qualitatively similar, we only present the results for the cases with $c_0 = c_1 = c_2 = 1$, $\rho_0 = 0.3$, $\rho_1 = 0.3$, $\mu_1 = 1$, $\mu_2 = 0.5$, with either (A) $\rho_2 = 0.2$ or (B) $\rho_2 = 0.5$, and varying μ_0. Throughout this section, we use the notation $N^\pi := \sum_{j=0}^{2} N_j^\pi$.

In Figure 4.1 we plot the total mean number of users under different policies as a function of μ_0 for cases A and B. We denote by $\pi(\alpha)$ the α-fair policy with parameter α and unit weights $w_j = 1$, $j = 0, \ldots, L$. The smallest total mean number of users among all α-fair policies, $\min_\alpha \mathbb{E}(N^{\pi(\alpha)})$, is labeled with “opt. α-fair”. The other curves correspond to the policies $\pi^* \in \Pi^* \cap \bar{\Pi}$, the optimal policy (“opt. policy”), and two α-fair bandwidth-sharing policies corresponding to $\alpha = 1$ (PF) and $\alpha = 2$. First of all, we note that for $\mu_0 \geq \mu_1 + \mu_2$ the queue length under π^* indeed coincides with that under the optimal allocation. Second, we see that the performance of the optimal α-fair policy compares well with that of the optimal policy in the moderately-loaded systems we consider. The gap does not exceed 20%. Apparently, α-fair policies succeed in *dynamically* adjusting the rate allocation in an efficient manner, without any knowledge of the service requirement parameters. Note that in this chapter we considered α-fair policies with unit weights. In Chapter 5 we compare the performance of weighted α-fair policies when the weights are chosen appropriately, and we will see that weighted α-fair policies are able to approximate the optimal policy rather well. In case of a heavily-loaded system however, the performance of α-fair policies can be arbitrarily worse compared to *anticipating* policies. This is the subject of Chapter 6.

In Figure 4.2 we plot the total mean number of users as a function of α for several values of μ_0, for cases A and B. A striking observation from the simulations is that as long as the value of α is not too small, the total mean number of users is fairly insensitive to the value of μ_0 (for fixed ρ_0). In addition, the differences within the class of α-fair policies are small: For $\alpha > 0.5$, the difference between the best and the worst α-fair policies is roughly 5% and 10%, in cases A and B, respectively.

As can be seen from Figure 4.2, in all cases the optimal value of α is either close to 0 (for small values of μ_0) or equals ∞ (for large values of μ_0). In fact, the performance under the α-fair policies appears to have monotonicity properties in α (either decreasing or increasing). This effect is further investigated in Chapter 7.

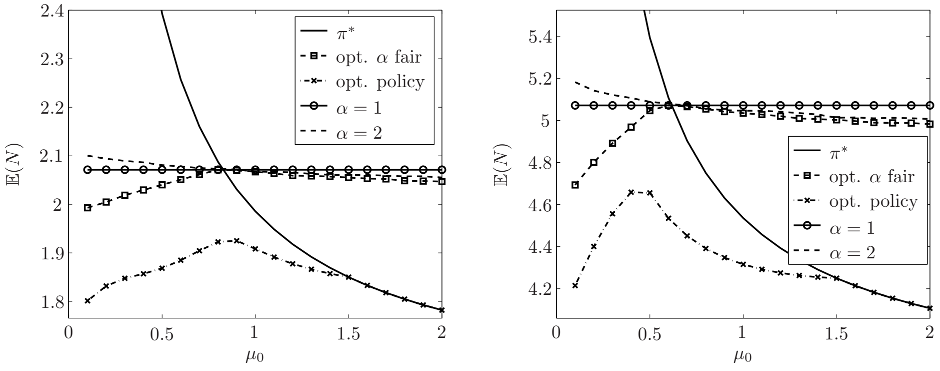
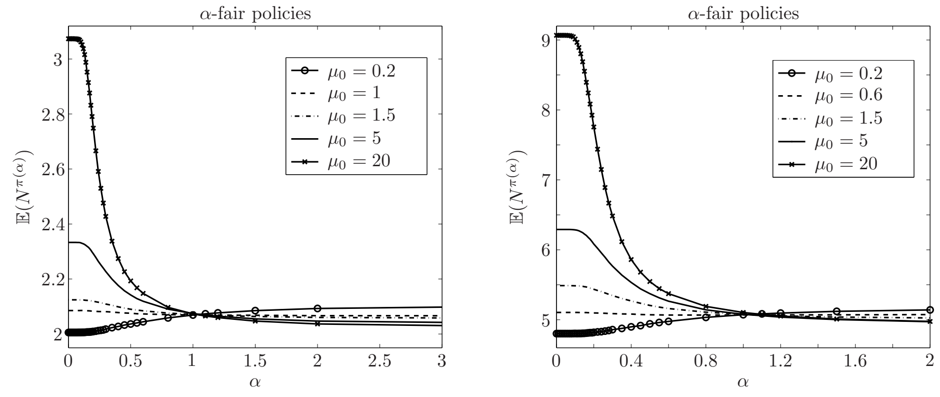


Figure 4.1: Total mean number of users in case A (left) and case B (right).


α-fair policies

α-fair policies

Figure 4.2: Total mean number of users in case A (left) and case B (right).

4.5 Concluding remarks

We have characterized optimal non-anticipating policies in a linear bandwidth-sharing network with exponential service requirements. These optimal scheduling policies require a high degree of coordination within the network as well as knowledge of the mean service requirements, which may prohibit actual implementation. As a benchmark, though, they are extremely useful to assess the effectiveness of other bandwidth-sharing policies, such as α-fair policies. In all our experiments we observed that for moderately loaded systems (i) the differences within the class of α-fair policies are not significant (as long as α is not too small), and (ii) these policies compare well with the optimal policies. A related result has been obtained in [161] for the special case of a star network with three links and three classes. The authors derive performance bounds for PF and for the optimal policy, and obtain that, in heavy traffic, the total mean number of users under PF is at most 3/2 times larger than the mean number of users under the optimal non-anticipating policy.

The optimality results obtained in this chapter concern rate allocation *across* classes, and do not involve the scheduling *within* classes. For exponential service requirements the performance is in fact independent of the employed non-anticipating intra-class policy. For general distributions however, this is not the case, and a well-chosen intra-class policy may significantly improve the performance. In Chapter 3 it was shown that standard size-based scheduling policies such as SRPT and LAS applied *across* all classes can cause instability effects. However, size-based scheduling *within* classes, i.e., size-based intra-class policies, may still produce substantial performance benefits, provided the rate allocation *across* classes is carefully arbitrated to avoid the above instability phenomena. Exactly how to combine size-based scheduling within classes, and what the potential gains might be, is a non-trivial issue. In Chapter 6 we investigate this for a linear network in heavy traffic.

Appendix

4.A Proofs of Lemmas 4.3.4, 4.3.5, and 4.3.7

In this section we prove the inequalities in Lemmas 4.3.4, 4.3.5, and 4.3.7 by induction on m. Obviously, for $Z = V_0$ they hold. The induction step consists in showing that when an inequality holds for $Z = V_m$ it holds for $Z = V_{m+1}$ as well. Define

$$\begin{aligned}\tilde{V}_{m+1}(\vec{x}) \quad &:= \quad V_{m+1}(\vec{x}) - C(\vec{x}) - \sum_{j=0}^{2} \lambda_j V_m(\vec{x} + \vec{e}_j) \\ &= \quad \min\{\mu_0 V_m((x_0-1)^+, x_1, x_2) + \mu_1 V_m(\vec{x}) + \mu_2 V_m(\vec{x}), \\ &\qquad \mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_2 V_m(x_0, x_1, (x_2-1)^+)\}.\end{aligned}$$

By assumption, $C(\vec{x}) + \sum_{j=0}^{2} \lambda_j V_m(\vec{x}+\vec{e}_j)$ satisfies the inequality. In order to prove that $V_{m+1}(\cdot)$ does so as well, it is therefore sufficient to show that $\tilde{V}_{m+1}(\cdot)$ does. This will be done in the remainder of this section.

Proof of Lemma 4.3.4: Assume inequality (4.11) holds for $Z = V_m$ with $i = 1$ and $j = 2$. We will prove that this holds as well for $Z = \tilde{V}_{m+1}$. (The proof for the case $i = 2$ and $j = 1$ follows similarly.)

Let $x_0, x_1 > 0$ and $x_2 \geq 0$. By definition, we have

$$\begin{aligned}&\mu_0 \tilde{V}_{m+1}(\vec{x} - \vec{e}_0) + \mu_1 \tilde{V}_{m+1}(\vec{x}) \\ &\leq \mu_0[\mu_0 V_m(\vec{x} - \vec{e}_0) + \mu_1 V_m(\vec{x} - \vec{e}_0 - \vec{e}_1) + \mu_2 V_m(x_0 - 1, x_1, (x_2-1)^+)] \\ &\quad + \mu_1[\mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x} - \vec{e}_1) + \mu_2 V_m(x_0, x_1, (x_2-1)^+)] \\ &= \mu_0[\mu_0 V_m(\vec{x} - \vec{e}_0) + \mu_1 V_m(\vec{x})] \\ &\quad + \mu_1[\mu_0 V_m(\vec{x} - \vec{e}_0 - \vec{e}_1) + \mu_1 V_m(\vec{x} - \vec{e}_1)] \\ &\quad + \mu_2[\mu_0 V_m(x_0 - 1, x_1, (x_2-1)^+) + \mu_1 V_m(x_0, x_1, (x_2-1)^+)]. \qquad (4.26)\end{aligned}$$

For the state $\vec{x}$ we do not know which action is the minimizer in $\tilde{V}_{m+1}(\vec{x})$. If serving class 0 is the minimizer, then

$$\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x}) = \tilde{V}_{m+1}(\vec{x}) - \mu_2 V_m(\vec{x}).$$

If serving classes 1 and 2 is the minimizer, then

$$\begin{aligned}\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x}) &\leq \mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x}-\vec{e}_1)\\ &= V_{m+1}(\vec{x}) - \mu_2 V_m(x_0, x_1, (x_2-1)^+),\end{aligned}$$

where the inequality follows since (4.11) holds for $V_m(\cdot)$. By Lemma 4.3.3, $V_m(\cdot)$ is non-decreasing, so we can conclude that

$$\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x}) \leq \tilde{V}_{m+1}(\vec{x}) - \mu_2 V_m(x_0, x_1, (x_2-1)^+). \quad (4.27)$$

Similarly for the state $\vec{x}-\vec{e}_1$ we deduce that

$$\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_1) \leq \tilde{V}_{m+1}(\vec{x}-\vec{e}_1) - \mu_2 V_m(x_0, x_1-1, (x_2-1)^+).$$

Together with (4.26) and (4.27) this gives

$$\begin{aligned}&\mu_0 \tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + \mu_1 \tilde{V}_{m+1}(\vec{x})\\ &\leq \mu_0 \tilde{V}_{m+1}(\vec{x}) + \mu_1 \tilde{V}_{m+1}(\vec{x}-\vec{e}_1) + \mu_2 \left[\mu_0 V_m(x_0-1, x_1, (x_2-1)^+)\right.\\ &\quad \left.+(\mu_1-\mu_0) V_m(x_0, x_1, (x_2-1)^+) - \mu_1 V_m(x_0, x_1-1, (x_2-1)^+)\right]\\ &\leq \mu_0 \tilde{V}_{m+1}(\vec{x}) + \mu_1 \tilde{V}_{m+1}(\vec{x}-\vec{e}_1),\end{aligned}$$

where the last inequality follows since (4.11) holds for $V_m(\cdot)$. This proves that $\tilde{V}_{m+1}(\cdot)$ satisfies (4.11). □

Proof of Lemma 4.3.5: Assume inequality (4.12) holds for $Z = V_m$. We prove that this holds as well for $Z = \tilde{V}_{m+1}$. Let $x_0, x_1, x_2 > 0$. We have

$$\begin{aligned}&\mu_0 \tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)\tilde{V}_{m+1}(\vec{x})\\ &\leq \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x}-\vec{e}_0-\vec{e}_1) + \mu_2 V_m(\vec{x}-\vec{e}_0-\vec{e}_2)]\\ &\quad +(\mu_1+\mu_2)[\mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 V_m(\vec{x}-\vec{e}_2)]\\ &= \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2) V_m(\vec{x})] && (4.28)\\ &\quad +\mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1) + (\mu_1+\mu_2) V_m(\vec{x}-\vec{e}_1)] && (4.29)\\ &\quad +\mu_2[\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) + (\mu_1+\mu_2) V_m(\vec{x}-\vec{e}_2)]. && (4.30)\end{aligned}$$

By (4.12), the expression (4.28) is equal to $\mu_0 \tilde{V}_{m+1}(\vec{x})$. Note that Lemma 4.3.4 implies that $V_m(\cdot)$ satisfies (4.11). If $x_1 = 1$, it follows from (4.11) that (4.29) is equal to $\mu_1 \tilde{V}_{m+1}(\vec{x}-\vec{e}_1)$. If $x_1 > 1$, it follows from (4.12) that (4.29) is equal to $\mu_1 \tilde{V}_{m+1}(\vec{x}-\vec{e}_1)$ as well. Similarly, the expression in (4.30) is equal to $\mu_2 \tilde{V}_{m+1}(\vec{x}-\vec{e}_2)$, because of (4.11) and (4.12). This implies that $Z = \tilde{V}_{m+1}$ satisfies (4.12). □

Proof of Lemma 4.3.7: Assume inequality (4.13) holds for $Z = V_m$. We prove that this holds as well for $Z = \tilde{V}_{m+1}$. Let $x_0, x_1, x_2 > 0$. We have

$$\begin{aligned}
&\mu_0 \tilde{V}_{m+1}(\vec{x}) + \mu_1 \tilde{V}_{m+1}(\vec{x} - \vec{e}_1) + \mu_2 \tilde{V}_{m+1}(\vec{x} - \vec{e}_2) \\
&\leq \mu_0[\mu_0 V_m(\vec{x} - \vec{e}_0) + \mu_1 V_m(\vec{x}) + \mu_2 V_m(\vec{x})] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x} - \vec{e}_0 - \vec{e}_1) + \mu_1 V_m(\vec{x} - \vec{e}_1) + \mu_2 V_m(\vec{x} - \vec{e}_1)] \\
&\quad + \mu_2[\mu_0 V_m(\vec{x} - \vec{e}_0 - \vec{e}_2) + \mu_1 V_m(\vec{x} - \vec{e}_2) + \mu_2 V_m(\vec{x} - \vec{e}_2)] \\
&= \mu_0[\mu_0 V_m(\vec{x} - \vec{e}_0) + \mu_1 V_m(\vec{x} - \vec{e}_0 - \vec{e}_1) + \mu_2 V_m(\vec{x} - \vec{e}_0 - \vec{e}_2)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x} - \vec{e}_1) + \mu_2 V_m(\vec{x} - \vec{e}_2)] \\
&\quad + \mu_2[\mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x} - \vec{e}_1) + \mu_2 V_m(\vec{x} - \vec{e}_2)].
\end{aligned} \tag{4.31}$$

Since $V_m(\cdot)$ satisfies (4.13) and is non-decreasing (by Lemma 4.3.3), the expression in (4.31) is equal to $\mu_0 \tilde{V}_{m+1}(\vec{x} - \vec{e}_0) + \mu_1 \tilde{V}_{m+1}(\vec{x}) + \mu_2 \tilde{V}_{m+1}(\vec{x})$. Hence, $Z = \tilde{V}_{m+1}$ satisfies (4.13). □

4.B Proof of Lemma 4.3.10

In this section we prove Lemma 4.3.10 by induction on m. For $Z = V_0$ it holds. Suppose we know that the inequalities (4.16)–(4.22) hold for $Z = V_m$. We have to show that they hold for $Z = V_{m+1}$. Since the inequalities hold for both $C(\cdot)$ and $V_m(\cdot)$, it is straightforward to check that $C(\vec{x}) + \sum_{j=0}^{2} \lambda_j V_m(\vec{x} + \vec{e}_j)$ satisfies these inequalities as well. In order to prove that $V_{m+1}(\cdot)$ satisfies (4.16)–(4.22), it is therefore sufficient to show that $\tilde{V}_{m+1}(\cdot)$ (as defined in Appendix 4.A) does.

The following observation can be made, which will be helpful in proving (4.16)–(4.22) for $\tilde{V}_{m+1}(\cdot)$. The inequalities being true for $Z = V_m$ implies that at time $m+1$ the optimal actions are of a switching curve structure. So for example, if at time $m+1$ it is optimal to serve class 1 when we are in state $\vec{x}$, this is also optimal if at time $m+1$ we are in state $\vec{x} + \vec{e}_1$. This property will be referred to as $\tilde{V}_{m+1}(\cdot)$ following a switching curve.

Proof of inequality (4.16): We have to show that $Z = \tilde{V}_{m+1}$ satisfies (4.16). In order to prove this, we need to distinguish between which actions are optimal in the states $\vec{x}, \vec{x} - \vec{e}_0 + \vec{e}_1, (x_0, (x_1 - 1)^+, x_2)$ and $\vec{x} + \vec{e}_1$ at $m+1$ steps from the horizon. In every state there are two possibilities, either serve classes 1 and 2, or serve class 0. Since $\tilde{V}_{m+1}(\cdot)$ follows a switching curve, only the following five combinations of optimal actions in the various states can occur: In situation 1 it is optimal to serve classes 1 and 2 in states $\vec{x}, \vec{x} - \vec{e}_0 + \vec{e}_1$ and $\vec{x} + \vec{e}_1$, and class 0 in state $(x_0, (x_1 - 1)^+, x_2)$. In situation 2 it is optimal to serve classes 1 and 2 in states $\vec{x} - \vec{e}_0 + \vec{e}_1$ and $\vec{x} + \vec{e}_1$, and class 0 in states $\vec{x}$ and $(x_0, (x_1 - 1)^+, x_2)$. In situation 3 it is optimal to serve classes 1 and 2 in state $\vec{x} - \vec{e}_0 + \vec{e}_1$ and serve class 0 in the other three states. In situation 4 it is optimal to serve class 0 in all four states and in situation 5 it is optimal to serve classes 1 and 2 in all four states. For each of the five possible situations we will show that $Z = \tilde{V}_{m+1}$ satisfies (4.16).

Situation 1: Let $x_0 > 0, x_1, x_2 \geq 0$. We can write

$$\begin{aligned}
&\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1\tilde{V}_{m+1}(\vec{x})\\
&\leq \mu_0[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_1 V_m(\vec{x}) + \mu_2 V_m(x_0, x_1+1, (x_2-1)^+)]\\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1, (x_1-1)^+, x_2) + \mu_2 V_m(x_0-1, x_1, (x_2-1)^+)]\\
&\quad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2) V_m(\vec{x})]\\
&\quad + \mu_1[\mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_2 V_m(x_0, x_1, (x_2-1)^+)]. \qquad (4.32)
\end{aligned}$$

For the terms in (4.32) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[\mu_0 V_m(x_0, x_1+1, (x_2-1)^+) + \mu_0 V_m(x_0-1, x_1, (x_2-1)^+)\\
&\quad + \mu_1 V_m(\vec{x}) + \mu_1 V_m(x_0, x_1, (x_2-1)^+)]\\
&\leq \mu_2[\mu_0 V_m(x_0, x_1, (x_2-1)^+) + \mu_0 V_m(x_0-1, x_1+1, (x_2-1)^+)\\
&\qquad + \mu_1 V_m(x_0, (x_1-1)^+, (x_2-1)^+) + \mu_1 V_m(x_0, x_1+1, (x_2-1)^+)\\
&\qquad - \mu_1 V_m(x_0, x_1, (x_2-1)^+) + \mu_1 V_m(\vec{x})]\\
&\leq \mu_2[\mu_0 V_m(x_0, x_1, (x_2-1)^+) + \mu_0 V_m(x_0-1, x_1+1, (x_2-1)^+)\\
&\qquad + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_1 V_m(x_0, x_1+1, (x_2-1)^+)], \qquad (4.33)
\end{aligned}$$

where the first inequality follows from (4.16) and the second from (4.22). By (4.16) the terms in (4.32) without a factor μ_2 are smaller than or equal to

$$\begin{aligned}
&\mu_0[\mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0, (x_1-1)^+, x_2)] + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_0)]\\
&+ \mu_1[\mu_0 V_m(x_0-1, (x_1-1)^+, x_2) + \mu_1 V_m(x_0, (x_1-1)^+, x_2)]\\
&+ \mu_1[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_1 V_m(\vec{x})].
\end{aligned}$$

Together with (4.32) and (4.33) this yields

$$\begin{aligned}
&\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1\tilde{V}_{m+1}(\vec{x})\\
&\leq \mu_0[\mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_2 V_m(x_0, x_1, (x_2-1)^+)]\\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_0) + \mu_2 V_m(x_0-1, x_1+1, (x_2-1)^+)]\\
&\quad + \mu_1[\mu_0 V_m(x_0-1, (x_1-1)^+, x_2) + (\mu_1+\mu_2) V_m(x_0, (x_1-1)^+, x_2)]\\
&\quad + \mu_1[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_1 V_m(\vec{x}) + \mu_2 V_m(x_0, x_1+1, (x_2-1)^+)]\\
&= \mu_0\tilde{V}_{m+1}(\vec{x}) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0+\vec{e}_1)\\
&\quad + \mu_1\tilde{V}_{m+1}(x_0, (x_1-1)^+, x_2) + \mu_1\tilde{V}_{m+1}(\vec{x}+\vec{e}_1),
\end{aligned}$$

which was to be proved. In the last step we used that in situation 1 it is optimal to serve classes 1 and 2 in states $\vec{x}, \vec{x}-\vec{e}_0+\vec{e}_1$ and $\vec{x}+\vec{e}_1$, and class 0 in state $(x_0, (x_1-1)^+, x_2)$ at $m+1$ steps from the horizon.

Situation 2: Let $x_0 > 0, x_1, x_2 \geq 0$. We can write

$$\begin{aligned}
&\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1\tilde{V}_{m+1}(\vec{x}) \\
&\leq \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + (\mu_1+\mu_2)V_m(\vec{x}+\vec{e}_1)] \\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1,(x_1-1)^+,x_2) + \mu_2 V_m(x_0-1,x_1,(x_2-1)^+)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)V_m(\vec{x})] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_2 V_m(x_0,x_1,(x_2-1)^+)]. \qquad (4.34)
\end{aligned}$$

For the terms in (4.34) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_0 V_m(x_0-1,x_1,(x_2-1)^+) \\
&\qquad + \mu_1 V_m(\vec{x}) + \mu_1 V_m(x_0,x_1,(x_2-1)^+)] \\
&\leq \mu_2[\mu_0 V_m(x_0-1,x_1,(x_2-1)^+) - \mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0,x_1,(x_2-1)^+) \\
&\qquad - \mu_1 V_m(\vec{x}) + \mu_0 V_m(\vec{x}) + \mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) \\
&\qquad + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_1 V_m(\vec{x}+\vec{e}_1)] \\
&\leq \mu_2[\mu_0 V_m(\vec{x}) + \mu_0 V_m(x_0-1,x_1+1,(x_2-1)^+) \\
&\qquad + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_1 V_m(x_0,x_1+1,(x_2-1)^+)],
\end{aligned}$$

where the first inequality follows from (4.16) and the second from (4.22). This inequality together with (4.34) yields

$$\begin{aligned}
&\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1\tilde{V}_{m+1}(\vec{x}) \\
&\leq \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)V_m(\vec{x})] \\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_0) + \mu_2 V_m(x_0-1,x_1+1,(x_2-1)^+)] \\
&\quad + \mu_1[\mu_0 V_m(x_0-1,(x_1-1)^+,x_2) + (\mu_1+\mu_2)V_m(x_0,(x_1-1)^+,x_2)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_1 V_m(\vec{x}) + \mu_2 V_m(x_0,x_1+1,(x_2-1)^+)] \\
&= \mu_0\tilde{V}_{m+1}(\vec{x}) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0+\vec{e}_1) \\
&\quad + \mu_1\tilde{V}_{m+1}(x_0,(x_1-1)^+,x_2) + \mu_1\tilde{V}_{m+1}(\vec{x}+\vec{e}_1),
\end{aligned}$$

which was to be proved. In the last step we used that in situation 2 it is optimal to serve classes 1 and 2 in states $\vec{x}-\vec{e}_0+\vec{e}_1$ and $\vec{x}+\vec{e}_1$, and class 0 in states $\vec{x}$ and $(x_0,(x_1-1)^+,x_2)$ at $m+1$ steps from the horizon.

Situation 3: Let $x_0 > 0, x_1, x_2 \geq 0$. We can write

$$\begin{aligned}
&\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1\tilde{V}_{m+1}(\vec{x}) \\
&\leq \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + (\mu_1+\mu_2)V_m(\vec{x}+\vec{e}_1)] \\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1,(x_1-1)^+,x_2) + \mu_2 V_m(x_0-1,x_1,(x_2-1)^+)] \\
&\quad + 2\mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)V_m(\vec{x})]. \qquad (4.35)
\end{aligned}$$

For the terms in (4.35) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[\mu_0 V_m(\vec{x}+\vec{e}_1) + \mu_0 V_m(x_0-1, x_1, (x_2-1)^+) + 2\mu_1 V_m(\vec{x})] \\
&\quad \le \mu_2[\mu_0 V_m(\vec{x}) + \mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_1 V_m(\vec{x}+\vec{e}_1) \\
&\qquad\quad + \mu_0 V_m(x_0-1, x_1, (x_2-1)^+) - \mu_0 V_m(\vec{x}-\vec{e}_0)] \\
&\quad \le \mu_2[\mu_0 V_m(\vec{x}) + \mu_0 V_m(x_0-1, x_1+1, (x_2-1)^+) \\
&\qquad\quad + \mu_1 V_m(x_0, (x_1-1)^+, x_2) + \mu_1 V_m(\vec{x}+\vec{e}_1)],
\end{aligned} \tag{4.36}$$

where the first inequality follows from (4.16) and the second from (4.22). By (4.16) the terms in (4.35) without a factor μ_2 are smaller than or equal to

$$\begin{aligned}
&\mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x})] + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_0)] \\
&\quad + \mu_1[\mu_0 V_m(x_0-1, (x_1-1)^+, x_2) + \mu_1 V_m(x_0, (x_1-1)^+, x_2)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}+\vec{e}_1)].
\end{aligned}$$

Together with (4.35) and (4.36) this yields

$$\begin{aligned}
&\mu_0 \tilde{V}_{m+1}(\vec{x}+\vec{e}_1) + \mu_0 \tilde{V}_{m+1}(\vec{x}-\vec{e}_0) + 2\mu_1 \tilde{V}_{m+1}(\vec{x}) \\
&\quad \le \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2) V_m(\vec{x})] \\
&\qquad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + \mu_1 V_m(\vec{x}-\vec{e}_0) + \mu_2 V_m(x_0-1, x_1+1, (x_2-1)^+)] \\
&\qquad + \mu_1[\mu_0 V_m(x_0-1, (x_1-1)^+, x_2) + (\mu_1+\mu_2) V_m(x_0, (x_1-1)^+, x_2)] \\
&\qquad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0+\vec{e}_1) + (\mu_1+\mu_2) V_m(\vec{x}+\vec{e}_1)] \\
&\quad = \mu_0 \tilde{V}_{m+1}(\vec{x}) + \mu_0 \tilde{V}_{m+1}(\vec{x}-\vec{e}_0+\vec{e}_1) \\
&\qquad + \mu_1 \tilde{V}_{m+1}(x_0, (x_1-1)^+, x_2) + \mu_1 \tilde{V}_{m+1}(\vec{x}+\vec{e}_1),
\end{aligned}$$

which was to be proved. In the last step we used that in situation 3 it is optimal to serve classes 1 and 2 in state $\vec{x}-\vec{e}_0+\vec{e}_1$ and serve class 0 in the other states, at $m+1$ steps from the horizon.
Situation 4: In this case, it is optimal to serve class 0 at $m+1$ steps from the horizon in the states $\vec{x}, \vec{x}-\vec{e}_0+\vec{e}_1, (x_0, (x_1-1)^+, x_2)$ and $\vec{x}+\vec{e}_1$. We can only be in this situation if $x_0 > 1$, since $V_m(\cdot)$ is non-decreasing in x_0, x_1 and x_2. Since (4.16) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.
Situation 5: In this case, it is optimal to serve classes 1 and 2 at $m+1$ steps from the horizon in the states $\vec{x}, \vec{x}-\vec{e}_0+\vec{e}_1, (x_0, (x_1-1)^+, x_2)$ and $\vec{x}+\vec{e}_1$. Since (4.16) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.

Proof of inequality (4.17): This goes along similar lines as the proof of (4.16).

Proof of inequality (4.18): We have to show that $Z = \tilde{V}_{m+1}$ satisfies (4.18). In order to prove this, we need to distinguish between which actions are optimal in the states $\vec{x}+\vec{e}_0, \vec{x}-\vec{e}_0, (x_0+1, (x_1-1)^+, x_2)$ and $\vec{x}$ at $m+1$ steps from the horizon. Since the optimal actions at time $m+1$ have a switching curve structure (as explained earlier), there are exactly five possible situations:

Situation 1: In this case it is optimal to serve classes 1 and 2 in states $\vec{x}+\vec{e}_0$, $\vec{x}-\vec{e}_0$, and $\vec{x}$, and serve class 0 in state $(x_0+1,(x_1-1)^+,x_2)$. Let $x_0>0, x_1,x_2\geq 0$. We can write

$$\begin{aligned}
&2\mu_0\tilde{V}_{m+1}(\vec{x})+\mu_1\tilde{V}_{m+1}(\vec{x}+\vec{e}_0)+\mu_1\tilde{V}_{m+1}(x_0,(x_1-1)^+,x_2)\\
&\leq 2\mu_0[\mu_0V_m(\vec{x})+\mu_1V_m(x_0,(x_1-1)^+,x_2)+\mu_2V_m(x_0,x_1,(x_2-1)^+)]\\
&\quad+\mu_1[\mu_0V_m(\vec{x}+\vec{e}_0)+\mu_1V_m(x_0+1,(x_1-1)^+,x_2)+\mu_2V_m(x_0+1,x_1,(x_2-1)^+)]\\
&\quad+\mu_1[\mu_0V_m(x_0-1,(x_1-1)^+,x_2)+(\mu_1+\mu_2)V_m(x_0,(x_1-1)^+,x_2)]. \qquad (4.37)
\end{aligned}$$

For the terms in (4.37) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[2\mu_0V_m(x_0,x_1,(x_2-1)^+)+\mu_1V_m(x_0+1,x_1,(x_2-1)^+)\\
&\quad+\mu_1V_m(x_0,(x_1-1)^+,x_2)]\\
&\leq\mu_2[\mu_0V_m(x_0+1,x_1,(x_2-1)^+)+\mu_0V_m(x_0-1,x_1,(x_2-1)^+)\\
&\qquad+\mu_1V_m(x_0+1,(x_1-1)^+,(x_2-1)^+)+\mu_1V_m(x_0,x_1,(x_2-1)^+)\\
&\qquad-\mu_1V_m(x_0,(x_1-1)^+,(x_2-1)^+)+\mu_1V_m(x_0,(x_1-1)^+,x_2)]\\
&\leq\mu_2[\mu_0V_m(x_0+1,x_1,(x_2-1)^+)+\mu_0V_m(x_0-1,x_1,(x_2-1)^+)\\
&\qquad+\mu_1V_m(x_0+1,(x_1-1)^+,x_2)+\mu_1V_m(x_0,x_1,(x_2-1)^+)], \qquad (4.38)
\end{aligned}$$

where the first inequality follows from (4.18) and the second from (4.20). In addition, by (4.18) we have

$$\begin{aligned}
&\mu_0[2\mu_0V_m(\vec{x})+\mu_1V_m(\vec{x}+\vec{e}_0)+\mu_1V_m(x_0,(x_1-1)^+,x_2)]\\
&\leq\mu_0[\mu_0V_m(\vec{x}+\vec{e}_0)+\mu_0V_m(\vec{x}-\vec{e}_0)+\mu_1V_m(x_0+1,(x_1-1)^+,x_2)+\mu_1V_m(\vec{x})].
\end{aligned}$$

Together with (4.38) this gives that (4.37) is not larger than

$$\begin{aligned}
&\mu_0[\mu_0V_m(\vec{x}+\vec{e}_0)+\mu_1V_m(x_0+1,(x_1-1)^+,x_2)+\mu_2V_m(x_0+1,x_1,(x_2-1)^+)]\\
&+\mu_0[\mu_0V_m(\vec{x}-\vec{e}_0)+\mu_1V_m(x_0-1,(x_1-1)^+,x_2)+\mu_2V_m(x_0-1,x_1,(x_2-1)^+)]\\
&+\mu_1[\mu_0V_m(x_0,(x_1-1)^+,x_2)+(\mu_1+\mu_2)V_m(x_0+1,(x_1-1)^+,x_2)]\\
&+\mu_1[\mu_0V_m(\vec{x})+\mu_1V_m(x_0,(x_1-1)^+,x_2)+\mu_2V_m(x_0,x_1,(x_2-1)^+)]\\
&=\mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_0)+\mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0)\\
&\qquad+\mu_1\tilde{V}_{m+1}(x_0+1,(x_1-1)^+,x_2)+\mu_1\tilde{V}_{m+1}(\vec{x}).
\end{aligned}$$

Situation 2: In this case it is optimal to serve classes 1 and 2 in states $\vec{x}-\vec{e}_0$ and $\vec{x}$, and serve class 0 in states $\vec{x}+\vec{e}_0$ and $(x_0+1,(x_1-1)^+,x_2)$. Let $x_0>0$, $x_1,x_2\geq 0$.

We can write

$$\begin{aligned}
&2\mu_0\tilde{V}_{m+1}(\vec{x}) + \mu_1\tilde{V}_{m+1}(\vec{x}+\vec{e}_0) + \mu_1\tilde{V}_{m+1}(x_0,(x_1-1)^+,x_2)\\
&\le \mu_0[\mu_0V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)V_m(\vec{x})]\\
&\quad + \mu_0[\mu_0V_m(\vec{x}) + \mu_1V_m(x_0,(x_1-1)^+,x_2) + \mu_2V_m(x_0,x_1,(x_2-1)^+)]\\
&\quad + \mu_1[\mu_0V_m(\vec{x}+\vec{e}_0) + \mu_1V_m(x_0+1,(x_1-1)^+,x_2)\\
&\qquad\quad + \mu_2V_m(x_0+1,x_1,(x_2-1)^+)]\\
&\quad + \mu_1[\mu_0V_m(x_0-1,(x_1-1)^+,x_2) + (\mu_1+\mu_2)V_m(x_0,(x_1-1)^+,x_2)]. \qquad (4.39)
\end{aligned}$$

For the terms in (4.39) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[\mu_0V_m(\vec{x}) + \mu_0V_m(x_0,x_1,(x_2-1)^+)\\
&\quad + \mu_1V_m(x_0+1,x_1,(x_2-1)^+) + \mu_1V_m(x_0,(x_1-1)^+,x_2)]\\
&\le \mu_2[\mu_0V_m(x_0,x_1,(x_2-1)^+) - \mu_0V_m(\vec{x}) + \mu_1V_m(x_0+1,x_1,(x_2-1)^+)\\
&\qquad - \mu_1V_m(\vec{x}+\vec{e}_0) + \mu_0V_m(\vec{x}+\vec{e}_0) + \mu_0V_m(\vec{x}-\vec{e}_0)\\
&\qquad + \mu_1V_m(x_0+1,(x_1-1)^+,x_2) + \mu_1V_m(\vec{x})]\\
&\le \mu_2[\mu_0V_m(\vec{x}+\vec{e}_0) + \mu_0V_m(x_0-1,x_1,(x_2-1)^+)\\
&\qquad + \mu_1V_m(x_0+1,(x_1-1)^+,x_2) + \mu_1V_m(x_0,x_1,(x_2-1)^+)],
\end{aligned}$$

where the first inequality follows from (4.18) and the second from (4.20). We can conclude that (4.39) is not larger than

$$\begin{aligned}
&\mu_0[\mu_0V_m(\vec{x}) + (\mu_1+\mu_2)V_m(\vec{x}+\vec{e}_0)]\\
&+ \mu_0[\mu_0V_m(\vec{x}-\vec{e}_0) + \mu_1V_m(x_0-1,(x_1-1)^+,x_2) + \mu_2V_m(x_0-1,x_1,(x_2-1)^+)]\\
&+ \mu_1[\mu_0V_m(x_0,(x_1-1)^+,x_2) + (\mu_1+\mu_2)V_m(x_0+1,(x_1-1)^+,x_2)]\\
&+ \mu_1[\mu_0V_m(\vec{x}) + \mu_1V_m(x_0,(x_1-1)^+,x_2) + \mu_2V_m(x_0,x_1,(x_2-1)^+)]\\
&= \mu_0\tilde{V}_{m+1}(\vec{x}+\vec{e}_0) + \mu_0\tilde{V}_{m+1}(\vec{x}-\vec{e}_0)\\
&\quad + \mu_1\tilde{V}_{m+1}(x_0+1,(x_1-1)^+,x_2) + \mu_1\tilde{V}_{m+1}(\vec{x}).
\end{aligned}$$

Situation 3: In this case it is optimal to serve classes 1 and 2 in state $\vec{x}-\vec{e}_0$, and serve class 0 in states $\vec{x}$, $\vec{x}+\vec{e}_0$ and $(x_0+1,(x_1-1)^+,x_2)$. Let $x_0>0$, $x_1,x_2\ge 0$. We can write

$$\begin{aligned}
&2\mu_0\tilde{V}_{m+1}(\vec{x}) + \mu_1\tilde{V}_{m+1}(\vec{x}+\vec{e}_0) + \mu_1\tilde{V}_{m+1}(x_0,(x_1-1)^+,x_2)\\
&\le \mu_0[\mu_0V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2)V_m(\vec{x})]\\
&\quad + \mu_0[\mu_0V_m(\vec{x}) + \mu_1V_m(x_0,(x_1-1)^+,x_2) + \mu_2V_m(x_0,x_1,(x_2-1)^+)]\\
&\quad + \mu_1[\mu_0V_m(\vec{x}) + (\mu_1+\mu_2)V_m(\vec{x}+\vec{e}_0)]\\
&\quad + \mu_1[\mu_0V_m(x_0-1,(x_1-1)^+,x_2) + (\mu_1+\mu_2)V_m(x_0,(x_1-1)^+,x_2)]. \qquad (4.40)
\end{aligned}$$

For the terms in (4.40) with a factor μ_2, we have

$$\begin{aligned}
&\mu_2[\mu_0 V_m(\vec{x}) + \mu_0 V_m(x_0, x_1, (x_2-1)^+) + \mu_1 V_m(\vec{x}+\vec{e}_0) + \mu_1 V_m(x_0, (x_1-1)^+, x_2)] \\
&\quad \le \mu_2[\mu_0 V_m(\vec{x}+\vec{e}_0) + \mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0+1, (x_1-1)^+, x_2) + \mu_1 V_m(\vec{x}) \\
&\qquad\quad - \mu_0 V_m(\vec{x}) + \mu_0 V_m(x_0, x_1, (x_2-1)^+)] \\
&\quad \le \mu_2[\mu_0 V_m(\vec{x}+\vec{e}_0) + \mu_0 V_m(x_0-1, x_1, (x_2-1)^+) \\
&\qquad\quad + \mu_1 V_m(x_0+1, (x_1-1)^+, x_2) + \mu_1 V_m(\vec{x})],
\end{aligned} \tag{4.41}$$

where the first inequality follows from (4.18) and the second from (4.20). By (4.18), the terms in (4.40) without a factor μ_2 are smaller than or equal to

$$\begin{aligned}
&\mu_0[\mu_0 V_m(\vec{x}) + \mu_1 V_m(\vec{x}+\vec{e}_0)] + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1, (x_1-1)^+, x_2)] \\
&\quad + \mu_1[\mu_0 V_m(x_0, (x_1-1)^+, x_2) + \mu_1 V_m(x_0+1, (x_1-1)^+, x_2)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(\vec{x})].
\end{aligned}$$

Together with (4.41) this gives that (4.40) is not larger than

$$\begin{aligned}
&\mu_0[\mu_0 V_m(\vec{x}) + (\mu_1+\mu_2) V_m(\vec{x}+\vec{e}_0)] \\
&\quad + \mu_0[\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1, (x_1-1)^+, x_2) + \mu_2 V_m(x_0-1, x_1, (x_2-1)^+)] \\
&\quad + \mu_1[\mu_0 V_m(x_0, (x_1-1)^+, x_2) + (\mu_1+\mu_2) V_m(x_0+1, (x_1-1)^+, x_2)] \\
&\quad + \mu_1[\mu_0 V_m(\vec{x}-\vec{e}_0) + (\mu_1+\mu_2) V_m(\vec{x})] \\
&\quad = \mu_0 \tilde{V}_{m+1}(\vec{x}+\vec{e}_0) + \mu_0 \tilde{V}_{m+1}(\vec{x}-\vec{e}_0) \\
&\qquad + \mu_1 \tilde{V}_{m+1}(x_0+1, (x_1-1)^+, x_2) + \mu_1 \tilde{V}_{m+1}(\vec{x}).
\end{aligned}$$

Situation 4: In this case it is optimal to serve class 0 in states $\vec{x}+\vec{e}_0$, $\vec{x}-\vec{e}_0$, $(x_0+1, (x_1-1)^+, x_2)$ and $\vec{x}$. We can only be in this situation if $x_0 > 1$. Since (4.18) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.
Situation 5: In this case it is optimal to serve classes 1 and 2 in states $\vec{x}+\vec{e}_0$, $\vec{x}-\vec{e}_0$, $(x_0+1, (x_1-1)^+, x_2)$ and $\vec{x}$. Since (4.18) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.

Proof of inequality (4.19): This goes along similar lines as the proof of (4.18).

Proof of inequality (4.20): We have to show that $Z = \tilde{V}_{m+1}$ satisfies (4.20), i.e.,

$$\begin{aligned}
&\tilde{V}_{m+1}(x_0, x_1, (x_2-1)^+) + \tilde{V}_{m+1}((x_0-1)^+, x_1, x_2) \\
&\quad \le \tilde{V}_{m+1}(\vec{x}) + \tilde{V}_{m+1}((x_0-1)^+, x_1, (x_2-1)^+).
\end{aligned} \tag{4.42}$$

For $x_0 = 0$ or $x_2 = 0$ this is trivially true. Hence, we assume that $x_0 > 0$ and $x_2 > 0$. The value of the right-hand side of (4.42) depends on which of the two actions are optimal in the states $\vec{x}$ and $\vec{x}-\vec{e}_0-\vec{e}_2$ at $m+1$ steps from the horizon. There are exactly four possible situations:

Situation 1: In this case it is optimal to serve class 0 in state $\vec{x}-\vec{e}_0-\vec{e}_2$, and serve classes 1 and 2 in state $\vec{x}$. Since $V_m(\cdot)$ is non-decreasing, this can never be the case when $x_0 = 1$. So we may assume that $x_0 > 1$, $x_1 \geq 0$, $x_2 > 0$. We can write

$$\begin{aligned}
&\tilde{V}_{m+1}(\vec{x}-\vec{e}_2) + \tilde{V}_{m+1}(\vec{x}-\vec{e}_0) \\
&\quad \leq \mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) + (\mu_1+\mu_2)V_m(\vec{x}-\vec{e}_2) \\
&\qquad + \mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_1 V_m(x_0-1,(x_1-1)^+,x_2) + \mu_2 V_m(\vec{x}-\vec{e}_0-\vec{e}_2). \qquad (4.43)
\end{aligned}$$

By (4.18), the terms in (4.43) with factors μ_0 and μ_1 are not larger than

$$\begin{aligned}
&-\mu_0 V_m(\vec{x}-\vec{e}_0) + \mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) + \mu_1 V_m(\vec{x}-\vec{e}_2) - \mu_1 V_m(\vec{x}) \\
&+ \mu_0 V_m(\vec{x}) + \mu_0 V_m(\vec{x}-2\vec{e}_0) + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_1 V_m(\vec{x}-\vec{e}_0) \\
&= \mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_0 V_m(\vec{x}-2\vec{e}_0-\vec{e}_2) + \mu_1 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) \\
&\quad + \mu_0 V_m(\vec{x}-2\vec{e}_0) + \mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) - \mu_0 V_m(\vec{x}-2\vec{e}_0-\vec{e}_2) - \mu_0 V_m(\vec{x}-\vec{e}_0) \\
&\quad + \mu_1 V_m(\vec{x}-\vec{e}_2) + \mu_1 V_m(\vec{x}-\vec{e}_0) - \mu_1 V_m(\vec{x}) - \mu_1 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) \\
&\leq \mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_0 V_m(\vec{x}-2\vec{e}_0-\vec{e}_2) + \mu_1 V_m(\vec{x}-\vec{e}_0-\vec{e}_2),
\end{aligned}$$

where the last inequality follows from (4.20). From the above we obtain that (4.43) is not larger than

$$\begin{aligned}
&\mu_0 V_m(\vec{x}) + \mu_1 V_m(x_0,(x_1-1)^+,x_2) + \mu_2 V_m(\vec{x}-\vec{e}_2) \\
&+ \mu_0 V_m(\vec{x}-2\vec{e}_0-\vec{e}_2) + \mu_1 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) + \mu_2 V_m(\vec{x}-\vec{e}_0-\vec{e}_2) \\
&= \tilde{V}_{m+1}(\vec{x}) + \tilde{V}_{m+1}(\vec{x}-\vec{e}_0-\vec{e}_2).
\end{aligned}$$

Situation 2: In this case it is optimal to serve class 0 in state $\vec{x}$, and serve classes 1 and 2 in state $\vec{x}-\vec{e}_0-\vec{e}_2$. First note that from (4.17) and (4.20) we have

$$2V_m(\vec{x}) \leq V_m(x_0,x_1,(x_2-1)^+) + V_m(\vec{x}+\vec{e}_2), \quad \text{for} \;\; x_0 > 0, x_1, x_2 \geq 0. \qquad (4.44)$$

Let $x_0 > 0$, $x_1 \geq 0$, $x_2 > 0$. For the terms in (4.43) with a factor μ_2, we can write

$$\begin{aligned}
&\mu_2[V_m(\vec{x}-\vec{e}_2) + V_m(\vec{x}-\vec{e}_0-\vec{e}_2)] \\
&\leq \mu_2[V_m(\vec{x}-\vec{e}_0-\vec{e}_2) - V_m(\vec{x}-\vec{e}_2) + V_m(x_0,x_1,(x_2-2)^+) + V_m(\vec{x})] \\
&\leq \mu_2[V_m(x_0-1,x_1,(x_2-2)^+) + V_m(\vec{x})], \qquad (4.45)
\end{aligned}$$

where the first inequality follows from (4.44) and the second from (4.20). Note that it is assumed that the cost functions satisfy (4.13). Hence, when $x_1 > 0$, it follows from Lemma 4.3.7 that the optimal action is to serve classes 1 and 2 in state $\vec{x}$, since $x_2 > 0$. In situation 2, class 0 is served in state $\vec{x}$, hence we can assume that $x_1 = 0$. In that case, the terms in the right-hand side of (4.43) with a factor μ_1 are: $\mu_1 V_m(x_0,0,x_2-1) + \mu_1 V_m(x_0-1,0,x_2)$. By (4.20) we have that this is smaller than or equal to $\mu_1 V_m(x_0,0,x_2) + \mu_1 V_m(x_0-1,0,x_2-1)$. Together with (4.45) this gives that (4.43) is not larger than

$$\begin{aligned}
&\mu_0 V_m(x_0-1,0,x_2) + (\mu_1+\mu_2)V_m(x_0,0,x_2) \\
&+ \mu_0 V_m(x_0-1,0,x_2-1) + \mu_1 V_m(x_0-1,0,x_2-1) + \mu_2 V_m(x_0-1,0,(x_2-2)^+), \\
&= \tilde{V}_{m+1}(x_0,0,x_2) + \tilde{V}_{m+1}(x_0-1,0,x_2-1).
\end{aligned}$$

Situation 3: In this case it is optimal to serve class 0 in states $\vec{x}$ and $\vec{x}-\vec{e}_0-\vec{e}_2$. Since (4.20) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.
Situation 4: In this case it is optimal to serve classes 1 and 2 in states $\vec{x}$ and $\vec{x}-\vec{e}_0-\vec{e}_2$. Since (4.20) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.

Proof of inequality (4.21): This goes along similar lines as the proof of (4.20).

Proof of inequality (4.22): We have to show that $Z=\tilde{V}_{m+1}$ satisfies (4.22). For $x_1=0$ or $x_2=0$, this is trivially true. Hence, we assume that $x_1>0$ and $x_2>0$. In order to prove relation (4.22), we need to distinguish between which of the two actions are optimal in the states $\vec{x}-\vec{e}_1$ and $\vec{x}-\vec{e}_2$ at $m+1$ steps from the horizon. There are exactly four possible combinations:
Situation 1: In this case it is optimal to serve class 0 in state $\vec{x}-\vec{e}_2$ and serve classes 1 and 2 in state $\vec{x}-\vec{e}_1$. When $x_0=0$, we can never be in this situation, since $V_m(\cdot)$ is non-decreasing. Therefore we assume $x_0>0$. We can write

$$\begin{aligned}
&\tilde{V}_{m+1}(\vec{x}-\vec{e}_1-\vec{e}_2)+\tilde{V}_{m+1}(\vec{x})\\
&\le \mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1-\vec{e}_2)+\mu_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2)+\mu_2 V_m(\vec{x}-\vec{e}_1-\vec{e}_2)\\
&\quad +\mu_0 V_m(\vec{x})+\mu_1 V_m(\vec{x}-\vec{e}_1)+\mu_2 V_m(\vec{x}-\vec{e}_2).
\end{aligned} \tag{4.46}$$

By (4.16), the terms in (4.46) with factor μ_0 or μ_1 are not larger than

$$\begin{aligned}
&-\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1)+\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1-\vec{e}_2)+\mu_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2)\\
&+(\mu_0-\mu_1)V_m(\vec{x}-\vec{e}_1)+\mu_0 V_m(\vec{x}-\vec{e}_0)+\mu_1 V_m(x_0,(x_1-2)^+,x_2)+\mu_1 V_m(\vec{x})\\
&=\mu_0 V_m(\vec{x}-\vec{e}_1)+\mu_1 V_m(x_0,(x_1-2)^+,x_2)+\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2)+\mu_1 V_m(\vec{x}-\vec{e}_2)\\
&\quad +\mu_1 V_m(\vec{x})+\mu_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2)-\mu_1 V_m(\vec{x}-\vec{e}_1)-\mu_1 V_m(\vec{x}-\vec{e}_2)\\
&\quad +\mu_0 V_m(\vec{x}-\vec{e}_0)+\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1-\vec{e}_2)-\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_1)\\
&\quad -\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2)\\
&\le \mu_0 V_m(\vec{x}-\vec{e}_1)+\mu_1 V_m(x_0,(x_1-2)^+,x_2)+\mu_0 V_m(\vec{x}-\vec{e}_0-\vec{e}_2)+\mu_1 V_m(\vec{x}-\vec{e}_2),
\end{aligned}$$

where the last inequality follows from (4.22). From the above we obtain that (4.46) is not larger than

$$\begin{aligned}
&\mu_0 V_m(\vec{x}-\vec{e}_1)+\mu_1 V_m(x_0,(x_1-2)^+,x_2)+\mu_2 V_m(\vec{x}-\vec{e}_1-\vec{e}_2)\\
&+\mu_0 V_m(\vec{x}_0-\vec{e}_0-\vec{e}_2)+\mu_1 V_m(\vec{x}-\vec{e}_2)+\mu_2 V_m(\vec{x}-\vec{e}_2),\\
&=\tilde{V}_{m+1}(\vec{x}-\vec{e}_1)+\tilde{V}_{m+1}(\vec{x}-\vec{e}_2).
\end{aligned}$$

Situation 2: In this case it is optimal to serve class 0 in state $\vec{x}-\vec{e}_1$ and serve classes 1 and 2 in state $\vec{x}-\vec{e}_2$. The proof is symmetric to situation 1.
Situation 3: In this case it is optimal to serve class 0 in states $\vec{x}-\vec{e}_1$ and $\vec{x}-\vec{e}_2$. Since (4.22) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well.
Situation 4: In this case it is optimal to serve classes 1 and 2 in states $\vec{x}-\vec{e}_1$ and $\vec{x}-\vec{e}_2$. Since (4.22) holds for $V_m(\cdot)$, it follows easily that it holds for $\tilde{V}_{m+1}(\cdot)$ as well. □

Chapter 5
Asymptotically optimal switching-curve policies

In Chapter 4 we used dynamic programming techniques to show that for a two-node linear network with exponential service requirements an optimal non-anticipating policy has certain structural properties: It is characterized by so-called switching curves. An explicit characterization of these curves is however in general not possible. To gain a better understanding of the functional form of the optimal switching curves, we therefore set out in this chapter to study these in asymptotic regimes. In particular, we study the system after a fluid or diffusion scaling.

We find that in many scenarios simple linear switching curves provide asymptotically fluid-optimal policies. These curves are obtained by solving the fluid control problem corresponding to the two-node linear network. When both nodes are equally congested, however, the fluid-based policy may not even ensure stability. In such cases we show that a diffusion scaling is appropriate, and that efficient policies have square-root shaped switching curves.

Through numerical experiments we conduct comparisons of the optimal policies (when numerically feasible), the class of weighted α-fair bandwidth-sharing policies and policies characterized by either linear, square-root or constant switching curves. We find that the fluid-based and diffusion-based policies give close-to-optimal performance. In addition, we confirm that weighted α-fair policies perform well among non-anticipating policies in moderately-loaded systems. In particular, we observe that the optimal policy can be approximated by a weighted α-fair policy when choosing the weights appropriately.

The remainder of this chapter is organized as follows. Section 5.1 describes the model, and presents new stability results. The fluid control model and asymptotically fluid-optimal policies are presented in Section 5.2. Section 5.3 considers the case when both nodes are equally congested and studies the network after a diffusion scaling. Numerical experiments can be found in Section 5.4. We conclude the chapter with a short summary and ideas for further research in Section 5.5.

5.1 Model and preliminaries

We consider a linear network with two nodes and three classes, where class i requires service at node i only, $i = 1, 2$, while class 0 requires service at both nodes simultaneously, see Figure 1.2 with $L = 2$. We assume each of the nodes to have unit capacity, so that a rate allocation is feasible when it belongs to the capacity region $S := \{(s_0, s_1, s_2) \in \mathbb{R}^3_+ : s_0 + s_i \leq 1,\ \text{for } i = 1, 2\}$, as depicted in Figure 1.5. Class-j users arrive according to independent Poisson processes of rate λ_j, and have exponentially distributed service requirements B_j with mean $1/\mu_j$, $j = 0, 1, 2$. Denote the traffic load of class j by $\rho_j := \lambda_j/\mu_j$. As in Chapter 4, the class of non-anticipating policies is denoted by $\bar{\Pi}$. For a given policy $\pi \in \bar{\Pi}$ denote by $N_j^\pi(t)$ and $W_j^\pi(t)$ the number of class-j users and the class-j workload, respectively, at time t. Define N_j^π and W_j^π as the random variables with the corresponding steady-state distributions (when they exist).

In Chapter 4 we characterized non-anticipating policies that minimize the average holding cost $\limsup_{T\to\infty} \mathbb{E}(\int_0^T \sum_{j=0}^2 c_j N_j(t)\mathrm{d}t)/T$, with c_j a non-negative cost associated with class j, $j = 0, 1, 2$. Simple priority rules were proved to be optimal when $c_0\mu_0 \geq \max(c_1\mu_1, c_2\mu_2)$, see Propositions 4.3.1 and 4.3.2. Otherwise, an optimal policy is characterized by a switching curve, see Proposition 4.3.9. In particular, when $c_0\mu_0 < \max(c_1\mu_1, c_2\mu_2)$, this policy prescribes that in states with $N_{3-i}(t) = 0$, class 0 is served when $N_i(t) < h_i(N_0(t))$, and class i is served otherwise, with $h_i(\cdot)$ the switching curve belonging to class i.

The above determines the structure of the optimal policy, but does not explicitly characterize the optimal switching curves. To gain some further understanding, let us compare two different policies, say policy $\pi^{(h)}$ with switching curves $h_1(\cdot)$ and $h_2(\cdot)$ and policy $\pi^{(g)}$ with switching curves $g_1(\cdot)$ and $g_2(\cdot)$, while $h_i(x_0) \leq g_i(x_0)$ for all x_0, $i = 1, 2$. Clearly, in the short run, a lower switching curve is better when $c_0\mu_0 < c_i\mu_i$, since the number of states that have a weighted departure rate equal to $c_i\mu_i$ will increase. In the long run, however, a higher curve may actually pay off: when starting in the same state, a higher curve empties the system faster (see Lemma 5.1.1 below) and has therefore less strict stability conditions (see Corollary 5.1.2 below). Clearly, an optimal switching curve strikes the right balance between these short- and long-run effects. Lemma 5.1.1 and its corollary are valid for generally distributed inter-arrival times. The proof of Lemma 5.1.1 may be found in Appendix 5.A.

Lemma 5.1.1. *(This lemma holds for generally distributed inter-arrival times.) Denote by $W_j^h(t)$ the workload of class j at time t under policy $\pi^{(h)}$. Let $h_i(x_0) \leq g_i(x_0)$ for all x_0, $i = 1, 2$. If $W_0^g(0) \leq_{st} W_0^h(0)$ and $W_0^g(0) + W_i^g(0) \leq_{st} W_0^h(0) + W_i^h(0)$, for $i = 1, 2$, then*

$$\{W_0^g(t)\}_{t\geq 0} \quad \leq_{st} \quad \{W_0^h(t)\}_{t\geq 0}, \tag{5.1}$$

$$\{W_0^g(t) + W_i^g(t)\}_{t\geq 0} \quad \leq_{st} \quad \{W_0^h(t) + W_i^h(t)\}_{t\geq 0}, \quad \textit{for } i = 1, 2. \tag{5.2}$$

Corollary 5.1.2. *(This lemma holds for generally distributed inter-arrival times.) Let $h_i(x_0) \leq g_i(x_0)$ for all x_0, $i = 1, 2$. If the system is empty under policy $\pi^{(h)}$, then* *it is empty under policy $\pi^{(g)}$ as well. In particular, if the empty state is positive recurrent under $\pi^{(h)}$ in the case of Poisson arrivals, then it is positive recurrent under $\pi^{(g)}$ as well.*

Define the policy π_l, for an $l = 1, 2$, as the policy that gives preemptive priority to class l, and when class l is empty, class 0 receives preemptive priority. Class $3-l$ is only served whenever there is capacity left unused. This strict priority rule has switching curves $h_l(x_0) = 0$ and $h_{3-l}(x_0) = \infty$. Under this policy, the total workload is almost surely finite, as is stated in the following lemma. In fact, the lemma considers even more general policies that are work-conserving in node l. The proof is provided in Appendix 5.B. It essentially uses that the behavior of classes 0 and l is determined by the dynamics within node l. In order for class $3 - l$ to grow unboundedly, it has to be the case that for a non-negligible portion of time class l is served while class $3 - l$ is not present. Obviously, the latter cannot be true.

Lemma 5.1.3. *For any Pareto-efficient policy which gives strict priority to class 0 over class $3 - l$ when class l is empty, for an $l = 1, 2$, the total workload in the system is almost surely finite when $\rho_0 + \rho_i < 1$, $i = 1, 2$.*

For general service requirement distributions, we can determine the mean workload for class l under policy π_l, $l = 1, 2$, from the Pollaczek-Khintchine formula: $\mathbb{E}(W_l^{\pi_l}) = \frac{\lambda_l \mathbb{E}({B_l}^2)}{2(1-\rho_l)}$. Class 0 sees its service being interrupted by busy periods of class l so that [133]:

$$\mathbb{E}(W_0^{\pi_l}) = \frac{\lambda_0 \mathbb{E}({B_0}^2) + \lambda_l \mathbb{E}({B_l}^2)}{2(1 - \rho_0 - \rho_l)} - \frac{\lambda_l \mathbb{E}({B_l}^2)}{2(1 - \rho_l)}.$$

Unfortunately, for class $3-l$ there are no expressions available for the mean workload. Determining these requires solving a boundary value problem [42].

5.2 Fluid analysis

In this section we consider the number of users under a fluid scaling and investigate close-to-optimal switching curves. It will be convenient to first study the related deterministic fluid control model. This will be done in Section 5.2.1. In contrast to the stochastic model, we obtain exact expressions for the optimal switching curves. In Section 5.2.2 we then show that these optimal switching curves provide asymptotically fluid-optimal policies in the stochastic model.

5.2.1 Optimal fluid control

The fluid control model arises from the original stochastic model by only taking into account the mean drifts. A fluid process for the two-node linear network is a solution $n(t) = (n_0(t), n_1(t), n_2(t))$ of the following equations:

$$n_j(t) = n_j + \lambda_j t - \mu_j U_j(t), \ j = 0, 1, 2, \tag{5.3}$$

$$n_j(t) \geq 0, \ j = 0, 1, 2. \tag{5.4}$$

Here $n = (n_0, n_1, n_2) \in \mathbb{R}^3_+$ and $U_j(t) = \int_0^t u_j(v)\mathrm{d}v$, such that for all $v \geq 0$,

$$u_0(v) + u_i(v) \leq 1, \; i = 1, 2, \tag{5.5}$$

$$u_j(v) \geq 0, \; j = 0, 1, 2, \tag{5.6}$$

(i.e., $u(v) \in S$) and the functions $u_j(\cdot)$ are measurable, $j = 0, 1, 2$. Note that $U_j(\cdot)$ is Lipschitz continuous with constant less than or equal to 1. Hence, it is absolutely continuous which implies that it is differentiable almost everywhere [112]. Thus, $n_j(\cdot)$ is differentiable almost everywhere as well, and

$$\frac{\mathrm{d}n_j(t)}{\mathrm{d}t} = \lambda_j - u_j(t)\mu_j, \; j = 0, 1, 2, \tag{5.7}$$

at regular points (a regular point is a value of t at which $n_j(t)$ is differentiable).

A policy π for the fluid control model is described by the control $u^\pi(t)$ (we also write $U_j^\pi(t)$). A corresponding trajectory is denoted by $n^\pi(t)$. Our aim is to derive an optimal clearing control for the fluid model, starting from any initial state. We use the following two definitions:

- A control is called path-wise optimal if it minimizes $\sum_{j=0}^2 c_j n_j^\pi(t)$ for all $t \geq 0$, with $(n^\pi(t), u^\pi(t))$ satisfying (5.3)–(5.6).

- A control is called average-cost optimal if it minimizes $\int_0^\infty \sum_{j=0}^2 c_j n_j^\pi(t)\mathrm{d}t$, with $(n^\pi(t), u^\pi(t))$ satisfying (5.3)–(5.6).

Path-wise optimal controls do not necessarily exist. However, if they exist they are automatically average-cost optimal. Before describing optimal controls, we first state two convenient lemmas.

Under the stability conditions $\rho_0 + \rho_i < 1$, $i = 1, 2$, the fluid model can be drained in finite time (and then remains empty if controlled optimally), see the next lemma.

Lemma 5.2.1. *If $\rho_0 + \rho_i < 1$, $i = 1, 2$, then the policy that serves class 0 whenever possible, drains the fluid model in finite time and keeps the system empty from that moment on.*

Proof: We focus on the policy that serves class 0 whenever possible. Hence, when $n_0(t) > 0$, we have $u_0(t) = 1$, so that $\frac{\mathrm{d}n_0(t)}{\mathrm{d}t} = \lambda_0 - \mu_0 < 0$. From this it follows that once class 0 hits zero, it will remain zero. Together with (5.7) this yields that $u_0(t) = \rho_0$ when $n_0(t) = 0$. Hence, if $n_i(t) > 0$ and $n_0(t) = 0$, then $\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = \mu_i(\rho_i - (1 - u_0(t))) = \mu_i(\rho_0 + \rho_i - 1) < 0$, $i = 1, 2$. We can conclude that the system empties in finite time. □

For the original stochastic model it was shown that when $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0$, it is optimal to serve classes 1 and 2 simultaneously, whenever both are present, see Proposition 4.3.9. The fluid model inherits this property. The proof may be found in Appendix 5.C.

Lemma 5.2.2. *Assume $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0$. For any policy $\tilde{\pi}$, there exists a policy π that does not do worse than $\tilde{\pi}$ (i.e., if $n^{\pi}(0) = n^{\tilde{\pi}}(0)$, then $\sum_{j=0}^{2} c_j n_j^{\pi}(t) \leq \sum_{j=0}^{2} c_j n_j^{\tilde{\pi}}(t)$, for all $t \geq 0$) and satisfies the following:*

$$u_1^{\pi}(t) = u_2^{\pi}(t) = 1, \ \textit{if} \ \ n_1(t), n_2(t) > 0, \tag{5.8}$$

$$u_i^{\pi}(t) = \rho_i, \ \textit{if} \ \ n_i(t) = 0 \ \ \textit{and} \ \ n_j(t) > 0, \ i \neq j, \ i, j = 1, 2, \tag{5.9}$$

$$u_i^{\pi}(t) = \rho_i, \ \textit{if} \ \ n_i(t) = n_j(t) = 0 \ \ \textit{and} \ \ \rho_i \leq \rho_j, \ i \neq j, \ i, j = 1, 2. \tag{5.10}$$

For the parameter settings where there is no conflict between maximizing the weighted departure rate and fully using all resources, one would expect that the fluid control model allows for a path-wise optimal policy, which is confirmed by the next proposition.

Proposition 5.2.3. *Assume $\rho_1 \leq \rho_2$ and $\rho_0 + \rho_2 < 1$.*
If $c_1\mu_1 + c_2\mu_2 \leq c_0\mu_0$ then a path-wise optimal control is:

- *$u_0^*(t) = 1$, if $n_0(t) > 0$,*
- *$u_0^*(t) = \rho_0$, if $n_0(t) = 0$.*

If $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0 \geq c_1\mu_1, c_2\mu_2$, then a path-wise optimal control is:

- *$u_0^*(t) = 0$, if $n_1(t), n_2(t) > 0$,*
- *$u_0^*(t) = 1 - \rho_2$, if $n_0(t), n_1(t) > 0, n_2(t) = 0$,*
- *$u_0^*(t) = 1 - \rho_1$, if $n_0(t) > 0$ and $n_1(t) = 0$,*
- *$u_0^*(t) = \rho_0$, otherwise.*

If $c_2\mu_2 \geq c_0\mu_0 \geq c_1\mu_1$, then a path-wise optimal control is:

- *$u_0^*(t) = 0$, if $n_2(t) > 0$,*
- *$u_0^*(t) = 1 - \rho_2$, if $n_0(t) > 0$ and $n_2(t) = 0$,*
- *$u_0^*(t) = \rho_0$, otherwise.*

In all cases, $u_i^*(t) = 1 - u_0^*(t)$, if $n_i(t) > 0$ and $u_i^*(t) = \min(\rho_i, 1 - u_0^*(t))$, if $n_i(t) = 0$, $i = 1, 2$.

The first two controls are similar to the stochastic policies $\pi^* \in \Pi^* \cap \bar{\Pi}$ and $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$ respectively, which are optimal in the stochastic model (Propositions 4.3.1 and 4.3.2). The third case corresponds to policy π_2, which, according to Lemma 5.1.3, keeps the total workload in the system finite a.s.

Proof of Proposition 5.2.3: First assume $c_1\mu_1 + c_2\mu_2 \leq c_0\mu_0$. Consider the control with $u_0^*(t) = 1$, if $n_0(t) > 0$, and $u_0^*(t) = \rho_0$, if $n_0(t) = 0$, and denote the corresponding trajectory by $n^*(\cdot)$. Obviously, the control $u^*(\cdot)$ minimizes the amount of class-0 fluid, i.e.,

$$n_0^*(t) \leq n_0^\pi(t), \quad \text{for all } t, \tag{5.11}$$

for any control π. In addition, $u^*(\cdot)$ is work-conserving in both nodes, i.e., $u_0^*(t) + u_i^*(t) = 1$ whenever $n_0(t) + n_i(t) > 0$, $i = 1, 2$. Hence, for a given policy π we have $U_0^*(t) + U_i^*(t) \geq U_0^\pi(t) + U_i^\pi(t)$, $i = 1, 2$. By (5.3), this implies

$$\frac{1}{\mu_0} n_0^*(t) + \frac{1}{\mu_i} n_i^*(t) \leq \frac{1}{\mu_0} n_0^\pi(t) + \frac{1}{\mu_i} n_i^\pi(t), \quad i = 1, 2, \quad \text{for all } t. \tag{5.12}$$

Multiplying (5.11) by $(c_0\mu_0 - c_1\mu_1 - c_2\mu_2)/\mu_0 \geq 0$, multiplying (5.12) by $c_i\mu_i$, for $i = 1, 2$, and summing these three inequalities gives that $\sum_{j=0}^2 c_j n_j^*(t) \leq \sum_{j=0}^2 c_j n_j^\pi(t)$, i.e., $u^*(\cdot)$ is path-wise optimal.

Now assume $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0 (\geq c_1\mu_1)$. Lemma 5.2.2 fully characterizes the optimal action in states where both classes 1 and 2 are backlogged, i.e., $u_1^*(t) = u_2^*(t) = 1$ whenever $n_1(t), n_2(t) > 0$. We therefore only need to consider the following two cases: no backlog of class 1 and no backlog of class 2.

First we consider states with $n_1(t) = 0$. Since $\rho_1 \leq \rho_2$, by Lemma 5.2.2 we have $u_1^*(t) = \rho_1$ and $u_2^*(t) \geq \rho_1$. Hence, the corresponding optimal trajectory keeps class 1 empty from then on. The time until reaching the origin is the same for every Pareto-efficient policy that keeps class 1 empty. Hence, the policy that allocates the remaining fraction $1 - \rho_1$ of capacity between classes 0 and 2 such that the weighted departure rate is maximized, minimizes the cost at any moment in time. Hence, if $c_0\mu_0 \leq c_2\mu_2$, then $u_2^*(t) = 1$ when $n_2(t) > 0$ and $u_2^*(t) = \rho_2$ when $n_2(t) = 0$. If $c_0\mu_0 \geq c_2\mu_2$, then $u_0^*(t) = 1 - \rho_1$ when $n_0(t) > 0$ and $u_0^*(t) = \rho_0$ when $n_0(t) = 0$.

Now consider states with $n_1(t) > 0$ and $n_2(t) = 0$. By Lemma 5.2.2, an optimal control satisfies $u_2^*(t) = \rho_2$ (and hence keeps class 2 empty) and $u_1^*(t) \geq \rho_2$ as long as $n_1(t) > 0$. We are left with finding the optimal way to allocate the remaining fraction $1 - \rho_2$ of capacity between classes 0 and 1, until class 1 is empty. Below we will show that the control $u^*(\cdot)$ as suggested in the statement of the proposition is path-wise optimal (we denote the corresponding trajectory by $n^*(\cdot)$). For states with $n_1(t) > 0$ and $n_2(t) = 0$ this control is $u_0^*(t) = 1 - \rho_2$ when $n_0(t) > 0$ and $u_0^*(t) = \rho_0$ when $n_0(t) = 0$. Let T^* be the first moment that class 1 is empty under control $u^*(t)$. In the interval $(t, T^*]$ the control $u^*(\cdot)$ is work-conserving in both nodes, hence for any control π (throughout we assume $n^\pi(t) = n^*(t)$), we have

$$\frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_i} n_i^*(u) \leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_i} n_i^\pi(u), \tag{5.13}$$

for $i = 1, 2$ and $t \leq u \leq T^*$. In order to conclude the proof, we need to distinguish between the two cases: $c_0\mu_0 \geq c_2\mu_2$ and $c_0\mu_0 \leq c_2\mu_2$.

First assume $c_0\mu_0 \geq c_2\mu_2$. As shown above, once class 1 empties, a path-wise optimal control keeps class 1 empty and allocates the remaining capacity to class 0. Since $\rho_1 \leq \rho_2$, this allocation is work-conserving in both nodes, hence (5.13) is valid for all $u \geq t$. Together with the fact that $\min(n_1^*(u), n_2^*(u)) = 0$, for all $u \geq t$, we obtain that for all $u \geq t$

$$\begin{aligned} \frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_1} n_1^*(u) + \frac{1}{\mu_2} n_2^*(u) &= \max_{i=1,2}\left(\frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_i} n_i^*(u)\right) \\ &\leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_1} n_1^\pi(u) + \frac{1}{\mu_2} n_2^\pi(u), \quad (5.14) \end{aligned}$$

for any control π. Multiplying (5.13) by $c_0\mu_0 - c_{3-i}\mu_{3-i} \geq 0$, for $i = 1, 2$, multiplying (5.14) by $\sum_{i=1}^2 c_i\mu_i - c_0\mu_0 \geq 0$, and summing these three inequalities we obtain $\sum_{j=0}^2 c_j n_j^*(u) \leq \sum_{j=0}^2 c_j n_j^\pi(u)$, for all $u \geq t$, i.e., $u^*(\cdot)$ is path-wise optimal.

Now assume $c_0\mu_0 \leq c_2\mu_2$. Recall that we start at time t in states with $n_1(t) > 0$ and $n_2(t) = 0$. Since $c_0\mu_0 \leq c_2\mu_2$, once class 1 empties, a path-wise optimal control keeps both classes 1 and 2 empty, and allocates the remaining fraction of capacity $1 - \rho_2$ to class 0 (as shown above). Since this allocation is work-conserving in node 2 we obtain, together with (5.13), that for all $u \geq t$

$$\frac{1}{\mu_0} n_0^*(u) = \frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_2} n_2^*(u) \leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_2} n_2^\pi(u). \quad (5.15)$$

Hence, for all $u > T^*$, $\frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_1} n_1^*(u) = \frac{1}{\mu_0} n_0^*(u) \leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_2} n_2^\pi(u) \leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_1} n_1^\pi(u) + \frac{1}{\mu_2} n_2^\pi(u)$. Together with (5.13) we obtain that for all $u \geq t$

$$\frac{1}{\mu_0} n_0^*(u) + \frac{1}{\mu_1} n_1^*(u) \leq \frac{1}{\mu_0} n_0^\pi(u) + \frac{1}{\mu_1} n_1^\pi(u) + \frac{1}{\mu_2} n_2^\pi(u). \quad (5.16)$$

Since class 2 is empty for all $u \geq t$, we have as well

$$0 \leq n_2^\pi(u). \quad (5.17)$$

Multiplying (5.15) by $c_0\mu_0 - c_1\mu_1 \geq 0$, multiplying (5.16) by $c_1\mu_1$, multiplying (5.17) by $(c_2\mu_2 - c_0\mu_0)/\mu_2 \geq 0$, and summing these inequalities we obtain $\sum_{j=0}^2 c_j n_j^*(u) = c_0 n_0^*(u) + c_1 n_1^*(u) \leq \sum_{j=0}^2 c_j n_j^\pi(u)$, for all $u \geq t$, i.e., $u^*(\cdot)$ is path-wise optimal. □

In case $c_1\mu_1 > c_0\mu_0$ (with $\rho_1 \leq \rho_2$), a path-wise optimal policy does not exist. We will derive average-cost optimal fluid controls instead. Before doing so, in the next lemma we prove that these indeed exist. The proof may be found in Appendix 5.D. More importantly, we obtain that if $u^*(t)$ is an average-cost optimal control, then it also minimizes a finite horizon cost whenever the horizon is large enough. This property will be useful to prove convergence of the stochastic model in Section 5.2.2.

Lemma 5.2.4. *If $\rho_0 + \rho_i < 1$, $i = 1, 2$, then there exists a control $u^*(t)$ that is average-cost optimal. Let $n^*(t)$ be the corresponding trajectory. In addition, there exists a function $H : \mathbb{R} \to \mathbb{R}$ such that,*

$$\min_{n(t) \ s.t. \ (5.3)-(5.6)} \int_0^D \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t = \int_0^D \sum_{j=0}^2 c_j n_j^*(t) \mathrm{d}t = \int_0^\infty \sum_{j=0}^2 c_j n_j^*(t) \mathrm{d}t,$$

for all $D \geq H(c_0 n_0 + c_1 n_1 + c_2 n_2)$, with n the initial state.

As in the stochastic model, there is no straightforward way to allocate the capacity in states with $n_2(t) = 0$ when $c_0\mu_0 < c_1\mu_1$. Giving full priority to class 1 maximizes the weighted departure rate. This, however, leaves a fraction $1-u_2^*(t) = 1-\rho_2$ of the capacity in node 2 unutilized. As soon as class 1 empties, we are faced with an unnecessarily high backlog in node 2. A trade-off between serving class 0 or 1 arises and it turns out to be optimal to give first priority to class 1 and then to switch to class 0. This is made precise in the next lemma.

Lemma 5.2.5. *Assume $\rho_1 \leq \rho_2$, $\rho_0 + \rho_2 < 1$, and $c_0\mu_0 \leq c_1\mu_1$. Consider a trajectory starting in $\tilde{n} \in \{(n_0, n_1, n_2) \in \mathbb{R}_+^3 : n_0 \geq 0, n_1 > 0, n_2 = 0\}$ with the following properties: (i) first $u_1(t) = 1$ during a contiguous period, and then (ii) we switch to $u_0(t) = 1 - \rho_2$ and $u_1(t) = u_2(t) = \rho_2$ during another contiguous period. Let $\hat{n}$ be the end point of this trajectory.*

Among all feasible trajectories that move from $\tilde{n}$ to $\hat{n}$ without coinciding with the $n_1 = 0$ axis, the trajectory described above minimizes $\sum_{j=0}^{2} c_j n_j(t)$ at all times (until reaching $\hat{n}$). In case $c_0\mu_0 < c_1\mu_1$ this is the unique optimal trajectory.

Proof: Since we consider only trajectories from $\tilde{n}$ to $\hat{n}$ that do not coincide with the $n_1 = 0$ axis, it follows from Lemma 5.2.2 that we can focus on trajectories that keep class 2 empty, i.e., $u_2(t) = \rho_2$. Hence, for any Pareto-efficient trajectory from $\tilde{n}$ to $\hat{n}$ the cost decreases at rate $\frac{\mathrm{d}(c_0 n_0(t)+c_1 n_1(t))}{\mathrm{d}t} = c_0\lambda_0 + c_1\lambda_1 - (1-u_1(t))c_0\mu_0 - u_1(t)c_1\mu_1$ with $u_1(t) \geq \rho_2$, and the cumulative amount of time it spends on serving class 1 is given by $(\hat{n}_1 - \tilde{n}_1)/(\lambda_1 - \mu_1)$. Since $c_0\mu_0 \leq c_1\mu_1$, first prioritizing class 1 maximizes the rate of decrease in cost, and hence minimizes $\sum_{j=0}^{2} c_j n_j(t)$ at all times (until reaching $\hat{n}$). □

Lemma 5.2.5 allows us to derive average-cost optimal controls for the cases where no path-wise optimal control could be found.

Proposition 5.2.6. *Assume $\rho_1 \leq \rho_2$ and $\rho_0 + \rho_2 < 1$.*
If $c_1\mu_1 \geq c_0\mu_0 \geq c_2\mu_2$, then an average-cost optimal control is:

- $u_0^*(t) = 0$, *if* $n_1(t) \geq \frac{c_2\mu_2}{c_1\mu_1 + c_2\mu_2 - c_0\mu_0} \cdot \frac{\mu_1}{\mu_0} \cdot \frac{\rho_2 - \rho_1}{1-\rho_0-\rho_2} n_0(t)$ *or if* $n_1(t), n_2(t) > 0$,
- $u_0^*(t) = \rho_0$, *if* $n_0(t) = n_1(t) = 0$,
- $u_0^*(t) = 1 - \rho_1$, *if* $n_0(t) > 0$ *and* $n_1(t) = 0$,
- $u_0^*(t) = 1 - \rho_2$, *otherwise.*

If $c_1\mu_1, c_2\mu_2 \geq c_0\mu_0$, then an average-cost optimal control is:

- $u_0^*(t) = 0$ *if* $n_1(t) \geq \frac{c_0}{c_1} \cdot \frac{\rho_2 - \rho_1}{1-\rho_0-\rho_2} n_0(t)$ *or if* $n_2(t) > 0$,
- $u_0^*(t) = 1 - \rho_2$, *otherwise.*

In all cases, $u_i^*(t) = 1 - u_0^*(t)$, if $n_i(t) > 0$ and $u_i^*(t) = \min(\rho_i, 1 - u_0^*(t))$, if $n_i(t) = 0$, $i = 1, 2$.

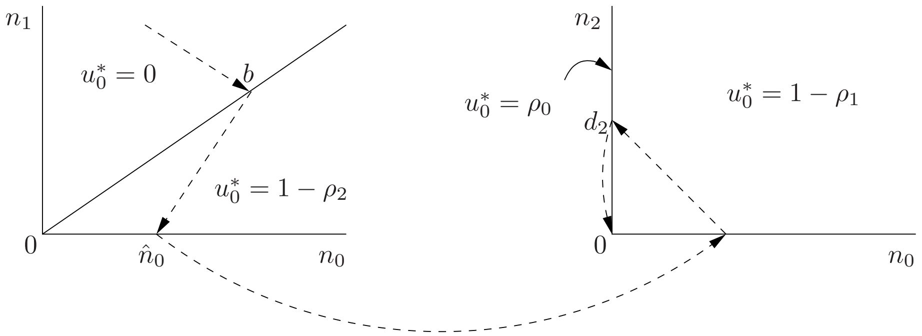


Figure 5.1: Optimal capacity allocation when $n_2(t) = 0$ (left) and $n_1(t) = 0$ (right); if $c_1\mu_1 \geq c_0\mu_0 \geq c_2\mu_2$ and $\rho_1 < \rho_2$.

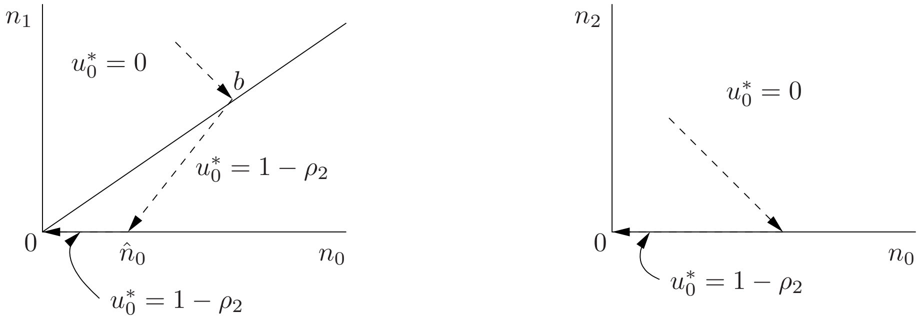


Figure 5.2: Optimal capacity allocation when $n_2(t) = 0$ (left) and $n_1(t) = 0$ (right); if $c_1\mu_1, c_2\mu_2 \geq c_0\mu_0$ and $\rho_1 < \rho_2$.

For illustration, the trajectories corresponding to the average-cost optimal controls are depicted in Figures 5.1 and 5.2 (the case when both $n_1(t) > 0$ and $n_2(t) > 0$ is not shown; in that case $u_0^*(t) = 0$). Figure 5.1 considers the case $c_1\mu_1 \geq c_0\mu_0 \geq c_2\mu_2$. According to Proposition 5.2.6, there is a linear switching curve above which class 1 must be served with full capacity when $n_2(t) = 0$, see the left plot of Figure 5.1. Below that curve, class 0 receives the fraction $1 - \rho_2$ that is left from keeping class 2 empty. Once the process hits the horizontal axis, class 0 receives $1 - \rho_1$, which forces class 2 to increase. The plane on the right shows that when class 1 is empty, it receives exactly its average load and hence remains empty. Class 2 increases while class 0 is emptied.

Figure 5.2 considers the case $c_1\mu_1, c_2\mu_2 \geq c_0\mu_0$. In states where class 2 is empty there is a linear switching curve, see the left plot of Figure 5.2. The plane on the right shows that once class 1 empties, it will remain empty. From that moment on, class 0 receives no capacity until class 2 is empty as well.

The policies in Proposition 5.2.6 may be described by linear switching curves $h_i(x_0) = k_i x_0$, $i = 1, 2$. When $c_1\mu_1 \geq c_0\mu_0 \geq c_2\mu_2$, we have $k_1 = \frac{c_2\mu_2}{c_1\mu_1+c_2\mu_2-c_0\mu_0} \cdot \frac{\mu_1}{\mu_0} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$ and $k_2 = \infty$. In this case k_1 depends on the traffic loads as well as the (weighted) departure rates. When $c_1\mu_1, c_2\mu_2 \geq c_0\mu_0$, we have $k_1 = \frac{c_0}{c_1} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$ and $k_2 = 0$. Now k_1 only depends on the traffic loads, c_0, and c_1. This can be explained as follows. Assume $c_2\mu_2 \geq c_0\mu_0$. When starting in a point with $n_2(t) = 0$, the value of k_1 determines the point where the trajectory hits the horizontal axis (where both classes 1 and 2 are empty). From that moment on, it is optimal to keep classes 1 and 2 empty. Hence class 2 remains empty, so that the precise value of $c_2\mu_2$ will have no impact on the cost or on the optimal value of k_1. In particular, one could set $c_2\mu_2 = c_0\mu_0$ and use the expression of k_1 for the case $c_2\mu_2 \leq c_0\mu_0$. Then indeed $k_1 = \frac{c_2\mu_2}{c_1\mu_1+c_2\mu_2-c_0\mu_0} \cdot \frac{\mu_1}{\mu_0} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2} = \frac{c_0}{c_1} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$.

Proof of Proposition 5.2.6: Lemma 5.2.2 fully characterizes the optimal policy in states where both classes 1 and 2 are backlogged. When we start in a state with no backlog of class 1, we can use the arguments in the proof of Proposition 5.2.3 to conclude that the actions as described in Proposition 5.2.6 (which keep class 1 empty) minimize the cost sample-path wise until the system is empty. We therefore only need to consider the case of no backlog in class 2.

Assume we start in a state with $n_1(t) > 0$ and $n_2(t) = 0$. By Lemma 5.2.2, an optimal control satisfies $u_2^*(t) = \rho_2$ (and hence keeps class 2 empty) and $u_1^*(t) \geq \rho_2$ as long as $n_1(t) > 0$. We are left with finding the optimal way to allocate the remaining fraction $1 - \rho_2$ of capacity between classes 0 and 1, until class 1 is empty.

At some point, an optimal trajectory will hit the $n_1 = 0$ axis for the first time. This point will be denoted by $\hat{n} = (\hat{n}_0, 0, 0)$, see the left plots in Figures 5.1 and 5.2. From this point on, an optimal trajectory from $\hat{n}$ to the origin is exactly known.

Since $c_0\mu_0 \leq c_1\mu_1$, an optimal trajectory from n to $\hat{n}$ first prioritizes class 1 and at some point switches to serving class 0 (while keeping class 2 empty), see Lemma 5.2.5. The turning point where the switch occurs is denoted by $b = (b_0, b_1, 0)$, see again the left plots in Figures 5.1 and 5.2. In order to obtain the average-cost optimal control, it is left to determine the optimal switching point. We do this by calculating the cost belonging to the trajectory that turns at b.

The time it takes to move from n to b is equal to $T(n, b) = \frac{b_0-n_0}{\lambda_0}$ during which the holding cost is on average $\frac{c_0(n_0+b_0)}{2} + \frac{c_1(n_1+b_1)}{2}$. The time it takes to move from b to $\hat{n}$ is equal to $T(b, \hat{n}) = \frac{b_1}{\mu_1\rho_2-\lambda_1}$ during which the holding cost is on average $\frac{c_0(b_0+\hat{n}_0)}{2} + \frac{c_1b_1}{2}$, with $\hat{n}_0 = b_0 - \frac{\mu_0(1-\rho_0-\rho_2)}{\mu_1(\rho_2-\rho_1)}b_1$. Let $K_n(b_0)$ be the cost of the fluid trajectory going from n to the origin when the turning point is b. Note that $b_1 = n_1 - \frac{b_0-n_0}{\lambda_0}(\mu_1 - \lambda_1)$, hence b_1 is uniquely determined by b_0 and n. The cost for switching point b can now be written as:

$$K_n(b_0) = T(n, b)(\frac{c_0(n_0 + b_0)}{2} + \frac{c_1(n_1 + b_1)}{2}) + T(b, \hat{n})(\frac{c_0(b_0 + \hat{n}_0)}{2} + \frac{c_1b_1}{2}) + K^*_{\hat{n}},$$

with $b_0 \in [n_0, n_0 + n_1\frac{\lambda_0}{\mu_1-\lambda_1}]$. The term $K^*_{\hat{n}}$ represents the minimum cost when going from $\hat{n}$ to the origin. Note that $\hat{n}_1 = 0$, hence the optimal trajectory starting in $\hat{n}$ is exactly known:

- If $c_0\mu_0 \geq c_2\mu_2$, then it is optimal to prioritize class 0, while keeping class 1 empty. An optimal trajectory corresponds to the right plot of Figure 5.1. Denote by $d = (0, 0, d_2)$ the point where the trajectory hits the vertical axis. From that point on, the trajectory stays on the vertical axis until it hits the origin. Hence, $K^*_{\hat{n}} = T(\hat{n}, d)\frac{c_0\hat{n}_0}{2} + T(\hat{n}, 0)\frac{c_2 d_2}{2}$, with $T(\hat{n}, d) = \frac{\hat{n}_0}{\mu_0(1-\rho_0-\rho_1)}$, $d_2 = T(\hat{n}, d)\mu_2(\rho_2 - \rho_1)$, and $T(\hat{n}, 0) = T(\hat{n}, d) + \frac{d_2}{\mu_2(1-\rho_0-\rho_2)}$.
- If $c_2\mu_2 \geq c_0\mu_0$, then it is optimal to keep both classes 1 and 2 empty. Hence, it takes $T(\hat{n}, 0) = \frac{\hat{n}_0}{\mu_0(1-\rho_0-\rho_2)}$ to reach the origin. The average holding cost of class 0 is $c_0\hat{n}_0/2$, so that $K^*_{\hat{n}} = T(\hat{n}, 0)\frac{c_0\hat{n}_0}{2}$.

It can be checked that when minimizing $K_n(\cdot)$ over b_0, the optimal b lies on the linear switching curve as stated in Proposition 5.2.6. □

5.2.2 Asymptotically fluid-optimal policies

The optimal switching curves for the stochastic model can be computed numerically by value iteration after truncating the state space. In Figure 5.3 we plotted the so-obtained optimal switching curves for scenarios with $\rho_1 \neq \rho_2$ and unit costs $c_j = 1$, $j = 0, 1, 2$. We also plotted the (shifted) switching curves obtained in the fluid control problem, as stated in Proposition 5.2.6, and observe that these give a good approximation for the optimal switching curves in the stochastic model. In this section we discuss the theoretical foundations that justify the use of the optimal fluid control in the stochastic model. In particular, we prove that the fluid-scaled

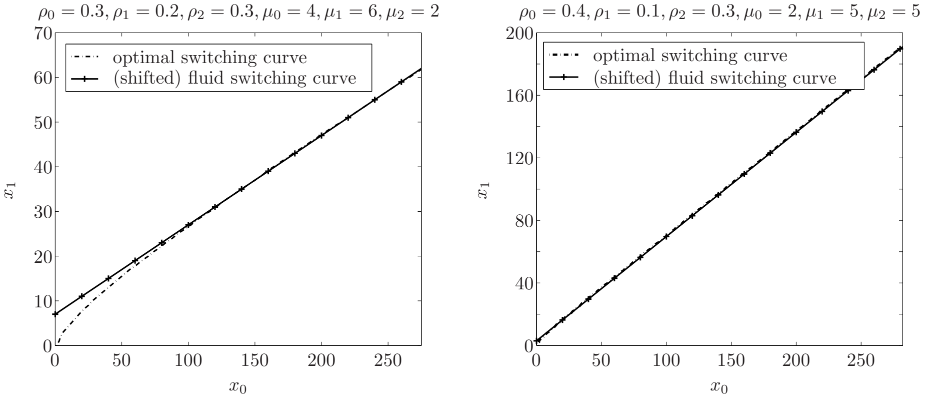


Figure 5.3: Optimal switching curves in the stochastic model when $\rho_1 < \rho_2$, and (shifted) optimal fluid switching curves, i.e., $x_1 = 7 + x_0 \cdot \mu_2/(\mu_1 + \mu_2 - \mu_0) \cdot \mu_1/\mu_0 \cdot (\rho_2 - \rho_1)/(1 - \rho_0 - \rho_2)$ (left) and $x_1 = 3 + x_0 \cdot (\rho_2 - \rho_1)/(1 - \rho_0 - \rho_2)$ (right).

numbers of users under certain switching-curve policies, converge to an optimal fluid trajectory $n^*(t)$ as determined in Section 5.2.1. Using the latter, we then show that these policies are asymptotically fluid-optimal in the stochastic model.

On a common probability space we construct a sequence of processes depending on the initial state. To be precise, for a given policy π we let $N_j^{\pi,r}(t)$ denote the number of class-j users at time t when the process starts in state $N^r(0) = r \cdot (n_0, n_1, n_2)$, with $r \in \mathbb{N}$. All processes $N^r(t)$ share the same sequence of arrivals and service requirements. For a given policy π, denote by $T_{I_i}^{\pi,r}(t)$ the cumulative amount of time during the interval $(0,t]$ that node i is idle, $i = 1,2$, and by $T_j^{\pi,r}(t)$ the cumulative amount of time that was spent on serving class j, $j = 0,1,2$. Then, $T_0^{\pi,r}(t) + T_i^{\pi,r}(t) + T_{I_i}^{\pi,r}(t) = t$, $i = 1,2$, and

$$N_j^{\pi,r}(t) = rn_j + E_j(t) - F_j(T_j^{\pi,r}(t)), \quad j = 0,1,2, \tag{5.18}$$

with $E_j(\cdot)$ a Poisson process with rate λ_j and $F_j(\cdot)$ a Poisson process with rate μ_j, [48]. We will be interested in the processes under the fluid scaling, i.e., both time and space are scaled linearly by the parameter r:

$$\overline{N}_j^{\pi,r}(t) := \frac{N_j^{\pi,r}(rt)}{r} \quad \text{and} \quad \overline{T}_l^{\pi,r}(t) := \frac{T_l^{\pi,r}(rt)}{r}, \; j = 0,1,2, \; l = 0,1,2,I_1,I_2.$$

Limit points for $\overline{N}_j^{\pi,r}(t)$ and $\overline{T}_l^{\pi,r}(t)$ are described in the next lemma.

Lemma 5.2.7. *For almost all sample paths ω there exists a subsequence r_k such that*

$$\lim_{k\to\infty} \overline{N}_j^{\pi,r_k}(t) = \overline{N}_j^{\pi}(t), \quad j = 0,1,2, \; \textit{u.o.c.},$$
$$\lim_{k\to\infty} \overline{T}_l^{\pi,r_k}(t) = \overline{T}_l^{\pi}(t), \quad l = 0,1,2,I_1,I_2, \quad \textit{u.o.c.}$$

Furthermore, $(\overline{N}^{\pi}, \overline{T}^{\pi})$ satisfies for $j = 0,1,2$, $i = 1,2$, $l = 0,1,2,I_1,I_2$,

$$\overline{N}_j^{\pi}(t) = n_j + \lambda_j t - \mu_j \overline{T}_j^{\pi}(t), \tag{5.19}$$

$\overline{N}_j(t) \geq 0$, $\overline{T}_l^{\pi}(0) = 0$, $\overline{T}_0^{\pi}(t) + \overline{T}_i^{\pi}(t) + \overline{T}_{I_i}^{\pi}(t) = t$, and $\overline{T}_l^{\pi}(t)$, are non-decreasing and Lipschitz continuous functions.

The notation u.o.c. stands for uniform convergence on compact sets. We call the processes $\overline{T}_l^{\pi}(t)$ and $\overline{N}_j^{\pi}(t)$ fluid limits for initial fluid level n and policy π.

Proof of Lemma 5.2.7: Making use of (5.18) and the fact that $\overline{T}_l^{\pi,r}(t)$ is Lipschitz continuous with a constant less than or equal to 1, the proof follows similarly as that of [44, Theorem 4.1]. Note that the Poisson assumptions are in fact not needed for the result of this lemma to hold. □

As cost in the stochastic model we take $\mathbb{E}\left(\int_0^D \sum_{j=0}^2 c_j N_j^{\pi,r}(t)\mathrm{d}t\right)$, with $D > 0$. As $r \to \infty$, this will tend to infinity. In order to obtain a non-trivial limit we divide the cost by r^2 and consider a horizon that grows linearly in r, i.e., we are interested in

$$\mathbb{E}\Big(\int_0^{r\cdot D}\frac{\sum_{j=0}^2 c_j N_j^{\pi,r}(t)}{r^2}\mathrm{d}t\Big)=\mathbb{E}\Big(\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r}(t)\mathrm{d}t\Big).$$

We have the following lower bound on the scaled cost.

Lemma 5.2.8. *For any policy π we have*

$$\liminf_{r\to\infty}\mathbb{E}\Big(\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r}(t)\mathrm{d}t\Big)\geq\int_0^D\sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t=\int_0^\infty\sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t,$$

whenever $D\geq H(c_0n_0+c_1n_1+c_2n_2)$, and where $n^*(t)$ represents an average-cost optimal trajectory of the fluid control problem for initial state n and $H(\cdot)$ is as defined in Lemma 5.2.4.

Proof: By applying Fatou's lemma, we obtain

$$\begin{aligned}\liminf_{r\to\infty}\mathbb{E}\Big(\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r}(t)\mathrm{d}t\Big) &\geq \mathbb{E}\Big(\liminf_{r\to\infty}\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r}(t)\mathrm{d}t\Big)\\ &= \mathbb{E}\Big(\lim_{k\to\infty}\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r_k}(t)\mathrm{d}t\Big),\end{aligned}$$

with the subsequence r_k (may depend on the sample path ω) corresponding to the lim inf-sequence. Lemma 5.2.7 states that for almost all sample paths ω, there exists a subsequence r_{k_l} of r_k such that $\lim_{l\to\infty}\overline{N}_j^{\pi,r_{k_l}}(t)=\overline{N}_j^{\pi}(t)$, u.o.c., $j=0,1,2$, with $\overline{N}_j^{\pi}(t)$ a fluid limit for initial fluid level n and policy π. Note that a fluid limit is an admissible trajectory for the fluid control problem. When we consider a finite horizon $D\geq H(\sum_{j=0}^2 c_jn_j)$, we obtain from Lemma 5.2.4 that

$$\begin{aligned}&\lim_{l\to\infty}\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r_{k_l}}(t)\mathrm{d}t=\int_0^D\lim_{l\to\infty}\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r_{k_l}}(t)\mathrm{d}t=\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi}(t)\mathrm{d}t\\ &\geq\min_{n(t)\ \text{s.t.}\ (5.3)-(5.6)}\int_0^D\sum_{j=0}^2 c_jn_j(t)\mathrm{d}t=\int_0^D\sum_{j=0}^2 c_jn_j^*(t)\mathrm{d}t=\int_0^\infty\sum_{j=0}^2 c_jn_j^*(t)\mathrm{d}t,\end{aligned}$$

with $n^*(t)$ an average-cost optimal trajectory. Note that in the first step we used uniform convergence of the functions $\overline{N}_j^{\pi,r_{k_l}}(t)$, $j=0,1,2$, on $[0,D]$, in order to interchange the limit and the integral. □

As described in Section 1.6.3, a policy is asymptotically fluid-optimal when the lower bound is obtained. Hence, policy π is asymptotically fluid-optimal when

$$\lim_{r\to\infty}\mathbb{E}\Big(\int_0^D\sum_{j=0}^2 c_j\overline{N}_j^{\pi,r}(t)\mathrm{d}t\Big)=\int_0^\infty\sum_{j=0}^2 c_jn_j^*(t)\mathrm{d}t,$$

with $D \geq H(\sum_{j=0}^{2} c_j n_j)$ and $n^*(t)$ an average-cost optimal trajectory of the fluid control problem for initial state n. In the remainder of this section, we use the optimal fluid controls as described in Propositions 5.2.3 and 5.2.6 to identify asymptotically fluid-optimal policies in the stochastic model. In case $\max(c_1\mu_1, c_2\mu_2) \leq c_0\mu_0$, optimal policies in the stochastic model were found in closed form, see Section 4.3.1. In this section we therefore focus on situations with $\max(c_1\mu_1, c_2\mu_2) > c_0\mu_0$.

We first consider the setting $c_1\mu_1 > c_0\mu_0$ and $\rho_1 < \rho_2$. In that case, an optimal control in the fluid model is characterized by a linear switching curve with a strictly positive slope, see Proposition 5.2.6. In the next lemma we describe the fluid limit of the stochastic model under such policies. The proof may be found in Appendix 5.E.

Lemma 5.2.9. *Assume $\rho_1 < \rho_2$. Denote by $\tilde{\pi}$ the policy with switching curves $h_i(x_0) = k_i x_0$, $i = 1, 2$, with $k_1 < \frac{\mu_1}{\mu_0}\frac{\rho_2 - \rho_1}{1 - \rho_0 - \rho_2}$ and $k_2 \in \{0, \infty\}$. The process $\overline{N}^{\tilde{\pi}}(t)$ is uniquely determined. In addition, the functions $\overline{T}_l^{\tilde{\pi}}(t)$ are differentiable almost everywhere, and for each regular point t it holds that:*
If $k_2 = \infty$, then

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \rho_2, \ \textit{if} \ \ \overline{N}_1^{\tilde{\pi}}(t) < k_1 \overline{N}_0^{\tilde{\pi}}(t), \ \overline{N}_1^{\tilde{\pi}}(t) > 0 \ \ \textit{and} \ \ \overline{N}_2^{\tilde{\pi}}(t) = 0, \tag{5.20}$$

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 0, \ \textit{if} \ \ \overline{N}_1^{\tilde{\pi}}(t), \overline{N}_2^{\tilde{\pi}}(t) > 0 \ \textit{or if} \ \overline{N}_1^{\tilde{\pi}}(t) \geq k_1 \overline{N}_0^{\tilde{\pi}}(t), \tag{5.21}$$

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_0, \ \textit{if} \ \ \overline{N}_0^{\tilde{\pi}}(t) = \overline{N}_1^{\tilde{\pi}}(t) = 0, \tag{5.22}$$

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \rho_1, \ \textit{if} \ \ \overline{N}_0^{\tilde{\pi}}(t) > 0 \ \ \textit{and} \ \ \overline{N}_1^{\tilde{\pi}}(t) = 0, \tag{5.23}$$

and if $k_2 = 0$, then

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \rho_2, \ \textit{if} \ \ \overline{N}_1^{\tilde{\pi}}(t) < k_1 \overline{N}_0^{\tilde{\pi}}(t) \ \ \textit{and} \ \overline{N}_2^{\tilde{\pi}}(t) = 0, \tag{5.24}$$

$$\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 0, \ \textit{if} \ \ \overline{N}_2^{\tilde{\pi}}(t) > 0 \ \textit{or if} \ \overline{N}_1^{\tilde{\pi}}(t) \geq k_1 \overline{N}_0^{\tilde{\pi}}(t). \tag{5.25}$$

In all cases $\frac{\mathrm{d}\overline{T}_i^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t}$, if $\overline{N}_i^{\tilde{\pi}}(t) > 0$, and $\frac{\mathrm{d}\overline{T}_i^{\tilde{\pi}}(t)}{\mathrm{d}t} = \min(\rho_i, 1 - \frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t})$, if $\overline{N}_i^{\tilde{\pi}}(t) = 0$, $i = 1, 2$.

In the next proposition we show that linear switching curves as obtained for the fluid control model provide asymptotically fluid-optimal policies for the original stochastic model.

Proposition 5.2.10. *Assume $\rho_1 < \rho_2$ and $\rho_0 + \rho_2 < 1$.*
If $c_1\mu_1 > c_0\mu_0 \geq c_2\mu_2$, then the policy with switching curves $h_1(x_0) = \frac{c_2\mu_2}{c_1\mu_1 + c_2\mu_2 - c_0\mu_0} \cdot \frac{\mu_1}{\mu_0} \cdot \frac{\rho_2 - \rho_1}{1 - \rho_0 - \rho_2} \cdot x_0$ and $h_2(x_0) = \infty$ is asymptotically fluid-optimal.
If $c_1\mu_1 > c_0\mu_0$ and $c_2\mu_2 \geq c_0\mu_0$, then the policy with switching curves $h_1(x_0) = \frac{c_0}{c_1} \cdot \frac{\rho_2 - \rho_1}{1 - \rho_0 - \rho_2} \cdot x_0$ and $h_2(x_0) = 0$ is asymptotically fluid-optimal.

Proof: Denote by $\tilde{\pi}$ the policy with switching curves $h_i(x_0) = k_i x_0$, $i = 1, 2$, with $k_1 = \frac{c_2\mu_2}{c_1\mu_1+c_2\mu_2-c_0\mu_0} \cdot \frac{\mu_1}{\mu_0} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$ and $k_2 = \infty$ if $c_0\mu_0 > c_2\mu_2$, and $k_1 = \frac{c_0}{c_1} \cdot \frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$ and $k_2 = 0$ if $c_2\mu_2 \geq c_0\mu_0$. Note that when $c_2\mu_2 = c_0\mu_0$, one may also choose $k_2 = \infty$.

Since $c_1\mu_1 > c_0\mu_0$ we have in both cases that $k_1 < \frac{\mu_1}{\mu_0}\frac{\rho_2-\rho_1}{1-\rho_0-\rho_2}$. Hence, the results in Lemma 5.2.9 hold for policy $\tilde{\pi}$, i.e., $\overline{N}^{\tilde{\pi}}(t)$ is uniquely determined and $\overline{T}_j^{\tilde{\pi}}(t)$, satisfies (5.20)–(5.23) when $k_2 = \infty$, and satisfies (5.24) and (5.25) when $k_2 = 0$. Using the correspondence $u_j^*(t) = \frac{\mathrm{d}\overline{T}_j^{\tilde{\pi}}(t)}{\mathrm{d}t}$, $j = 0, 1, 2$, with $u^*(t)$ the average-cost optimal control as defined in Proposition 5.2.6, it follows from (5.19) that $\overline{N}_j^{\tilde{\pi}}(t) = n_j^*(t)$, $j = 0, 1, 2$, with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$.

For a given sample path ω, let r_k be a subsequence such that

$$\liminf_{r\to\infty} \int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r}(t)\mathrm{d}t = \lim_{k\to\infty} \int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r_k}(t)\mathrm{d}t.$$

From Lemma 5.2.7 it follows that for almost all ω there exists a subsequence r_{k_l} of r_k such that $\lim_{l\to\infty} \overline{N}^{\tilde{\pi},r_{k_l}}(t) = \overline{N}^{\tilde{\pi}}(t)$ and $\lim_{l\to\infty} \overline{T}^{\tilde{\pi},r_{k_l}}(t) = \overline{T}^{\tilde{\pi}}(t)$, u.o.c. Since the functions $\overline{N}_j^{\tilde{\pi},r_{k_l}}(t)$, $j = 0, 1, 2$, converge uniformly on the set $[0, D]$, we can interchange the limit and the integral, so that

$$\liminf_{r\to\infty} \int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r}(t)\mathrm{d}t = \lim_{l\to\infty} \int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r_{k_l}}(t)\mathrm{d}t = \int_0^D \sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t.$$

The same holds for the lim sup and we can conclude that for almost all ω,

$$\lim_{r\to\infty} \int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r}(t)\mathrm{d}t = \int_0^D \sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t. \tag{5.26}$$

We have that $\int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r}(t)\mathrm{d}t$ is uniformly integrable. This follows from the same argument as in the proof of [44, Lemma 4.5]. Here we state it briefly. Note that $\overline{N}_j^{\pi,r}(t) \leq (n_j^r + E_j(rt))/r$, with $E_j(\cdot)$ a Poisson process with rate λ_j, $j = 0, 1, 2$. Since $\lim_{r\to\infty} E_j(rt)/r = \lambda_j t$ a.s. (see Lemma 5.2.7) and $\mathbb{E}(E_j(rt)/r) = \lambda_j t$, we obtain from [27, Theorem 3.6] that $E_j(rt)/r$ is uniformly integrable. Since $D < \infty$, uniform integrability of $\int_0^D \sum_{j=0}^2 c_j E_j(rt)/r\mathrm{d}t$ follows as well. Hence, by definition of uniform integrability it is immediate that also the function $\int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\pi,r}(t)\mathrm{d}t$ is uniformly integrable.

We obtain

$$\limsup_{r\to\infty} \mathbb{E}\Big(\int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r}(t)\mathrm{d}t\Big) = \lim_{m\to\infty} \mathbb{E}\Big(\int_0^D \sum_{j=0}^2 c_j \overline{N}_j^{\tilde{\pi},r_m}(t)\mathrm{d}t\Big)$$

$$= \mathbb{E}\Big(\int_0^D \sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t\Big) = \int_0^D \sum_{j=0}^2 c_j n_j^*(t)\mathrm{d}t, \tag{5.27}$$

where in the second step we used (5.26) and uniform integrability to interchange the limit and expectation (see [27, Theorem 3.5]). The subsequence r_m corresponds to the lim sup.

Equation (5.27) holds in particular for $D > H(\sum_{j=0}^{2} c_j n_j)$. Together with Lemma 5.2.8 we can conclude that $\tilde{\pi}$ is asymptotically fluid-optimal. □

We now consider the case $c_2\mu_2 > c_0\mu_0 \geq c_1\mu_1$ and $\rho_1 \leq \rho_2$. In that setting, an optimal control in the fluid model coincides with the policy π_2 (as defined in Section 5.1), see Proposition 5.2.3. In the next proposition we show that π_2 is asymptotically-fluid optimal in the stochastic model.

Proposition 5.2.11. *Assume $\rho_0 + \rho_2 < 1$ and $\rho_1 \leq \rho_2$. If $c_2\mu_2 > c_0\mu_0 \geq c_1\mu_1$, then policy π_2 is asymptotically fluid-optimal and the fluid limit $\overline{N}^{\pi_2}(t)$ is uniquely determined.*

Proof: Note that the functions $\overline{T}_l^{\pi_2}(\cdot)$ are differentiable almost everywhere. In Appendix 5.F we show that for each regular point t it holds that

$$\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} = 0, \text{ if } \overline{N}_2^{\pi_2}(t) > 0, \tag{5.28}$$

$$\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} = 1-\rho_2, \text{ if } \overline{N}_0^{\pi_2}(t) > 0 \text{ and } \overline{N}_2^{\pi_2}(t) = 0, \tag{5.29}$$

$$\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} = \rho_0, \text{ if } \overline{N}_0^{\pi_2}(t) = \overline{N}_2^{\pi_2}(t) = 0. \tag{5.30}$$

In all cases $\frac{\mathrm{d}\overline{T}_i^{\pi_2}(t)}{\mathrm{d}t} = 1 - \frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t}$, if $\overline{N}_i^{\pi_2}(t) > 0$, and $\frac{\mathrm{d}\overline{T}_i^{\pi_2}(t)}{\mathrm{d}t} = \min(\rho_i, 1 - \frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t})$, if $\overline{N}_i^{\pi_2}(t) = 0$, $i = 1, 2$.

From (5.19) and (5.28)–(5.30), it follows that $\overline{N}_j^{\pi_2}(t)$ is uniquely determined. Using the correspondence $u_j^*(t) = \frac{\mathrm{d}\overline{T}_j^{\pi_2}(t)}{\mathrm{d}t}$, $j = 0, 1, 2$, with $u^*(t)$ as defined in Proposition 5.2.3, it follows that $\overline{N}_j^{\pi_2}(t) = n_j^*(t)$, $j = 0, 1, 2$, with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$. The remainder of the proof is similar to the proof of Proposition 5.2.10. □

5.3 Diffusion scaling for $\rho_1 = \rho_2$

In Section 5.2.1 we derived that linear switching curves provide optimal fluid controls. When either $\rho_1 = \rho_2$ and $\min(c_1\mu_1, c_2\mu_2) \leq c_0\mu_0$ or $\rho_1 \neq \rho_2$, these curves approximate the optimal switching curves in the stochastic model very well, and provide asymptotically fluid-optimal policies, see Section 5.2.2. However, when $\rho_1 = \rho_2$ and $c_0\mu_0 < \min(c_1\mu_1, c_2\mu_2)$, this is not the case. In that setting, the optimal switching curves in the fluid control model are both equal to zero, i.e., it is optimal to serve class 0 only if there is work of neither class 1 nor class 2. In the stochastic model, giving classes 1 and 2 preemptive priority leads unnecessarily to an unstable system when $\rho_0 > (1 - \rho_1)(1 - \rho_2)$, see Proposition 3.2.1, and is therefore certainly

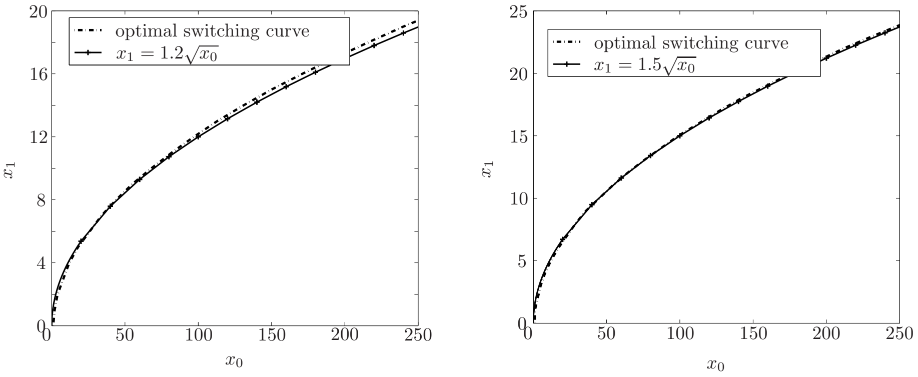


Figure 5.4: Optimal switching curves and square-root approximations, when $\rho_1 = \rho_2$: $\rho_0 = 0.4$, $\rho_1 = \rho_2 = 0.2$ (left) and $\rho_0 = \rho_1 = \rho_2 = 0.3$ (right).

not close-to-optimal. In the fluid control model we have no instability, since we can keep classes 1 and 2 simultaneously empty, while in the stochastic model there can be stochastic fluctuations that cause instability effects.

In Figure 5.4 we plotted the optimal switching curves for parameters that satisfy $\rho_1 = \rho_2$, together with a function that provides a good approximation of the curve. We chose $\mu_0 = 2, \mu_1 = \mu_2 = 5$, $c_0 = c_1 = c_2 = 1$, and plotted $h_1(\cdot)$. (By symmetry, the switching curve $h_2(\cdot)$ is identical.) The figures indicate that the switching curve has a sub-linear shape, and in fact is close to the square-root function. In the previous section, we scaled the processes $N_0(t), N_1(t)$ and $N_2(t)$ identically. Due to its sub-linear shape, the switching curve therefore collapsed on the horizontal axis after taking the limit. This motivates the choice for a different scaling when $\rho_1 = \rho_2$: We need to scale the system such that the switching curves remain observable.

In the remainder of this section we assume $c_0\mu_0 < \min(c_1\mu_1, c_2\mu_2)$. In the case that classes 1 and 2 are both strictly positive, we know that it is optimal to serve both classes 1 and 2 until one of them empties. Without loss of generality, we can therefore concentrate on initial states with $N_2(0) = 0$.

We generically consider switching curves of the shape $h_i(\cdot) = k_i f(\cdot)$, $k_i \geq 0$, for $i = 1, 2$. The function $f(\cdot)$ is not specified for now. Again we consider the sequence of processes indexed by a superscript r, where the workload and number of users in class j at time t are denoted by $W_j^r(t)$ and $N_j^r(t)$, respectively, $j = 0, 1, 2$. The initial state depends on r and is chosen in accordance with the above observations: $N^r(0) = (rn_0, \sqrt{r}m_1, 0)$, with $n_0, m_1 > 0$. We investigate the fluid-scaled processes $\overline{N}_j^r(t)$ and $\overline{W}_j^r(t)$. We will see that $\lim_{r\to\infty} \overline{N}_j^r(t) = \lim_{r\to\infty} \overline{W}_j^r(t) = 0$ for $j = 1, 2$. Therefore, we are also interested in the diffusion-scaled processes $\frac{N_j^r(rt)}{\sqrt{r}}$ and $\frac{W_j^r(rt)}{\sqrt{r}}$, $j = 1, 2$.

Remark 5.3.1. The workload and number of users present in the system under the fluid and diffusion scaling can be related in the following way: $\lim_{r\to\infty} \overline{N}_j^r(t) \overset{d}{=}$

$\lim_{r\to\infty} \mu_j \overline{W}_j^r(t)$ and $\lim_{r\to\infty} \frac{N_j^r(rt)}{\sqrt{r}} \overset{d}{=} \lim_{r\to\infty} \mu_j \frac{W_j^r(rt)}{\sqrt{r}}$, assuming these limits exist. This can be seen as follows. Due to the exponentially distributed service requirements and the fact that we consider non-anticipating policies, we have

$$W_j^r(rt) \overset{d}{=} \frac{\sum_{k=1}^{N_j^r(rt)} \mathrm{Exp}_k(\mu_j)}{N_j^r(rt)} \cdot N_j^r(rt), \tag{5.31}$$

$j = 0, 1, 2$, with $\mathrm{Exp}_k(\mu_j)$, $k = 1, 2, \ldots$, i.i.d. exponential random variables with mean $1/\mu_j$. In addition, it can be proved that $\lim_{r\to\infty} \frac{\sum_{k=1}^{N_j^r(rt)} \mathrm{Exp}_k(\mu_j)}{N_j^r(rt)} \cdot \frac{N_j^r(rt)}{r} \overset{d}{=} \mu_j \lim_{r\to\infty} \frac{N_j^r(rt)}{r}$ and $\lim_{r\to\infty} \frac{\sum_{k=1}^{N_j^r(rt)} \mathrm{Exp}_k(\mu_j)}{N_j^r(rt)} \cdot \frac{N_j^r(rt)}{\sqrt{r}} \overset{d}{=} \mu_j \lim_{r\to\infty} \frac{N_j^r(rt)}{\sqrt{r}}$. Combining this with (5.31) the statements follow.

In Sections 5.3.1 and 5.3.2 we describe the free processes corresponding to the behavior above and below the switching curve. This is used in Section 5.3.3 to explain the square-root shape of the switching curves when $\rho_1 = \rho_2$.

5.3.1 Free process above the switching curve

Class i is given preemptive priority in states above the switching curve $h_i(\cdot)$, $i = 1, 2$. Hence, the free process that corresponds to the behavior above the switching curve $h_i(\cdot)$ is the process that gives class i priority, regardless of the number of class-0 and class-$(3-i)$ users present. This implies that under fluid-scaling the free process is linearly decreasing in class i and linearly increasing in class 0, while keeping class $3-i$ empty.

5.3.2 Free process below the switching curve

We now consider the free process that corresponds to the behavior of the stochastic process below the switching curve. Hence, in the free process, classes 1 and 2 are served during (short) excursions when both of them are positive, or whenever there are no class-0 users present.

We reflect the fact that we look at the free process by adding the symbol $\sim$ to the notation. In the following proposition, it is stated that the free process has two different types of components: the component corresponding to class 0 behaves as a deterministic fluid component, while classes 1 and 2 show random fluctuations of the order $\sqrt{r}$ in a time span r, i.e., their workloads remain of the order $\sqrt{r}$ a.s.

Proposition 5.3.2. *Consider the free process that serves classes 1 and 2 whenever both are present, and otherwise serves class 0. Assume $\rho_1 = \rho_2$. Let $N^r(0) = (rn_0, \sqrt{r}m_1, 0)$ and let τ_0 be the first moment that class 0 is empty. For all $t < \tau_0$, we have*

$$\lim_{r\to\infty} \frac{\tilde{W}_0^r(rt)}{r} \quad = \quad \frac{n_0}{\mu_0} - (1 - \rho_0 - \rho_1)t,$$

and $\lim_{r\to\infty} \frac{\tilde W_i^r(rt)}{r} = 0$ *for* $i = 1, 2$. *In addition, for all* $t < \tau_0$*,*

$$\begin{aligned}\lim_{r\to\infty} \frac{\tilde W_1^r(rt)}{\sqrt{r}} &= \lim_{n\to\infty} \mathbf{1}_{(\tilde W_1^r(rt)\geq \tilde W_2^r(rt))} \frac{\tilde W_1^r(rt) - \tilde W_2^r(rt)}{\sqrt{r}} \\ &\overset{d}{=} \mathbf{1}_{(BM(t)+\frac{m_1}{\mu_1}\geq 0)} \Big(BM(t) + \frac{m_1}{\mu_1}\Big), \qquad (5.32)\end{aligned}$$

$$\begin{aligned}\lim_{r\to\infty} \frac{-\tilde W_2^r(rt)}{\sqrt{r}} &= \lim_{r\to\infty} \mathbf{1}_{(\tilde W_2^r(rt)\geq \tilde W_1^r(rt))} \frac{\tilde W_1^r(rt) - \tilde W_2^r(rt)}{\sqrt{r}} \\ &\overset{d}{=} \mathbf{1}_{(BM(t)+\frac{m_1}{\mu_1}\leq 0)} \Big(BM(t) + \frac{m_1}{\mu_1}\Big), \qquad (5.33)\end{aligned}$$

with $BM(t)$ *a zero-mean Brownian motion with variance* $\theta^2 := \lambda_1/\mu_1^2 + \lambda_2/\mu_2^2$.

Proof: Denote by $A_i(0,t)$ the amount of class-i work that arrived in the interval $(0,t]$ and by $\tilde B_i(0,t)$, the cumulative amount of capacity that is given to class i in the free process in the interval $(0,t]$. Let $t < \tau_0$. For $i = 1, 2$, we have

$$\begin{aligned}\tilde W_i^r(t) &= W_i^r(0) + A_i(0,t) - \tilde B_i(0,t), \qquad (5.34)\\ \tilde B_1(0,t) &= \tilde B_2(0,t), \qquad (5.35)\end{aligned}$$

where the last equality holds since classes 1 and 2 are served only when both of them are positive. Using (5.34) and (5.35), we obtain

$$\tilde W_1^r(rt) - \tilde W_2^r(rt) = A_1(0,rt) - A_2(0,rt) + W_1^r(0) - W_2^r(0).$$

From Remark 5.3.1, the functional central limit theorem [40, 80] and the fact that we have Poisson arrivals, we conclude that

$$\begin{aligned}\lim_{r\to\infty} \frac{1}{\sqrt{r}}\Big(\tilde W_1^r(rt) - \tilde W_2^r(rt)\Big) &= \lim_{r\to\infty} \frac{1}{\sqrt{r}}\Big(A_1(0,rt) - A_2(0,rt) + W_1^r(0) - W_2^r(0)\Big)\\ &\overset{d}{=} BM(t) + \frac{m_1}{\mu_1}, \qquad (5.36)\end{aligned}$$

where $BM(t)$ is a zero-mean Brownian motion with variance θ^2.

Define $s^* := \inf\{s \leq t :$ classes 1 and 2 are continuously backlogged in $(s,t]\}$. Since classes 1 and 2 are served at rate 1 in the interval $(s^*,t]$, this implies $\tilde W_i^r(t) = \tilde W_i^r(s^*) + A_i(s^*,t) - (t - s^*)$, $i = 1, 2$. Denote by $\tilde W_{min}^r(t) := \min(\tilde W_1^r(t), \tilde W_2^r(t))$ the minimum workload in classes 1 and 2. We have

$$\begin{aligned}\tilde W_{min}^r(t) &\leq \min(\tilde W_1^r(s^*), \tilde W_2^r(s^*)) + \max(A_1(s^*,t), A_2(s^*,t)) - (t - s^*)\\ &\leq \min(\tilde W_1^r(s^*), \tilde W_2^r(s^*)) + \sup_{s\leq t}\{\hat A(s,t) - (t-s)\}, \qquad (5.37)\end{aligned}$$

where $\hat A(s,t) := \max(A_1(s,t), A_2(s,t))$. Using the fact that $\hat A(s,t)/(t-s) \to \rho_1 < 1$ as $t - s \to \infty$, we have from [9, Corollary III.7.2] that the right-hand side of (5.37) converges (as $t \to \infty$) to a random variable that is finite almost surely. Consequently, $\lim_{r\to\infty} \tilde{W}^r_{min}(rt)$ is bounded from above by a non-defective random variable, which implies $\lim_{r\to\infty} \frac{\tilde{W}^r_{min}(rt)}{\sqrt{r}} = 0$, a.s. Together with (5.36) and

$$\tilde{W}^r_i(t) = (\tilde{W}^r_i(t) - \tilde{W}^r_{3-i}(t) + \tilde{W}^r_{min}(t))\mathbf{1}_{(\tilde{W}^r_i(t)\geq\tilde{W}^r_{3-i}(t))} + \tilde{W}^r_{min}(t)\mathbf{1}_{(\tilde{W}^r_i(t)<\tilde{W}^r_{3-i}(t))},$$

we obtain (5.32) and (5.33).

As long as class 0 is not empty, both nodes are work-conserving, so that

$$\tilde{W}^r_0(rt) + \tilde{W}^r_1(rt) = W^r_0(0) + W^r_1(0) + A_0(0,rt) + A_1(0,rt) - rt.$$

Since $\lim_{r\to\infty} \frac{1}{r}\tilde{W}^r_1(rt) = 0$, this gives $\lim_{r\to\infty} \frac{\tilde{W}^r_0(rt)}{r} = \frac{n_0}{\mu_0} + (\rho_0 + \rho_1 - 1)t$. □

5.3.3 Shape of switching curve

In this section we intuitively explain the square-root shape of the optimal switching curves when $\rho_1 = \rho_2$ and $\min(c_1\mu_1, c_2\mu_2) > c_0\mu_0$. From the fluid control model, we learned that a switching curve in the stochastic model should lie close to the horizontal axis (the optimal switching curve in the fluid control model in fact coincides with the horizontal axis, see Proposition 5.2.6). Letting the switching curve be too close to the horizontal axis, however, poses the risk of significant capacity loss: capacity is lost in node 2 if there are no class-2 users and the process is in a state above the switching curve $h_1(\cdot)$ and, vice versa, capacity is lost in node 1 if there are no class-1 users and the process is in a state above the switching curve $h_2(\cdot)$. The switching curve must therefore be high enough to make it sufficiently unlikely for the process to reach it from below. But it should not be impossible to reach the switching curve, because above the switching curve the weighted departure rate is higher. Remark 5.3.1 and Proposition 5.3.2 give that the free processes $\tilde{N}_1(t)$ and $\tilde{N}_2(t)$ below the switching curves have zero drift whose fluctuations in linear time O(r) are of the order O$(\sqrt{r})$. This indicates that square-root switching curves, i.e., $h_i(x_0) = k_i\sqrt{x_0}$, $k_i \geq 0$, $i = 1, 2$, are able to strike the right balance between the above-described effects. For comparison: a linear switching curve would be impossible to reach, therefore the policy would not profit from serving the fast class 1 or class 2 even if there is a lot of work of it. On the other hand, a threshold policy (i.e., a constant switching curve) can quickly give instability problems as, at large states, it is too easy to move up to the switching curve, thus risking considerable capacity loss.

In fact, we believe that square-root shaped switching curves provide asymptotically fluid-optimal policies, however, we do not formally prove this. One approach to do this would be to determine the fluid limits. The latter requires further investigation of the reflection of the process on the switching curves and demands calculating the first-passage probabilities of the zero-mean Brownian motion to the square-root switching curves. This is not trivial, see for example [4, 18, 103]. For the same reason, finding the best value for the coefficient k_i is not straightforward. In the remainder of this section we therefore numerically illustrate the impact of the choice for k_i.

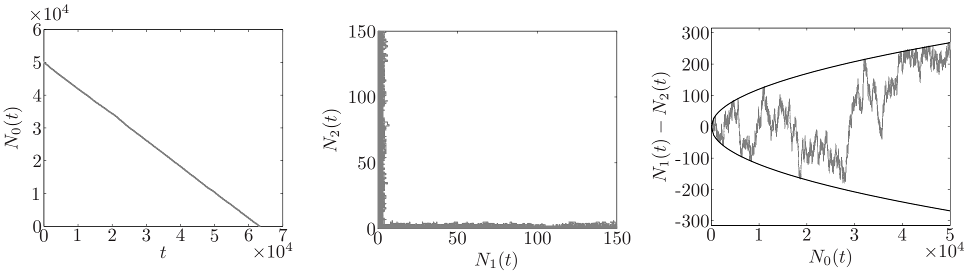


Figure 5.5: Trajectories of the processes $N_0(t), N_1(t), N_2(t)$ and $N_1(t) - N_2(t)$ under a policy with switching curves $h_i(x_0) = 6/5\sqrt{x_0}$, $i = 1, 2$.

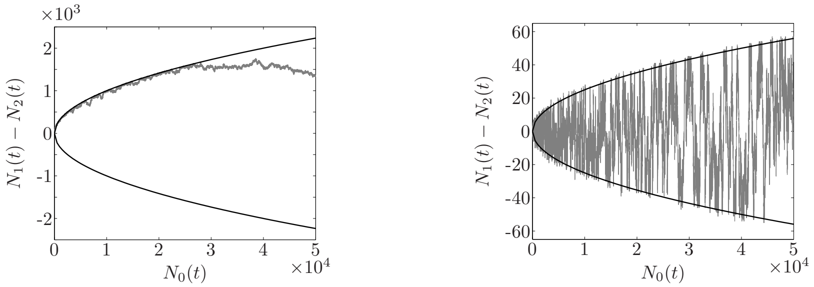


Figure 5.6: Trajectory of the process $N_1(t) - N_2(t)$ under the policy with switching curves $h_i(x_0) = 10\sqrt{x_0}$, $i = 1, 2$ (left), and $h_i(x_0) = 1/4\sqrt{x_0}$, $i = 1, 2$ (right).

We set $c_0 = c_1 = c_2 = 1$, $\rho_0 = 0.4, \rho_1 = \rho_2 = 0.2$ and $\mu_0 = 2, \mu_1 = \mu_2 = 5$. In the first simulation we chose the switching curves $h_i(x_0) = k_i\sqrt{x_0}$ with $k_i = 6/5$, for $i = 1, 2$. In Figure 5.5 we see that the number of class-0 users indeed decreases linearly in time (left graph), while the minimum of the number of class-1 and class-2 users is typically very small (middle graph). The right most graph shows the trajectory of the difference between the number of class-1 and class-2 users. In addition, both switching curves are plotted. Recall from Proposition 5.3.2 (and Remark 5.3.1) that, after diffusion scaling, $\tilde{N}_1(t) - \tilde{N}_2(t)$ represents $\tilde{N}_1(t)$ when it is positive, and $-\tilde{N}_2(t)$ when it is negative. We see that as the number of class-0 users decreases, the trajectory stays mostly between the two switching curves, making some excursions between the switching curves in both planes.

Taking k_i very large implies that for points that lie (for example) just below the switching curve $h_1(\cdot)$, the probability of emptying the work in class 1 and hitting the switching curve $h_2(\cdot)$ becomes almost zero. See Figure 5.6 (left) where a trajectory is plotted for $k_i = 10$. Therefore, the policy focuses too much on being work-conserving in node 1. On the other hand, taking k_1 too small, we see that we switch too often between the two planes, and we loose unnecessarily a considerable amount of capacity, see Figure 5.6 (right) where $k_1 = k_2 = 1/4$.

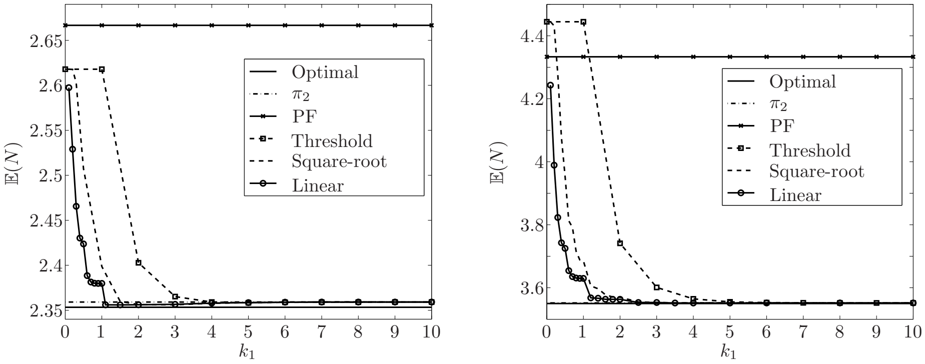


Figure 5.7: Total mean number of users under the optimal policy, π_2, PF, and switching-curve policies when $\rho_0 = 0.4, \rho_1 = 0.1, \rho_2 = 0.3$ (left) and $\rho_0 = 0.4, \rho_1 = 0.1, \rho_2 = 0.4$ (right).

5.4 Numerical evaluation

In this section we numerically compare the performance of weighted α-fair bandwidth-sharing policies (as defined in Section 1.4.1) and fluid-based and diffusion-based switching-curve policies with that of the true optimal policy. The latter is determined using value iteration after a suitable state space truncation. Throughout this section we assume $c_0 = c_1 = c_2 = 1$, i.e., we focus on minimizing the total number of users. In Section 5.4.1 we assess the effectiveness of different switching-curve policies and in Section 5.4.2 we investigate whether an optimal policy can be approximated with a weighted α-fair policy. Throughout this section, we use the notation $N^\pi := \sum_{j=0}^{2} N_j^\pi$.

5.4.1 Switching-curve policies

We have conducted a large set of simulation experiments to assess the effectiveness of different switching-curve policies. Under these policies, classes 1 and 2 are served whenever both are present. When $N_{3-i}(t) = 0$ for an $i = 1, 2$, class 0 is served when $N_i(t) < h_i(N_0(t))$, and class i is served otherwise. We consider switching curves of the shape $h_i(x_0) = k_i f(x_0)$, $i = 1, 2$, where the function $f(\cdot)$ is either a square-root, linear or is equal to one. The latter is referred to as a threshold policy. The value of k_i is varied to assess its impact. We let $\mu_0 = 2$, $\mu_1 = \mu_2 = 5$. We simulate in the order of 10^6 busy periods and the obtained total mean number of users under the different policies are compared with the optimal policy and with PF. PF falls within the class of weighted α-fair policies by setting $\alpha = 1$ and $w_j = 1$, $j = 0, 1, 2$. The mean numbers of users of the various classes under PF are as given in (4.23) and (4.24).

In Figure 5.7 we considered $\rho_1 \neq \rho_2$, chose $k_2 = 0$, and let k_1 vary. We ob-

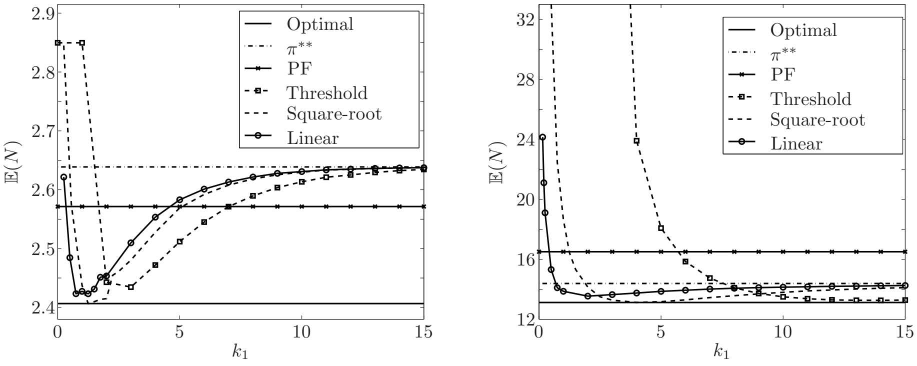


Figure 5.8: Total mean number of users under the optimal policy, π^{**}, PF, and switching-curve policies when $\rho_0 = \rho_1 = \rho_2 = 0.3$ (left) and $\rho_0 = 0.6, \rho_1 = \rho_2 = 0.3$ (right).

serve that a policy with a linear switching curve attains the value of the optimal policy provided that the best coefficient k_1 is chosen. This is in accordance with Proposition 5.2.10 which stated that when $c_1\mu_1, c_2\mu_2 \geq c_0\mu_0$, the asymptotically fluid-optimal policy has a linear switching curve for class 1, $h_1(x_0) = (\rho_2 - \rho_1)/(1 - \rho_0 - \rho_2)x_0$, and gives preemptive priority to class 2. The slope of the linear curve equals 2/3 for the parameter setting of the left plot and 3/2 for the parameter setting of the right plot. We observe that these slopes are already close to optimal. In addition, we observe that the square-root policy also performs very well. Furthermore, note that when k_1 grows large (and $k_2 = 0$), the behavior of the system under the considered switching curve policies converges to that of policy π_2. Policy π_2 is already close to optimal. This is not surprising, since policy π_2 is asymptotically fluid-optimal when node 2 is heavily loaded ($\rho_0 + \rho_2 \approx 1$ and $\rho_0 + \rho_1 < 1$), see Proposition 5.2.10.

In Figure 5.8 we considered $\rho_1 = \rho_2$ and chose $k_1 = k_2$, i.e., the switching curves for both classes are identical. We observe that the square-root policy attains the value of the optimal policy provided that the best coefficient k_1 is chosen. This agrees with the discussion of the shape of the switching curve in Section 5.3.3. Also note that the setting in the left graph in Figure 5.8 corresponds to the right graph in Figure 5.4. The approximation we found there for the switching curve was $h_i(x_0) = 1.5\sqrt{x_0}$ which indeed is close to optimal. When k_1 grows large, the behavior of the system resembles that of policy $\pi^{**} \in \Pi^{**} \cap \bar{\Pi}$, as defined in Section 4.1.

As described in Section 5.3.3, we did not derive an estimate for an efficient preconstant k_1 when $\rho_1 = \rho_2$. Therefore, in Figures 5.9 and 5.10 we test the impact of the value for k_1 on the square-root, linear, and threshold policies for several combinations of the loads with $\rho_1 = \rho_2$. We compare the total mean number of users under the switching-curve policies with PF. We set $k_2 = k_1$. We observe that square-root switching curves perform best, provided that the value of k_1 is chosen

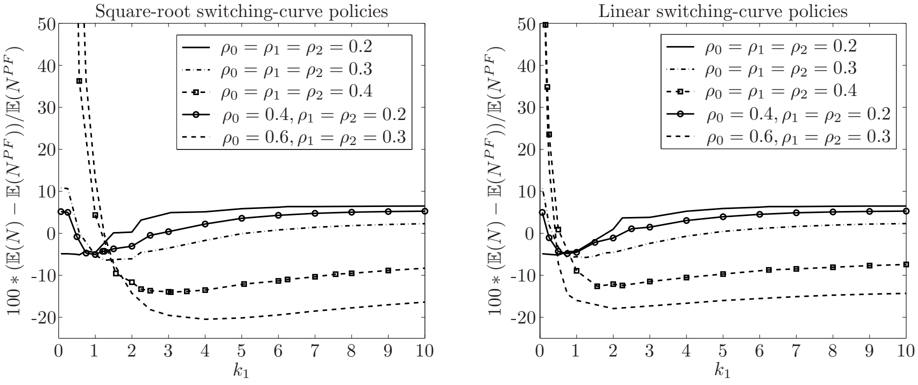


Figure 5.9: Comparison of PF with square-root (left) and linear (right) switching-curve policies.

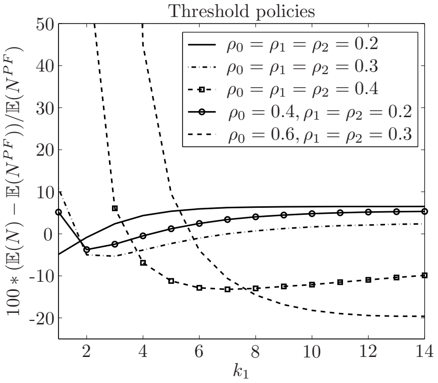


Figure 5.10: Comparison of PF with threshold policies.

optimally. The linear switching curves perform surprisingly well, although they are less efficient than the square-root policies. The square-root and linear policies are not that sensitive to the actual value of k_1, as long as its value is not too small. The threshold policies are more sensitive to the value of k_1 and run a higher risk of being unstable. Overall, we observe that the best choice among these policies only gives a modest improvement over PF (5-20%).

5.4.2 Weighted α-fair policies

For completeness, in Figure 5.11 we plot the relative improvement of α-fair policies (with unit weights) over the total mean number of users under PF ($\alpha = 1$). We chose $\mu_0 = 2$, $\mu_1 = \mu_2 = 5$ and different settings of the loads. We notice that α-fair policies seem to be rather insensitive to the value of α, as long as α is not too small.

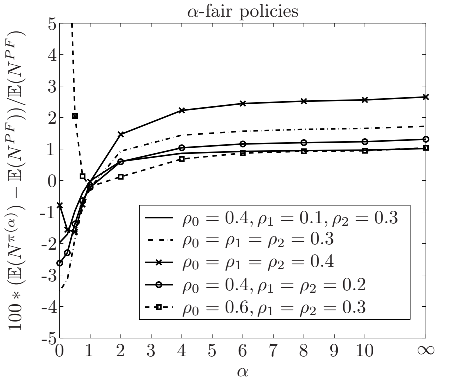


Figure 5.11: Comparison of α-fair policies with PF for different settings of the loads.

We next test the scope for improvement of α-fair policies by adding weights to the various classes, i.e., we focus on *weighted* α-fair bandwidth-sharing policies. We numerically investigate the effect of changing the weights and check whether we can approximate the optimal policy with a weighted α-fair policy. Without loss of generality, we fix $w_0 = 1$. In the numerical examples we choose $\mu_0 = 2$ and $\mu_1 = \mu_2 = 5$, but the observations hold more generally for $\mu_0 < \mu_1, \mu_2$.

In Figure 5.12, with $\alpha = 1$ corresponding to PF, and Figure 5.13, with $\alpha = 2$, we compare weighted α-fair policies with the optimal policy. We note that the gap between the best weighted α-fair policy and the optimal policy is at most 5%. From the left graphs in Figures 5.12 and 5.13 we observe that when $\rho_1 < \rho_2$, choosing w_1 close to zero and $w_2 = \infty$ approximates the optimal policy very well. In fact, when choosing these weights the policy resembles π_2, which, as observed in Section 5.4.1, is close to optimal. From the right graphs in Figures 5.12 and 5.13 we observe that when $\rho_1 = \rho_2$, choosing one of the two weights equal to ∞ and the other weight strictly positive, approximates the optimal policy very well. More precisely, for PF these weights are $w_i = 1/2$, $w_{3-i} = \infty$, $i = 1, 2$ and for $\alpha = 2$ these weights are $w_i = 1/8, w_{3-i} = \infty$, $i = 1, 2$. This can be explained as follows. From Proposition 4.3.9 we know that when both class-1 and class-2 users are present, the optimal allocation gives the full capacity to classes 1 and 2. Having one of the weights equal to ∞, say w_2, guarantees that the weighted α-fair policy does this as well. Now when there are no class-2 users present, there exists a switching curve that determines the optimal trade-off between serving class 0 or class 1, see Proposition 4.3.9. In the case of a weighted α-fair policy, when there are no class-2 users present the allocated capacity to class 0 is $s_0(t) = \frac{N_0(t)}{N_0(t)+w_1^{1/\alpha}N_1(t)}$. There exists a $0 \leq w_1 < \infty$ that strikes the right balance to share the capacity between class 0 and class 1. Note that $w_1 = 0$ ($w_1 = \infty$) implies that class 0 (class 1) is given strict priority when class 2 is not present.

Similarly, efficient weights for the remaining cases can be obtained. When

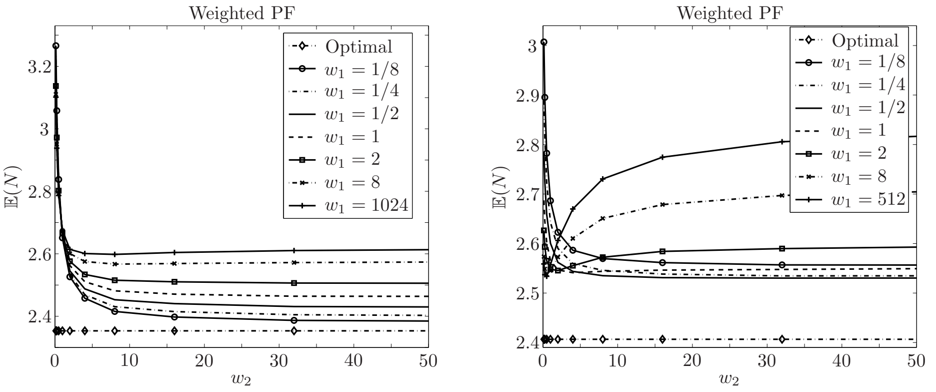


Figure 5.12: Total mean number of users under the optimal policy and weighted PF ($\alpha = 1$) for different choices of the weights: (left) $\rho_0 = 0.4, \rho_1 = 0.1, \rho_2 = 0.3$, (right) $\rho_0 = \rho_1 = \rho_2 = 0.3$.

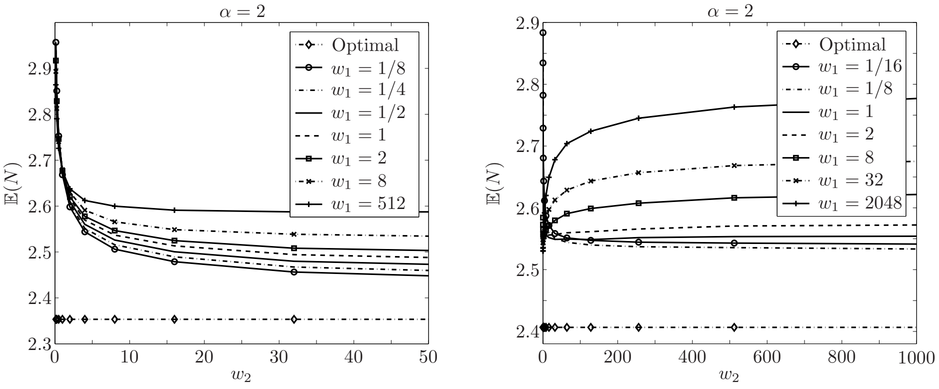


Figure 5.13: Total mean number of users under the optimal policy and the weighted α-fair policy with $\alpha = 2$ for different choices of the weights: (left) $\rho_0 = 0.4, \rho_1 = 0.1, \rho_2 = 0.3$, (right) $\rho_0 = \rho_1 = \rho_2 = 0.3$.

$\mu_0 > \mu_1 + \mu_2$, an optimal policy gives preemptive priority to class 0 (Proposition 4.3.6). This policy can be approximated by the weighted α-fair policy by setting the weights w_1 and w_2 equal to zero. When $\mu_1, \mu_2 < \mu_0 < \mu_1 + \mu_2$, an optimal policy gives preemptive priority to classes 1 and 2 whenever both are present and otherwise serves class 0 (Proposition 4.3.8). We expect that an α-fair policy with weights w_1 and w_2 strictly positive, but smaller than $w_0 = 1$, will approximate this optimal policy. When $\mu_2 < \mu_0 < \mu_1$ and $\rho_1 \geq \rho_2$, the weights $w_1 = \infty$ and $w_2 = 0$ approximate the corresponding asymptotically fluid-optimal policy π_1 (Proposition 5.2.11). However, when $\rho_1 < \rho_2$, the asymptotically fluid optimal policy is described by a switching curve in states where there are no class-2 users present (Proposition 5.2.10). In that case we numerically observed that a weighted α-fair policy with w_1 non-degenerate and $w_2 = 0$, is close to optimal.

5.5 Concluding remarks

Using scaling approaches, we determined accurate approximations to optimal policies in a linear bandwidth-sharing network. These policies were shown to provide sensible benchmarks for assessing the performance of bandwidth-sharing policies. We showed that weighted α-fair policies performed well in all our experiments, and, when the weights are chosen appropriately, are within a few percent from the theoretical optimum.

Despite its simplicity, the two-node network already illustrates the essential complexity of the scheduling problem, and serves as a basis for the analysis of more general networks. For linear networks with $L > 2$ nodes we expect that an optimal fluid control can again be described by switching curves: When class k empties, with $k = 1, \ldots, L$ such that $\rho_k = \min(\rho_1, \ldots, \rho_L)$, an optimal fluid trajectory will keep this class empty from that moment on. Hence, in that case the fluid trajectories can be reduced to typical paths in a network with one node less. However, when class i, $i \neq 0$, with $\rho_i > \rho_k$ empties, while class k is still present, one needs to determine optimal switching points by calculating the corresponding costs (as was done in the proof of Proposition 5.2.6 for the case of two nodes).

As a final remark, we note that the best multiplicative preconstant of the square-root switching curve (for the case $\rho_1 = \rho_2$) has so far been determined numerically. We saw in all our experiments that the optimum can, indeed, be attained for a specific choice of the multiplier. The computation time of this procedure is virtually negligible compared with numerically determining the true optimal policy. In order to *analytically* characterize the optimal value of the preconstant, further investigation of the reflection of the process on the switching curve is required.

Appendix

5.A Proof of Lemma 5.1.1

We couple the systems that arise under the two policies by taking the same arrival time and service requirement sequences. We will show that (5.1) and (5.2) hold on each sample path. Since the service requirements are exponentially distributed, the scheduling within classes does not influence the stochastic behavior of the system (recall that we restrict the attention to size-oblivious policies). For our coupling arguments it is convenient to assume that FCFS is applied within each class. As a consequence, if $W_i^g(t) \leq W_i^h(t)$, then this immediately translates to the same inequality in terms of the numbers of users. Define $s := \inf\{t > 0 : (5.1) \text{ or } (5.2) \text{ is violated}\}$. An inequality can only be violated immediately after time s when it holds with equality at time s. In the proof we use $f(t^+) > f'(t^+)$ to denote that there exists a sufficiently small $\delta > 0$ such that $f(u) > f'(u)$ for all $u \in (t, t+\delta]$.

First, assume that immediately after time s, equation (5.1) is violated, that is $W_0^g(s^+) > W_0^h(s^+)$ while $W_0^g(s) = W_0^h(s)$. From (5.2) we have $W_i^g(s) \leq W_i^h(s)$, $i = 1, 2$. To ensure that $W_0^g(s^+) > W_0^h(s^+)$, policy g must serve classes 1 and/or 2 while policy h serves class 0 at time s. Since $W_i^g(s) \leq W_i^h(s)$ (and hence $N_i^g(s) \leq N_i^h(s)$), $i = 1, 2$, serving classes 1 and/or 2 under policy g implies that also under policy h classes 1 and/or 2 are served (since $h_i(n_0) \leq g_i(n_0)$ and $N_0^g(s) = N_0^h(s)$), which yields a contradiction.

Now, assume equation (5.2) for $i = 2$ is the first to be violated immediately after time s. Hence, $W_0^g(s) + W_2^g(s) = W_0^h(s) + W_2^h(s)$, $W_0^g(s^+) + W_2^g(s^+) > W_0^h(s^+) + W_2^h(s^+)$, $W_0^g(s) \leq W_0^h(s)$. This implies that at time s, policy g serves class 1 and there is no work of class 2 present ($W_2^g(s) = 0$), while policy h serves either class 0 or class 2. We can conclude from the above that $W_2^g(s) \geq W_2^h(s)$. But $W_2^g(s) = 0$, so that $W_2^h(s) = 0$ as well, and hence $W_0^g(s) = W_0^h(s)$. By (5.2) we now obtain $W_1^g(s) \leq W_1^h(s)$. Since $h_i(n_0) \leq g_i(n_0)$ and since policy g serves class 1, policy h serves class 1 as well, which contradicts the initial assumption. □

5.B Proof of Lemma 5.1.3

We set $l = 2$. Node 2 is work-conserving under policy π_2, hence classes 0 and 2 are stable when $\rho_0 + \rho_2 < 1$. In particular, the workload in classes 0 and 2 is finite a.s. In the remainder of the proof we show that the class-1 workload is finite a.s. as well. Let $A_j(s,t)$ denote the amount of class-j work that arrived in the interval $(s,t]$, and let $B_j(s,t)$ denote the cumulative amount of capacity that is given to class j in the interval $(s,t]$. Define $s_1 := \sup\{u \leq t : W_1(u) = 0\}$ and $s := \sup\{u \leq s_1 : W_0(u) + W_2(u) = 0\}$. Then,

$$\begin{aligned}
&W_0(t) + W_1(t) = W_0(t) + A_1(s_1,t) - B_1(s_1,t)\\
&= W_0(t) + A_1(s_1,t) - (t - s_1) + B_0(s_1,t)\\
&= W_0(t) + A_1(s_1,t) - (t - s_1) + W_0(s_1) - W_0(t) + A_0(s_1,t)\\
&\leq A_1(s_1,t) - (t - s_1) + W_0(s_1) + W_2(s_1) + A_0(s_1,t)\\
&= A_1(s_1,t) - (t - s_1) + A_0(s_1,t) + A_0(s,s_1) + A_2(s,s_1) - (s_1 - s)\\
&= A_1(s_1,t) + A_0(s,t) + A_2(s,t) - A_2(s_1,t) - (t - s)\\
&= A_1(s_1,t) - (\rho_1 + \epsilon)(t - s_1) + A_0(s,t) - (\rho_0 + \epsilon)(t - s)\\
&\quad + A_2(s,t) - (\rho_2 + \epsilon)(t - s) + (\rho_2 - \epsilon)(t - s_1) - A_2(s_1,t) + R,
\end{aligned} \tag{5.38}$$

with $\epsilon = \frac{1-\rho_0-\max(\rho_1,\rho_2)}{4}$ and $R = (\rho_1+\epsilon)(t-s_1) + (\rho_0+\epsilon)(t-s) + (\rho_2+\epsilon)(t-s) - (\rho_2-\epsilon)(t-s_1) - (t-s)$. The fourth equation follows from the fact that node 2 is work-conserving, i.e., when node 2 is backlogged, the work is served at full rate.

For $\rho_2 \geq \rho_1$, we can bound R from above as follows:

$$\begin{aligned} R &\leq (\rho_2+\epsilon)(t-s_1)+(\rho_0+\epsilon)(t-s)+(\rho_2+\epsilon)(t-s)-(\rho_2-\epsilon)(t-s_1) \\ &\quad -(t-s) \\ &= (\rho_0+\rho_2-1)(t-s)+\epsilon(4t-2s_1-2s) \leq (\rho_0+\rho_2+4\epsilon-1)(t-s)=0. \end{aligned}$$

For $\rho_2 \leq \rho_1$, we have $\rho_1-\rho_2+2\epsilon \geq 0$ and we bound R from above as follows:

$$\begin{aligned} R &= t(\rho_0+\rho_1+4\epsilon-1)-s(\rho_0+\rho_2+2\epsilon-1)-s_2(\rho_1-\rho_2+2\epsilon) \\ &\leq t(\rho_0+\rho_1+4\epsilon-1)-s(\rho_0+\rho_2+2\epsilon-1)-s(\rho_1-\rho_2+2\epsilon) \\ &= (\rho_0+\rho_1+4\epsilon-1)(t-s)=0. \end{aligned}$$

Denote by $\hat{W}_i^c(t)$, the workload at time t in a reference system with class-i traffic only, service rate c, and with $\hat{W}_i^c(0)=0$. Define $U_j^d(t) := \sup_{0\leq s\leq t}\{d(t-s)-A_j(s,t)\}$. Since $R \leq 0$, we have from (5.38) that

$$\begin{aligned} & W_0(t)+W_1(t) \\ & \leq \sup_{0\leq s\leq t}\{A_1(s,t)-(\rho_1+\epsilon)(t-s)\} + \sup_{0\leq s\leq t}\{A_0(s,t)-(\rho_0+\epsilon)(t-s)\} \\ & \quad + \sup_{0\leq s\leq t}\{A_2(s,t)-(\rho_2+\epsilon)(t-s)\} + \sup_{0\leq s\leq t}\{(\rho_2-\epsilon)(t-s)-A_2(s,t)\} \\ & = \hat{W}_1^{\rho_1+\epsilon}(t)+\hat{W}_0^{\rho_0+\epsilon}(t)+\hat{W}_2^{\rho_2+\epsilon}(t)+U_2^{\rho_2-\epsilon}(t). \end{aligned} \tag{5.39}$$

The first three terms in (5.39) represent workloads in stable queues, since the service rate is larger than the offered loads. Hence, the first three terms are finite a.s, [9]. By [9, Corollary III.7.2] we obtain that the fourth term in (5.39) converges to the supremum of a random walk with drift $\rho_2-\epsilon-\rho_2<0$. Since the drift is negative, we obtain in particular that $U_2^{\rho_2-\epsilon}(t)<\infty$ a.s. Hence the workload in node 1 can be bounded from above by four terms that are finite a.s. □

5.C Proof of Lemma 5.2.2

Without loss of generality, assume that $\tilde{\pi}$ is Pareto-efficient, that is, it does not leave capacity unnecessarily unused. Assume there exists a $\Delta>0$ such that policy $\tilde{\pi}$ does not satisfy (5.8)–(5.10) for all $t\in(0,\Delta)$ (without loss of generality, we assume that such an interval starts at time 0). This implies in particular that $n_1^{\tilde{\pi}}(t), n_2^{\tilde{\pi}}(t)>0$ for all $t\in(0,\Delta)$, and hence $u_0^{\tilde{\pi}}(t)>0$, for all $t\in(0,\Delta)$.

We construct policy π as follows. For all $0\leq t\leq \tilde{T}$ we set

- $u_0^{\pi}(t)=0$, if $n_1^{\pi}(t), n_2^{\pi}(t)>0$ or if $n_0^{\pi}(t)=0$,
- $u_0^{\pi}(t)=1-\rho_i$, if $n_0^{\pi}(t)>0$, $n_i^{\pi}(t)=0$, and $n_j^{\pi}(t)>0$, $i\neq j$, $i,j=1,2$,
- $u_0^{\pi}(t)=1-\rho_i$, if $n_0^{\pi}(t)>0$ and $n_i^{\pi}(t)=n_j^{\pi}(t)=0$, $\rho_i\leq\rho_j$, $i\neq j$, $i,j=1,2$,

and $u_i^\pi(t) = 1 - u_0^\pi(t)$, if $n_i(t) > 0$, and $u_i^\pi(t) = \min(\rho_i, 1 - u_0^\pi(t))$, if $n_i(t) = 0$, $i = 1, 2$. The time $\tilde{T}$ is defined as $\min(\tilde{T}_0, \tilde{T}_{12})$, with $\tilde{T}_0 := \inf\{t > 0 : n_0^{\tilde{\pi}}(t) = n_0^\pi(t)\}$ and $\tilde{T}_{12} := \inf\{t > 0 : n_1^{\tilde{\pi}}(t) = 0 \text{ or } n_2^{\tilde{\pi}}(t) = 0\} > 0$. After time $\tilde{T}$, policy π takes the same decisions as policy $\tilde{\pi}$ (whenever possible), and otherwise idles.

We assumed $n^\pi(0) = n^{\tilde{\pi}}(0)$. Since $\tilde{\pi}$ does not satisfy (5.8)–(5.10) and by definition of π, it follows that for all $\epsilon > 0$ small enough, $U_0^\pi(\epsilon) < U_0^{\tilde{\pi}}(\epsilon)$, so that $n_0^\pi(\epsilon) > n_0^{\tilde{\pi}}(\epsilon)$ and $\tilde{T}_0 > 0$. In particular, this yields $n_0^\pi(t) > 0$ for all $0 < t < \tilde{T}$, so that, by definition of policy π, we obtain $u_0^\pi(t) + u_i^\pi(t) = 1$, for $i = 1, 2$. Since $n_i^{\tilde{\pi}}(t) > 0$, for all $0 < t \leq \tilde{T}$, we have as well $u_0^{\tilde{\pi}}(t) + u_i^{\tilde{\pi}}(t) = 1$, $i = 1, 2$. This implies

$$U_0^\pi(t) + U_i^\pi(t) = U_0^{\tilde{\pi}}(t) + U_i^{\tilde{\pi}}(t), \quad \text{for all } 0 < t \leq \tilde{T}. \tag{5.40}$$

Since $\tilde{T} \leq \tilde{T}_0$ and $U_0^\pi(\epsilon) < U_0^{\tilde{\pi}}(\epsilon)$ for $\epsilon > 0$ small enough, it holds in particular that $U_0^\pi(t) \leq U_0^{\tilde{\pi}}(t)$, for all $t \leq \tilde{T}$. Hence, we obtain that for all $t \leq \tilde{T}$,

$$\begin{aligned}\sum_{j=0}^{2} c_j n_j^\pi(t) &= \sum_{j=0}^{2} c_j n_j(0) + t \cdot \sum_{j=0}^{2} \lambda_j - c_0\mu_0 U_0^\pi(t) - (c_1\mu_1 + c_2\mu_2)(t - U_0^\pi(t)) \\ &\leq \sum_{j=0}^{2} c_j n_j(0) + t \cdot \sum_{j=0}^{2} \lambda_j - c_0\mu_0 U_0^{\tilde{\pi}}(t) - (c_1\mu_1 + c_2\mu_2)(t - U_0^{\tilde{\pi}}(t)) = \sum_{j=0}^{2} c_j n_j^{\tilde{\pi}}(t),\end{aligned}$$

where we used that $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0$. Hence, policy π does better than policy $\tilde{\pi}$, for all $0 \leq t \leq \tilde{T}$.

We conclude the proof by showing that $U_j^\pi(\tilde{T}) = U_j^{\tilde{\pi}}(\tilde{T})$, $j = 0, 1, 2$, which implies $n^\pi(\tilde{T}) = n^{\tilde{\pi}}(\tilde{T})$. In order to do this, we distinguish between the two possible values of $\tilde{T}$: (i) Assume $\tilde{T} = \tilde{T}_0$. Then, by definition, $U_0^\pi(\tilde{T}) = U_0^{\tilde{\pi}}(\tilde{T})$, and hence by (5.40), $U_i^\pi(\tilde{T}) = U_i^{\tilde{\pi}}(\tilde{T})$, $i = 1, 2$. (ii) Assume $\tilde{T} = \tilde{T}_{12}$. Then $n_i^{\tilde{\pi}}(\tilde{T}) = 0$, for an $i = 1, 2$. From (5.40) and $U_0^\pi(\tilde{T}) \leq U_0^{\tilde{\pi}}(\tilde{T})$, we obtain $U_i^\pi(\tilde{T}) \geq U_i^{\tilde{\pi}}(\tilde{T})$, so that $n_i^\pi(\tilde{T}) \leq n_i^{\tilde{\pi}}(\tilde{T}) = 0$. This yields $U_i^\pi(\tilde{T}) = U_i^{\tilde{\pi}}(\tilde{T})$, and hence by (5.40), $U_j^\pi(\tilde{T}) = U_j^{\tilde{\pi}}(\tilde{T})$, $j = 0, 1, 2$. □

5.D Proof of Lemma 5.2.4

By the Filippov-Cesari theorem [122, Chapter 2.8], there exists an optimal control $u^{*D}(t)$ and a corresponding optimal trajectory $n^{*D}(t)$ for the problem $\min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^D (\sum_{j=0}^2 c_j n_j(t)) \mathrm{d}t$, for any $D \geq 0$.

For the moment, assume there exists a function $H(\cdot)$ such that

$$\sum_{j=0}^{2} n_j^{*D}(t) = 0, \qquad \text{for all } t \geq H(\sum_{j=0}^{2} c_j n_j), \tag{5.41}$$

with n denoting the initial state. The proof of (5.41) will be given later on.

From (5.41) we obtain

$$\min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^\infty \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t \geq \min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^D \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t \tag{5.42}$$
$$= \int_0^D \sum_{j=0}^2 c_j n_j^{*D}(t) \mathrm{d}t = \int_0^\infty \sum_{j=0}^2 c_j n_j^{*D}(t) \mathrm{d}t \geq \min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^\infty \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t,$$

for all $D \geq H(\sum_{j=0}^2 c_j n_j)$. Hence, $(u^{*D}(t), n^{*D}(t))$ is an average-cost optimal solution. In particular, this implies the existence result. In addition, from (5.42) we obtain that for any average-cost optimal trajectory $n^*(t)$, it holds that

$$\min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^\infty \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t = \int_0^\infty \sum_{j=0}^2 c_j n_j^*(t) \mathrm{d}t \geq \int_0^D \sum_{j=0}^2 c_j n_j^*(t) \mathrm{d}t$$
$$\geq \min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^D \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t = \min_{n(t) \text{ s.t. } (5.3)-(5.6)} \int_0^\infty \sum_{j=0}^2 c_j n_j(t) \mathrm{d}t,$$

for all $D \geq H(\sum_{j=0}^2 c_j n_j)$. This proves the lemma under the condition that there indeed exists a function $H(\cdot)$ satisfying (5.41). The latter will be shown in the remainder of the proof. We use similar arguments as in [88, Proposition 6.1].

Denote by $\pi^{(0)}$ the policy that serves class 0 whenever possible. Let $n^{(0)}(t)$ be the trajectory that corresponds to policy $\pi^{(0)}$. Under the stability conditions we know that $n^{(0)}(t)$ hits zero after a finite time and then remains empty, see Lemma 5.2.1. Denote by $T^{(0)}(\tilde{n}, 0)$ the time it takes for policy $\pi^{(0)}$ to empty the system, when starting in state $\tilde{n}$. It can be written as follows

$$T^{(0)}(\tilde{n}, 0) = T_0^{(0)}(\tilde{n}, 0) + \max_{i=1,2} \left(\frac{\tilde{n}_i + \lambda_i T_0^{(0)}(\tilde{n}, 0)}{\mu_i(1-\rho_0) - \lambda_i} \right), \tag{5.43}$$

where $T_0^{(0)}(\tilde{n}, 0) = \frac{\tilde{n}_0}{\mu_0 - \lambda_0}$ is the time it takes until class 0 hits zero. It is clear that the depletion time scales as follows: $T^{(0)}(a \cdot \tilde{n}, 0) = a \cdot T^{(0)}(\tilde{n}, 0)$, $a \geq 0$.

Let $0 < \zeta < 1$ be fixed, and $x > 0$. We now have the following upper bound for all initial states n with $\sum_{j=0}^2 c_j n_j = x$:

$$\int_0^D \sum_{j=0}^2 c_j n_j^{*D}(t) \mathrm{d}t \leq \int_0^D \sum_{j=0}^2 c_j n_j^{(0)}(t) \mathrm{d}t$$
$$\leq \sup_{0 \leq t \leq D} \{ \sum_{j=0}^2 c_j n_j^{(0)}(t) \} \cdot T^{(0)}(n, 0) \leq x \cdot \zeta \cdot (1-\zeta) \cdot H(x). \tag{5.44}$$

The function $H(\cdot)$ is defined as

$$H(x) := \frac{\beta}{\zeta \cdot (1-\zeta)} \cdot \sup_{l : \sum_{j=0}^2 c_j l_j = x} \{ T^{(0)}(l, 0) \}, \ x \geq 0,$$

with the constant

$$\beta := 1 + \max(0, \frac{-c_0\mu_0 + \sum_{j=0}^{2} c_j\lambda_j}{\mu_0 - \lambda_0}),$$

so that for all initial states n with $\sum_{j=0}^{2} c_j n_j = x$ we have $\sup_{0\leq t\leq D}\{\sum_{j=0}^{2} c_j n_j^{(0)}(t)\}$ $= \max\left(x,\ x + T_0^{(0)}(n,0)\cdot(-c_0\mu_0 + \sum_{j=0}^{2} c_j\lambda_j)\right) \leq \beta \cdot x$. (In the first equation we used that once class 0 is empty, the total number of users will decrease.)

From (5.43) it easily follows that $T^{(0)}(l,0)$ is continuous in l. Hence $\sup_{l:\sum_{j=0}^{2} c_j l_j = x} T^{(0)}(l,0) < \infty$ and in particular $H(x) < \infty$ for all $x > 0$. Assume $D \geq H(x)$ (in particular, $D \geq (1-\zeta)\cdot H(x)$). By (5.44) we obtain that

$$\tau(x) := \arg\min_{t\geq 0}\{\sum_{j=0}^{2} c_j n_j^{*D}(t) \leq x\cdot\zeta\} \leq (1-\zeta)\cdot H(x), \tag{5.45}$$

for all initial states n with $\sum_{j=0}^{2} c_j n_j = x$.

From continuity of $n^{*D}(t)$ it follows that $\sum_{j=0}^{2} c_j n_j^{*D}(\tau(x)) = x\cdot\zeta$. Hence, if $n^{*D}(0) = n$, then $n^{*D}\left(\sum_{m=1}^{\infty}\tau(\zeta^{m-1}\sum_{j=0}^{2} c_j n_j)\right) = 0$. Together with (5.45) and $H(a\cdot x) = a\cdot H(x)$, $a \geq 0$, we obtain $\sum_{m=1}^{\infty}\tau(\zeta^{m-1}\sum_{j=0}^{2} c_j n_j) \leq \sum_{m=1}^{\infty}\zeta^{m-1}(1-\zeta)\cdot H(\sum_{j=0}^{2} c_j n_j) = H(\sum_{j=0}^{2} c_j n_j) < \infty$. Hence, relation (5.41) holds. □

5.E Proof of Lemma 5.2.9

Let $\overline{N}_j^{\tilde{\pi}}(t)$, $j = 0,1,2$, $\overline{T}_l^{\tilde{\pi}}(t)$, $l = 0,1,2,I_1,I_2$, be a fluid limit of policy $\tilde{\pi}$. So the functions $\overline{N}_j^{\tilde{\pi}}(t)$, satisfy (5.19), and the functions $\overline{T}_l^{\tilde{\pi}}(\cdot)$, are absolutely continuous (follows from Lipschitz continuity), and hence are differentiable almost everywhere. Fix a sample path ω such that there is a subsequence r_k with $\lim_{r_k\to\infty}\overline{N}_j^{\tilde{\pi},r_k}(t) = \overline{N}_j^{\tilde{\pi}}(t)$, $j = 0,1,2$, u.o.c., and $\lim_{r_k\to\infty}\overline{T}_l^{\tilde{\pi},r_k}(t) = \overline{T}_l^{\tilde{\pi}}(t)$, $l = 0,1,2,I_1,I_2$, u.o.c.. Further, let $t > 0$ be a regular point of $\overline{T}_l^{\tilde{\pi}}(t)$ for all l.

First assume $\overline{N}_1^{\tilde{\pi}}(t) > 0$ and $\overline{N}_2^{\tilde{\pi}}(t) > 0$. Then there is an $\epsilon > 0$ such that $\overline{N}_1^{\tilde{\pi}}(s) > 0$ and $\overline{N}_2^{\tilde{\pi}}(s) > 0$ for all $s \in [t-\epsilon, t+\epsilon]$. By the uniform convergence of $\overline{N}_j^{\tilde{\pi},r_k}(\cdot)$ to $\overline{N}_j^{\tilde{\pi}}(\cdot)$, $j = 0,1,2$, on $[t-\epsilon,t+\epsilon]$, we have $N_1^{\tilde{\pi},r_k}(r_k s) > 0$ and $N_2^{\tilde{\pi},r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon,t+\epsilon]$. Hence, under policy $\tilde{\pi}$, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ classes 1 and 2 are served, so that $\frac{\mathrm{d}\overline{T}_i^{\tilde{\pi},r_k}(t+\epsilon)}{\mathrm{d}t} - \frac{\mathrm{d}\overline{T}_i^{\tilde{\pi},r_k}(t-\epsilon)}{\mathrm{d}t} = 2\epsilon$, $i = 1,2$. Letting $r_k \to \infty$ and $\epsilon \downarrow 0$, we obtain $\frac{\mathrm{d}\overline{T}_i^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1$, $i = 1,2$.

Now assume $\overline{N}_1^{\tilde{\pi}}(t) < k_1\overline{N}_0^{\tilde{\pi}}(t)$, $\overline{N}_1^{\tilde{\pi}}(t) > 0$ and $\overline{N}_2^{\tilde{\pi}}(t) = 0$. Then there is an $\epsilon > 0$ such that $N_1^{\tilde{\pi},r_k}(r_k s) < k_1 N_0^{\tilde{\pi},r_k}(r_k s)$ and $N_1^{\tilde{\pi},r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon,t+\epsilon]$. Hence, under policy $\tilde{\pi}$, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ class 2 is served whenever present, and otherwise class 0 is served, so that $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1$.

Note that if $\overline{N}_2^{\tilde{\pi}}(t+\delta) > 0$, for all $0 < \delta < \Delta$, then $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t+\delta)}{\mathrm{d}t} = 1$. Since $\rho_2 < 1$, from $\frac{\mathrm{d}\overline{N}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = \lambda_2 - \mu_2 \frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t}$ (follows from (5.19)) we obtain that class 2 will stay empty, and thus $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_2$ (and $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \rho_2$).

Assume $\overline{N}_1^{\tilde{\pi}}(t) > k_1 \overline{N}_0^{\tilde{\pi}}(t)$. Then there is an $\epsilon > 0$ such that $N_1^{\tilde{\pi},r_k}(r_k s) > k_1 N_0^{\tilde{\pi},r_k}(r_k s)$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. Hence, under policy $\tilde{\pi}$, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ class 0 receives no service, so that $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 0$.

Assume $0 < \overline{N}_1^{\tilde{\pi}}(t) = k_1 \overline{N}_0^{\tilde{\pi}}(t)$ and $\overline{N}_2^{\tilde{\pi}}(t) = 0$. Then there is an $\epsilon > 0$ such that $N_1^{\tilde{\pi},r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. Hence, class 2 is served whenever present. Similar as before, this implies that $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_2$, so that $\frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(t)}{\mathrm{d}t} \geq \rho_2$. In addition, the full capacity in node 1 is used, hence $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1$. Together with (5.19), this implies

$$k_1 \frac{\mathrm{d}\overline{N}_0^{\tilde{\pi}}(s)}{\mathrm{d}s} - \frac{\mathrm{d}\overline{N}_1^{\tilde{\pi}}(s)}{\mathrm{d}s} = k_1 \left(\lambda_0 - \mu_0 \frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(s)}{\mathrm{d}s} \right) - \lambda_1 + \mu_1 \frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(s)}{\mathrm{d}s} \tag{5.46}$$

$$= \frac{\mu_0}{\mu_1} k_1 \left(\rho_0 - 1 + \frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(s)}{\mathrm{d}s} \right) - \rho_1 + \frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(s)}{\mathrm{d}s} \geq \frac{\mu_0}{\mu_1} k_1 (\rho_0 + \rho_2 - 1) - \rho_1 + \rho_2 > 0,$$

whenever $s \in [t-\epsilon, t+\epsilon]$ is a regular point. In the last step we used that $\rho_0 + \rho_2 - 1 < 0$ and $k_1 < \frac{\mu_1}{\mu_0} \frac{\rho_2 - \rho_1}{1 - \rho_0 - \rho_2}$. Equation (5.46) implies that if at a certain time $\overline{N}^{\tilde{\pi}}$ lies above the switching curve, then it moves towards the switching curve and if $\overline{N}^{\tilde{\pi}}$ lies on or below the switching curve, it will move away from (and below) the switching curve. Since at time t we are in a state on the switching curve, we have $\overline{N}_1^{\tilde{\pi}}(s) > k_1 \overline{N}_0^{\tilde{\pi}}(s)$ for $s \in [t-\epsilon, t)$ and $\overline{N}_1^{\tilde{\pi}}(s) < k_1 \overline{N}_0^{\tilde{\pi}}(s)$ for $s \in (t, t+\epsilon]$. Note that $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t-)}{\mathrm{d}t} = 0$, while $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t+)}{\mathrm{d}t} = 1 - \rho_2$, so that the point t itself is not a regular point.

Assume $\overline{N}_0^{\tilde{\pi}}(t) = \overline{N}_1^{\tilde{\pi}}(t) = 0$ and $\overline{N}_2^{\tilde{\pi}}(t) > 0$. Then there is an $\epsilon > 0$ such that $N_2^{\tilde{\pi},r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. If $k_2 = \infty$, under policy $\tilde{\pi}$, class 1 is served whenever present (since class 2 is continuously present), and otherwise class 0 is served. Hence, $\frac{\mathrm{d}\overline{T}_j^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_j$, $j = 0, 1$. If $k_2 = 0$, then class 2 is given full priority, hence $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1$.

Assume $\overline{N}_0^{\tilde{\pi}}(t) > 0$ and $\overline{N}_1^{\tilde{\pi}}(t) = \overline{N}_2^{\tilde{\pi}}(t) = 0$. There is an $\epsilon > 0$ such that $0 < N_0^{\tilde{\pi},r_k}(r_k s)$ and $N_1^{\tilde{\pi},r_k}(r_k s) < k_1 N_0^{\tilde{\pi},r_k}(r_k s)$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. If $k_2 = 0$, then in this interval class 2 is served whenever present, and otherwise class 0 is served. This implies $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_2$ and $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1 - \rho_2$. If $k_2 = \infty$, then classes 1 and 2 are served whenever both are present, and otherwise class 0 is served in the interval $[t-\epsilon, t+\epsilon]$. Hence, $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(s)}{\mathrm{d}s} + \frac{\mathrm{d}\overline{T}_i^{\tilde{\pi}}(s)}{\mathrm{d}s} = 1$, $i = 1, 2$, and

$$\frac{\mathrm{d}(\overline{N}_1^{\tilde{\pi}}(s)/\mu_1 - \overline{N}_2^{\tilde{\pi}}(s)/\mu_2)}{\mathrm{d}s} = \rho_1 - \frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(s)}{\mathrm{d}s} - \rho_2 + \frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(s)}{\mathrm{d}s} = \rho_1 - \rho_2 < 0, \tag{5.47}$$

whenever $s \in [t-\epsilon, t+\epsilon]$ is a regular point. Together with $\overline{N}_0^{\tilde{\pi}}(t) > 0$ and $\overline{N}_1^{\tilde{\pi}}(t) = \overline{N}_2^{\tilde{\pi}}(t) = 0$, this yields $\overline{N}_0^{\tilde{\pi}}(s) > 0$ and $\overline{N}_1^{\tilde{\pi}}(s)\frac{\mu_2}{\mu_1} > \overline{N}_2^{\tilde{\pi}}(s)$ for $s \in [t-\epsilon, t)$. Hence, $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t-)}{\mathrm{d}t}$ is either $1-\rho_2$ (when $\overline{N}_2^{\tilde{\pi}}(t-) = 0$) or 0 (when $\overline{N}_2^{\tilde{\pi}}(t-) > 0$). In addition, from (5.47) we obtain $\overline{N}_0^{\tilde{\pi}}(s) > 0$ and $\overline{N}_2^{\tilde{\pi}}(s) > \overline{N}_1^{\tilde{\pi}}(s)\frac{\mu_2}{\mu_1} \geq 0$ for $s \in (t, t+\epsilon]$. Hence, class 1 is served whenever present, which implies $\overline{N}_1^{\tilde{\pi}}(s) = 0$ for $s \in (t, t+\epsilon]$, so that $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t+)}{\mathrm{d}t} = 1-\rho_1$. Thus, $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t-)}{\mathrm{d}t} \neq \frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t+)}{\mathrm{d}t}$, and hence t is not a regular point.

Assume $\overline{N}_0^{\tilde{\pi}}(t) > 0$, $\overline{N}_1^{\tilde{\pi}}(t) = 0$ and $\overline{N}_2^{\tilde{\pi}}(t) > 0$. Then there is an $\epsilon > 0$ such that $N_j^{\tilde{\pi}, r_k}(r_k s) > 0$, $j = 0, 2$, for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. If $k_2 = \infty$, then class 1 is served whenever present (since class 2 is continuously present), and otherwise class 0 is served in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$. It follows that $\frac{\mathrm{d}\overline{T}_1^{\tilde{\pi}}(t)}{\mathrm{d}t} = \rho_1$, and $\frac{\mathrm{d}\overline{T}_0^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1-\rho_1$. If $k_2 = 0$, then class 2 is continuously served in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$. Hence, $\frac{\mathrm{d}\overline{T}_2^{\tilde{\pi}}(t)}{\mathrm{d}t} = 1$.

From (5.19), together with either (5.20)–(5.23), or (5.24) and (5.25), it follows that $\overline{N}_j^{\tilde{\pi}}(t)$ is uniquely determined. □

5.F Proof of relations (5.28)–(5.30)

Let $\overline{N}_j^{\pi_2}(t)$, $j = 0, 1, 2$, $\overline{T}_l^{\pi_2}(t)$, $l = 0, 1, 2, I_1, I_2$, be a fluid limit of policy π_2. So the functions $\overline{N}_j^{\pi_2}(t)$, satisfy (5.19), and the functions $\overline{T}_l^{\pi_2}(\cdot)$, are absolutely continuous (follows from Lipschitz continuity), and hence are differentiable almost everywhere. Fix a sample path ω such that there is a subsequence r_k with $\overline{N}_j^{\pi_2}(t) = \lim_{r_k \to \infty} \overline{N}_j^{\pi_2, r_k}(t)$, u.o.c., $j = 0, 1, 2$, and $\overline{T}_l^{\pi_2}(t) = \lim_{r_k \to \infty} \overline{T}_l^{\pi_2, r_k}(t)$, u.o.c., $l = 0, 1, 2, I_1, I_2$. Further, let $t > 0$ be a regular point of $\overline{T}_l^{\pi_2}(t)$ for all l.

First assume $\overline{N}_2^{\pi_2}(t) > 0$. Then there is an $\epsilon > 0$ such that $\overline{N}_2^{\pi_2}(s) > 0$ for all $s \in [t-\epsilon, t+\epsilon]$. By the uniform convergence of $\overline{N}_2^{\pi_2, r_k}(\cdot)$ to $\overline{N}_2^{\pi_2}(\cdot)$ on $[t-\epsilon, t+\epsilon]$, we have $N_2^{\pi_2, r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. Hence, under policy π_2, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ class 2 is served, so that $\frac{\mathrm{d}\overline{T}_2^{\pi_2, r_k}(t+\epsilon)}{\mathrm{d}t} - \frac{\mathrm{d}\overline{T}_2^{\pi_2, r_k}(t-\epsilon)}{\mathrm{d}t} = 2\epsilon$. Letting $r_k \to \infty$ and $\epsilon \downarrow 0$, we obtain $\frac{\mathrm{d}\overline{T}_2^{\pi_2}(t)}{\mathrm{d}t} = 1$.

Now assume $\overline{N}_0^{\pi_2}(t) > 0$ and $\overline{N}_2^{\pi_2}(t) = 0$. Hence, under policy π_2, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ class 2 is served whenever present, and otherwise class 0 is served, so that $\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_2^{\pi_2}(t)}{\mathrm{d}t} = 1$. Note that if $\overline{N}_2^{\pi_2}(t+\delta) > 0$, for all $0 < \delta < \Delta$, then $\frac{\mathrm{d}\overline{T}_2^{\pi_2}(t+\delta)}{\mathrm{d}t} = 1$. Since $\rho_2 < 1$, from $\frac{\mathrm{d}\overline{N}_2^{\pi_2}(t)}{\mathrm{d}t} = \lambda_2 - \mu_2 \frac{\mathrm{d}\overline{T}_2^{\pi_2}(t)}{\mathrm{d}t}$ (follows from (5.19)) we obtain that class 2 will stay empty, and thus $\frac{\mathrm{d}\overline{T}_2^{\pi_2}(t)}{\mathrm{d}t} = \rho_2$ (and $\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} = 1-\rho_2$).

Finally assume $\overline{N}_0^{\pi_2}(t) = \overline{N}_2^{\pi_2}(t) = 0$ and $\overline{N}_1^{\pi_2}(t) > 0$. Note that if $\overline{N}_0^{\pi_2}(t+\delta) + \overline{N}_2^{\pi_2}(t+\delta) > 0$, for all $0 < \delta < \Delta$, then $\frac{\mathrm{d}\overline{T}_0^{\pi_2}(t+\delta)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_2^{\pi_2}(t+\delta)}{\mathrm{d}t} = 1$. Since $\rho_0 + \rho_2 < 1$, from $\frac{\mathrm{d}\overline{N}_0^{\pi_2}(t)/\mu_0}{\mathrm{d}t} + \frac{\mathrm{d}\overline{N}_2^{\pi_2}(t)/\mu_2}{\mathrm{d}t} = \rho_0 + \rho_2 - \frac{\mathrm{d}\overline{T}_0^{\pi_2}(t)}{\mathrm{d}t} - \frac{\mathrm{d}\overline{T}_2^{\pi_2}(t)}{\mathrm{d}t}$ (follows from (5.19)) we obtain that classes 0 and 2 will stay empty, and thus $\frac{\mathrm{d}\overline{T}_j^{\pi_2}(t)}{\mathrm{d}t} = \rho_j$, $j = 0, 2$. □

Chapter 6
Heavy-traffic analysis of size-based bandwidth-sharing policies

In Chapters 4 and 5 we characterized optimal scheduling policies for a linear bandwidth-sharing network among all non-anticipating policies in the case of exponentially distributed service requirements. We found that it is only possible to explicitly identify optimal policies in a few limited cases. In the present chapter we consider the linear network with generally distributed service requirements, for which optimal policies may be significantly more complicated or even totally intractable. Rather than aiming for strictly optimal policies, we investigate a class of relatively simple size-based priority policies in heavy-traffic conditions.

Nearly all size-based scheduling results concern single-server settings that do not exhibit the potential capacity loss that may occur in scenarios with concurrent resource possession as encountered in bandwidth-sharing networks. In Chapter 3 it was shown that size-based scheduling policies such as SRPT and LAS may in fact unnecessarily fail to achieve stability in network settings (even at arbitrarily low loads) when size-based scheduling is applied across the various classes. However, size-based scheduling within classes may still produce substantial performance benefits, provided the rate allocation across classes is carefully chosen to avoid the above instability phenomena.

In this chapter we focus on certain work-conserving policies for which stability of the system is guaranteed for general service requirements. Within a class we then separate the large service requirements from the small ones, which as noted above, might considerably improve the performance. More precisely, we introduce a fairly simple size-based intra-class policy: Within a class, all users with a service requirement above a certain threshold are given low priority and, in particular, cannot be served when there are users of size smaller than the threshold present in this class.

For generally distributed service requirements, we examine the performance of such size-based priority policies in heavy-traffic conditions, i.e., each of the links is near-critically loaded. We show these policies to be asymptotically optimal in heavy traffic for service requirement distributions with bounded support. Although the link utilization may not always be that high, a heavy-traffic regime is relevant to consider, because at low load the performance will tend to be satisfactory no matter what. In addition, even when the typical link utilization is relatively low, the load might fluctuate over time and exhibit significant surges, causing severe congestion periods or even temporary overload conditions.

In addition, we compare the performance of the size-based priority policies with that of Proportional Fair (PF) as the prototypical α-fair bandwidth-sharing policy, and demonstrate that the mean delay may be reduced by an arbitrarily large factor when the load is sufficiently high. We recall that in Chapters 4 and 5 an opposite observation was made. For exponentially distributed service requirements, α-fair policies are close to optimal provided α is not too small. Notice however, that in Chapters 4 and 5 a moderately-loaded system was considered, while in this chapter we consider heavy traffic. In addition, the optimal policy was found within the set of non-anticipating policies, while in this chapter we also allow anticipating policies.

The remainder of the chapter is organized as follows. In Section 6.1 we provide a detailed model description and gather some useful preliminaries. In Section 6.2 we develop a heavy-traffic analysis of a single-node system in order to illuminate the key observations and mathematical constructs in the simplest possible context. In Section 6.3 we then turn the attention to linear bandwidth-sharing networks. Section 6.3.1 deals with the case where all the flows on the long route are granted priority over the large flows on the short routes. In Section 6.3.2 we address the case where the flows on the short routes, when simultaneously present, are favored over the large flows on the long route. In Section 6.4 we present the numerical experiments that we conducted to validate the analytical findings and in particular compare the performance of the above strategies with that of PF. These numerical experiments indicate that even at fairly moderate load values the performance gains can be significant. Section 6.5 concludes the chapter with concluding remarks.

6.1 Model and preliminaries

We consider a linear network with L nodes and $L+1$ classes, where class i requires service at node i only, $i = 1, \ldots, L$, while class 0 requires service at all L nodes simultaneously, see Figure 1.2. For convenience, we assume each of the nodes to have a unit service rate. Class-j users arrive according to independent Poisson processes of rate λ_j, and have generally distributed service requirements B_j with distribution function $B_j(x) = \mathbb{P}(B_j < x)$, $j = 0, 1, \ldots, L$. We assume $\mathbb{E}(B_j^2) < \infty$. Define

$$M_j := \sup\{x : B_j(x) < 1\}$$

as the maximum possible value of B_j, with $M_j = \infty$ in case B_j has infinite support. Denote by $p_j := \lambda_j/\lambda$ the fraction of class-j users, with $\lambda = \sum_{j=0}^{L} \lambda_j$ the total arrival rate. Let the traffic load of class j be $\rho_j := \lambda_j \mathbb{E}(B_j)$. Throughout, we assume that the maximum stability conditions are satisfied, i.e., $\rho_0 + \rho_i < 1$ for all $i = 1, \ldots, L$.

For a policy $\pi \in \Pi$, denote by $N_j^{\pi}(t)$ the number of class-j users at time t and by $W_j^\pi(t)$ their total residual amount of work. Define $N^\pi(t) := \sum_{j=0}^L N_j^\pi(t)$ as the total number of users in the system at time t. Denote by $N_{j,<x_j}^\pi(t)$ and $N_{j,\geq x_j}^\pi(t)$ the number of class-j users with original service requirement smaller than x_j and larger than or equal to x_j, respectively. Similarly, we define $W_{j,<x_j}^\pi(t)$ and $W_{j,\geq x_j}^\pi(t)$ as the amount of work consisting of class-j users with original service requirement smaller than x_j and larger than or equal to x_j, respectively. Denote by

$$\rho_j(x_j) := \lambda_j \mathbb{P}(B_j < x_j)\mathbb{E}(B_j | B_j < x_j) = \lambda_j \int_0^{x_j^-} y \mathrm{d}B_j(y),$$

the load composed of class-j users with original service requirement smaller than x_j, $j = 0, \ldots, L$. We further define N_j^π, W_j^π, N^π, $N_{j,<x_j}^\pi$, $N_{j,\geq x_j}^\pi$, $W_{j,<x_j}^\pi$ and $W_{j,\geq x_j}^\pi$ as random variables with the corresponding steady-state distributions (when they exist).

In Chapter 4 we introduced the classes of policies $\bar{\Pi}$, Π^*, and Π^{**}. In this chapter we define an additional class of policies $\hat{\Pi} \subseteq \Pi$ containing all work-conserving policies: $\hat{\pi} \in \hat{\Pi}$ if $\hat{\pi}$ utilizes the full service rate at any node i, $i = 1, \ldots, L$, that is backlogged.

Observe that under any policy $\hat{\pi} \in \hat{\Pi}$ the total workload in any node i behaves as that of a single work-conserving server offered traffic from classes 0 and i. It immediately follows that any policy $\hat{\pi} \in \hat{\Pi}$ ensures stability under the maximum stability conditions. In addition, the Pollaczek-Khintchine formula gives

$$\mathbb{E}(W_0^{\hat{\pi}}) + \mathbb{E}(W_i^{\hat{\pi}}) = \frac{\lambda_0 \mathbb{E}({B_0}^2) + \lambda_i \mathbb{E}({B_i}^2)}{2(1 - \rho_0 - \rho_i)}, \quad i = 1, \ldots, L, \tag{6.1}$$

for any policy $\hat{\pi} \in \hat{\Pi}$. It further follows that any policy $\hat{\pi} \in \hat{\Pi}$ minimizes the total workload in any node i at every point in time. More specifically, if $W_0^{\hat{\pi}}(0) + W_i^{\hat{\pi}}(0) \leq_{st} W_0^\pi(0) + W_i^\pi(0)$ for some arbitrary policy $\pi \in \Pi$, then

$$\{W_0^{\hat{\pi}}(t) + W_i^{\hat{\pi}}(t)\}_{t \geq 0} \leq_{st} \{W_0^\pi(t) + W_i^\pi(t)\}_{t \geq 0}, \quad i = 1, \ldots, L. \tag{6.2}$$

Since $\Pi^*, \Pi^{**} \subseteq \hat{\Pi}$, all policies in these two classes satisfy (6.2) for all $i = 1, \ldots, L$.

Recall that policies in $\Pi^* \cap \bar{\Pi}$ and $\Pi^{**} \cap \bar{\Pi}$ exhibit optimality properties in the case of exponentially and hyperexponentially distributed service requirements provided that the service requirements of class 0 are not too large, see Section 4.3.1. These results provide a strong notion of optimality, but involve correspondingly stringent assumptions on the service requirements. In the next sections, we seek policies, possibly anticipating, that are optimal under significantly milder conditions, although only in a heavy traffic sense.

6.2 Single-server system in heavy traffic

Although the issue of concurrent resource possession only arises in network scenarios, we first present a heavy-traffic analysis of a single-server system in order to illustrate the key concepts and insights in the simplest possible context. In the next section we will return to the linear network.

Consider a single-server system, where users arrive according to a Poisson process of rate λ and have service requirements B with $B(x) := \mathbb{P}(B \leq x)$ and $\mathbb{E}(B^2) < \infty$. Let $M := \sup\{x : B(x) < 1\}$. Denote the load by $\rho := \lambda\mathbb{E}(B) < 1$ and define $\rho(x) := \lambda \int_0^{x^-} y \mathrm{d}B(y)$. For every policy π, the mean workload in the system obeys the lower bound $\mathbb{E}(W^\pi) \geq \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)}$, with equality when policy π is work-conserving. We analyze a heavy-traffic regime where the system is critically loaded, i.e., $\lambda \uparrow \lambda^* := \frac{1}{\mathbb{E}(B)}$ (so $\rho \uparrow 1$, since ρ implicitly depends on λ).

In order to improve the overall user performance, we exploit the variability in service demands, and give precedence to small users over large ones. Specifically, we introduce a class of anticipating policies $\Pi_x \subseteq \hat{\Pi}\backslash\bar{\Pi}$ which use a simple threshold x to determine whether a user is small or large, and give preemptive priority to users with (original) service requirement smaller than x. Among users with an (original) service requirement larger than x, service is non-preemptive, i.e., the service of a user of size larger than x cannot be preempted by the service of another user of size larger than x. Motivated by the classical heavy-traffic scaling for non-preemptive policies, we consider throughout this chapter the workload and number of users scaled by $1-\rho$.

For policies in the class Π_x, the small users do not notice the presence of the large users, and experience similar performance as in a system without any large users. Now observe that the load in the latter system is $\rho(x)$, and remains bounded away from 1, even when the load ρ approaches 1 as $\lambda \uparrow \lambda^*$ (assuming that there are in fact large users, i.e., $\mathbb{P}(B < x) < 1$). Hence, the small users are "shielded" from the heavy-traffic conditions, as is formalized in the next proposition, which shows that the number of small users remains bounded as the load approaches the capacity.

Proposition 6.2.1. *For a policy $\pi_x \in \Pi_x$ with $\mathbb{P}(B < x) < 1$, it holds that $\mathbb{E}(N_{<x}^{\pi_x}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: When $\mathbb{P}(B < x) = 0$, the statement is trivial. In the remainder of the proof we therefore assume $\mathbb{P}(B < x) > 0$. Consider a policy $\pi_x \in \Pi_x$. Users of size smaller than x do not notice the presence of users of size larger than x. Therefore, $\mathbb{E}(W_{<x}^{\pi_x}) = \frac{\lambda\mathbb{P}(B<x)\mathbb{E}(B^2|B<x)}{2(1-\rho(x))} \leq \frac{\lambda x^2}{2(1-\rho(x))}$. The condition $\mathbb{P}(B < x) < 1$ guarantees that $\lim_{\lambda\uparrow\lambda^*} \rho(x) < 1$. Hence, $\mathbb{E}(W_{<x}^{\pi_x}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$. Now suppose that service is non-preemptive among users of size smaller than x as well (this assumption is not essential, see Remark 6.2.2 below). Then $(\mathbb{E}(N_{<x}^{\pi_x}) - 1)\mathbb{E}(B|B < x) \leq \mathbb{E}(W_{<x}^{\pi_x})$, which implies $\mathbb{E}(N_{<x}^{\pi_x}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$. □

Remark 6.2.2. The assumption in the proof of Proposition 6.2.1 that service is non-preemptive among users of size smaller than x, is not crucial. Instead, we could use the fact that the preemptive Longest Remaining Processing Time (LRPT) policy maximizes sample-path wise the number of users among all work-conserving policies. This follows from the fact that under LRPT all users leave together at the end of the busy period. Since users of size smaller than x receive preemptive priority under policy π_x, their number is smaller than the number of users under LRPT in a system with only users of size smaller than x. The latter has mean $\mathbb{E}(N_{<x}^{LRPT}) = \lambda\mathbb{P}(B < x)(\frac{\mathbb{E}(B|B<x)}{1-\rho(x)} + \frac{\lambda\mathbb{P}(B<x)\mathbb{E}(B^2|B<x)}{2(1-\rho(x))^2})$, see [60]. The result now follows by noting that $\lim_{\lambda\uparrow\lambda^*} \rho(x) < 1$.

Proposition 6.2.1 implies that the scaled mean number of small users tends to zero in heavy traffic. The number of large users can be bounded in terms of the total workload in the system, which results in an upper bound for the scaled total mean number of users in the system, as provided in the next proposition.

Proposition 6.2.3. *For a policy $\pi_x \in \Pi_x$ with $\mathbb{P}(B < x) < 1$, it holds that* $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_x}) \leq \frac{\lambda^*\mathbb{E}(B^2)}{2x}$.

Proof: Consider a policy $\pi_x \in \Pi_x$. Note that $\mathbb{E}(W_{\geq x}^{\pi_x}) \geq x(\mathbb{E}(N_{\geq x}^{\pi_x}) - 1)$, because service is non-preemptive among users of size larger than x. Proposition 6.2.1 implies in particular that the scaled mean number of users smaller than x converges to zero. Together, this yields $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_x}) = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_{\geq x}^{\pi_x}) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W_{\geq x}^{\pi_x})}{x} \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W^{\pi_x})}{x} = \frac{\lambda^*\mathbb{E}(B^2)}{2x}$. □

6.2.1 Comparison with processor sharing

The next proposition provides a comparison of the policies in the class $\cup_x\Pi_x$ with the processor-sharing (PS) policy, which corresponds to the PF policy in a single-server system.

Proposition 6.2.4. *Let $\pi_x \in \Pi_x$. When B has infinite support, we have*

$$\lim_{x\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_x})}{\mathbb{E}(N^{PS})} = 0.$$

When B has finite support, we have

$$\lim_{x\uparrow M}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_x})}{\mathbb{E}(N^{PS})} \leq \frac{\lambda^*\mathbb{E}(B^2)}{2M}.$$

Proof: It is well-known that $\mathbb{E}(N^{PS}) = \rho/(1-\rho)$, so $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{PS}) = 1$. Invoking Proposition 6.2.3, we obtain $\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_x})}{\mathbb{E}(N^{PS})} \leq \frac{\lambda^*\mathbb{E}(B^2)}{2x}$ for any x such that $\rho(x) < \rho$, which proves both assertions. □

In case B has infinite support, it may be deduced that a policy from the class $\cup_x\Pi_x$ can outperform PS by an arbitrarily large factor. In case B has finite support, the ratio $\lambda^*\mathbb{E}(B^2)/M$ can be arbitrarily small for a wide range of service requirement distributions since $\mathbb{E}(B^2) \leq \frac{1}{\lambda}\Big[k\rho(k) + M(1-\rho(k))\Big]$. These two findings may be intuitively explained as follows. Under the PS policy, the total workload is distributed across users of various sizes, in proportion to their share in the total load, and hence the total number of users grows linearly with the workload as $\lambda \uparrow \lambda^*$. In contrast, under policies in the class $\cup_x \Pi_x$ the overwhelming fraction of the workload is contributed by users of size larger than x as $\lambda \uparrow \lambda^*$. Thus, as the value of x increases, the entire workload is concentrated in fewer and fewer users compared to PS.

Remark 6.2.5. The assumption that service is non-preemptive among users of size larger than x under policies in the class Π_x, is not essential for Proposition 6.2.4. For example, for the first statement of Proposition 6.2.4 to hold, it would be sufficient to have that $\lim_{x\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(W)}{\mathbb{E}(N^{\pi_x}_{\geq x})} = \infty$, with W the workload in a work-conserving queue. It then easily follows that

$$\lim_{x\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi_x}_{\geq x})}{\mathbb{E}(N^{PS})} = \lim_{x\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi_x}_{\geq x})}{\mathbb{E}(W)} \frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)} = 0,$$

where we used that $\mathbb{E}(W) = \frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)}\mathbb{E}(N^{PS})$. Together with Proposition 6.2.1 this yields Proposition 6.2.4 in case B has infinite support.

Remark 6.2.6. Note that policies in Π_x rely on knowledge of the service requirements, which is not always easy to obtain. Instead, we could consider policies that give preemptive priority to users with *attained* service less than x. Let $\tilde{\pi}_x$ be such a non-anticipating policy. In addition, we assume that under policy $\tilde{\pi}_x$ service of a user with attained service larger than x cannot be preempted by the service of another user with attained service larger than x. Let B have infinite support. For $x < \infty$, denote by

$$\tilde{\rho}(x) = \lambda \int_0^{x^-} y \mathrm{d}B(y) + \lambda x \mathbb{P}(B \geq x) < \rho,$$

the load due to users truncated at size x (users larger than or equal to x contribute an amount x, rather than zero as in $\rho(x)$). Let $\tilde{N}^{\tilde{\pi}_x}_{<x}$ and $\tilde{N}^{\tilde{\pi}_x}_{\geq x}$ denote the number of users with attained service less than x and larger than or equal to x, respectively. We define $\tilde{W}^{\tilde{\pi}_x}_{<x}$ as the amount of work in the system consisting of users with attained service smaller than x, with their service requirement truncated at size x. Furthermore, let $\tilde{W}^{\tilde{\pi}_x}_{\geq x} = W^{\tilde{\pi}_x} - \tilde{W}^{\tilde{\pi}_x}_{<x}$.

Since users with attained service less than x do not notice the presence of users that have attained more than x, we can upper bound the former by considering a system where users have service requirement $\min(B, x)$ and where we apply the LRPT policy. This gives, $\mathbb{E}(\tilde{N}^{\tilde{\pi}_x}_{<x}) \leq \lambda(\frac{\mathbb{E}(\min(B,x))}{1-\tilde{\rho}(x)} + \frac{\lambda\mathbb{E}((\min(B,x))^2)}{(1-\tilde{\rho}(x))^2})$. Using the fact that $\lim_{\lambda\uparrow\lambda^*} \tilde{\rho}(x) < 1$ for $x < \infty$, we obtain $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(\tilde{N}^{\tilde{\pi}_x}_{<x}) = 0$. Furthermore,

$$\begin{aligned} \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(\tilde{N}^{\tilde{\pi}_x}_{\geq x}) &= \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(\tilde{W}^{\tilde{\pi}_x}_{\geq x})}{\mathbb{E}(B|B>x)-x} \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W)}{\mathbb{E}(B|B>x)-x} \\ &= \frac{\mathbb{P}(B>x)}{\int_x^\infty \mathbb{P}(B>y)\mathrm{d}y} \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W), \end{aligned}$$

where we used in the last step that $\mathbb{E}(B|B > x) = \frac{\int_x^\infty y\mathrm{d}B(y)}{\mathbb{P}(B>x)} = x + \frac{\int_x^\infty \mathbb{P}(B>y)\mathrm{d}y}{\mathbb{P}(B>x)}$ (follows from integration by parts). For service requirement distributions with

$$\lim_{x\to\infty} \frac{\mathbb{P}(B > x)}{\int_x^\infty \mathbb{P}(B > y)\mathrm{d}y} = 0, \tag{6.3}$$

we then obtain $\lim_{x\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\tilde{\pi}_x})}{\mathbb{E}(N^{PS})} = \lim_{x\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(\tilde{N}^{\tilde{\pi}_x}_{>x})}{\mathbb{E}(N^{PS})} = 0$, i.e., the non-anticipating policy $\tilde{\pi}_x$ can outperform PS by an arbitrarily large factor.

An important class of service requirements that satisfy condition (6.3) are long-tailed distributions, i.e., $\lim_{y\to\infty} \frac{\mathbb{P}(B>y+z)}{\mathbb{P}(B>y)} = 1$ for any z, that, in addition, have a decreasing failure rate (DFR). (For example, the Pareto distribution belongs to this class.) Under the DFR assumption, the function $\frac{\mathbb{P}(B>x+z)}{\mathbb{P}(B>x)}$ is non-decreasing in x, see [101, Theorem 1.8.2]. From the monotone convergence theorem we then obtain

$$\begin{aligned}\lim_{x\to\infty} \frac{\int_x^\infty \mathbb{P}(B > y)\mathrm{d}y}{\mathbb{P}(B > x)} &= \lim_{x\to\infty} \int_0^\infty \frac{\mathbb{P}(B > x + z)}{\mathbb{P}(B > x)}\mathrm{d}z \\ &= \int_0^\infty \lim_{x\to\infty} \frac{\mathbb{P}(B > x + z)}{\mathbb{P}(B > x)}\mathrm{d}z = \infty,\end{aligned}$$

where the last step follows from the long-tailed assumption. Hence, (6.3) is indeed satisfied for long-tailed service requirements with a decreasing failure rate.

6.2.2 Optimality properties

The next proposition shows that for any policy π, there exists a policy in $\cup_x \Pi_x$ that performs at least as well as π in heavy traffic. In other words, the class of policies $\cup_x \Pi_x$ is asymptotically optimal in heavy traffic. This may be heuristically interpreted as follows. As mentioned above, under policies in the class $\cup_x \Pi_x$ the vast bulk of the workload is concentrated in users of size larger than x, while at the same time the total workload is minimal. In case B has finite support and the value of x is close to M, it is not possible to achieve a smaller number of users for the given workload. (When B has infinite support, it may be possible to reduce the number of users for a given workload yet further, by allowing preemptive service among large users.)

Proposition 6.2.7. *Let $\pi_x \in \Pi_x$. If B has finite support, then for any policy $\pi \in \Pi$,*

$$\lim_{x\uparrow M} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi_x})}{\mathbb{E}(N^\pi)} \leq 1.$$

Proof: Let W be the workload in a work-conserving queue. For any policy $\pi \in \Pi$,

$$\mathbb{E}(N^\pi) \geq \frac{\mathbb{E}(W^\pi)}{M} \geq \frac{\mathbb{E}(W)}{M} = \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)M}. \tag{6.4}$$

Applying Proposition 6.2.3 to $x < M$, we obtain

$$\lim_{\lambda\uparrow\lambda^*} (1-\rho)\mathbb{E}(N^{\pi_x}) \leq \frac{\lambda^*\mathbb{E}(B^2)}{2x}. \tag{6.5}$$

Comparing (6.4) and (6.5), and letting $x \uparrow M$ yields the assertion. □

6.3 Linear network in heavy traffic

In case of exponential service requirements, with relatively small class-0 users, policies in either class $\Pi^* \cap \bar{\Pi}$ or $\Pi^{**} \cap \bar{\Pi}$ are optimal among all non-anticipating policies. In this section we explore whether, in a heavy-traffic regime, these results extend to more general service requirement distributions, now also allowing anticipating policies.

As described in Section 6.1, the linear network consists of L nodes and $L+1$ classes of users. We impose that $p_1\mathbb{E}(B_1) = \ldots = p_L\mathbb{E}(B_L)$, so that $\rho_1 = \ldots = \rho_L$. We analyze a heavy-traffic regime where each node is critically loaded, i.e.,

$$\rho_0 + \rho_i =: \rho \uparrow 1, \quad \text{for all} \quad i = 1, \ldots, L.$$

This is equivalent to $\lambda \uparrow \lambda^* := (p_0\mathbb{E}(B_0) + p_i\mathbb{E}(B_i))^{-1}$. We consider the workload and number of users scaled by $1-\rho$.

Just like for the single-server system in Section 6.2, we focus on simple size-based priority policies, i.e., within a class small service requirements are prioritized over large service requirements. We study two classes of work-conserving inter-class policies. In Section 6.3.1 we analyze a class of policies where class 0 is favored, while in Section 6.3.2 policies are studied which simultaneously favor classes $1, \ldots, L$.

6.3.1 Favoring class 0

We first consider policies that serve either class-i users of size smaller than x_i simultaneously, $i = 1, \ldots, L$, whenever at least one such user of each class is present, or serve class-0 users. If that is not possible, then classes $i = 1, \ldots, L$ are served, with class-i users with service requirement smaller than x_i receiving priority. Other than that, the priority structure within each of the classes is not essential for the analysis. Service is non-preemptive among class-i users of original size larger than x_i, i.e., the service of a class-i user of size larger than x_i cannot be preempted by the service of another class-i user of size larger than x_i. We denote this class of anticipating policies by $\Pi^*_{\boldsymbol{x}}$, where $\boldsymbol{x} = (x_1, \ldots, x_L)$. We adopt the notation $\boldsymbol{x} \uparrow \boldsymbol{M}$ and $\boldsymbol{x} \to \infty$ to indicate that $x_i \uparrow M_i$ for all $i = 1, \ldots, L$ and $x_i \to \infty$ for all $i = 1, \ldots, L$, respectively (order is irrelevant).

Under policies in the class $\Pi^*_{\boldsymbol{x}}$, class-0 users and small class-i users, $i = 1, \ldots, L$, do not notice the presence of the large users in the classes $1, \ldots, L$, and experience similar performance as in a system without any large class-i users, $i = 1, \ldots, L$. Now observe that the load at node i in the latter system is $\rho_0 + \rho_i(x_i)$, and remains bounded away from 1, even when the load ρ approaches 1 as $\lambda \uparrow \lambda^*$ (provided that there are in fact large class-i users, i.e., $\mathbb{P}(B_i < x_i) < 1$). Hence, the class-0 users and small class-i users are 'immune' from the heavy-traffic conditions, as is proved in the next proposition, which shows that the number of class-0 users and small class-i users remains bounded as the load approaches the capacity.

Proposition 6.3.1. *For a policy $\pi^*_{\boldsymbol{x}} \in \Pi^*_{\boldsymbol{x}}$ with $\mathbb{P}(B_i < x_i) < 1$ for all $i = 1, \ldots, L$, it holds that $\mathbb{E}(N_0^{\pi^*_{\boldsymbol{x}}} + N_{i,<x_i}^{\pi^*_{\boldsymbol{x}}}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: Without loss of generality we assume $\mathbb{P}(B_i < x_i) > 0$, $i = 1, \ldots, L$. Consider a policy $\pi^*_{\boldsymbol{x}} \in \Pi^*_{\boldsymbol{x}}$. Class-$i$ users of size smaller than x_i, $i = 1, \ldots, L$, and class-0 users do not notice the presence of users from the classes $1, \ldots, L$ with size larger than x_i. Policy $\pi^*_{\boldsymbol{x}}$ is work-conserving, and therefore

$$\mathbb{E}(W_0^{\pi^*_{\boldsymbol{x}}}) + \mathbb{E}(W_{i,<x_i}^{\pi^*_{\boldsymbol{x}}}) = \lambda \frac{p_0 \mathbb{E}(B_0^2) + p_i \mathbb{P}(B_i < x_i) \mathbb{E}(B_i^2 | B_i < x_i)}{2(1 - \rho_0 - \rho_i(x_i))}.$$

The condition $\mathbb{P}(B_i < x_i) < 1$ implies that $\lim_{\lambda \uparrow \lambda^*} 1 - \rho_0 - \rho_i(x_i) > 0$. Hence, we conclude that

$$\mathbb{E}(W_0^{\pi^*_{\boldsymbol{x}}}) + \mathbb{E}(W_{i,<x_i}^{\pi^*_{\boldsymbol{x}}}) = \mathrm{O}(1), \text{ as } \lambda \uparrow \lambda^*. \tag{6.6}$$

Now suppose that service among class-0 users and class-i users of size smaller than x_i, $i = 1, \ldots, L$, is non-preemptive as well (this assumption is not essential, see Remark 6.3.2 below). Then $(\mathbb{E}(N_0^{\pi^*_{\boldsymbol{x}}}) - 1)\mathbb{E}(B_0) \leq \mathbb{E}(W_0^{\pi^*_{\boldsymbol{x}}})$ and $(\mathbb{E}(N_{i,<x_i}^{\pi^*_{\boldsymbol{x}}}) - 1)\mathbb{E}(B_i | B_i < x_i) \leq \mathbb{E}(W_{i,<x_i}^{\pi^*_{\boldsymbol{x}}})$. Together with (6.6) this proves the proposition. □

Remark 6.3.2. In a similar way as in the single-server case, Proposition 6.3.1 can also be proved without the non-preemptive assumption with regard to class-0 users and class-i users smaller than x_i, $i = 1, \ldots, L$. Under policy $\pi^*_{\boldsymbol{x}}$, these users do not notice the presence of class-i users of size larger than x_i. Since each node is work-conserving we can therefore upper bound $\mathbb{E}(N_0^{\pi^*_{\boldsymbol{x}}}) + \mathbb{E}(N_{i,<x_i}^{\pi^*_{\boldsymbol{x}}})$ by the mean number of users in a system with only class-0 users and class-i users smaller than x_i under the LRPT policy. This gives that $\mathbb{E}(N_0^{\pi^*_{\boldsymbol{x}}}) + \mathbb{E}(N_{i,<x_i}^{\pi^*_{\boldsymbol{x}}})$ is less than or equal to

$$\lambda(p_0 + p_i \mathbb{P}(B_i < x_i)) \Big(\frac{\mathbb{E}(\bar{B})}{1 - \rho_0 - \rho_i(x_i)} + \frac{\lambda(p_0 + p_i \mathbb{P}(B_i < x_i)) \mathbb{E}(\bar{B}^2)}{2(1 - \rho_0 - \rho_i(x_i))^2} \Big),$$

with $\mathbb{E}(\bar{B}^k) = \frac{p_0}{p_0 + p_i \mathbb{P}(B_i < x_i)} \mathbb{E}(B_0^k) + \frac{p_i \mathbb{P}(B_i < x_i)}{p_0 + p_i \mathbb{P}(B_i < x_i)} \mathbb{E}(B_i^k | B_i < x_i)$, $k = 1, 2$. The result now follows by noting that $\lim_{\lambda \uparrow \lambda^*} 1 - \rho_0 - \rho_i(x_i) > 0$.

Proposition 6.3.1 implies that the scaled mean number of class-0 users and small class-i users tends to zero in heavy traffic. The number of large class-i users can be bounded in terms of the total workload at node i, which results in an upper bound for the scaled total mean number of users in the system, as provided in the next proposition.

Proposition 6.3.3. *For a policy $\pi^*_{\boldsymbol{x}} \in \Pi^*_{\boldsymbol{x}}$ with $\mathbb{P}(B_i < x_i) < 1$ for all $i = 1, \ldots, L$, it holds that*

$$\lim_{\lambda \uparrow \lambda^*} (1 - \rho) \mathbb{E}(N^{\pi^*_{\boldsymbol{x}}}) \leq \lambda^* \sum_{i=1}^{L} \frac{p_0 \mathbb{E}(B_0^2) + p_i \mathbb{E}(B_i^2)}{2 x_i}.$$

Proof: It follows from Proposition 6.3.1 that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_{\boldsymbol{x}}^*}) = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\sum_{i=1}^{L}\mathbb{E}(N_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*})$. Since service is non-preemptive among class-i users of original size larger than x_i, we have $\mathbb{E}(W_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*}) \geq x_i(\mathbb{E}(N_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*})-1)$. Hence, we obtain $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*}) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*})}{x_i} \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W_i^{\pi_{\boldsymbol{x}}^*})+\mathbb{E}(W_0^{\pi_{\boldsymbol{x}}^*})}{x_i} = \lambda^*\frac{p_0\mathbb{E}(B_0^2)+p_i\mathbb{E}(B_i^2)}{2x_i}$, for $i=1,\dots,L$, which proves the statement. □

Comparison with proportional fairness

We now compare the performance of the policies in the class $\cup_{\boldsymbol{x}}\Pi_{\boldsymbol{x}}^*$ with that of PF as a natural extension of PS.

Proposition 6.3.4. *Let $\pi_{\boldsymbol{x}}^* \in \Pi_{\boldsymbol{x}}^*$. When $B_1,\dots,B_L$ have infinite support, we have*

$$\lim_{\boldsymbol{x}\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_{\boldsymbol{x}}^*})}{\mathbb{E}(N^{PF})} = 0.$$

When $B_1,\dots,B_L$ have finite support, we have

$$\lim_{\boldsymbol{x}\uparrow\boldsymbol{M}}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_{\boldsymbol{x}}^*})}{\mathbb{E}(N^{PF})} \leq \frac{\lambda^*}{L}\sum_{i=1}^{L}\frac{p_0\mathbb{E}(B_0^2)+p_i\mathbb{E}(B_i^2)}{2M_i}.$$

Proof: From (4.23) and (4.24), we obtain

$$\begin{aligned}\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{PF}) &= \lim_{1-\rho_0-\rho_i\downarrow 0}(1-\rho_0-\rho_i)\mathbb{E}(N^{PF}) = \lim_{1-\rho_0-\rho_i\downarrow 0}\frac{\sum_{i=1}^{L}\rho_i}{1-\rho_0}\\ &= \frac{L(1-\rho_0)}{1-\rho_0} = L. \end{aligned} \tag{6.7}$$

Together with Proposition 6.3.3, this proves the assertion. □

We deduce that when $B_1,\dots,B_L$ have infinite support, there exists a policy $\pi_{\boldsymbol{x}}^* \in \cup_{\boldsymbol{x}}\Pi_{\boldsymbol{x}}^*$ that outperforms PF by an arbitrarily large factor in a heavy-traffic regime. This may be intuitively explained as follows. Under the PF policy, the total workload is distributed across users of various sizes, and hence the total number of users grows linearly with the workload as $\lambda\uparrow\lambda^*$. In contrast, under policies in the class $\cup_{\boldsymbol{x}}\Pi_{\boldsymbol{x}}$ the dominant fraction of the workload is contributed by class-i users of size larger than x_i as $\lambda\uparrow\lambda^*$. Thus, as the value of x_i increases, the entire workload is concentrated in fewer and fewer users compared to PF.

Comparing Propositions 6.2.4 and 6.3.4 we observe that the relative improvement over the PF policy achieved by policies in the class $\pi_{\boldsymbol{x}}^*$ is equal to the average relative improvement that would have been obtained over the PS policy by policies in the class $\Pi_{\boldsymbol{x}}$ in each of the L nodes separately.

Optimality properties

Assume that B_j has finite support for all classes $j = 0, \ldots, L$, with $\sum_{i=1}^{L} \frac{1}{M_i} \leq \frac{1}{M_0}$. The next proposition shows that for any policy $\pi \in \Pi$, there exists a policy in $\cup_{\boldsymbol{x}} \Pi_{\boldsymbol{x}}^*$ that performs at least as well in heavy-traffic conditions. This may be heuristically interpreted as follows. As mentioned above, under policies in the class $\cup_{\boldsymbol{x}} \Pi_{\boldsymbol{x}}^*$ the lion share of the workload is composed of class-i users of size larger than x_i, while at the same time the total workload in each node is minimized. In case $\sum_{i=1}^{L} \frac{1}{M_i} \leq \frac{1}{M_0}$, and the value of x_i is close to M_i, it turns out that it is not possible to achieve a smaller total number of users for the given workload than attained under policies in $\cup_{\boldsymbol{x}} \Pi_{\boldsymbol{x}}^*$.

Proposition 6.3.5. *Let $\pi_{\boldsymbol{x}}^* \in \Pi_{\boldsymbol{x}}^*$. Assume $M_j < \infty$ for $j = 0, \ldots, L$, and $\sum_{i=1}^{L} \frac{1}{M_i} \leq \frac{1}{M_0}$. Then for any policy $\pi \in \Pi$,*

$$\lim_{\boldsymbol{x} \uparrow \boldsymbol{M}} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi_{\boldsymbol{x}}^*})}{\mathbb{E}(N^{\pi})} \leq 1.$$

Proof: Policy $\pi_{\boldsymbol{x}}^* \in \Pi_{\boldsymbol{x}}^*$ is work-conserving in all nodes. Therefore we have for any policy $\pi \in \Pi$,

$$\mathbb{E}(W_0^{\pi_{\boldsymbol{x}}^*}) + \mathbb{E}(W_i^{\pi_{\boldsymbol{x}}^*}) \leq \mathbb{E}(W_0^{\pi}) + \mathbb{E}(W_i^{\pi}). \tag{6.8}$$

Proposition 6.3.1 implies that $\lim_{\lambda \uparrow \lambda^*} (1-\rho)\mathbb{E}(N_{i,<x_i}^{\pi_{\boldsymbol{x}}^*}) = 0$. In conjunction with $(\mathbb{E}(N_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*}) - 1)x_i \leq \mathbb{E}(W_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*})$, this yields that $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N_i^{\pi_{\boldsymbol{x}}^*})x_i \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(W_{i,\geq x_i}^{\pi_{\boldsymbol{x}}^*})$. It also follows from Proposition 6.3.1 that $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^*}) = 0$. Furthermore, we have $\mathbb{E}(W_i^{\pi}) \leq \mathbb{E}(N_i^{\pi})M_i$ for every policy $\pi \in \Pi$. Together with (6.8), the above implies that for any policy $\pi \in \Pi$,

$$(1-\rho)\Big(\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^*})M_0 + \mathbb{E}(N_i^{\pi_{\boldsymbol{x}}^*})x_i\Big) \leq (1-\rho)\Big(\mathbb{E}(N_0^{\pi})M_0 + \mathbb{E}(N_i^{\pi})M_i\Big) + \mathrm{o}(1-\rho), \tag{6.9}$$

for $i = 1, \ldots, L$. In addition, $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^*}) = 0$ yields that

$$(1-\rho)\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^*})M_0 \leq (1-\rho)\mathbb{E}(N_0^{\pi})M_0 + \mathrm{o}(1-\rho). \tag{6.10}$$

Multiplying (6.9) by $\frac{1}{M_i}$ for all $i = 1, \ldots, L$, and (6.10) by $\frac{1}{M_0} - \sum_{i=1}^{L} \frac{1}{M_i} \geq 0$, and summing these $L+1$ inequalities, gives

$$(1-\rho)\Big(\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^*}) + \sum_{i=1}^{L} \mathbb{E}(N_i^{\pi_{\boldsymbol{x}}^*})\frac{x_i}{M_i}\Big) \leq (1-\rho)\sum_{j=0}^{L} \mathbb{E}(N_j^{\pi}) + \mathrm{o}(1-\rho).$$

Letting $x_i \uparrow M_i$, $i = 1, \ldots, L$, concludes the proof. □

6.3.2 Favoring classes $i = 1, \dots, L$ simultaneously

We now consider policies that serve all classes $1, \dots, L$ simultaneously whenever possible or serve class-0 users of size smaller than x_0. If that is not feasible, then class-0 users of size larger than x_0 are served. Otherwise, classes $1, \dots, L$ are served. Within class j, users of size smaller than x_j receive priority, $j = 0, \dots, L$. Other than that, the priority structure within each of the classes is irrelevant for the analysis. Service is non-preemptive among class-j users of size larger than x_j. We denote this class of anticipating policies by $\Pi^{**}_{\boldsymbol{x}}$, where $\boldsymbol{x} = (x_0, \dots, x_L)$. As before, we use the notation $\boldsymbol{x} \uparrow \boldsymbol{M}$ and $\boldsymbol{x} \to \infty$ to indicate that $x_j \uparrow M_j$ for all $j = 0, \dots, L$ and $x_j \to \infty$ for all $j = 0, \dots, L$, respectively (order is irrelevant).

Under policies in the class $\Pi^{**}_{\boldsymbol{x}}$, the number of small class-0 users as well as the number of small class-i users remain bounded as the load approaches the capacity, see Proposition 6.3.7. Compared to Proposition 6.3.1, this is far more difficult to prove. While these users indeed receive some degree of preferred treatment, it is no longer the case that they do not notice the presence of the large users. Observe that simultaneous service of large class-i users can have precedence over service of small class-0 users, and that small class-i users must be simultaneously present in order to receive priority over large class-0 users. Hence, in order to prove the above assertion, we need elaborate arguments as provided in Lemma 6.3.6 below. The proof of Lemma 6.3.6 may be found in Appendix 6.A. Denote by $\hat{W}^c_j(t)$ the workload at time t in a reference system with class-j traffic only, service rate c, and with $\hat{W}^c_j(0) = 0$. Define $U^d_j(t) := \sup_{0 \le s \le t}\{d(t-s) - A_j(s,t)\}$.

Lemma 6.3.6. *Let $\pi^{**}_{\boldsymbol{x}} \in \Pi^{**}_{\boldsymbol{x}}$ and let $\delta > 0$ be such that $\delta < 1 - \rho_i$ for all $i = 1, \dots, L$. Assume $W^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}(0) = W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i}(0) = 0$, for a certain $i \in \{1, \dots, L\}$. Then at time $t \ge 0$, there exists a $j^* \in \{1, \dots, L\}$, such that*

$$W^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}(t) + W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i}(t) \le \hat{W}^{\rho_0-\delta}_{0,<x_0}(t) + \hat{W}^{\rho_i-\delta}_{i,<x_i}(t) + \hat{W}^{\rho_{j^*}+\delta}_{j^*}(t) + U^{\rho_{j^*}-\delta}_{j^*}(t). \qquad (6.11)$$

Proposition 6.3.7. *For a policy $\pi^{**}_{\boldsymbol{x}} \in \Pi^{**}_{\boldsymbol{x}}$ with $\mathbb{P}(B_j < x_j) < 1$, $j = 0, \dots, L$, it holds that $\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}) = \mathrm{O}(1)$ and $\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_{j,<x_j}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: Using Lemma 6.3.6, we obtain that

$$\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}) + \mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i}) \le \mathbb{E}(\hat{W}^{\rho_0-\delta}_{0,<x_0}) + \mathbb{E}(\hat{W}^{\rho_i-\delta}_{i,<x_i}) + \sum_{j=1}^{L} \mathbb{E}(\hat{W}^{\rho_j+\delta}_j) + \sum_{j=1}^{L} \mathbb{E}(U^{\rho_j-\delta}_j).$$

For δ small enough, $\mathbb{E}(\hat{W}^{\rho_j-\delta}_{j,<x_j}) = \lambda \frac{p_j \mathbb{P}(B_j<x_j)\mathbb{E}(B_j^2|B_j<x_j)}{2(\rho_j-\delta-\rho_j(x_j))}$ and $\mathbb{E}(\hat{W}^{\rho_j+\delta}_j) = \lambda \frac{p_j \mathbb{E}(B_j^2)}{2\delta}$, which implies $\mathbb{E}(\hat{W}^{\rho_j-\delta}_{j,<x_j}) = \mathbb{E}(\hat{W}^{\rho_j+\delta}_j) = \mathrm{O}(1)$. Furthermore, $U^{\rho_j-\delta}_j(t)$ converges in distribution to $\sup_{T \ge 0}\{(\rho_j - \delta)T - A_j(0,T)\}$ as $t \to \infty$ (see [9, Corollary III.7.2]). The latter is the supremum of a random walk with drift $\rho_j - \delta - \rho_j = -\delta < 0$, and has a finite mean by [9, Theorem X.2.1]. Since the drift is independent of λ, this implies that $\mathbb{E}(U^{\rho_j-\delta}_j) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$. Together, this gives $\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}) + \mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i}) = \mathrm{O}(1)$.

Using similar arguments as in the proof of Proposition 6.3.1 yields the assertion. □

Proposition 6.3.7 implies that the scaled mean number of small class-0 and small class-i users tends to zero in heavy traffic. The number of large class-0 and large class-i users can be bounded in terms of the total workload at node i, which results in an upper bound for the scaled total mean number of users in the system, as provided in the next proposition.

Proposition 6.3.8. *For a policy $\pi_{\boldsymbol{x}}^{**} \in \Pi_{\boldsymbol{x}}^{**}$ with $\mathbb{P}(B_j < x_j) < 1$ for all $j = 0, \ldots, L$, it holds that* $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N^{\pi_{\boldsymbol{x}}^{**}}) \leq \lambda^* \frac{Lp_0\mathbb{E}(B_0^2)+\sum_{i=1}^L p_i\mathbb{E}(B_i^2)}{2\min(x_0,x_1,\ldots,x_L)}$.

Proof: Proposition 6.3.7 indicates that the mean number of class-j users smaller than x_j, $j = 0, \ldots, L$, under policy $\pi_{\boldsymbol{x}}^{**}$ remains bounded as $\lambda \uparrow \lambda^*$. Since service is non-preemptive among class-j users of size larger than x_j, $j = 0, \ldots, L$, it follows that $(1-\rho)(\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^{**}}) + \mathbb{E}(N_i^{\pi_{\boldsymbol{x}}^{**}})) \leq (1-\rho)\frac{\mathbb{E}(W_0^{\pi_{\boldsymbol{x}}^{**}})+\mathbb{E}(W_i^{\pi_{\boldsymbol{x}}^{**}})}{\min(x_0,x_i)} + o(1-\rho)$ as $\lambda \uparrow \lambda^*$. Hence, the scaled total mean number of users can be upper bounded: $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N^{\pi_{\boldsymbol{x}}^{**}}) \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\Big(L\mathbb{E}(N_0^{\pi_{\boldsymbol{x}}^{**}}) + \sum_{i=1}^L \mathbb{E}(N_i^{\pi_{\boldsymbol{x}}^{**}})\Big) \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\Big(\frac{\sum_{i=1}^L(\mathbb{E}(W_0^{\pi_{\boldsymbol{x}}^{**}})+\mathbb{E}(W_i^{\pi_{\boldsymbol{x}}^{**}}))}{\min(x_0,x_1,\ldots,x_L)}\Big) = \lambda^* \frac{Lp_0\mathbb{E}(B_0^2)+\sum_{i=1}^L p_i\mathbb{E}(B_i^2)}{2\min(x_0,x_1,\ldots,x_L)}$. □

Comparison with proportional fairness

We now compare the performance of the policies in the class $\cup_{\boldsymbol{x}}\Pi_{\boldsymbol{x}}^{**}$ with that of PF.

Proposition 6.3.9. *Let $\pi_{\boldsymbol{x}}^{**} \in \Pi_{\boldsymbol{x}}^{**}$. When $B_0, B_1, \ldots, B_L$ have infinite support, we have*

$$\lim_{\boldsymbol{x}\to\infty} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi_{\boldsymbol{x}}^{**}})}{\mathbb{E}(N^{PF})} = 0.$$

When $B_0, B_1, \ldots, B_L$ have finite support, we have

$$\lim_{\boldsymbol{x}\uparrow \boldsymbol{M}} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi_{\boldsymbol{x}}^{**}})}{\mathbb{E}(N^{PF})} = \frac{\lambda^*}{L} \frac{Lp_0\mathbb{E}(B_0^2) + \sum_{i=1}^L p_i\mathbb{E}(B_i^2)}{2\min(M_0, M_1, \ldots, M_L)}.$$

Proof: Proposition 6.3.8, together with (6.7), gives the result. □

As before, we conclude that when $B_0, B_1, \ldots, B_L$ have infinite support, there exists a policy $\pi_{\boldsymbol{x}}^{**} \in \cup_{\boldsymbol{x}}\Pi_{\boldsymbol{x}}^{**}$ that outperforms PF by an arbitrarily large factor in heavy-traffic conditions.

Optimality properties

We now assume that B_j has finite support for all classes, with $\sum_{i=1}^L \frac{1}{M_i} \geq \frac{1}{M_0}$ and $\frac{1}{M_0} \geq \sum_{j=1,j\neq i}^L \frac{1}{M_j}$, for all $i = 1, \ldots, L$. The next proposition shows that for any policy $\pi \in \Pi$, there exists a policy in $\cup_{\boldsymbol{x}} \Pi^{**}_{\boldsymbol{x}}$ that performs at least as well in heavy-traffic conditions. As before, these policies manage to simultaneously minimize the workload in each of the nodes and concentrate the entire workload in users of maximum size.

Proposition 6.3.10. *Assume $M_j < \infty$ for all $j = 0, \ldots, L$, $\sum_{i=1}^{L} \frac{1}{M_i} \geq \frac{1}{M_0}$, and $\frac{1}{M_0} \geq \sum_{j=1, j\neq i}^{L} \frac{1}{M_j}$ for all $i = 1, \ldots, L$. Let $\pi^{**}_{\boldsymbol{x}} \in \Pi^{**}_{\boldsymbol{x}}$. Then for any policy $\pi \in \Pi$,*

$$\lim_{\boldsymbol{x} \uparrow \boldsymbol{M}} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}})}{\mathbb{E}(N^{\pi})} \leq 1.$$

The proof may be found in Appendix 6.B. The idea of the proof can be described as follows. Instead of proving the result for an arbitrary policy in $\Pi^{**}_{\boldsymbol{x}}$, we first focus on a policy $\pi^p \in \Pi^{**}_{\boldsymbol{x}} \cap \Pi^{**}$. Lemma 4.2.1 holds for π^p, and states that the aggregate workload in at least one pair of nodes is minimized under policy π^p. As will be shown in Appendix 6.B, this allows to prove the optimality of π^p. Since the scaled workloads for policies in the class $\Pi^{**}_{\boldsymbol{x}}$ turn out to be identical in heavy traffic, the optimality result holds for any policy in this class.

6.4 Numerical evaluation

In the previous sections we compared the performance of policies in classes $\Pi^*_{\boldsymbol{x}}$ and $\Pi^{**}_{\boldsymbol{x}}$ with that of PF in a heavy-traffic regime. In this section we focus on a subset of the class $\Pi^*_{\boldsymbol{x}}$, and conduct numerical experiments to illustrate the analytical findings and assess the scope for performance gains. We specifically examine those policies $\pi \in \Pi^*_{\boldsymbol{x}} \cap \Pi^*$ that serve class-i users of original size smaller than x_i, $i \neq 0$ and class-0 users in a non-preemptive fashion. Because of the non-preemptive feature, the following upper bounds hold:

$$\begin{aligned}
\mathbb{E}(N^{\pi}_0) &\leq 1 + \frac{\mathbb{E}(W^{\pi}_0)}{\mathbb{E}(B_0)}, \\
\mathbb{E}(N^{\pi}_{i,<x_i}) &\leq 1 + \frac{\mathbb{E}(W^{\pi}_{i,<x_i})}{\mathbb{E}(B_i | B_i < x_i)}, \quad i = 1, \ldots, L, \\
\mathbb{E}(N^{\pi}_{i,\geq x_i}) &\leq 1 + \frac{\mathbb{E}(W^{\pi}_{i,\geq x_i})}{\mathbb{E}(B_i | B_i \geq x_i)}, \quad i = 1, \ldots, L.
\end{aligned}$$

In case of exponentially distributed service requirements, the "1" in the right-hand side of the first equation may in fact be omitted. Since class 0 receives preemptive priority ($\pi \in \Pi^*$), its mean workload is

$$\mathbb{E}(W^{\pi}_0) = \frac{\lambda p_0 \mathbb{E}(B_0^2)}{2(1-\rho_0)}.$$

Class-0 and class-i users of size smaller than x_i, $i \neq 0$, are served in a work-conserving manner, and therefore

$$\mathbb{E}(W^{\pi}_0) + \mathbb{E}(W^{\pi}_{i,<x_i}) = \frac{\lambda(p_0 \mathbb{E}(B_0^2) + p_i \mathbb{P}(B_i < x_i)\mathbb{E}(B_i^2 | B_i < x_i))}{2(1 - \rho_0 - \rho_i(x_i))}.$$

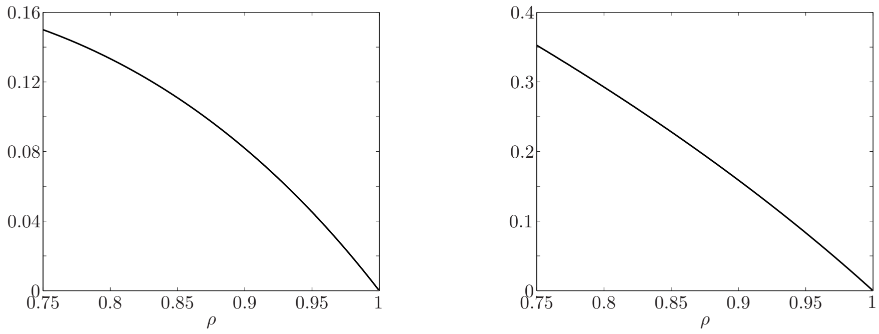


Figure 6.1: Upper bound for $(1-\rho)\mathbb{E}(N_0^\pi)$, for exponential service requirements (left) and Pareto service requirements (right).

Policy π is work-conserving in each node, hence

$$\mathbb{E}(W_{i,\geq x_i}^\pi) = \frac{\lambda(p_0\mathbb{E}(B_0^2)+p_i\mathbb{E}(B_i^2))}{2(1-\rho_0-\rho_i)} - \mathbb{E}(W_{i,<x_i}^\pi) - \mathbb{E}(W_0^\pi).$$

Using the formulas above, we calculated upper bounds for the mean number of users under policies in the class $\Pi_{\boldsymbol{x}}^* \cap \Pi^*$. We considered a system with two nodes, and set $p_0 = 0.5$, $p_1 = 0.25$, $p_2 = 0.25$ and $x_1 = x_2 = x$, and studied both exponentially and Pareto distributed service requirements. In the former case, we took $\mu_0 = 2$, $\mu_1 = 1$, $\mu_2 = 1$, while in the latter case we chose $a_0 = 10$, $a_1 = 3$, $a_2 = 3$ where $\mathrm{d}B_i(y) = a_i y^{-(a_i+1)}\mathrm{d}y$ for $y \geq 1$, $i = 0, 1, 2$.

In Figure 6.1 we plotted the upper bound for the scaled mean number of class-0 users as a function of ρ. Note that the scaled mean number of class-0 users does not depend on x and as ρ increases it converges to zero. In Figure 6.2 we plotted the upper bound for the scaled mean number of class-1 users smaller than x_1 as a function of ρ. Again, as ρ increases, it converges to zero. Furthermore, for a fixed ρ, we observe a horizontal asymptote as x grows large. This asymptote of the upper bound can be found by interchanging the order of limits, i.e.,

$$\begin{aligned}\lim_{\lambda\uparrow\lambda^*}\lim_{x\to\infty}(1-\rho)\mathbb{E}(N_{1,<x_1}^\pi) &\leq \lim_{\lambda\uparrow\lambda^*}\lim_{x\to\infty}(1-\rho)\Big(1+\frac{\mathbb{E}(W_{1,<x_1}^\pi)}{\mathbb{E}(B_1|B_1<x_1)}\Big)\\ &= \lambda^*\frac{p_0\mathbb{E}(B_0^2)+p_1\mathbb{E}(B_1^2)}{2\mathbb{E}(B_1)}.\end{aligned}$$

In Figure 6.3 we plotted the upper bound for the scaled mean number of class-1 users larger than x_1.

The three bounds specified above provide an upper bound for the total mean number of users. In Figure 6.4 we plot the ratio between this upper bound and the total mean number of users under PF as a function of ρ for exponentially and Pareto distributed service requirements. In both cases, the policy with threshold $x = 2$ gives already a substantial performance improvement for a load of 0.85.

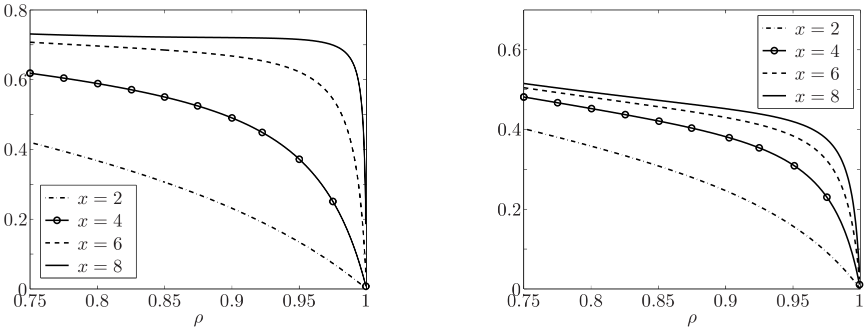


Figure 6.2: Upper bound for $(1-\rho)\mathbb{E}(N^{\pi}_{1,<x_1})$, for exponential service requirements (left) and Pareto service requirements (right).

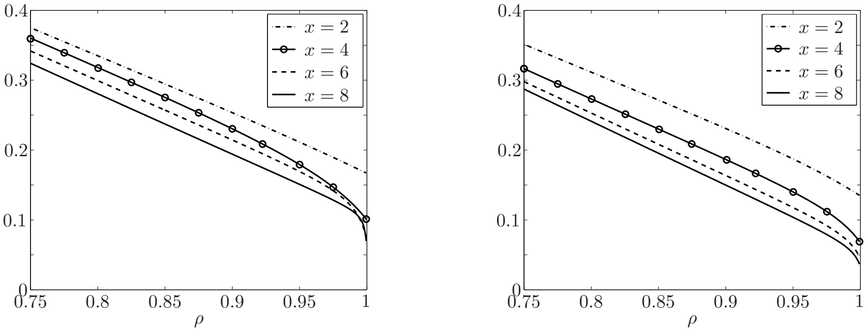


Figure 6.3: Upper bound for $(1-\rho)\mathbb{E}(N^{\pi}_{1,\geq x_1})$, for exponential service requirements (left) and Pareto service requirements (right).

6.5 Concluding remarks

We have demonstrated that size-based priority policies can be asymptotically optimal in heavy traffic for service requirements with bounded support. In addition, these policies obtain provable performance gains over PF, the prototypical α-fair policy. In particular, for service requirements with unbounded support, the total mean number of users may be reduced by an arbitrarily large factor when the load is sufficiently high. It is worth observing here that we have pursued deliberately simple policies in order to obtain provable asymptotic performance guarantees. There are clearly more sophisticated policies conceivable that will typically achieve larger gains, but may be too complex to allow any explicit performance guarantees. The results in this chapter indicate however that PF can be substantially improved upon, and hence studying possible implementable policies is promising.

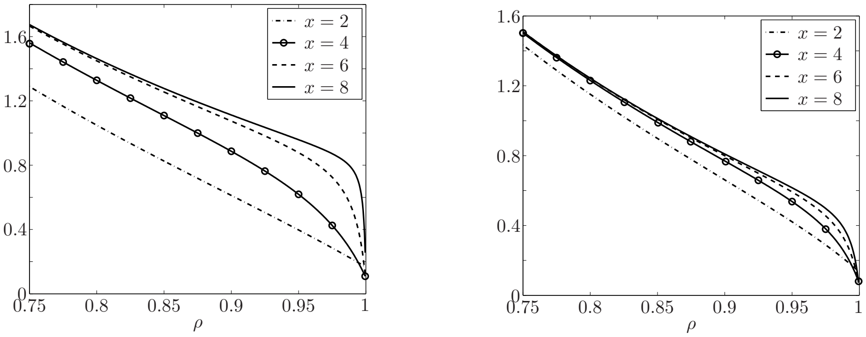


Figure 6.4: Upper bound for $\mathbb{E}(N^{\pi})/\mathbb{E}(N^{PF})$, for exponential service requirements (left) and Pareto service requirements (right).

Appendix

6.A Proof of Lemma 6.3.6

Let $\pi_{\boldsymbol{x}}^{**} \in \Pi_{\boldsymbol{x}}^{**}$ and take $\delta > 0$ such that $\delta < 1 - \rho_j$ for all $j = 1, \ldots, L$. Let $W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(0) = W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(0) = 0$, for a certain $i \in \{1, \ldots, L\}$. We will prove that at time $t \geq 0$, there is a $j^* \in \{1, \ldots, L\}$, such that (6.11) holds. For convenience, we assume that among class-i users of size smaller than x_i service is non-preemptive, although this is not essential in any way for the assertion to hold.

Define $s_1 := \sup\{s \leq t : W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(s) + \min(W_1^{\pi_{\boldsymbol{x}}^{**}}(s), \ldots, W_L^{\pi_{\boldsymbol{x}}^{**}}(s)) = 0\}$ and $s_2 := \sup\{s \leq t : W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(s) = 0\}$. Note that $W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(s_1) = 0$ and $W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(s_2) = 0$. In addition, there exists a $j^* \in \{1, \ldots, L\}$ such that $W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_1) = 0$. Denote by $B_i(s,t)$ the total amount of service given to class-i users during the time interval $(s,t]$, and denote by $B_{i,<x_i}(s,t)$ the portion of service that is given to class-i users of size smaller than x_i. Then,

$$B_{0,<x_0}(s,t) + B_{i,<x_i}(s,t) = t - s, \tag{6.12}$$

with $s := \max(s_1, s_2)$.

We distinguish between two cases: $s_1 \leq s_2$ and $s_1 \geq s_2$. If $s_1 \leq s_2$, then from (6.12) we obtain

$$B_{0,<x_0}(s_2,t) + B_{i,<x_i}(s_2,t) = t - s_2.$$

By definition, of s_1 and s_2 we have

$$\begin{aligned}
W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(t) &= W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(s_2) + A_{0,<x_0}(s_2,t) - B_{0,<x_0}(s_2,t), \\
W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(s_2) &= A_{0,<x_0}(s_1,s_2) - B_{0,<x_0}(s_1,s_2), \\
W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t) &= A_{i,<x_i}(s_2,t) - B_{i,<x_i}(s_2,t).
\end{aligned}$$

In the interval (s_1,t) there is continuously work present of class-0 users of size smaller than x_0 and work of class-j users, $j=1,\dots,L$. Therefore, under policy $\pi_{\boldsymbol{x}}^{**}$, for all $j\in\{1,\dots,L\}$,

$$B_{0,<x_0}(s_1,s_2)+B_j(s_1,s_2)=s_2-s_1.$$

Furthermore,

$$B_{j^*}(s_1,s_2)=W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_1)+A_{j^*}(s_1,s_2)-W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_2)=A_{j^*}(s_1,s_2)-W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_2).$$

Combining the above equations, we obtain

$$\begin{aligned}
&W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(t)+W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t)\\
&=A_{0,<x_0}(s_1,t)+A_{i,<x_i}(s_2,t)+A_{j^*}(s_1,s_2)-(t-s_1)-W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_2)\\
&\le A_{0,<x_0}(s_1,t)+A_{i,<x_i}(s_2,t)+A_{j^*}(s_1,s_2)-(t-s_1)\\
&\le A_{0,<x_0}(s_1,t)+A_{i,<x_i}(s_2,t)+A_{j^*}(s_1,t)-A_{j^*}(s_2,t)\\
&\quad-(\rho_0-\delta)(t-s_1)-(\rho_i+\delta)(t-s_1)-(\rho_i-\delta)(t-s_2)+(\rho_i-\delta)(t-s_2)\\
&\le \sup_{s\le t}\{A_{0,<x_0}(s,t)-(\rho_0-\delta)(t-s)\}+\sup_{s\le t}\{A_{i,<x_i}(s,t)-(\rho_i-\delta)(t-s)\}\\
&\quad+\sup_{s\le t}\{A_{j^*}(s,t)-(\rho_i+\delta)(t-s)\}+\sup_{s\le t}\{(\rho_i-\delta)(t-s)-A_{j^*}(s,t)\}.
\end{aligned}$$

Since $\rho_1=\dots,\rho_L$, this implies (6.11).

Now assume $s_1\ge s_2$. From (6.12) we obtain

$$B_{0,<x_0}(s_1,t)+B_{i,<x_i}(s_1,t)=t-s_1.$$

Furthermore,

$$\begin{aligned}
W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(t)&=A_{0,<x_0}(s_1,t)-B_{0,<x_0}(s_1,t),\\
W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t)&=A_{i,<x_i}(s_2,t)-B_{i,<x_i}(s_2,t),
\end{aligned}$$

and

$$B_{j^*}(s_2,s_1)=W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_2)+A_{j^*}(s_2,s_1)-W_{j^*}^{\pi_{\boldsymbol{x}}^{**}}(s_1)\ge A_{j^*}(s_2,s_1).$$

There is continuously work present of class-i users of size smaller than x_i in the interval (s_2,t). Hence, for all $j\in\{1,\dots,L\}$,

$$B_{i,<x_i}(s_2,s_1)\ge B_j(s_2,s_1).$$

Combining the above equations, we obtain

$$\begin{aligned}
&W_{0,<x_0}^{\pi_{\boldsymbol{x}}^{**}}(t) + W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t)\\
&= A_{0,<x_0}(s_1,t) + A_{i,<x_i}(s_2,t) - B_{i,<x_i}(s_2,s_1) - (t-s_1)\\
&\leq A_{0,<x_0}(s_1,t) + A_{i,<x_i}(s_2,t) - A_{j^*}(s_2,s_1) - (t-s_1)\\
&\leq A_{0,<x_0}(s_1,t) + A_{i,<x_i}(s_2,t) + A_{j^*}(s_1,t) - A_{j^*}(s_2,t)\\
&\quad -(\rho_0-\delta)(t-s_1) - (\rho_i+\delta)(t-s_1) + (\rho_i-\delta)(t-s_2) - (\rho_i-\delta)(t-s_2)\\
&\leq \sup_{s\leq t}\{A_{0,<x_0}(s,t) - (\rho_0-\delta)(t-s)\} + \sup_{s\leq t}\{A_{i,<x_i}(s,t) - (\rho_i-\delta)(t-s)\}\\
&\quad + \sup_{s\leq t}\{A_{j^*}(s,t) - (\rho_i+\delta)(t-s)\} + \sup_{s\leq t}\{(\rho_i-\delta)(t-s) - A_{j^*}(s,t)\}.
\end{aligned}$$

Since $\rho_1 = \ldots, \rho_L$, this implies (6.11). □

6.B Proof of Proposition 6.3.10

Take $\pi^p \in \Pi_{\boldsymbol{x}}^{**} \cap \Pi^{**}$. In Lemma 4.2.1 it is proved that for every policy $\pi \in \Pi$, there are at time t classes $j,k \in \{1,\ldots,L\}$, $j \neq k$, such that

$$W_0^{\pi^p}(t) + W_j^{\pi^p}(t) + W_k^{\pi^p}(t) \leq W_0^{\pi}(t) + W_j^{\pi}(t) + W_k^{\pi}(t). \tag{6.13}$$

Furthermore, π^p is work-conserving in all nodes. Therefore,

$$W_0^{\pi^p}(t) + W_i^{\pi^p}(t) \leq W_0^{\pi}(t) + W_i^{\pi}(t),\ i = 1,\ldots,L. \tag{6.14}$$

Multiplying (6.13) by $\sum_{i=1}^L \frac{1}{M_i} - \frac{1}{M_0} \geq 0$ and (6.14) by $\frac{1}{M_0} - \sum_{l=1,l\neq i}^L \frac{1}{M_l} \geq 0$ for $i = j,k$ and by $\frac{1}{M_i}$ for all $i = 1,\ldots,L$ with $i \neq j,k$, and summing these $L+1$ inequalities results in $\sum_{i=0}^L \frac{1}{M_i} W_i^{\pi^p}(t) \leq \sum_{i=0}^L \frac{1}{M_i} W_i^{\pi}(t)$, hence

$$\sum_{i=0}^L \frac{1}{M_i}\mathbb{E}(W_i^{\pi^p}) \leq \sum_{i=0}^L \frac{1}{M_i}\mathbb{E}(W_i^{\pi}). \tag{6.15}$$

We now extend this result to an arbitrary policy $\pi_{\boldsymbol{x}}^{**} \in \Pi_{\boldsymbol{x}}^{**}$ by analyzing the (scaled) workloads. We first show that

$$W_0^{\pi_{\boldsymbol{x}}^{**}}(t) + W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t) = W_0^{\pi^p}(t) + W_{i,<x_i}^{\pi^p}(t). \tag{6.16}$$

We prove this by contradiction. Assume at time t it holds, but immediately after time t the equality is violated. Hence, for example $W_0^{\pi_{\boldsymbol{x}}^{**}}(u) + W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(u) > W_0^{\pi^p}(u) + W_{i,<x_i}^{\pi^p}(u)$, for $u \in (t, t+\delta)$, with $\delta > 0$ small enough. In order for this to happen, we need that at time t, $W_0^{\pi_{\boldsymbol{x}}^{**}}(t) + W_{i,<x_i}^{\pi_{\boldsymbol{x}}^{**}}(t) = 0$, since otherwise policy $\pi_{\boldsymbol{x}}^{**}$ would be serving either class 0 or class-i users with service requirements strictly smaller than x_i. But this implies that $W_0^{\pi^p}(t) + W_{i,<x_i}^{\pi^p}(t) = 0$ as well, and hence $W_0^{\pi_{\boldsymbol{x}}^{**}}(u) +$

$W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i}(u) = 0 = W^{\pi^p}_0(u) + W^{\pi^p}_{i,<x_i}(u)$, for all $u \geq t$ until a new user arrives. Hence, we obtain a contradiction.

Since policies $\pi^{**}_{\boldsymbol{x}}$ and π^p are work-conserving in each node, we have by (6.16) that for $i = 1, \ldots, L$,

$$W^{\pi^{**}_{\boldsymbol{x}}}_{i,\geq x_i}(t) = W^{\pi^p}_{i,\geq x_i}(t). \tag{6.17}$$

We obtain from Proposition 6.3.7 that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)(\mathbb{E}(W^{\pi}_{0,<x_0}) + \mathbb{E}(W^{\pi}_{i,<x_i})) = 0$ for $\pi \in \{\pi^{**}_{\boldsymbol{x}}, \pi^p\}$. Together with (6.17) and the fact that $\pi^{**}_{\boldsymbol{x}}$ and π^p are work-conserving, this implies that

$$\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{0,\geq x_0}) = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W^{\pi^p}_{0,\geq x_0}). \tag{6.18}$$

Using (6.17), (6.18) and the fact that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\Big(\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{0,<x_0}) + \mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{i,<x_i})\Big) = 0$, we obtain from (6.15) that

$$\lim_{\lambda\uparrow\lambda^*}(1-\rho)\sum_{i=0}^{L}\frac{1}{M_i}\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_i) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\sum_{i=0}^{L}\frac{1}{M_i}\mathbb{E}(W^{\pi}_i), \tag{6.19}$$

for every policy $\pi \in \Pi$. Class-i users of size larger than x_i are served in a non-preemptive way, which implies $(\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_{i,\geq x_i}) - 1)x_i \leq \mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_{i,\geq x_i}) \leq \mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_i)$. Proposition 6.3.7 shows that under policy $\pi^{**}_{\boldsymbol{x}}$, all scaled class-$i$ work is composed of users of size x_i or larger, $i = 0, \ldots, L$, hence $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_i)x_i = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_{i,\geq x_i})x_i \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W^{\pi^{**}_{\boldsymbol{x}}}_i)$. Together with $\mathbb{E}(W^{\pi}_i) \leq \mathbb{E}(N^{\pi}_i)M_i$, we obtain from (6.19) that

$$(1-\rho)\sum_{i=0}^{L}\mathbb{E}(N^{\pi^{**}_{\boldsymbol{x}}}_i)\frac{x_i}{M_i} \leq (1-\rho)\sum_{i=0}^{L}\mathbb{E}(N^{\pi}_i) + \mathrm{o}(1-\rho).$$

Letting $\boldsymbol{x} \uparrow \boldsymbol{M}$, this concludes the proof. □

Chapter 7
Monotonicity properties for multi-class queueing systems

In this chapter we study a general multi-class queueing system where the capacity allocated to a class may depend on the numbers of users present in all classes. The linear bandwidth-sharing network under weighted α-fair bandwidth sharing policies can be seen as a special case of this framework. We compare policies in terms of their stability conditions, the workloads, and the numbers of users present in the various classes.

There is a wide range of literature on the ordering of random processes, see for example [101, 126]. In the seminal paper [92] (see also [85]) necessary and sufficient conditions on the transition rates are given for the existence of a stochastic ordering between two Markov processes starting from any two ordered initial states. It turns out that these conditions are often too strong in order to compare the behavior of different policies in a queueing system. In particular, they are not satisfied for weighted α-fair policies. In this chapter we take a different approach to compare the stochastic processes corresponding to different policies. By restricting the initial states and considering the same realizations of the arrival processes and service requirements, we obtain sufficient conditions on two policies in order to compare sample-path wise their workloads and number of users of the various classes. Since our result is a pure sample-path comparison, it holds for arbitrary arrival processes and service requirements. For exponential service requirements, these workload relations allow for a comparison of the mean holding cost as well.

In this chapter special attention is paid to the class of weighted α-fair bandwidth-sharing policies in the linear network. In particular, we obtain stability results and, for exponential service requirements with relatively small class-0 users, we show that the mean holding cost is decreasing in the fairness parameter α and the relative weights. The latter matches with the observation we made in Section 4.4. To cover all service requirement parameters, we consider a two-node linear network in a heavy-traffic regime and obtain further monotonicity results based on a conjecture in [67, 68]. In particular, in the case of relatively large class-0 users the heavy-traffic result suggests the following: As α increases, i.e., the policy becomes more fair, the holding cost increases as well. Hence, in that case there is a trade-off between achieving fairness and obtaining good performance.

Finally, we extend the framework to cover the multi-class single-server system for which we are especially interested in weighted time-sharing policies such as Discriminatory Processor Sharing (DPS) and Generalized Processor Sharing (GPS). For a single-server system with two classes of users with exponentially distributed service requirements we find that the mean holding cost under DPS or GPS is monotone with respect to the ratio of the weights. A similar result was obtained in Proposition 2.6.4 for the DPS multi-class single-server system in a heavy-traffic setting.

The remainder of the chapter is organized as follows. In Section 7.1 the model is introduced and Section 7.2 describes the comparison results for the general framework. We apply this framework to the linear network in Section 7.3 and we focus on weighted α-fair bandwidth-sharing policies in Section 7.4. In Section 7.5 we consider the multi-class single-server queue. Concluding remarks can be found in Section 7.6.

7.1 Model description

We consider a multi-class queueing system with $L+1$ classes of users. Class-j users arrive according to a renewal process with mean inter-arrival time $1/\lambda_j$, and have service requirements B_j with mean $1/\mu_j$, $j = 0, \ldots, L$. Let $\rho_j = \frac{\lambda_j}{\mu_j}$ represent the offered work of class j per time unit. The inter-arrival times and service requirements are mutually independent random variables.

For a given scheduling policy π, denote by $N_j^\pi(t)$ the number of class-j users in the system at time t and let $\vec{N}^\pi(t) = (N_0^\pi(t), N_1^\pi(t), \ldots, N_L^\pi(t))$. Let $W_j^\pi(t)$ denote the workload in class j at time t. We assume the processes $N_j^\pi(t)$ and $W_j^\pi(t)$ to be right-continuous with left limits. We further define N_j^π and W_j^π as random variables with the corresponding steady-state distributions (when they exist).

For a given policy π, denote by $s_j^\pi(t, \vec{n})$ the capacity received by class j at time t when the system is in state $\vec{n} = (n_0, n_1, \ldots, n_L)$. Hence the allocation given to class j can only depend on the time and on the number of users present in the system. We assume that $s_j^\pi(t, \vec{n}) = 0$ when $n_j = 0$. In addition, the allocation vector $\vec{s}^\pi(t, \vec{n}) = (s_0^\pi(t, \vec{n}), \ldots, s_L^\pi(t, \vec{n}))$ has to lie in a certain capacity region $R(t) \subset \mathbb{R}_+^{L+1}$ which may depend on the time t but not on the state $\vec{n}$ itself, that is $\vec{s}^\pi(t, \vec{n}) \in R(t)$. In the remainder of the chapter we suppress the dependence on t and write $\vec{s}^\pi(\vec{n})$ instead of $\vec{s}^\pi(t, \vec{n})$.

For a given policy π, denote by

$$S_j^\pi(t) := \int_0^t s_j^\pi(\vec{N}^\pi(u)) \mathrm{d}u$$

the cumulative amount of service received by class j during the time interval $(0, t]$. Let $A_j(0, t)$ be the amount of class-j work that arrived in the time interval $(0, t]$.

Then the workload of class j at time t can be written as

$$W_j^\pi(t) = W_j^\pi(0) + A_j(0,t) - S_j^\pi(t). \tag{7.1}$$

In order to characterize the evolution of the number of users we need information on the intra-class policy. The intra-class policy prescribes how the capacity allocated to class j, $s_j^\pi(\vec{n})$, is distributed among the class-j users. In this chapter we assume that the intra-class policy is FCFS.

Remark 7.1.1. When the service requirements are exponentially distributed, for any non-anticipating intra-class policy, the stochastic behavior of the workloads and the numbers of users of the various classes is determined completely by the allocation vector $\vec{s}^\pi(\vec{n})$ and does not depend on the intra-class policy used. This implies that for exponential service requirements, the results we obtain (by assuming FCFS) are also valid for any non-anticipating intra-class policy, e.g. PS and LAS.

Our goal in this chapter is to compare the performance of a multi-class queueing system under different policies. First of all, we will be interested in whether a policy can achieve stability. Another important performance measure we consider is the holding cost, $\sum_{j=0}^{L} c_j N_j^\pi(t)$, where c_j is an arbitrary nonnegative cost associated with class j, $j = 0, \ldots, L$.

In Sections 7.3 and 7.4 we focus on the linear network, as depicted in Figure 1.2, which is a particular example of a multi-class queueing system. It might be convenient for the reader to bear this network in mind when reading Section 7.2. In particular, in Sections 7.3 and 7.4 we focus on a linear network consisting of L nodes with time-varying capacity. Hence the capacity region corresponding to the linear network is equal to

$$R(t) = \{(s_0, s_1, \ldots, s_L) \in \mathbb{R}^{L+1} : s_0 + s_i \leq C_i(t),\ \forall i = 1, \ldots, L\},$$

where $C_i(t)$, $i = 1, \ldots, L$, denotes the capacity of node i at time t. When $C_i(t) = C$ for all $i = 1, \ldots, L$, we refer to it as a symmetric linear network.

7.2 Comparison of policies

In this section we consider the behavior of the queueing system under two different policies for the same realizations of the arrival processes and service requirements. The next property states conditions that will allow us to compare two policies.

Property 7.2.1. *Let π and $\tilde{\pi}$ be two policies such that*

(i) $s_0^\pi(\vec{n}^\pi) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$*, when* $n_0^\pi = n_0^{\tilde{\pi}}$ *and* $n_i^\pi \geq n_i^{\tilde{\pi}}, \forall i = 1, \ldots, L$*, and,*

(ii) $s_0^\pi(\vec{n}^\pi) + s_i^\pi(\vec{n}^\pi) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) + s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$*,* $i = 1, \ldots, L$*, for all states* $\vec{n}^\pi$ *and* $\vec{n}^{\tilde{\pi}}$ *that satisfy one of the following conditions:*

- $n_0^\pi > 0$*,* $n_0^\pi \geq n_0^{\tilde{\pi}}$*,* $0 < n_i^{\tilde{\pi}}$ *and* $n_i^\pi \leq n_i^{\tilde{\pi}}$*.*

- $n_0^\pi = n_0^{\tilde{\pi}} = 0$, $0 < n_i^\pi = n_i^{\tilde{\pi}}$ *and* $n_j^\pi \geq n_j^{\tilde{\pi}}$ *for all* $j \neq 0, i$.

Under Property 7.2.1 we establish a sample-path comparison result for the number of class-0 users and for the workload in the system. This result will play a key role in the remainder of this chapter.

Proposition 7.2.2. *Let* π *and* $\tilde{\pi}$ *be two policies that satisfy Property 7.2.1 and consider the same realizations of the arrival processes and service requirements. Assume* $W_0^\pi(0) \geq W_0^{\tilde{\pi}}(0)$ *and* $W_0^\pi(0)+W_i^\pi(0) \geq W_0^{\tilde{\pi}}(0)+W_i^{\tilde{\pi}}(0)$ *for all* $i = 1, \ldots, L$. *It holds that for all* $t \geq 0$,

(i) $S_0^\pi(t) - W_0^\pi(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$,

(ii) $S_0^\pi(t)-W_0^\pi(0)+S_i^\pi(t)-W_i^\pi(0) \leq S_0^{\tilde{\pi}}(t)-W_0^{\tilde{\pi}}(0)+S_i^{\tilde{\pi}}(t)-W_i^{\tilde{\pi}}(0), \quad i = 1, \ldots, L,$

and hence

(iii) $N_0^\pi(t) \geq N_0^{\tilde{\pi}}(t)$, $W_0^\pi(t) \geq W_0^{\tilde{\pi}}(t)$,

(iv) $W_0^\pi(t) + W_i^\pi(t) \geq W_0^{\tilde{\pi}}(t) + W_i^{\tilde{\pi}}(t), \quad i = 1, \ldots, L.$

We like to emphasize that because of the FCFS assumption and the same realizations of the arrival processes and service requirements, we implicitly assume that at time 0 the k-th most recently arrived class-j user has the same service requirement under both policies, $j = 0, 1, \ldots, L$, $k = 1, \ldots, \min(N_j^\pi(0), N_j^{\tilde{\pi}}(0))$. Hence, the condition in Proposition 7.2.2 always holds when both processes start in the same state $\vec{N}^\pi(0) = \vec{N}^{\tilde{\pi}}(0)$, where at time $t = 0$ each user has the same (remaining) service requirement under both policies.

In the proof of Proposition 7.2.2 we use $f(t^+) > g(t^+)$ to denote that there exists a sufficiently small $\delta > 0$ such that $f(u) > g(u)$ for all $u \in (t, t+\delta]$. Since $\{N_i(t)\}_{t\geq 0}$ is a piece-wise constant right-continuous process, this ensures that an inequality for $N_i^\pi(t)$ and $N_i^{\tilde{\pi}}(t)$ immediately translates to the same inequality for $N_i^\pi(t^+)$ and $N_i^{\tilde{\pi}}(t^+)$. This property is used throughout the proof.

Proof of Proposition 7.2.2: From (7.1) we obtain that inequality (i) implies $W_0^\pi(t) \geq W_0^{\tilde{\pi}}(t)$ and inequality (ii) implies inequality (iv). Also note that $W_0^\pi(t) \geq W_0^{\tilde{\pi}}(t)$ implies $N_0^\pi(t) \geq N_0^{\tilde{\pi}}(t)$, since the intra-class policy is FCFS and we assume the same realizations of the arrival and service requirements under both policies. Therefore, it suffices to prove that inequalities (i) and (ii) hold.

We prove (i) and (ii) by contradiction. Suppose they do not hold sample-path wise. Let t be the first time epoch at which one of the two inequalities is violated.

First assume that inequality (i) is the first one to be violated, i.e., $S_0^\pi(t) - W_0^\pi(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $s_0^\pi(\vec{N}^\pi(t^+)) > s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ (with strict inequality), but $S_0^\pi(t) - W_0^\pi(0) + S_i^\pi(t) - W_i^\pi(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ for all $i = 1, \ldots, L$. Hence, from (7.1) we obtain $W_0^\pi(t) = W_0^{\tilde{\pi}}(t)$ and $W_i^\pi(t) \geq W_i^{\tilde{\pi}}(t)$ for all $i = 1, \ldots, L$. Since the k-th most recently arrived class-j user before the current time t has the same (original) service requirement under both policies and the intra-class policy is FCFS, we have as well

$$N_0^{\pi}(t) = N_0^{\tilde{\pi}}(t) \quad \text{and} \quad N_i^{\pi}(t) \geq N_i^{\tilde{\pi}}(t) \quad \text{for all} \quad i = 1, \dots, L. \tag{7.2}$$

The process $\{N_i(t)\}_{t \geq 0}$ is a piece-wise constant and right-continuous process, hence (7.2) remains true at time t^+. Together with Property 7.2.1 this gives $s_0^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption.

Next, assume that inequality (ii) is violated at time t, i.e., $S_0^{\pi}(t) - W_0^{\pi}(0) + S_i^{\pi}(t) - W_i^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ and $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) > s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ (with strict inequality), but $S_0^{\pi}(t) - W_0^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $S_0^{\pi}(t) - W_0^{\pi}(0) + S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for all $j \neq 0, i$. Hence $W_0^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)$ and $W_i^{\pi}(t) \leq W_i^{\tilde{\pi}}(t)$, from which (as before) we can conclude that $N_0^{\pi}(t^+) \geq N_0^{\tilde{\pi}}(t^+)$ and $N_i^{\pi}(t^+) \leq N_i^{\tilde{\pi}}(t^+)$. We now distinguish between the following possibilities:

- If $N_i^{\tilde{\pi}}(t^+) > 0$ and $N_0^{\pi}(t^+) > 0$, then by Property 7.2.1 (ii) it follows that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption.

- If $N_i^{\tilde{\pi}}(t^+) > 0$ and $N_0^{\pi}(t^+) = 0$, then $N_0^{\tilde{\pi}}(t^+) = 0$ and hence $S_0^{\pi}(t) - W_0^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ which implies $S_i^{\pi}(t) - W_i^{\pi}(0) = S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ and $S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for $j \neq 0, i$. So $0 = N_0^{\pi}(t^+) = N_0^{\tilde{\pi}}(t^+)$, $N_i^{\pi}(t^+) = N_i^{\tilde{\pi}}(t^+) > 0$, and $N_j^{\pi}(t^+) \geq N_j^{\tilde{\pi}}(t^+)$ for all $j \neq 0, i$. By Property 7.2.1 (ii) it follows that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption.

- If $N_i^{\tilde{\pi}}(t^+) = 0$, then $N_i^{\pi}(t^+) = 0$ as well, and hence $S_i^{\pi}(t) - W_i^{\pi}(0) = S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$. This implies $S_0^{\pi}(t) - W_0^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for all j, implying $W_0^{\pi}(t) = W_0^{\tilde{\pi}}(t)$ and $W_j^{\pi}(t) \geq W_j^{\tilde{\pi}}(t)$. As before, we obtain that $N_0^{\pi}(t^+) = N_0^{\tilde{\pi}}(t^+)$ and $N_j^{\pi}(t^+) \geq N_j^{\tilde{\pi}}(t^+)$ for all $j \neq 0$. By virtue of Property 7.2.1 this means that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$. Since $N_i^{\tilde{\pi}}(t^+) = N_i^{\pi}(t^+) = 0$, we also have that $s_i^{\pi}(\vec{N}^{\pi}(t^+)) = s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) = 0$, and hence $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption. □

Remark 7.2.3. Proposition 7.2.2 is a sample-path result and does not require any distributional or independence assumptions with respect to the inter-arrival times and service requirements. The only assumption required is that the arrival characteristics are independent of the state of the system, since in Proposition 7.2.2 we use the same realizations of the arrival processes and service requirements when comparing the policies.

Proposition 7.2.2 (iii) states in fact a sample-path wise pre-ordering on two continuous-time processes $\{\vec{N}^{\pi}(t)\}_{t \geq 0}$ and $\{\vec{N}^{\tilde{\pi}}(t)\}_{t \geq 0}$ starting from ordered initial states. There is a broad range of literature on the existence of orderings of stochastic processes. In particular, let $X(t)$ and $Y(t)$ be two continuous-time Markov processes. In [92, Theorem 5.3] and [85, Theorem 2] necessary and sufficient conditions on the transition rates are given in order for a coupling $(X'(t), Y'(t))$ to exist that is order-preserving ($X(t) \overset{d}{=} X'(t)$, $Y(t) \overset{d}{=} Y'(t)$ and $\mathbb{P}(X'(t) \leq Y'(t), \forall t \geq 0) = 1$) for any ordered initial states ($X(0) \leq Y(0)$). So if the processes X and Y are initially ordered, then the order is maintained at all times. Here $\leq$ denotes a pre-order relation on the state space. In particular, from [85, 92] it follows that, in a Markovian setting, the necessary and sufficient conditions on the policies π and $\tilde{\pi}$ in order to obtain

$$\{N_0^{\pi}(t)\}_{t\geq 0} \geq_{st} \{N_0^{\tilde{\pi}}(t)\}_{t\geq 0}, \quad \text{for any} \quad N_0^{\pi}(0) \geq_{st} N_0^{\tilde{\pi}}(0), \tag{7.3}$$

are

$$s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \;\text{ whenever }\; n_0^{\pi} = n_0^{\tilde{\pi}}. \tag{7.4}$$

(The pre-ordering relation used here for the $L+1$-dimensional process $\vec{N}(t)$ is defined by the number of class-0 users.) The sufficient condition in Property 7.2.1 for the sample-path comparison of Proposition 7.2.2 to hold, and the necessary and sufficient condition in (7.4) for the stochastic comparison in (7.3) to hold, are not directly comparable. Given two policies, it is possible that either only Property 7.2.1 is satisfied, or only (7.4) is satisfied. Note that the stochastic ordering result in (7.3) holds for any two ordered initial states, $N_0^{\pi}(0) \geq N_0^{\tilde{\pi}}(0)$. In Proposition 7.2.2 the initial states are ordered as well, but we assume that at time $t = 0$ we have additional knowledge on the service requirements of the users present under policy π and $\tilde{\pi}$. So in this respect we would expect Property 7.2.1 to be weaker than (7.4). On the other hand, in Proposition 7.2.2 the coupling is specified in advance, namely the two processes are coupled by their arrival processes and service requirements, while in (7.3) any coupling is allowed to obtain the desired order-preserving result. So in this respect we would expect (7.4) to be weaker than Property 7.2.1.

In a queueing context, condition (7.4) is rather strong. One often encounters examples where $s_0(\vec{n}) \to 0$ as $n_i \to \infty$, $i \neq 0$. If this is the case for policy $\tilde{\pi}$, then (7.4) will not be satisfied. In Sections 7.4 and 7.5 we consider settings for which Property 7.2.1 is satisfied, while (7.4) does not hold. In addition, Proposition 7.2.2 is not restricted to Markov processes, hence it applies as well for general arrival processes, service requirements and time-varying capacity regions.

In the remainder of this section, Proposition 7.2.2 is used to derive results for the stability and mean holding cost.

7.2.1 Stability

The sample-path comparison in Proposition 7.2.2 does not require the system to be stable. In particular, Proposition 7.2.2 (iv) implies the following result.

Corollary 7.2.4. *Assume policies π and $\tilde{\pi}$ satisfy Property 7.2.1. If the system is stable under policy π, then it is stable under policy $\tilde{\pi}$ as well, in the sense that the system is empty under policy $\tilde{\pi}$ whenever it is empty under policy π.*

In particular, if the empty state has a finite mean recurrence time under policy π in the case of Poisson arrivals, then it has a finite mean recurrence time under policy $\tilde{\pi}$ as well.

Proof: The first statement follows by noting that if $\sum_{j=0}^{L} W_j^{\pi}(t) = 0$, then we obtain from Proposition 7.2.2 (iv) that $\sum_{j=0}^{L} W_j^{\tilde{\pi}}(t) = 0$. The second assertion is a direct implication of the first one. □

7.2.2 Mean holding cost

In case the service requirements are exponentially distributed with $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$, the sample-path comparison established in Proposition 7.2.2 allows us to compare the mean holding cost.

Proposition 7.2.5. *Assume the service requirements are exponentially distributed. Let π and $\tilde{\pi}$ be two policies that satisfy Property 7.2.1 and assume policy π gives a stable system. If $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$, then*

$$\sum_{j=0}^{L} c_j \mathbb{E}(N_j^{\pi}(t)) \geq \sum_{j=0}^{L} c_j \mathbb{E}(N_j^{\tilde{\pi}}(t)), \quad \forall\ t \geq 0.$$

Proof: Assume at time $t = 0$ the conditions as stated in Proposition 7.2.2 are satisfied (for example, assume both policies π and $\tilde{\pi}$ start with an empty system). From Proposition 7.2.2 (iii) we have that $N_0^{\pi}(t) \geq N_0^{\tilde{\pi}}(t)$ for all $t \geq 0$, and hence

$$\mathbb{E}(N_0^{\pi}(t)) \geq \mathbb{E}(N_0^{\tilde{\pi}}(t)), \quad \text{for all } t \geq 0. \tag{7.5}$$

From Proposition 7.2.2 (iv) we derive that $W_0^{\pi}(t)+W_i^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)+W_i^{\tilde{\pi}}(t)$, so that $\mathbb{E}(W_0^{\pi}(t))+\mathbb{E}(W_i^{\pi}(t)) \geq \mathbb{E}(W_0^{\tilde{\pi}}(t))+\mathbb{E}(W_i^{\tilde{\pi}}(t))$ for all $i = 1, \ldots, L$. Since the policy is non-anticipating and the service requirements are exponentially distributed, and thus memoryless, we obtain $\mathbb{E}(W_i^{\pi}(t)) = \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t))$ and hence for all $i = 1, \ldots, L$,

$$\frac{1}{\mu_0}\mathbb{E}(N_0^{\pi}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t)) \geq \frac{1}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\tilde{\pi}}(t)), \quad \text{for all } t \geq 0. \tag{7.6}$$

Inequalities (7.5) and (7.6) together with $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$ give

$$\begin{aligned}
&\sum_{j=0}^{L} c_j \mathbb{E}(N_j^{\pi}(t)) \\
&= \frac{c_0\mu_0 - \sum_{i=1}^{L} c_i\mu_i}{\mu_0}\mathbb{E}(N_0^{\pi}(t)) + \sum_{i=1}^{L} c_i\mu_i \left(\frac{1}{\mu_0}\mathbb{E}(N_0^{\pi}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t))\right) \\
&\geq \frac{c_0\mu_0 - \sum_{i=1}^{L} c_i\mu_i}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \sum_{i=1}^{L} c_i\mu_i \left(\frac{1}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\tilde{\pi}}(t))\right) \\
&= \sum_{j=0}^{L} c_i \mathbb{E}(N_j^{\tilde{\pi}}(t)),
\end{aligned}$$

for all $t \geq 0$. □

Note that by Remark 7.1.1, Proposition 7.2.5 holds for any non-anticipating intra-class policy, so not only for FCFS.

Remark 7.2.6. We only obtain a comparison result in terms of the *mean* holding cost, while we start from a sample-path comparison as stated in Proposition 7.2.2. The derivation of stochastic ordering results remains as a challenging topic for further research.

When $\vec{N}^{\pi}(t)$ and $\vec{N}^{\tilde{\pi}}(t)$ are two Markov processes, the necessary and sufficient conditions in order to obtain $\sum_{j=0}^{L} N_j^{\pi}(t) \geq_{st} \sum_{j=0}^{L} N_j^{\tilde{\pi}}(t)$, for any ordered initial states with $\sum_{j=0}^{L} N_j^{\pi}(0) \geq \sum_{j=0}^{L} N_j^{\tilde{\pi}}(0)$, are $\sum_{j=0}^{L} \mu_j s_j^{\pi}(\vec{n}^{\pi}) \leq \sum_{j=0}^{L} \mu_j s_j^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$ for all states with $\sum_{j=0}^{L} n_j^{\pi} = \sum_{j=0}^{L} n_j^{\tilde{\pi}}$, [85, 92]. In a queueing context this condition is rather strong. In Sections 7.3 and 7.5 we will see settings for which this condition is not satisfied.

7.3 Linear network

In this section we apply the results obtained in Section 7.2 to the linear network with time-varying capacity as introduced at the end of Section 7.1. Throughout this chapter we focus on Pareto-efficient policies. A policy π is said to be Pareto-efficient if it does not leave any capacity unnecessarily unused. For the linear network this implies that $s_i^{\pi}(\vec{n}) = C_i(t) - s_0^{\pi}(\vec{n})$ when $n_i > 0$, $i = 1\ldots, L$, and $s_0^{\pi}(\vec{n}) = \min_{i=1,\ldots,L} C_i(t)$ when $n_i = 0$, for all $i = 1, \ldots, L$. It can be shown that any policy that leaves capacity unused, can be improved sample-path wise (in terms of the workload and the number of users of the various classes) by a Pareto-efficient policy. However, a Pareto-efficient policy is *not* sufficient to ensure a stable system under the maximum stability conditions, as explained in Section 1.4.

Condition (ii) in Property 7.2.1 is always satisfied for a Pareto-efficient policy $\tilde{\pi}$, since $s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) + s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = C_i(t)$ whenever $n_i^{\tilde{\pi}} > 0$. Hence, in the specific case of a linear network, Property 7.2.1 simplifies as follows.

Property 7.3.1. *Let π and $\tilde{\pi}$ be two Pareto-efficient policies such that $s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$, when $n_0^{\pi} = n_0^{\tilde{\pi}}$ and $n_i^{\pi} \geq n_i^{\tilde{\pi}}$ for all $i = 1, \ldots, L$.*

In particular, Property 7.3.1 is implied by the following property.

Property 7.3.1'. *Let π and $\tilde{\pi}$ be two Pareto-efficient policies such that $s_0^{\pi}(\vec{n}) \leq s_0^{\tilde{\pi}}(\vec{n})$, and either $s_0^{\pi}(\vec{n})$ or $s_0^{\tilde{\pi}}(\vec{n})$ is non-increasing with respect to n_i for all $i \neq 0$.*

In order to see this, assume that Property 7.3.1' is satisfied with (for example) $s_0^{\tilde{\pi}}(\vec{n})$ non-increasing with respect to n_i for all $i \neq 0$. Then we have $s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$, with $n_i^{\pi} \geq n_i^{\tilde{\pi}}$ for all $i \neq 0$ and $n_0^{\pi} = n_0^{\tilde{\pi}}$. This is exactly Property 7.3.1.

Assume policies π and $\tilde{\pi}$ satisfy either Property 7.3.1 or 7.3.1'. This basically means that higher priority is given to class 0 under policy $\tilde{\pi}$ compared to π. From Section 7.2 we then obtain the following results. Under policy $\tilde{\pi}$ the number of class-0 users is less than under policy π (Proposition 7.2.2 (iii)) and the stability conditions are less strict for policy $\tilde{\pi}$ (Corollary 7.2.4). These results arise from the fact that when class 0 is served, it simultaneously uses capacity in all nodes. Hence, giving more preference to class 0 makes better use of the available capacity and hence makes the workload in each node smaller, i.e., $W_0^\pi(t) + W_i^\pi(t) \geq W_0^{\tilde{\pi}}(t) + W_i^{\tilde{\pi}}(t),\ i = 1, \ldots, L$ (Proposition 7.2.2 (iv)). When in addition $c_0\mu_0 \geq \sum_{i=1}^L c_i\mu_i$, that is the maximum weighted departure rate is obtained when class 0 is served, giving higher priority to class 0 decreases the mean holding cost $\sum_{j=0}^L c_j\mathbb{E}(N_j(t))$ as well (Proposition 7.2.5).

In a linear network, one natural choice for the weights c_j could be to relate them to the number of links each class uses. For example, take $c_0 = L$ and $c_i = 1$, $i = 1, \ldots, L$. In this case the result of Proposition 7.2.5 is valid under the intuitively appealing condition $\frac{1}{L}\sum_{i=1}^L \mu_i \leq \mu_0$, i.e., the departure rate of class 0 is larger than or equal to the average departure rate for classes $1, \ldots, L$.

Remark 7.3.2. Assume $\vec{N}^\pi(t)$ and $\vec{N}^{\tilde{\pi}}(t)$ are two Markov processes for any two policies π and $\tilde{\pi}$. When Property 7.3.1 is satisfied, a sample-path comparison for the number of class-0 users in a linear network holds. The condition (7.4) is a necessary and sufficient condition for a stochastic ordering relation for the number of class-0 users to exist as in the framework of [85, 92]. It can be immediately seen that Property 7.3.1 is a weaker condition than (7.4). Interestingly, for applications as will be given later in this chapter, the policies do satisfy Property 7.3.1, but not (7.4).

When $\mu_0 \geq \sum_{i=1}^L \mu_i$ and Property 7.3.1 is satisfied, it is possible to compare the total mean number of users in a linear network under the two policies. As mentioned in Remark 7.2.6, in a queueing context the sufficient and necessary conditions to stochastically order the total number of users for any ordered initial states, are rather strong. For the special case of a linear network they are even never satisfied. When choosing the states such that $\vec{n}^\pi = (0, 1, \ldots, 1)$ and $\vec{n}^{\tilde{\pi}} = (L, 0, \ldots, 0)$, it is needed that $\sum_{i=1}^L \mu_i \leq \mu_0$, but when choosing the states such that $\vec{n}^\pi = (1, 0, \ldots, 0)$ and $\vec{n}^{\tilde{\pi}} = (0, \ldots, 0, 1, 0, \ldots, 0)$, it is needed that $\mu_0 \leq \mu_i$, $i = 1, \ldots, L$, see Remark 7.2.6. Hence, we see that there does not exist any combination of the variables $\mu_0, \ldots, \mu_L$, for which these conditions are satisfied, and a stochastic ordering relation for the total number of users as in the framework of [85, 92] does not hold.

Recall that in Proposition 4.3.1 we obtained a policy that minimizes the mean holding cost for the linear network with unit capacities. Proposition 7.2.5 allows to readily extend this to time-varying capacities.

Corollary 7.3.3. *Consider a linear network with time-varying capacities. Assume the service requirements are exponentially distributed. Let policy π^* be the policy that serves class 0 at maximum rate, i.e.,*

$$s_0^{\pi^*}(\vec{n}) = \min_i C_i(t) \quad \textit{if} \quad n_0 > 0 \quad \textit{and} \quad s_0^{\pi^*}(\vec{n}) = 0 \quad \textit{otherwise.}$$

Classes $1,\ldots,L$ *obtain what is left, i.e.,*

$$s_i^{\pi^*}(\vec{n}) = C_i(t) - s_0^{\pi^*}(\vec{n}) \;\; \textit{if} \;\; n_i > 0 \;\; \textit{and} \;\; s_i^{\pi^*}(\vec{n}) = 0 \;\; \textit{otherwise.}$$

If $\sum_{i=1}^{L} c_i\mu_i \le c_0\mu_0$, *then policy* π^* *minimizes the mean holding cost* $\sum_{j=0}^{L} c_j\mathbb{E}(N_j(t))$, *for all* $t \ge 0$, *among all non-anticipating policies.*

Proof: Note that $s_0^{\pi^*}(\vec{n})$ is constant with respect to n_i, $i \neq 0$. In addition, $s_0^{\pi^*}(\vec{n}) \ge s_0^{\pi}(\vec{n})$ for any policy π. Hence, Property 7.3.1' is satisfied and from Proposition 7.2.5 we obtain $\sum_{j=0}^{L} c_j\mathbb{E}(N_j^{\pi}(t)) \ge \sum_{j=0}^{L} c_j\mathbb{E}(N_j^{\pi^*}(t))$ for all $t \ge 0$. □

Proposition 7.2.2 and Property 7.3.1 are stated in order to compare two different *policies*. However, they also allow us to evaluate the impact of removing a node from the linear network on the performance of class 0, i.e., compare two different *networks* under the same policy. In the next corollary we show that the number of class-0 users is reduced when a node (and hence the corresponding cross traffic) is removed.

Corollary 7.3.4. *Let* π *be a policy in a linear network with* L *nodes that satisfies the following property:*

$$s_0^{\pi}(n_0, n_1, \ldots, n_L) \le s_0^{\pi}(n_0, m_1, \ldots, m_{L-1}, 0)$$

for all $n_i \ge m_i, i = 1, \ldots, L-1$.

Also consider the linear network where node L *is removed (and hence has* $L-1$ *nodes) and apply the same policy* π *in the following way:* $s_0^{\pi}(n_0, \ldots, n_{L-1}) := s_0^{\pi}(n_0, \ldots, n_{L-1}, 0)$.

If $W_0^{\pi,L}(0) \ge W_0^{\pi,L-1}(0)$ *and* $W_0^{\pi,L}(0) + W_i^{\pi,L}(0) \ge W_0^{\pi,L-1}(0) + W_i^{\pi,L-1}(0)$, *then*

$$N_0^{\pi,L}(t) \ge N_0^{\pi,L-1}(t)$$

and for $i = 1, \ldots, L-1$

$$W_0^{\pi,L}(t) + W_i^{\pi,L}(t) \ge W_0^{\pi,L-1}(t) + W_i^{\pi,L-1}(t),$$

with $N_i^{\pi,l}(t)$ *and* $W_i^{\pi,l}(t)$ *the number of class-i users and the class-i workload, respectively, at time t under policy* π *in a linear network with l nodes.*

Proof: Policy π in a linear network with $L-1$ nodes can be seen as a policy in a linear network with L nodes by ignoring the class-L users. Denote this policy by $\tilde{\pi}$. So for all $x \ge 0$, $s_0^{\tilde{\pi}}(n_0, n_1, \ldots, n_{L-1}, x) := s_0^{\pi}(n_0, n_1, \ldots, n_{L-1})$. Hence

$$\begin{aligned} s_0^{\pi}(n_0, n_1, \ldots, n_{L-1}, n_L) &\le s_0^{\pi}(n_0, m_1, \ldots, m_{L-1}, 0) \\ &= s_0^{\pi}(n_0, m_1, \ldots, m_{L-1}) \\ &= s_0^{\tilde{\pi}}(n_0, m_1, \ldots, m_{L-1}, x) \end{aligned}$$

for all x and all $n_i \ge m_i$, $i = 1, \ldots, L-1$. This implies that policies π and $\tilde{\pi}$ satisfy Property 7.3.1 and from Proposition 7.2.2 the result follows. □

7.4 Weighted α-fair policies

In this section we focus on weighted α-fair policies as defined in Section 1.4.1. We denote the weighted α-fair policy with weights $w = (w_0, w_1, \ldots, w_L)$ and parameter α by $\pi(\alpha, w)$ and the corresponding allocation vector by $\vec{s}^{\,\pi(\alpha,w)}(\vec{n})$. Recall that the latter is the solution to the following optimization problem:

$$\begin{cases} \max_{\vec{s}\in R(t)} \sum_{j=0}^{L} w_j n_j \left(\frac{s_j}{n_j}\right)^{1-\alpha} / (1-\alpha) & \text{if } \alpha > 0,\ \alpha \neq 1, \\ \max_{\vec{s}\in R(t)} \sum_{j=0}^{L} w_j n_j \log(\frac{s_j}{n_j}) & \text{if } \alpha = 1, \end{cases} \tag{7.7}$$

and that the intra-class policy of $\pi(\alpha, w)$ is PS. In Section 7.1 we assumed however that the intra-class policy is FCFS. Throughout this section we consider exponentially distributed service requirements, thus, the results we obtain will also be valid if the intra-class policy is PS, see Remark 7.1.1.

In order to compare two α-fair policies we only need to check whether Property 7.3.1' holds. In [30] it was shown that for a symmetric linear network with unit capacities the weighted α-fair allocation equals

$$s_0^{\pi(\alpha,w)}(\vec{n}) = \frac{(w_0 n_0^{\alpha})^{1/\alpha}}{(w_0 n_0^{\alpha})^{1/\alpha} + (\sum_{i=1}^{L} w_i n_i^{\alpha})^{1/\alpha}} \tag{7.8}$$

and $s_i^{\pi(\alpha,w)}(\vec{n}) = 1 - s_0^{\pi(\alpha,w)}(\vec{n})$ for all i with $n_i > 0$. Using (7.8), it can be checked that for this allocation Property 7.3.1' is satisfied when comparing policies $\pi(\beta, w)$ and $\pi(\gamma, \tilde{w})$ with $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \ldots, L$ (see also [81, Proposition 6.1]). For an asymmetric network we have no expression for the weighted α-fair allocation available. However, in that case the optimization problem (7.7) allows us to prove that Property 7.3.1' is satisfied. This is stated in the next lemma and the proof may be found in Appendix 7.A.

Lemma 7.4.1. *The following results hold in a linear network:*

(i) $s_0^{\pi(\alpha,w)}(\vec{n})$ *is non-increasing in* n_i, $i = 1, \ldots, L$.

(ii) If $\beta \leq \gamma$, *then* $s_0^{\pi(\beta,w)}(\vec{n}) \leq s_0^{\pi(\gamma,w)}(\vec{n})$ *for all* $\vec{n}$.

(iii) If $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \ldots, L$, *then* $s_0^{\pi(\alpha,w)}(\vec{n}) \leq s_0^{\pi(\alpha,\tilde{w})}(\vec{n})$ *for all* $\vec{n}$.

Since Property 7.3.1' holds for weighted α-fair policies, the comparison results in Proposition 7.2.2 apply. This allows us to gain insights into the performance of such policies in linear networks, see Sections 7.4.1 and 7.4.2.

The stochastic comparison results in [85, Theorem 2] and [92, Theorem 5.3] are not applicable to the weighted α-fair policies. As we already mentioned in Remark 7.3.2, such an ordering is not possible for the total number of users present in the system. Also, an ordering for the number of class-0 users for any ordered initial states is not possible, since equation (7.4) is not satisfied for the class of weighted α-fair policies in linear networks. Consider for example the simple symmetric linear network and choose states such that $n_0^\pi = n_0^{\tilde{\pi}}$, $n_1^\pi = 1$ and $n_1^{\tilde{\pi}} = m$ with π and $\tilde{\pi}$ two α-fair policies. From (7.8) we see that if m tends to ∞ then $s_0^{\pi(\alpha,w)}(\vec{n}^{\tilde{\pi}})$ tends to 0. Hence (7.4) cannot hold for any pair of α-fair policies.

In [33] the authors consider a network of processor-sharing queues. The capacity of the various queues is variable and depends on the number of users present in all the queues. Stochastic bounds for the number of users present in each queue are obtained for policies that satisfy the monotonicity property (removing a user from any queue, increases the capacity allocated to every other user). This property fails to hold for a linear network under α-fair policies, as also indicated in [33]. For example, removing a class-1 user implies that class 1 gets less capacity and class 0 gets more. This however implies that classes $i = 2, \ldots, L$ obtain less capacity as well and hence a class-i user gets less capacity, $i = 2, \ldots, L$. A requirement in Property 7.3.1' is that removing a class-i user, $i \neq 0$, increases the capacity allocated to the class-0 users. As shown in Lemma 7.4.1, this holds under natural conditions on the parameters of weighted α-fair policies.

Remark 7.4.2. From Lemma 7.4.1 and Corollary 7.3.4 we obtain that under a weighted α-fair policy, the number of class-0 users in a linear network with L nodes is larger than in a linear network with $L - 1$ nodes.

In Section 7.4.1 the stability results are presented and in Section 7.4.2 monotonicity of the mean holding cost with respect to the fairness parameter and the relative weights is established. In order to broaden the comparison result, in Section 7.4.3 we investigate a heavy-traffic regime and in Section 7.4.4 we perform numerical experiments. In Section 7.4.5 we describe a time-scale separation (the dynamics of class-0 users are infinitely faster than those of classes $1, \ldots, L$) and derive approximations for the mean number of users.

7.4.1 Stability

In [30] it is proved that for Poisson arrivals and exponentially distributed service requirements, any weighted α-fair policy, $\alpha > 0$, in a bandwidth-sharing network with *fixed* capacity gives a stable system, in the sense that the queue length process is positive recurrent, under the maximum stability conditions. Corollary 7.2.4 and Lemma 7.4.1 allow us to derive stability results for a linear network with *time-varying* capacities.

Corollary 7.4.3. *Consider a linear network with time-varying capacities. Let the service requirements be exponentially distributed. Assume $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \ldots, L$. If the network is stable under policy $\pi(\beta, w)$, then it is stable under policy $\pi(\gamma, \tilde{w})$ as well, in the sense that the system is empty under policy $\pi(\gamma, \tilde{w})$ whenever it is empty under policy $\pi(\beta, w)$.*

Proof: The α-fair policies have PS as intra-class policy. However, since we assume that the service requirements are exponentially distributed, the stochastic behavior of the network does not depend on which non-anticipating intra-class policy is being used. Therefore we can assume that we have a FCFS intra-class policy. From

Lemma 7.4.1 we obtain that Property 7.3.1 is satisfied, hence the result in Corollary 7.2.4 applies. □

A related stability result is obtained in [82], where the authors consider systems with a time-varying general capacity region under an α-fair policy with unit weights. They assume that the capacity region can be in a finite number of states according to a stationary and ergodic process. For Poisson arrivals and exponentially distributed service requirements they characterize the stability conditions under which the process is positive recurrent, and show that the stability region is non-increasing in the value of α. Interestingly, Corollary 7.4.3 indicates that the stability region is in fact also non-decreasing in the value of α in the setting of a linear network. We therefore obtain the following result.

Corollary 7.4.4. *Assume Poisson arrivals and exponentially distributed service requirements. Consider a linear network and assume the set of all the possible capacity vectors $(C_1(t),\dots,C_L(t))$ can be in a finite number of states and evolves as a stationary and ergodic process. Let $\overline{C}_i$ be the average of the process $C_i(t)$.*

Policy $\pi(\alpha,w)$ with $\alpha>0$ and $w_i\leq w_0, i=1,\dots,L$, gives a stable system (positive recurrent) under the necessary stability conditions $\rho_0+\rho_i<\overline{C}_i$, $i=1,\dots,L$.

Proof: In [82] it is shown that for α-fair policies with $\alpha>0$ and unit weights $(w_j=1, j=0,\dots,L)$ the necessary stability conditions are given by $\rho_0+\rho_i<\overline{C}_i$, $i=1,\dots,L$. Moreover, it is established that these conditions are sufficient as well for the policy $\pi(\alpha,\vec{1})$ when $\alpha\downarrow 0$. On the other hand, Corollary 7.4.3 states that the stability conditions become less strict when α increases. This proves that $\pi(\alpha,\vec{1})$ is stable under the necessary stability conditions, for all $\alpha>0$. From Corollary 7.4.3 we can then conclude that the same holds for policy $\pi(\alpha,w)$ with $w_i\leq w_0, i=1,\dots,L$. □

7.4.2 Mean holding cost

We are now ready to derive a monotonicity result for the mean holding cost for weighted α-fair policies in a time-varying linear network. When $\sum_{i=1}^L c_i\mu_i\leq c_0\mu_0$, the instantaneous weighted departure rate of class 0 is relatively large, hence, it will be attractive to give preference to class-0 users, either by increasing the relative weight given to class 0, w_0/w_i, or by increasing the parameter α, see Lemma 7.4.1. At the same time this makes better use of the available capacity of the nodes, see Proposition 7.2.2 (iv). In the next corollary we prove that the mean holding cost indeed decreases when more preference is given to class 0. More precisely, the mean holding cost is non-increasing in α and in $\frac{w_0}{w_i}$, $i=1,\dots,L$.

Corollary 7.4.5. *Consider a linear network with time-varying capacities. Assume exponentially distributed service requirements with $\sum_{i=1}^L c_i\mu_i\leq c_0\mu_0$. If $\beta\leq\gamma$ and $\frac{w_0}{w_i}\leq\frac{\tilde{w}_0}{\tilde{w}_i}$, $i=1,\dots,L$, then*

$$\sum_{j=0}^{L} c_j\mathbb{E}(N_j^{\pi(\beta,w)}(t))\geq\sum_{j=0}^{L} c_j\mathbb{E}(N_j^{\pi(\gamma,\tilde{w})}(t)),\quad\forall\ t\geq 0.$$

Proof: From Lemma 7.4.1 we obtain that $\pi(\beta, w)$ and $\pi(\gamma, \tilde{w})$ satisfy Property 7.3.1'. The result then follows from Proposition 7.2.5. □

When $\sum_{i=1}^{L} c_i\mu_i > c_0\mu_0$ the analysis is more difficult. For example, in a two-node linear network ($L = 2$) with $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, it is beneficial to give more preference to classes 1 and 2 (and hence less preference to class 0) since that will maximize the total instantaneous weighted departure rate. From Lemma 7.4.1 we see that this can be done by choosing α small. In the case of exponentially distributed service requirements and a *heavily loaded* system, the mean holding cost is indeed strictly increasing in α, as we will see in Section 7.4.3. For a *normally loaded* system this is however not the case (see the simulations in Section 7.4.4). Then the effect that a smaller α uses the available capacity in each node less efficiently becomes more apparent.

7.4.3 Heavy-traffic regime

In this section we compare α-fair policies in a heavy-traffic scenario for a two-node linear network with fixed capacities C_1 and C_2. Below we briefly state heavy-traffic results from [67, 68] specialized to the two-node linear network under α-fair policies with unit weights. We refer to [67, 68] for the full details. We assume exponential distributed service requirements, as is the case in [67, 68].

Assume the heavy-traffic setting $\rho_i + \rho_0 = C_i$, $i = 1, 2$. Define the diffusion scaled processes as follows. For $j = 0, 1, 2$,

$$\hat{N}_j^{k,\pi(\alpha)}(t) := \frac{N_j^{\pi(\alpha,\vec{1})}(kt)}{\sqrt{k}},$$

and for $i = 1, 2$,

$$\hat{V}_i^{k,\pi(\alpha)}(t) := \frac{N_0^{\pi(\alpha,\vec{1})}(kt)/\mu_0 + N_i^{\pi(\alpha,\vec{1})}(kt)/\mu_i}{\sqrt{k}} = \frac{\hat{N}_0^{k,\pi(\alpha)}(t)}{\mu_0} + \frac{\hat{N}_i^{k,\pi(\alpha)}(t)}{\mu_i}.$$

Here $\hat{V}_i^{k,\pi(\alpha)}(t)$ can be seen as the total workload in node i under the diffusion scaling. In [68, Conjecture 5.1] it is conjectured that the diffusion scaled workload process $\vec{\hat{V}}^{k,\pi(\alpha)}(t)$ converges in distribution to $\vec{\hat{V}}^{\pi(\alpha)}(t)$ as $k \to \infty$, where $\vec{\hat{V}}^{\pi(\alpha)}(t)$ is a semimartingale reflecting Brownian motion (with a covariance matrix independent of α) living in a workload cone. For $\alpha = 1$ this conjecture is proved in [67, 68]. In addition, it is mentioned that this result can be extended to $\alpha \neq 1$. Throughout this section we assume that the conjecture holds for the two-node linear network.

The workload cone for an α-fair policy with unit weights is given by

$$\{\vec{v} : v_i = \frac{\rho_0}{\mu_0}(q_1 + q_2)^{\frac{1}{\alpha}} + \frac{\rho_i}{\mu_i} q_i^{\frac{1}{\alpha}},\ q_1, q_2 \geq 0,\ i = 1, 2\} \tag{7.9}$$

$$= \{\vec{v} : v_1 \geq 0,\quad v_1 \frac{\rho_0/\mu_0}{(C_1 - \rho_0)/\mu_1 + \rho_0/\mu_0} \leq v_2 \leq v_1 \frac{(C_2 - \rho_0)/\mu_2 + \rho_0/\mu_0}{\rho_0/\mu_0}\}, \tag{7.10}$$

which is independent of the parameter α. Hence, the workload process $\vec{\hat{V}}^{\pi(\alpha)}(t)$ is independent of α as well. The diffusion scaled number of users, $\vec{\hat{N}}^{k,\pi(\alpha)}(t)$, converges in distribution as $k \to \infty$ to some process $\vec{\hat{N}}^{\pi(\alpha)}(t)$, which does depend on α (this process is specified in Appendix 7.B).

Since the process of the total workload in a node does not depend on α, we are able to derive monotonicity results for the mean holding cost over the whole range of the parameter μ_0. We can express the scaled holding cost as follows:

$$
\begin{aligned}
&\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)}(t) \\
&= \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0} \cdot \hat{N}_0^{\pi(\alpha)}(t) + \sum_{i=1}^{2} c_i\mu_i \cdot (\frac{1}{\mu_0}\hat{N}_0^{\pi(\alpha)}(t) + \frac{1}{\mu_i}\hat{N}_i^{\pi(\alpha)}(t)) \\
&\overset{d}{=} \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0} \cdot \hat{N}_0^{\pi(\alpha)}(t) + \sum_{i=1}^{2} c_i\mu_i \hat{V}_i^{\pi(\alpha)}(t). \qquad (7.11)
\end{aligned}
$$

From Proposition 7.2.2 we know that $N_0^{\pi(\alpha,\vec{1})}(t)$ is decreasing in α, and hence $\hat{N}_0^{\pi(\alpha)}(t)$ is decreasing in α as well. Since $\hat{V}_i^{\pi(\alpha)}(t)$ is independent of α, and by taking expectations in (7.11), we obtain that if $c_1\mu_1 + c_2\mu_2 \le c_0\mu_0$ or $c_1\mu_1 + c_2\mu_2 \ge c_0\mu_0$, then $\mathbb{E}(\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)}(t))$ is non-increasing or non-decreasing in α, respectively.

When in addition we use the characterization of $\vec{\hat{N}}^{\pi(\alpha)}(t)$, we are able to derive a stronger monotonicity result. The proof may be found in Appendix 7.B.

Proposition 7.4.6. *Consider a linear network with fixed capacities C_1 and C_2. Assume that the inter-arrival times and service requirements are exponentially distributed, $\rho_i + \rho_0 = C_i$ for $i = 1, 2$, and that the conjecture in [68] is valid.*

- *If $c_1\mu_1 + c_2\mu_2 < c_0\mu_0$, then $\mathbb{E}(\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)}(t))$ is strictly decreasing in α.*
- *If $c_1\mu_1 + c_2\mu_2 = c_0\mu_0$, then $\mathbb{E}(\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)}(t))$ is constant in α.*
- *If $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, then $\mathbb{E}(\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)}(t))$ is strictly increasing in α.*

7.4.4 Numerical results

In this section we present numerical experiments to provide further insight into the performance of α-fair policies. We consider a two-node linear network where both nodes have unit capacity. We assume Poisson arrivals and exponentially distributed service requirements and fix $\mu_1 = 1, \mu_2 = 0.5$, $\rho_1 = \rho_2$ and $w_j = c_j = 1$, $j = 0, 1, 2$. Throughout this section, we use the notation $N^\pi := \sum_{j=0}^{2} N_j^\pi$.

In Figure 7.1 and Figure 7.2 (left) we let α vary on the horizontal axis and plot the corresponding total mean number of users for various values of μ_0. As expected from Corollary 7.4.5, we observe that for $\mu_0 \ge \mu_1 + \mu_2 = 1.5$ the total mean number of users is decreasing with respect to the value of α. When $\mu_0 < \mu_1 + \mu_2 = 1.5$, we observe that the total mean number of users is monotone (either decreasing or increasing) in α as well in the range $\alpha \in [1, \infty)$. However, when $\alpha \in (0, 1)$ and $\mu_0 < \mu_1 + \mu_2 = 1.5$, it is possible that the total mean number of users is not monotone in α. This fact may be explained as follows. Since $\mu_0 < \mu_1 + \mu_2 = 1.5$, it is attractive to give more preference to classes 1 and 2 when they are both present (hence less preference to class 0). This corresponds to a small value for α. However, an α-fair policy with a small α uses the available capacity less efficiently, see Proposition 7.2.2 (iv) and Lemma 7.4.1 (ii). These two opposite effects might cause the total mean number of users to not be monotone in α. Note that for

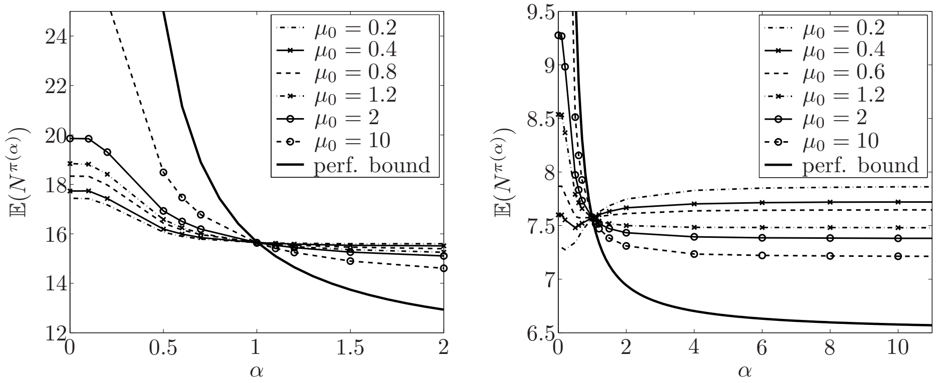


Figure 7.1: Total mean number of users under α-fair policies in a two-node linear network with $\mu_1 = 1$, $\mu_2 = 0.5$ and $\rho_0 = 0.7, \rho_1 = \rho_2 = 0.2$ (left), and $\rho_0 = 0.3, \rho_1 = \rho_2 = 0.5$ (right).

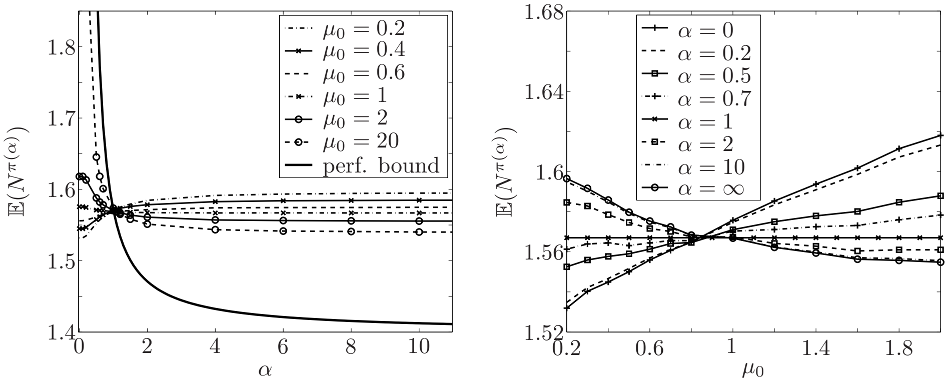


Figure 7.2: Total mean number of users under α-fair policies in a two-node linear network with $\mu_1 = 1$, $\mu_2 = 0.5$, $\rho_0 = 0.3, \rho_1 = \rho_2 = 0.2$.

the heavy-traffic regime as considered in Section 7.4.3, the workload in a node was independent of the parameter α and hence every value for α had the same efficiency. Therefore, there was no trade-off and we were able to prove in that setting the monotonicity result for $\mu_0 < \mu_1 + \mu_2$.

In Figure 7.2 (right) we let μ_0 vary on the horizontal axis and plot the corresponding total mean number of users for various values of α. We observe that the total mean number of users is mostly increasing in μ_0 when $\alpha < 1$ and decreasing in μ_0 when $\alpha > 1$, respectively. This can be explained as follows. First of all, if $\alpha = 1$, the policy reduces to PF. In the case of unit weights, PF is insensitive to the service requirement distributions apart from their respective means (see [94]). Hence, its total mean number of users is independent of the parameters μ_0, μ_1 and μ_2 for given values of ρ_0, ρ_1 and ρ_2. When $\alpha > 1$, from Lemma 7.4.1 (ii) we observe that class 0 is treated preferentially over classes 1 and 2 (compared to PF). Under an α-fair policy that gives preference to class 0, it is likely that the total mean number of users decreases when the class-0 users become smaller, i.e., when μ_0 increases, while $\mu_1, \mu_2, \rho_0, \rho_1$ and ρ_2 are kept fixed. Similarly, when $\alpha < 1$, classes 1 and 2 are treated preferentially over class 0 (compared to PF). When μ_0 becomes larger (while $\mu_1, \mu_2, \rho_0, \rho_1$ and ρ_2 are kept fixed), class-1 and 2 users become relatively larger. Under an α-fair policy that gives preference to classes 1 and 2, it is likely that the total mean number of users increases when μ_0 increases.

7.4.5 Time-scale separation

In [33] the authors introduce the so-called quasi-stationary and fluid-limit regimes (see also [74]). In these regimes, the flow dynamics of the various classes occur on separate time scales, which can greatly simplify the analysis. It was conjectured in [33] that these limiting regimes provide performance bounds. For the symmetric linear network with unit weights, Poisson arrivals and generally distributed service requirements, we refer to the quasi-stationary and fluid regimes when $\mu_0 \to \infty$ and $\mu_0 \to 0$, respectively, and keeping $\mu_1, \ldots, \mu_L$ and $\rho_0, \ldots, \rho_L$ fixed. From our simulation results for a linear network it seems that these limiting regimes can indeed be performance bounds, see Figure 7.2 (right). When $\alpha > 1$, the quasi-stationary regime ($\mu_0 \to \infty$) is a lower bound on the total mean number of users and the fluid regime ($\mu_0 \to 0$) an upper bound on the total mean number of users, and when $\alpha < 1$ vice versa. A similar observation was made in [74] for a DPS queue.

We develop here an approximate analysis of the quasi-stationary regime. The approximate formulae might be useful in assessing the performance of α-fair policies, since exact closed-form formulae are not available.

In the quasi-stationary regime, $\mu_0 \to \infty$, the dynamics of class 0 will 'average out' on the relevant time scale for class i, $i = 1, \ldots, L$. Hence, we can say that class 0 takes away a constant service rate ρ_0. Class i behaves as in a PS system with capacity $1 - \rho_0$, which implies that the number of class-i users in the system is geometrically distributed with mean $\frac{\rho_i}{1-\rho_0-\rho_i}$ [69]. Hence, $\lim_{\mu_0 \to \infty} \mathbb{E}(N_i^{\pi(\alpha,w)}) = \frac{\rho_i}{1-\rho_0-\rho_i}$, which is independent of α and $\frac{w_0}{w_i}$. Note that for $\alpha = 1$ this approximation is the correct expression, see (4.24).

The time scale of class 0 is infinitely faster than that of classes $1,\ldots,L$. Thus on the time scale of class 0, the dynamics of classes $1,\ldots,L$ almost vanish. It can be assumed that for a given number of class-i users, $i=1,\ldots,L$, class 0 will reach some sort of statistical equilibrium. We recall from (7.8) that $s_0^{\pi(\alpha,w)}(\vec{n}) = \frac{n_0}{n_0+c}$, with $c = c(n_1,\ldots,n_L) = (\sum_{i=1}^L \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha}$. Thus, given a population $\vec{n}$, class 0 behaves like a PS system with c permanent users. The mean number of users in such a system is $\frac{\rho_0}{1-\rho_0}(1+c)$. Unconditioning and noting that $N_i^{\pi(\alpha,w)}$ is in the limit geometrically distributed with mean $\frac{\rho_i}{1-\rho_0-\rho_i}$, $i=1,\ldots,L$, we get that approximately

$$\begin{aligned}
&\lim_{\mu_0\to\infty} \mathbb{E}(N_0^{\pi(\alpha,w)}) \\
&= \lim_{\mu_0\to\infty} \sum_{n_1,\ldots,n_L} \mathbb{E}(N_0^{\pi(\alpha,w)} | N_i^{\pi(\alpha,w)} = n_i, i\neq 0)\cdot \mathbb{P}(N_i^{\pi(\alpha,w)} = n_i, i\neq 0) \\
&= \lim_{\mu_0\to\infty} \sum_{n_1,\ldots,n_L} \frac{\rho_0}{1-\rho_0}\cdot\left(1+(\sum_{i=1}^L \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha}\right)\cdot \mathbb{P}(N_i^{\pi(\alpha,w)} = n_i, i\neq 0) \\
&\approx \frac{\rho_0}{1-\rho_0}\cdot\left(1+\left(\sum_{i=1}^L \frac{w_i}{w_0}(\frac{\rho_i}{1-\rho_0-\rho_i})^\alpha\right)^{1/\alpha}\right). \qquad (7.12)
\end{aligned}$$

We ignored here the non-linearity induced by the parameter α. We see that the performance of class 0 does depend on α and the weights w_i, and using similar arguments as in the proof of Lemma 7.4.1, it can be checked that the approximation for the mean number of class-0 users as given in (7.12) indeed decreases when α or $\frac{w_0}{w_i}$ increases (as was proved in Proposition 7.2.2). In addition, note that for $\alpha=1$ the approximation in (7.12) is the correct expression, see (4.23).

In Figures 7.1 and Figure 7.2 (left) we plotted the above obtained approximation for the total mean number of users against α (denoted in the figures by "perf. bound"). We observe that this approximation provides indeed an upper bound on the performance when $\alpha<1$, and a lower bound when $\alpha>1$. Even for moderate values of μ_0, the bound is quite tight and not off by more than 10% as long as the value of α is not too small or too large.

Unfortunately, it does not seem possible to derive an approximation for the fluid regime. When $\mu_0\to 0$, the dynamics of classes $1,\ldots,L$ 'average out' on the relevant time scale of class 0. Thus, class 0 sees a system with capacity $1-\max(\rho_1,\ldots,\rho_L)$. The time scales of classes $1,\ldots,L$ are infinitely faster than that of class 0, hence on the relevant time scale of classes $1,\ldots,L$, the dynamics of class 0 nearly vanish. Thus, given a certain number of class-0 users, class i obtains capacity $s_i^{\pi(\alpha,w)}(\vec{n}) = (\sum_{i=1}^L \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha}/(n_0+(\sum_{i=1}^L \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha})$, where n_0 can be considered fixed. From this equation we cannot approximate the behavior of classes $1,\ldots,L$ by any known queueing system unless $\alpha=1$.

7.5 Multi-class single-server system

In Sections 7.3 and 7.4 we have focused on a linear network. In this section we turn our attention to the multi-class single-server queue with time-varying capacity $C(t)$. There are K classes of users, where class-i users arrive according to a general arrival process with rate λ_i, and have generally distributed service requirements with mean $1/\mu_i$, $i = 1, \ldots, K$. Let $\rho = \sum_{i=1}^K \rho_i$, with $\rho_i = \lambda_i/\mu_i$. The inter-arrival times and the service requirements are mutually independent random variables. We consider allocation policies that are work-conserving, i.e., if $\sum_{i=1}^K n_i > 0$ then $\sum_{i=1}^K s_i(\vec{n}) = C(t)$, and if $n_i = 0$ then $s_i(\vec{n}) = 0$. The intra-class policy is FCFS.

In Section 7.5.1 we consider two weighted time-sharing policies and, using the general results from Section 7.2, we obtain monotonicity properties in the case of two classes of users. In Section 7.5.2 we derive a framework (similar to the one derived in Section 7.2) for a multi-class single-server system (with an arbitrary number of classes) under work-conserving policies.

7.5.1 GPS and DPS policies

The policies we are particularly interested in are GPS and DPS, two popular non-anticipating policies for a multi-class single-server system. Let $GPS(w)$ ($DPS(w)$) denote a GPS (DPS) policy that assigns weight w_j to class j, with $\sum_{j=1}^K w_j = 1$.

As described in Section 1.3.2, the GPS allocation is given by

$$s_i^{GPS(w)}(\vec{n}) = C(t)\frac{w_i}{\sum_{j=1}^K w_j \mathbf{1}_{(n_j>0)}}, \quad i = 1, \ldots, K,$$

for $\sum_{i=1}^K n_i > 0$. We take as intra-class policy in GPS the FCFS policy.

The DPS allocation is given by

$$s_i^{DPS(w)}(\vec{n}) = C(t)\frac{w_i n_i}{\sum_{j=1}^K w_j n_j}, \quad i = 1, \ldots, K,$$

for $\sum_{i=1}^K n_i > 0$. The allocated capacity to class i is shared equally among all class-i users, hence the intra-class policy in DPS is PS.

Assume the service requirements are exponentially distributed with $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_K\mu_K$. The $c\mu$-rule, which gives preemptive priority to the class with the highest $c_i\mu_i$, minimizes the mean holding cost among all non-anticipating policies, see Section 1.3.3. For both GPS and DPS, a class is given more preference when its weight is increased. Hence, it seems plausible that giving relatively more weight to classes with a high $c_i\mu_i$ will decrease the mean holding cost. For a single-server system with only two classes of users ($K = 2$) we can indeed prove this: Such a system is equivalent to a linear network with one node ($L = 1$). When $w_1 < \tilde{w}_1$, the policies $GPS(w)$ and $GPS(\tilde{w})$ ($DPS(w)$ and $DPS(\tilde{w})$) satisfy Property 7.3.1'. Hence, we can use the results of Section 7.2 to obtain monotonicity results.

Proposition 7.5.1. *Consider a single-server system with two classes of users and time-varying capacity. Let $w_1 < \tilde{w}_1$. Assume $W_1^{\pi}(0) \geq W_1^{\tilde{\pi}}(0)$, $W_2^{\pi}(0) \leq W_2^{\tilde{\pi}}(0)$*

and $W_1^{\pi}(0)+W_2^{\pi}(0) = W_1^{\tilde{\pi}}(0)+W_2^{\tilde{\pi}}(0)$*, where either* $\pi = GPS(w)$ *and* $\tilde{\pi} = GPS(\tilde{w})$*, or* $\pi = DPS(w)$ *and* $\tilde{\pi} = DPS(\tilde{w})$*. We consider the same realizations of the arrival processes and service requirements for both processes.*
For generally distributed service requirements it holds that

$$W_1^{GPS(w)}(t) \geq W_1^{GPS(\tilde{w})}(t) \quad \textit{and} \quad N_1^{GPS(w)}(t) \geq N_1^{GPS(\tilde{w})}(t), \quad \forall\ t \geq 0. \tag{7.13}$$

The opposite inequalities hold for class 2.
For exponentially distributed service requirements it holds that

$$\begin{aligned} \{W_1^{DPS(w)}(t)\}_{t\geq 0} &\geq_{st} \{W_1^{DPS(\tilde{w})}(t)\}_{t\geq 0}, \\ \{N_1^{DPS(w)}(t)\}_{t\geq 0} &\geq_{st} \{N_1^{DPS(\tilde{w})}(t)\}_{t\geq 0}. \end{aligned} \tag{7.14}$$

The opposite inequalities hold for class 2.
If the service requirements are exponentially distributed with $c_1\mu_1 \geq c_2\mu_2$ *and the system can be made stable, then*

$$\sum_{i=1}^{2} c_i \mathbb{E}(N_i^{GPS(w)}(t)) \geq \sum_{i=1}^{2} c_i \mathbb{E}(N_i^{GPS(\tilde{w})}(t)), \quad \forall\ t \geq 0, \tag{7.15}$$

$$\sum_{i=1}^{2} c_i \mathbb{E}(N_i^{DPS(w)}(t)) \geq \sum_{i=1}^{2} c_i \mathbb{E}(N_i^{DPS(\tilde{w})}(t)), \quad \forall\ t \geq 0. \tag{7.16}$$

Proof: Since the respective pair of policies satisfy Property 7.3.1', equation (7.13) follows directly from Propositions 7.2.2, and equations (7.15) and (7.16) follow directly from Proposition 7.2.5. For exponentially distributed service requirements, the stochastic behavior is independent of the used intra-class policy, see Remark 7.1.1. For DPS we consider exponentially distributed service requirements. Hence, the sample-path comparison in Proposition 7.2.2 obtained for FCFS, allows us to obtain the stochastic comparison result in (7.14) for DPS when the intra-class policy is PS. □

Inequalities (7.13) and (7.14) are rather natural, but to the best of our knowledge have not been obtained previously. In particular, the comparison results from [85] and [92] do not allow for such a comparison, as explained later in Remark 7.5.10. The result for GPS is particularly interesting. GPS is used to model the queueing delay experienced by packets in packet networks. An important body of research on GPS is devoted to the characterization of the workload when there are two classes of users, see for example [35, 109].

Inequalities (7.15) and (7.16) show that for two classes, the mean holding cost under DPS or GPS is monotone in the whole range $w_1 \in [0, \infty)$, where one extreme corresponds to giving preemptive priority to class 2 ($w_1 = 0$) and the other extreme to preemptive priority to class 1 ($w_1 = 1$). To the best of our knowledge, this kind of monotonicity result is new for GPS. In the case of a two-class single-server DPS-system *with* fixed capacity and Poisson arrivals, this result can also be obtained from the analysis in [51] (see [12] for more details).

For an arbitrary number of classes little is known on monotonicity results for GPS and DPS. As mentioned before, motivated by the optimality of the $c\mu$-rule, one would expect that giving relatively more weight to classes with a high $c_i\mu_i$, will decrease the mean holding cost. One of the most relevant results is obtained in [75]. The authors consider a single server with fixed capacity, Poisson arrivals and exponentially distributed service requirements with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Using the results of [51] they prove that if $\tilde{w}_1 \geq \tilde{w}_2 \geq \cdots \geq \tilde{w}_K$, then $\mathbb{E}(\sum_{i=1}^K N_i^{PS}) \geq \mathbb{E}(\sum_{i=1}^K N_i^{DPS(\tilde{w})})$. Note that PS is equivalent to a DPS policy with weights $w_i = 1$, for all i. In general, we expect the following results to hold (a similar conjecture for the steady-state distribution of DPS has been made in [75]).

Conjecture 7.5.2. *Consider a single-server system with K classes of users. Assume the service requirements are exponentially distributed with $c_1\mu_1 \geq \ldots \geq c_K\mu_K$. If $w_i/w_{i+1} \leq \tilde{w}_i/\tilde{w}_{i+1}$, for all $i = 1, \ldots, K-1$, then*

$$\sum_{j=1}^K c_j \mathbb{E}(N_j^{GPS(w)}(t)) \geq \sum_{j=1}^K c_j \mathbb{E}(N_j^{GPS(\tilde{w})}(t)), \qquad \textit{for all} \ \ t \geq 0,$$

and

$$\sum_{j=1}^K c_j \mathbb{E}(N_j^{DPS(w)}(t)) \geq \sum_{j=1}^K c_j \mathbb{E}(N_j^{DPS(\tilde{w})}(t)), \qquad \textit{for all} \ \ t \geq 0.$$

Note that the conjecture for DPS is valid in the heavy-traffic limit: In Proposition 2.6.4 it was proved that the scaled holding cost under DPS are monotone in the relative weights for a heavily-loaded system with phase-type distributed service requirements. In the next example we perform numerical experiments that support Conjecture 7.5.2 for a single-server system with three classes.

Example 7.5.3 (Numerical experiments for GPS and DPS). We consider a single server with fixed unit capacity, Poisson arrivals, and three classes of users with exponentially distributed service requirements. We consider both GPS and DPS with weights $w_i(r) = \Omega(r) \cdot r^{K-i}$, $r \geq 1$, and $\Omega(r) = 1/(\sum_{i=0}^{K-1} r^i)$ a normalization constant. Note that $w_i/w_{i+1} = r$, $i = 1, \ldots, K$. Hence, as the parameter r increases, class i obtains relatively a larger weight compared to class $i+1$. We choose $\mu_1 = 2$, $\mu_2 = 1$ and $\mu_3 = 0.5$, hence we expect that the functions $\mathbb{E}(\sum_{i=1}^K N_i^{GPS(w(r))})$ and $\mathbb{E}(\sum_{i=1}^K N_i^{DPS(w(r))})$ are decreasing in r. When $r \to \infty$, both $GPS(r)$ and $DPS(r)$ become a priority rule that gives preemptive priority to class 1, and if class 1 is empty, serves class 2. Since $\mu_1 > \mu_2 > \mu_3$, this policy minimizes the total mean number of users present in the system (follows from the optimality of the $c\mu$-rule).

For GPS with weights $w_i(r)$ we simulated the system and Figure 7.3 (left) plots the total mean number of users as a function of the parameter r. We observe that the total mean number of users indeed reduces as r increases.

In Figure 7.3 (right) we plot the total mean number of users under $DPS(r)$ as a function of the parameter r. The total mean number of users was obtained by solving a system of linear equations as given in [51]. When $r = 1$, the policy reduces

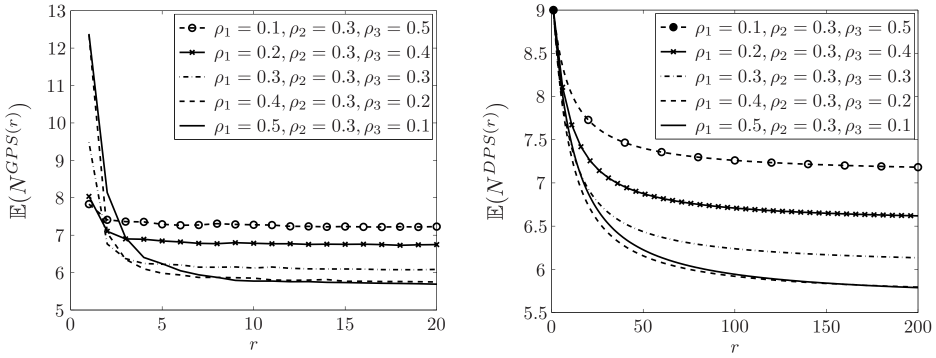


Figure 7.3: Total mean number of users under GPS policies (left) and under DPS policies (right), with $\mu_1 = 2, \mu_2 = 1, \mu_3 = 0.5$.

to PS, hence $\mathbb{E}(\sum_{i=1}^K N_i^{DPS(w(1))}) = \mathbb{E}(\sum_{i=1}^K N_i^{PS}) = \frac{\rho_1+\rho_2+\rho_3}{1-(\rho_1+\rho_2+\rho_3)}$. We observe that the mean total number of users is again decreasing in r.

For a single server with more than two classes, the framework and results as developed in Section 7.2 and in particular Property 7.2.1 are not applicable. Therefore, in Section 7.5.2 we develop a similar analysis as in Section 7.2, but now for a single-server system with an arbitrary number of classes. Unfortunately, this sample-path framework does not allow a full comparison of either two DPS or two GPS policies for more than two classes. This will be explained as well in the next section.

7.5.2 Comparison of policies

In this section we derive comparison results for a single-server system with K classes of users, $K \geq 2$, similar to the ones obtained in Section 7.2. We focus on Pareto-efficient policies, which in a single-server scenario are equivalent to work-conserving policies. The following property states sufficient conditions on two policies in order to compare them sample-path wise.

Property 7.5.4. *Let π and $\tilde{\pi}$ be two work-conserving policies such that for any $k = 1, \ldots, K-1$, we have that*

$$\sum_{i=1}^{k} s_i^{\pi}(\vec{n}^{\pi}) \leq \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}), \tag{7.17}$$

for all states $\vec{n}^{\tilde{\pi}}$ and $\vec{n}^{\pi}$ that satisfy the following three conditions:

- *$n_1^{\pi} \geq n_1^{\tilde{\pi}}$, $n_k^{\pi} \leq n_k^{\tilde{\pi}}$, $n_{k+1}^{\pi} \geq n_{k+1}^{\tilde{\pi}}$ and $n_K^{\pi} \leq n_K^{\tilde{\pi}}$.*
- *If for an $m \geq 0$ it holds that $n_{k-i}^{\tilde{\pi}} = 0$, for all $i = 0, \ldots, m$, then in addition $n_{k-i-1}^{\pi} \leq n_{k-i-1}^{\tilde{\pi}}$, for all $i = 0, \ldots, m$.*

- *If for an $m \geq 1$ it holds that $n^{\pi}_{k+i} = 0$, for all $i = 1, \ldots, m$, then in addition $n^{\pi}_{k+i+1} \geq n^{\tilde{\pi}}_{k+i+1}$, for all $i = 1, \ldots, m$.*

Equation (7.17) represents a weak notion of priority, with strict priority as a special case. When two policies satisfy Property 7.5.4, we can derive the following sample-path comparison result.

Proposition 7.5.5. *Let π and $\tilde{\pi}$ be two policies that satisfy Property 7.5.4 and consider the same realizations of the arrival processes and service requirements. If $\sum_{i=1}^{m} W_i^{\pi}(0) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(0), m = 1, \ldots, K-1$, and $\sum_{i=1}^{K} W_i^{\pi}(0) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(0)$, then for all $t \geq 0$*

$$\sum_{i=1}^{m} \left(S_i^{\pi}(t) - W_i^{\pi}(0)\right) \leq \sum_{i=1}^{m} \left(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)\right), \quad m = 1, \ldots, K. \tag{7.18}$$

For $m = K$, (7.18) holds with equality.

In particular we have

$$N_1^{\pi}(t) \geq N_1^{\tilde{\pi}}(t), \quad N_K^{\pi}(t) \leq N_K^{\tilde{\pi}}(t), \tag{7.19}$$

and

$$\sum_{i=1}^{m} W_i^{\pi}(t) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(t), \quad m = 1, \ldots, K. \tag{7.20}$$

For $m = K$, (7.20) holds with equality.

Proof: Equation (7.20) follows from (7.1) and (7.18). The first relation in (7.19) follows from (7.20) with $m = 1$, since the intra-class policy is FCFS and the k-th most recently arrived class-1 user before time t has the same (original) service requirement under both policies. Similarly, the second relation in (7.19) follows from (7.20) with $m = K - 1$ and $\sum_{i=1}^{K} W_i^{\pi}(t) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(t)$. Therefore, it suffices to prove (7.18).

The policies are work-conserving, so $\sum_{i=1}^{K} W_i^{\pi}(0) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(0)$ gives that $\sum_{i=1}^{K} S_i^{\pi}(t) = \sum_{i=1}^{K} S_i^{\tilde{\pi}}(t)$, and hence (7.18) holds with equality for $m = K$. Equation (7.18) for $m < K$ is proved by contradiction. Let t be the first time epoch at which (7.18) is violated for some k, $1 \leq k \leq K-1$. So we have $\sum_{i=1}^{k}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$ and $\sum_{i=1}^{k} s_i^{\pi}(\vec{N}^{\pi}(t^+)) > \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, but $\sum_{i=1}^{m}(S_i^{\pi}(t) - W_i^{\pi}(0)) \leq \sum_{i=1}^{m}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$, for $m \neq k$. Hence,

$$\begin{aligned} S_1^{\pi}(t) - W_1^{\pi}(0) &\leq S_1^{\tilde{\pi}}(t) - W_1^{\tilde{\pi}}(0), & S_k^{\pi}(t) - W_k^{\pi}(0) &\geq S_k^{\tilde{\pi}}(t) - W_k^{\tilde{\pi}}(0), \\ S_{k+1}^{\pi}(t) - W_{k+1}^{\pi}(0) &\leq S_{k+1}^{\tilde{\pi}}(t) - W_{k+1}^{\tilde{\pi}}(0), & S_K^{\pi}(t) - W_K^{\pi}(0) &\geq S_K^{\tilde{\pi}}(t) - W_K^{\tilde{\pi}}(0) \end{aligned}$$

Together with (7.1), we obtain $W_1^{\pi}(t) \geq W_1^{\tilde{\pi}}(t), W_k^{\pi}(t) \leq W_k^{\tilde{\pi}}(t), W_{k+1}^{\pi}(t) \geq W_{k+1}^{\tilde{\pi}}(t)$ and $W_K^{\pi}(t) \leq W_K^{\tilde{\pi}}(t)$. Since the k-th class-j user under both policies has the same (original) service requirement and the intra-class policy is FCFS, we have as well

$$N_1^{\pi}(t) \geq N_1^{\tilde{\pi}}(t), \quad N_k^{\pi}(t) \leq N_k^{\tilde{\pi}}(t), \quad N_{k+1}^{\pi}(t) \geq N_{k+1}^{\tilde{\pi}}(t) \text{ and } N_K^{\pi}(t) \leq N_K^{\tilde{\pi}}(t).$$

Since $\{N_i(t)\}_{t\geq 0}$ is a piece-wise constant process and is right-continuous, we have as well $N_1^{\pi}(t^+) \geq N_1^{\tilde{\pi}}(t^+), N_k^{\pi}(t^+) \leq N_k^{\tilde{\pi}}(t^+), N_{k+1}^{\pi}(t^+) \geq N_{k+1}^{\tilde{\pi}}(t^+)$ and $N_K^{\tilde{\pi}}(t^+) \leq N_K^{\tilde{\pi}}(t^+)$.

Note that if $N_k^{\tilde{\pi}}(t^+) = 0$, then $S_k^{\pi}(t) - W_k^{\pi}(0) = S_k^{\tilde{\pi}}(t) - W_k^{\tilde{\pi}}(0)$ and hence $\sum_{i=1}^{k-1}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k-1}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$. So $S_{k-1}^{\pi}(t) - W_{k-1}^{\pi}(0) \geq S_{k-1}^{\tilde{\pi}}(t) - W_{k-1}^{\tilde{\pi}}(0)$ and by (7.1) we obtain $N_{k-1}^{\pi}(t^+) \leq N_{k-1}^{\tilde{\pi}}(t^+)$. Now if also $N_{k-1}^{\tilde{\pi}}(t^+) = 0$, then we obtain in the same way that $N_{k-2}^{\pi}(t^+) \leq N_{k-2}^{\tilde{\pi}}(t^+)$, etc. Also note that if $N_{k+1}^{\pi}(t^+) = 0$, then $S_{k+1}^{\pi}(t) - W_{k+1}^{\pi}(0) = S_{k+1}^{\tilde{\pi}}(t) - W_{k+1}^{\tilde{\pi}}(0)$ and hence $\sum_{i=1}^{k+1}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k+1}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$. So $S_{k+2}^{\pi}(t) - W_{k+2}^{\pi}(0) \leq S_{k+2}^{\tilde{\pi}}(t) - W_{k+2}^{\tilde{\pi}}(0)$ and by (7.1) we obtain $N_{k+2}^{\pi}(t^+) \geq N_{k+2}^{\tilde{\pi}}(t^+)$. Now if also $N_{k+2}^{\pi}(t^+) = 0$, then we obtain in the same way that $N_{k+3}^{\pi}(t^+) \geq N_{k+3}^{\tilde{\pi}}(t^+)$, etc.

So at time t^+ we are in states $\vec{N}^{\pi}(t^+)$ and $\vec{N}^{\tilde{\pi}}(t^+)$ that satisfy Property 7.5.4 and hence $\sum_{i=1}^{k} s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$. This contradicts the initial assumption. □

Every work-conserving policy gives a stable system whenever possible. However, for a subset of the classes, the stability conditions can still depend on the policy being employed. We have the following result:

Corollary 7.5.6. *Assume π and $\tilde{\pi}$ satisfy Property 7.5.4. If classes $1, 2, \ldots, m$ are stable under policy π, then these classes are stable under policy $\tilde{\pi}$ as well, in the sense that the system is empty under policy $\tilde{\pi}$ whenever it is empty under policy π.*

Proof: If $\sum_{i=1}^{m} W_i^{\pi}(t) = 0$, then we obtain from Proposition 7.5.5 that $\sum_{i=1}^{m} W_i^{\tilde{\pi}}(t) = 0$. □

The following proposition states the analogous version of Proposition 7.2.5.

Proposition 7.5.7. *Assume the service requirements are exponentially distributed. Let π and $\tilde{\pi}$ be two policies that satisfy Property 7.5.4 and assume the system is stable. If $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$, then*

$$\sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\pi}(t)) \geq \sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\tilde{\pi}}(t)), \quad \forall\, t \geq 0.$$

Proof: Assume at time $t = 0$ the conditions as stated in Proposition 7.5.5 are satisfied. From Proposition 7.5.5 we obtain $\sum_{i=1}^{m} W_i^{\pi}(t) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(t)$. Since π is non-anticipating and the service requirements are exponentially distributed, we have

$$\sum_{i=1}^{m} \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t)) \geq \sum_{i=1}^{m} \frac{1}{\mu_i}\mathbb{E}(N_i^{\tilde{\pi}}(t)) \tag{7.21}$$

for $m \leq K$. Define $P_m^{\pi}(t) := \sum_{i=1}^{m} \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t))$. So $P_m^{\pi}(t) \geq P_m^{\tilde{\pi}}(t),\ m = 1, \ldots, K,$

and hence

$$\begin{aligned}\sum_{i=1}^{K} c_i\mathbb{E}(N_i^{\pi}(t)) &= (c_1\mu_1 - c_2\mu_2)P_1^{\pi}(t) + (c_2\mu_2 - c_3\mu_3)P_2^{\pi}(t) + \ldots + c_K\mu_K P_K^{\pi}(t)\\ &\geq (c_1\mu_1 - c_2\mu_2)P_1^{\tilde{\pi}}(t) + (c_2\mu_2 - c_3\mu_3)P_2^{\tilde{\pi}}(t) + \ldots + c_K\mu_K P_K^{\tilde{\pi}}(t)\\ &= \sum_{i=1}^{K} c_i\mathbb{E}(N_i^{\tilde{\pi}}(t)),\end{aligned}$$

where we used that $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$. □

Example 7.5.8 (**Optimality of the** $c\mu$**-rule).** As mentioned before, for exponentially distributed service requirements, the $c\mu$-rule, i.e., the policy that gives preemptive priority to the class i with the maximum $c_i\mu_i$, minimizes the mean holding cost among all non-anticipating policies. For a time-varying multi-class single-server system, this was shown in [102]. In fact this also follows from Proposition 7.5.7. Assume $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$. Denote the $c\mu$-rule by $\tilde{\pi}$, and consider an arbitrary non-anticipating policy π. Whenever $\sum_{i=1}^{k} n_i^{\tilde{\pi}} > 0$ we have that $\sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = C(t)$ and hence (7.17) is satisfied for these states. Now assume $\sum_{i=1}^{k} n_i^{\tilde{\pi}} = 0$. Since $n_i^{\tilde{\pi}} = 0$ for all $i \leq k$, the corresponding states $\vec{n}^{\pi}$ we have to consider in Property 7.5.4 should satisfy $n_i^{\pi} \leq n_i^{\tilde{\pi}} = 0$ for all $i \leq k$, that is $\sum_{i=1}^{k} n_i^{\pi} = 0$ as well. But then (7.17) is by definition satisfied. Hence Property 7.5.4 is satisfied and the optimality of the $c\mu$-rule follows now from Proposition 7.5.7.

Proposition 7.5.7, combined with Property 7.5.4 gives sufficient conditions in order to compare the mean holding cost under two policies. When $K = 2$, Property 7.5.4 reduces to the rather natural condition $s_1^{\pi}(\vec{n}) \leq s_1^{\tilde{\pi}}(\vec{n})$ for all states $\vec{n}$. This is for example satisfied by either two DPS policies or two GPS policies when $w_1 \leq \tilde{w}_1$. Unfortunately, for more than two classes Property 7.5.4 fails to hold for any two DPS policies. For a GPS system, Property 7.5.4 is satisfied under more stringent conditions than the ones stated in Conjecture 7.5.2. For example, for the case of three classes it can be checked that for two GPS policies, GPS(w) and GPS($\tilde{w}$), Property 7.5.4 is equivalent to

$$\frac{w_1}{w_1+w_2} \leq \tilde{w}_1, \qquad \frac{w_1}{w_1+w_3} \leq \frac{\tilde{w}_1}{\tilde{w}_1+\tilde{w}_3} \qquad \text{and} \qquad w_3 \geq \frac{\tilde{w}_3}{\tilde{w}_2+\tilde{w}_3}. \tag{7.22}$$

Hence, (7.22) is a sufficient condition to compare the mean holding cost under GPS(w) and GPS($\tilde{w}$). If we choose as weights $w_i(r) = \Omega(r) \cdot r^{K-i}$, $r > 1$ (as considered in Example 7.5.3), equation (7.22) is equivalent to $1 \leq r$ and $\tilde{r} \geq r + r^2$. We would expect the comparison result already to hold for all $\tilde{r} \geq r$, so this shows that there is still a gap of length r^2. For an arbitrary number of classes, the sufficient conditions in order for Conjecture 7.5.2 to hold for GPS can be obtained as well, however, the derivations become very cumbersome.

In this section we used sample-path inequalities as given in (7.18) in order to compare the mean holding cost under two different policies. Property 7.5.4 is a sufficient (but not necessary) condition for these sample-path inequalities to hold. For

DPS and GPS, this property is not (always) satisfied. In fact, the counterexample below illustrates for the case of three classes that the sample-path inequalities (7.18) do not need to hold for either two DPS policies or two GPS policies that satisfy the conditions of Conjecture 7.5.2. This indicates that for more than two classes, Conjecture 7.5.2 cannot be fully proved using such sample-path arguments and requires a different kind of approach.

Example 7.5.9 (Counterexamples for DPS and GPS). We give a counterexample for the inequality (7.18) that is valid for both DPS and GPS. Consider a system with three classes, and consider the two policies with weight vectors $w = (2, 1, 1)$ and $\tilde{w} = (\infty, 1, 1)$, respectively. It is easy to verify that the vectors w and $\tilde{w}$ satisfy the condition of Conjecture 7.5.2. Assume that at time $t = 0$ there is one user in every class, that is, $N^{\pi}(0) = N^{\tilde{\pi}}(0) = (1, 1, 1)$ and their service requirements are respectively 4, 10 and 1 under both policies π and $\tilde{\pi}$. At time $t = 6$ a class-3 user arrives with a strictly positive service requirement. Let us analyze the evolution under both policies over time:

- Policy π: In the interval $[0, 4)$ all users share the capacity according to the weights. At time $t = 4$ the class-3 user departs the system and the remaining service requirements of the class-1 and the class-2 user are 2 and 9, respectively. In the interval $[4, 6)$ the class-1 and class-2 users will share the capacity according to their weights, thus at time $t = 6$ the remaining service requirements of the class-1 and class-2 users are $\frac{2}{3}$ and $\frac{25}{3}$, respectively. It follows that $S_1^{\pi}(6) + S_2^{\pi}(6) = 4 + 10 - \frac{2}{3} - \frac{25}{3} = 5$.

- Policy $\tilde{\pi}$: In the interval $[0, 4)$ only class 1 will be served and it departs at time $t = 4$. In the interval $[4, 6)$ the class-2 and class-3 users will equally share the capacity. At time $t = 6$ the class-3 user departs and the class-2 user has a remaining service requirement of 9. It follows that $S_1^{\tilde{\pi}}(6) + S_2^{\tilde{\pi}}(6) = 4 + 10 - 9 = 5$.

Due to the new arrival at $t = 6$ it follows that $s_1^{\pi}(\vec{N}^{\pi}(6^+)) + s_2^{\pi}(\vec{N}^{\pi}(6^+)) = \frac{3}{4}$ whereas $s_1^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(6^+)) + s_2^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(6^+)) = \frac{1}{2}$. This together with the fact that $S_1^{\pi}(6) + S_2^{\pi}(6) = S_1^{\tilde{\pi}}(6) + S_2^{\tilde{\pi}}(6)$ implies that $S_1^{\pi}(6^+) + S_2^{\pi}(6^+) > S_1^{\tilde{\pi}}(6^+) + S_2^{\tilde{\pi}}(6^+)$, which contradicts (7.18) for $m = 2$.

In the following remark we explain that Conjecture 7.5.2 does not follow either from results in [85, 92] and hence that a novel approach is needed.

Remark 7.5.10. Assume $\{\vec{N}^{\pi}(t)\}_{t\geq 0}$ and $\{\vec{N}^{\tilde{\pi}}(t)\}_{t\geq 0}$ are two continuous-time Markov processes. From Remark 7.2.6 we readily see that the conditions on the policies π and $\tilde{\pi}$ in order to obtain $\{\sum_{i=1}^{K} N_i^{\pi}(t)\}_{t\geq 0} \geq_{st} \{\sum_{i=1}^{K} N_i^{\tilde{\pi}}(t)\}_{t\geq 0}$ for any initial states with $\sum_{i=1}^{K} N_i^{\pi}(0) \geq_{st} \sum_{i=1}^{K} N_i^{\tilde{\pi}}(0)$, are only satisfied when $\mu_i = \mu$ for all i. Consider for example the two states $\vec{n}^{\pi} = \vec{e}_k$ and $\vec{n}^{\tilde{\pi}} = \vec{e}_j$, where $\vec{e}_j$ denotes the j-th unit vector. Then the condition as stated in Remark 7.2.6 becomes $\sum_{i=1}^{K} \mu_i s_i^{\pi}(\vec{n}^{\pi}) = \mu_k \leq \sum_{i=1}^{K} \mu_i s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = \mu_j$, see also [85, 92]. However, for the states $\vec{n}^{\pi} = \vec{e}_j$ and $\vec{n}^{\tilde{\pi}} = \vec{e}_k$ we obtain similarly that we need $\mu_j \leq \mu_k$. So only when $\mu_i = \mu$ for all i, the conditions are satisfied, but this is not very interesting.

The necessary and sufficient conditions in order to obtain a similar comparison result as in Proposition 7.5.5, i.e., $\{N_1^{\pi}(t)\}_{t\geq 0} \geq_{st} \{N_1^{\tilde{\pi}}(t)\}_{t\geq 0}$ and $\{N_K^{\pi}(t)\}_{t\geq 0} \leq_{st} \{N_K^{\tilde{\pi}}(t)\}_{t\geq 0}$ given that $N_1^{\pi}(0) \geq N_1^{\tilde{\pi}}(0)$ and $N_K^{\pi}(0) \leq N_K^{\tilde{\pi}}(0)$, are

$$s_1^{\pi}(\vec{n}^{\pi}) \leq s_1^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \quad \text{for all} \quad n_1^{\pi} = n_1^{\tilde{\pi}} \text{ and } n_K^{\pi} \leq n_K^{\tilde{\pi}}, \tag{7.23}$$

$$s_K^{\pi}(\vec{n}^{\pi}) \geq s_K^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \quad \text{for all} \quad n_1^{\pi} \geq n_1^{\tilde{\pi}} \text{ and } n_K^{\pi} = n_K^{\tilde{\pi}}, \tag{7.24}$$

see [85, 92]. In a queueing context this can only be satisfied when policy $\tilde{\pi}$ gives preemptive priority to class 1 (see equation (7.23) with $n_2^{\tilde{\pi}} = \ldots = n_K^{\pi} = 0$) and policy π gives preemptive priority to class K (see equation (7.24) with $n_1^{\tilde{\pi}} = \ldots = n_{K-1}^{\tilde{\pi}} = 0$). In particular, for any two GPS policies or two DPS policies (with non-degenerate weights) the inequalities (7.23) and (7.24) do not hold.

7.6 Concluding remarks

In this chapter we have studied multi-class queueing systems and, using sample-path arguments, we have obtained comparison results for the performance under two different policies in terms of stability and mean holding cost. The results could naturally be applied to a linear network and a two-class single-server system. It might be interesting to consider different types of networks, like a star or grid network, and use the same approach in order to compare the performance of different policies.

For the linear network we proved monotonicity results for the mean holding cost under α-fair policies with respect to the parameter α and the relative weights. In the numerical section, we observed an additional monotonicity property: The total mean number of users shows monotone behavior in μ_0 for given load ρ_0, when the other parameters are kept fixed. There is no hope that this latter property can be proved using sample-path arguments, since this requires the same realizations for the service requirements. When we compare the two stochastic processes for different values of μ_0, this can no longer be done.

For the multi-class single-server system with exponential service requirements it is reasonable to expect that for weighted time-sharing policies like DPS and GPS, monotonicity results for the mean holding cost hold under natural conditions on the weights. We were able to prove this for some special cases using a sample-path argument. The other cases remain as a challenging topic for further research.

Appendix

7.A Proof of Lemma 7.4.1

For a given state $\vec{n}$, the α-fair allocation $\vec{s}^{\pi(\alpha,w)}(\vec{n})$ is the solution of the optimization problem (7.7). (For ease of notation, we drop the dependence on t in $C_i(t)$.)

Obviously, $s_i^{\pi(\alpha,w)}(\vec{n})$ equals $C_i - s_0^{\pi(\alpha,w)}(\vec{n})$ when $n_i > 0$, and equals 0 otherwise. Without loss of generality, we assume that $n_i > 0$, and hence, we can set $s_i = C_i - s_0$ in (7.7), for all i. It can be checked that the objective function in (7.7) expressed in terms of s_0 is concave. Taking the derivative with respect to s_0 and setting it equal to zero, we obtain that $s_0^{\pi(\alpha,w)}(\vec{n})$ satisfies

$$w_0 \cdot n_0^\alpha \cdot (s_0^{\pi(\alpha,w)}(\vec{n}))^{-\alpha} = \sum_{i=1}^{L} w_i \cdot n_i^\alpha \cdot (C_i - s_0^{\pi(\alpha,w)}(\vec{n}))^{-\alpha},$$

or equivalently

$$1 = \sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\alpha,w)}(\vec{n})}{C_i - s_0^{\pi(\alpha,w)}(\vec{n})} \Big)^\alpha. \tag{7.25}$$

The function $\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0}{C_i - s_0} \Big)^\alpha$ is non-decreasing in s_0. Hence, when either n_i or $\frac{w_i}{w_0}$ increases, by (7.25) the corresponding value of s_0 must decrease. Statements (i) and (iii) follow now immediately. Statement (ii) deserves some more elaboration. Let $\beta < \gamma$ and define $r := \gamma/\beta > 1$. By (7.25) we have

$$\begin{aligned} 1 &= \Big(\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i - s_0^{\pi(\beta,w)}(\vec{n})} \Big)^\beta \Big)^{\frac{1}{\beta}} = \Big(\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i - s_0^{\pi(\beta,w)}(\vec{n})} \Big)^\beta \Big)^{\frac{r}{r\beta}} \\ &\geq \Big(\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i - s_0^{\pi(\beta,w)}(\vec{n})} \Big)^{r\beta} \Big)^{\frac{1}{r\beta}} = \Big(\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i - s_0^{\pi(\beta,w)}(\vec{n})} \Big)^{\gamma} \Big)^{\frac{1}{\gamma}}. \end{aligned}$$

By (7.25) we also have that $\Big(\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\gamma,w)}(\vec{n})}{C_i - s_0^{\pi(\gamma,w)}(\vec{n})} \Big)^\gamma \Big)^{\frac{1}{\gamma}} = 1$. Together with the above, this implies $\sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i - s_0^{\pi(\beta,w)}(\vec{n})} \Big)^\gamma \leq 1 = \sum_{i=1}^{L} \frac{w_i}{w_0} \Big(\frac{n_i}{n_0} \frac{s_0^{\pi(\gamma,w)}(\vec{n})}{C_i - s_0^{\pi(\gamma,w)}(\vec{n})} \Big)^\gamma$ and hence $s_0^{\pi(\beta,w)}(\vec{n}) \leq s_0^{\pi(\gamma,w)}(\vec{n})$. □

7.B Proof of Proposition 7.4.6

By the conjecture of [68], the diffusion-scaled workload in node i converges in distribution to $\hat{V}_i^{\pi(\alpha)}(t)$, which is independent of α. In addition, it is stated that the diffusion-scaled number of users, $\hat{\vec{N}}^{k,\pi(\alpha)}(t)$, converges in distribution to $\hat{\vec{N}}^{\pi(\alpha)}(t)$. The latter process can be written as

$$\hat{N}_0^{\pi(\alpha)}(t) \overset{d}{=} \rho_0 (q_1(\alpha,t) + q_2(\alpha,t))^{\frac{1}{\alpha}} \quad \text{and} \quad \hat{N}_i^{\pi(\alpha)}(t) \overset{d}{=} \rho_i q_i(\alpha,t)^{\frac{1}{\alpha}}, \quad i = 1,2. \tag{7.26}$$

where $(q_1(\alpha,t), q_2(\alpha,t)) \in \mathbb{R}_+^2$, [68]. In the remainder of this proof we fix t and drop the dependence on t. The scaled workload in a node is independent of α, hence we can assume that there is a $\hat{v}_i$ with $\hat{V}_i^{\pi(\alpha)} = \hat{v}_i$, for all $\alpha > 0$, $i = 1, 2$. By (7.26), the definition of $\hat{V}^{\pi(\alpha)}$, and $\rho_0 + \rho_i = C_i$, we have

$$(C_i - \rho_0)\frac{1}{\mu_i} q_i(\alpha)^{1/\alpha} + \rho_0 \frac{1}{\mu_0}(q_1(\alpha) + q_2(\alpha))^{1/\alpha} \overset{d}{=} \hat{v}_i. \tag{7.27}$$

Together with (7.26), the diffusion-scaled holding cost can now be written as

$$\begin{aligned}\sum_{j=0}^{2} c_j \hat{N}_j^{\pi(\alpha)} &\overset{d}{=} c_0\rho_0(q_1(\alpha)+q_2(\alpha))^{\frac{1}{\alpha}} + c_1(C_1-\rho_0)q_1(\alpha)^{\frac{1}{\alpha}} + c_2(C_2-\rho_0)q_2(\alpha)^{\frac{1}{\alpha}} \\ &= c_1\Big((C_1-\rho_0)q_1(\alpha)^{1/\alpha} + \rho_0\frac{\mu_1}{\mu_0}(q_1(\alpha)+q_2(\alpha))^{1/\alpha}\Big) \\ &\quad + c_2\Big((C_2-\rho_0)q_2(\alpha)^{1/\alpha} + \rho_0\frac{\mu_2}{\mu_0}(q_1(\alpha)+q_2(\alpha))^{1/\alpha}\Big) \\ &\quad + \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0}\rho_0(q_1(\alpha)+q_2(\alpha))^{\frac{1}{\alpha}} \\ &\overset{d}{=} c_1\mu_1\hat{v}_1 + c_2\mu_2\hat{v}_2 + \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0}\rho_0(1+f(\alpha)^\alpha)^{\frac{1}{\alpha}} q_2(\alpha)^{\frac{1}{\alpha}},\end{aligned} \tag{7.28}$$

where $f(\alpha) := \left(\frac{q_1(\alpha)}{q_2(\alpha)}\right)^{\frac{1}{\alpha}}$. In the remainder of the proof we derive monotonicity properties for the term $(1+f(\alpha)^\alpha)^{\frac{1}{\alpha}} q_2(\alpha)^{\frac{1}{\alpha}}$.

Let $\alpha_1, \alpha_2 > 0$. From equation (7.27) with $i = 2$ we obtain

$$q_2(\alpha_1)^{1/\alpha_1} = q_2(\alpha_2)^{1/\alpha_2} \frac{C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1+f(\alpha_2)^{\alpha_2})^{1/\alpha_2}}{C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1+f(\alpha_1)^{\alpha_1})^{1/\alpha_1}}. \tag{7.29}$$

From (7.29) we conclude that

$$(1+f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}} q_2(\alpha_1)^{\frac{1}{\alpha_1}} < \ (=)\ (1+f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} q_2(\alpha_2)^{\frac{1}{\alpha_2}} \tag{7.30}$$

if and only if

$$\frac{(1+f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}}{(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1+f(\alpha_1)^{\alpha_1})^{1/\alpha_1}} < \ (=)\ \frac{(1+f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}}}{(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1+f(\alpha_2)^{\alpha_2})^{1/\alpha_2}},$$

if and only if

$$(1+f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}} < \ (=)\ (1+f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}}. \tag{7.31}$$

Let b be such that $\hat{v}_1 = b\hat{v}_2$. Assume without loss of generality $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} \leq b \leq 1$. (For states with $b > 1$ the analysis is the same, with only the roles of nodes 1 and 2 interchanged.) Note that when $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$ we are on the edge of the cone as described in (7.10). In Lemma 7.B.1 (see below) we prove that $(1+f(\alpha)^\alpha)^{\frac{1}{\alpha}}$ is strictly decreasing in α when $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$ and is constant when $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$, the edge of the cone. Assuming the probability mass is not all concentrated on the edge of the cone, we conclude from (7.28) and the equivalence between (7.30) and (7.31) that the diffusion-scaled holding cost is strictly decreasing (strictly increasing) in α when $c_1\mu_1 + c_2\mu_2 < c_0\mu_0$ ($c_1\mu_1 + c_2\mu_2 > c_0\mu_0$). □

The following lemma is used in the proof of Proposition 7.4.6.

Lemma 7.B.1. *The function* $(1 + f(\alpha)^\alpha)^{1/\alpha}$ *with* $f(\alpha) = \left(\frac{q_1(\alpha)}{q_2(\alpha)}\right)^{\frac{1}{\alpha}}$*, is strictly decreasing in* α *when* $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$ *and is constant in* α *when* $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$*, with* $b = \hat{v}_1/\hat{v}_2$*.*

Proof: From $\hat{v}_1 = b\hat{v}_2$ and (7.27) we obtain the relation

$$(C_1 - \rho_0)\frac{1}{\mu_1}q_1(\alpha_i)^{\frac{1}{\alpha_i}} + (1-b)\rho_0\frac{1}{\mu_0}(q_1(\alpha_i) + q_2(\alpha_i))^{\frac{1}{\alpha_i}} \quad = \quad b(C_2 - \rho_0)\frac{1}{\mu_2}q_2(\alpha_i)^{\frac{1}{\alpha_i}}.$$

Dividing both sides by $q_2(\alpha_i)^{\frac{1}{\alpha_i}}$, we obtain

$$(C_1 - \rho_0)\frac{1}{\mu_1}f(\alpha_i) + (1-b)\rho_0\frac{1}{\mu_0}(1 + f(\alpha_i)^{\alpha_i})^{\frac{1}{\alpha_i}} \quad = \quad b(C_2 - \rho_0)\frac{1}{\mu_2}. \tag{7.32}$$

By (7.32) we have that $f(\alpha) = 0$ if and only if $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$. Hence, the function $(1 + f(\alpha)^\alpha)^{\frac{1}{\alpha}}$ is constant when $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$.

Now assume $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$. Take $\alpha_1 < \alpha_2$ and let $r > 1$ be such that $\alpha_2 = r\alpha_1$. Then

$$(1+f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} = (1^{r\alpha_1}+f(r\alpha_1)^{r\alpha_1})^{\frac{1}{r\alpha_1}} < (1^{\alpha_1}+f(r\alpha_1)^{\alpha_1})^{\frac{r}{r\alpha_1}} = (1+f(r\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}, \tag{7.33}$$

since $1+f(r\alpha_1) > 1$. Suppose $f(\alpha_2) = f(r\alpha_1) \leq f(\alpha_1)$. From (7.33), we then obtain $(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} < (1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}$. However, from (7.32) we know that if $f(\alpha_2) \leq f(\alpha_1)$, then $(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} \geq (1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}$, hence we have a contradiction. We conclude that $f(\alpha_2) > f(\alpha_1)$, and hence $f(\alpha)$ is strictly increasing in α and from (7.32) it then follows that $(1 + f(\alpha)^\alpha)^{\frac{1}{\alpha}}$ is strictly decreasing in α. □

Chapter 8
Optimal scheduling in a parallel two-server model

The focus of Chapters 2–7 was on the single-server model and the linear network. In this last chapter we shift our attention to the parallel two-server model with two traffic classes. Both servers can be simultaneously allocated to one of the classes, or both classes can be served in parallel, each by a dedicated server. We seek policies that minimize in some sense the holding cost within the class of non-anticipating policies. As described in Section 1.5, determining the optimal policy in explicit form has so far proved infeasible. In this chapter our main goals are to study the structural properties of optimal scheduling policies and to determine computable approximations that are close to optimality. We assume exponentially distributed service requirements.

We primarily focus on the situation where the highest service capacity is achieved when serving both classes in parallel. For some special cases the optimal policy can be determined exactly, but this is not possible in general. In a similar setting, [17] states that switching-curve policies are optimal. Numerical experiments included for illustration in the present chapter indeed support this optimality. In order to find computable approximations for the optimal policies, we perform a fluid analysis similar to that of Chapter 5 for the two-node linear network: We first study the fluid control model for which we show that the optimal control is described by a linear switching curve. Using this result, we derive that policies characterized by either linear or exponential switching curves are asymptotically fluid-optimal in the original stochastic model. Our analysis is partly inspired by that in [53, 54] where a multi-class tandem-network is studied.

By simulations we compare the fluid-based policies with threshold-based [19, 20] and Max-Weight policies [89, 132], which are known to be optimal in heavy traffic. We show that the fluid-based and threshold-based policies give good performance in general, while significant improvements over Max-Weight policies can be achieved. For threshold-based policies, however, finding reasonable values for the thresholds is not trivial since performance as well as stability can be quite sensitive to the threshold values. This contrasts with the fluid-based policies for which the switching curve is explicitly given.

This chapter is organized as follows. In Section 8.1 we describe the model and state some preliminary results. Section 8.2 contains optimality results for the stochastic model. The fluid control model and asymptotically fluid-optimal policies are presented in Section 8.3. For certain choices of the parameters, exponential-shaped switching curves provide an asymptotically fluid-optimal policy. This is discussed in more detail in Section 8.4. In Section 8.5 we focus on another setting and argue that in that case an efficient policy has a quadratic switching curve. For comparison we briefly discuss optimal policies in heavy traffic using the results of [19, 20] and [89, 132] in Section 8.6. Numerical experiments and concluding remarks can be found in Sections 8.7 and 8.8, respectively.

8.1 Model and preliminaries

We consider a parallel two-server model with two classes of users, as depicted in Figure 1.6 (left). Class-i users, $i = 1, 2$, arrive according to independent Poisson processes with rate λ_i, and have exponentially distributed service requirements. We assume that both servers can work simultaneously on the same user. As described in Section 1.5, we can equivalently formulate the parallel-server model as a system with two classes of users having a capacity region S, defined as the convex hull of the set $\{(0, 0), (1, 0), (0, 1), (c_1, c_2)\}$ (see Figure 1.6 (right) with $C_1 = C_2 = 1$ in the case $c_1 + c_2 > 1$). Hence, at any time either one class can be served individually with capacity 1, or both classes 1 and 2 can be served in parallel with capacities c_1 and c_2 respectively, $c_i \leq 1$, or the system is idling (not serving any class), or any convex combination of these four. The latter representation of the parallel-server model will be considered throughout the chapter.

Let $1/\mu_i$ represent the mean class-i service requirement, $i = 1, 2$, and define the traffic load of class i as $\rho_i := \frac{\lambda_i}{\mu_i}$. For a given policy π, denote by $N_i^\pi(t)$ the number of class-i users at time t, and define N_i^π as the random variable with the corresponding equilibrium distribution (when it exists). Let $\vec{N}^\pi(t) = (N_1^\pi(t), N_2^\pi(t))$.

We assume that the numbers of users in the two classes are observable to a policy. For a given policy π, denote by $s_i^\pi(t)$ the service capacity devoted to class i at time t. We assume that $s_i^\pi(t) = 0$ when $N_i^\pi(t) = 0$. In addition, we assume the process $s_i^\pi(t)$ to be right-continuous with left limits, and the vector $\vec{s}^\pi(t) := (s_1^\pi(t), s_2^\pi(t))$ to lie in the capacity region S. Note that the total (service) capacity $s_1^\pi(t) + s_2^\pi(t)$, that is, the speed at which the total amount of backlogged work in the system decreases, is not constant in time. Depending on the decision taken at time t, it may vary between 0 and $\max(1, c_1 + c_2)$.

Let $S_i^\pi(t) := \int_0^t s_i^\pi(u)\mathrm{d}u$ denote the cumulative amount of capacity devoted to class i in the time interval $(0, t]$ under policy π. Let $A_i(u, t)$ be the amount of class-i work that arrived during the time interval $(u, t]$. Then, the workload in class i at time t can be written as

$$W_i^\pi(t) := W_i(0) + A_i(0, t) - S_i^\pi(t). \tag{8.1}$$

The objective of this chapter is to identify scheduling policies that in some appropriate sense minimize the holding cost $d_1N_1(t)+d_2N_2(t)$, with $d_i \geq 0$ some cost associated with class i, within the class of (possibly preemptive) non-anticipating policies. We denote this class of policies by $\bar{\Pi}$. Throughout the chapter, we assume $d_1\mu_1 \geq d_2\mu_2$.

In this chapter we investigate the case $c_1 + c_2 > 1$. However, before proceeding let us briefly consider the situation $c_1 + c_2 \leq 1$. In the latter case, the policy that gives preemptive priority to class 1 minimizes the mean holding cost. (In fact, this result holds for any shape of the capacity region where the points $(1, 0)$ and $(0, 1)$ are not dominated by any other element in the capacity region.) Denote by $\pi^{(1)}$ the policy that gives preemptive priority to class 1, and let $\pi \in \bar{\Pi}$. Consider the same realizations of the arrival processes and service requirements for policies $\pi^{(1)}$ and π. If at time $t = 0$ the workloads satisfy

$$\begin{aligned} W_1^{\pi^{(1)}}(t) &\leq W_1^{\pi}(t), &(8.2)\\ W_1^{\pi^{(1)}}(t)+W_2^{\pi^{(1)}}(t) &\leq W_1^{\pi}(t)+W_2^{\pi}(t), &(8.3)\end{aligned}$$

then the same is true for all $t \geq 0$. Multiplying (8.2) by $d_1\mu_1 - d_2\mu_2 \geq 0$ and (8.3) by $d_2\mu_2$ and adding the two inequalities gives that $d_1\mu_1W_1^{\pi^{(1)}}(t) + d_2\mu_2W_2^{\pi^{(1)}}(t) \leq d_1\mu_1W_1^{\pi}(t) + d_2\mu_2W_2^{\pi}(t)$. Since we have exponentially distributed service requirements and we consider only non-anticipating policies, we obtain $\mathbb{E}(W_i(t)) = \frac{1}{\mu_i}\mathbb{E}(N_i(t))$, so that $d_1\mathbb{E}(N_1^{\pi^{(1)}}(t)) + d_2\mathbb{E}(N_2^{\pi^{(1)}}(t)) \leq d_1\mathbb{E}(N_1^{\pi}(t)) + d_2\mathbb{E}(N_2^{\pi}(t))$, for all $t \geq 0$, and in particular, policy $\pi^{(1)}$ is average-cost optimal.

As mentioned before, in the remainder of the chapter we focus on the unsolved case $c_1 + c_2 > 1$. In this case, the total service capacity is largest when both classes are served in parallel. For application in wireless networks, this represents the joint capacity when both base stations transmit in parallel, and in computer scheduling it corresponds to dedicated specialized servers.

Stability conditions when $c_1 + c_2 > 1$

For a given policy π, the system is called stable when the process $N^{\pi}(t)$ is positive Harris recurrent. When $c_1+c_2 > 1$, the policy that serves classes 1 and 2 in parallel, whenever possible, minimizes the total workload in the system at every moment in time. Hence, this policy will keep the system stable whenever possible. Under this policy, the model becomes a coupled-processors model for which the stability conditions are

$$\min\left(\frac{\rho_1}{c_1}, \frac{\rho_2}{c_2}\right) < 1 \quad \text{and} \tag{8.4}$$

$$\text{if} \quad \frac{\rho_i}{c_i} < 1 \quad \text{then} \quad \rho_j + \frac{\rho_i}{c_i}(1 - c_j) < 1, \; i \neq j, \tag{8.5}$$

see [42, 50]. Conditions (8.4) and (8.5) are therefore the maximum stability conditions. Note that the load vectors (ρ_1, ρ_2) that satisfy the maximum stability conditions are exactly those vectors that lie in the interior of the capacity region S depicted in Figure 1.6 (with $C_1 = C_2 = 1$).

8.2 Optimality results

For a standard multi-class single-server queue it is well known that the $c\mu$-rule minimizes the mean holding cost if users of the various classes have exponentially distributed service requirements, see Example 7.5.8. In the case of unit costs, an even stronger result exists. The policy that gives preemptive priority to the class i with the highest departure rate μ_i, *stochastically* minimizes the total number of users [114]. One might expect a similar rule to be optimal in our model as well. This rule would amount to choosing the allocation $\vec{s}(t)$ that maximizes the weighted user departure rate, $d_1\mu_1 s_1(t) + d_2\mu_2 s_2(t)$, at any time t. Unfortunately, the total service capacity, $s_1(t)+s_2(t)$, depends on the chosen allocation as well. For example, serving class i only decreases the total amount of work at rate 1, while serving both classes in parallel gives a decrease of the workload at rate $c_1 + c_2 > 1$. Therefore, the objective to maximize the user departure rate may be conflicting with that of maximizing the total service capacity used. The latter will minimize the total time needed to empty the system, which is advantageous in the long run, while the former is better in the short run.

Recall that we chose $d_1\mu_1 \geq d_2\mu_2$. If, in addition, $d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$, then there is no trade-off and it is intuitively clear that the policy that always serves classes 1 and 2 in parallel (whenever both are backlogged) is optimal, since this maximizes both the workload depletion rate and the weighted departure rate. This is made precise in Section 8.2.1.

When $d_1\mu_1 \geq d_1\mu_1 c_1 + d_2\mu_2 c_2$, the highest weighted departure rate is obtained when serving class 1 individually. It may therefore be better to sometimes serve class 1 individually, even if that does not maximize the rate at which the total work in the system decreases. Hence as the number of users varies, the system should dynamically switch between different allocations. The general structure of an optimal policy is discussed in Section 8.2.2.

8.2.1 Priority rule and optimality

In this section we show that when $(d_2\mu_2 \leq)\, d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$, the priority rule that serves both classes in parallel (whenever possible) minimizes the mean holding cost. In case of unit costs, $d_1 = d_2 = 1$, this policy in fact *stochastically* minimizes the total number of users. These results are proved using a dynamic programming approach.

We consider the uniformized Markov chain, that is, transition epochs are generated by a Poisson process of uniform rate $\nu = \lambda_1 + \lambda_2 + \mu_1(1+c_1) + \mu_2(1+c_2)$. Since ν is finite, we may assume $\nu = 1$ without loss of generality. We then focus on the discrete-time Markov chain embedded at transition epochs and, for transparency of notation, denote the number of class-i users after t steps by $N_i(t),\ i = 1, 2$. We define the value functions $V_m(\cdot) : \mathbb{Z}_+^2 \to \mathbb{R}$, $m = 0, 1, \ldots$, as follows. Let $\vec{x} = (x_1, x_2) \in \mathbb{Z}_+^2$.

Then, $V_0(\vec{x}) := \tilde{C}(\vec{x})$, with $\tilde{C}(\cdot) : \mathbb{Z}_+^2 \to \mathbb{R}$ a terminal cost, and for $m = 1, 2, \ldots$,

$$\begin{aligned} V_{m+1}(\vec{x}) &:= \lambda_1 V_m(\vec{x}+\vec{e}_1) + \lambda_2 V_m(\vec{x}+\vec{e}_2) \\ &\quad + \min_{\vec{s}\in S}\Big\{\sum_{i=1,2} \mathbf{1}_{(x_i>0)}\mu_i s_i V_m(\vec{x}-\vec{e}_i) + (1-\lambda_1-\lambda_2-\sum_{i=1,2}\mathbf{1}_{(x_i>0)}\mu_i s_i)V_m(\vec{x})\Big\} \\ &= \lambda_1 V_m(\vec{x}+\vec{e}_1) + \lambda_2 V_m(\vec{x}+\vec{e}_2) + (\mu_1(1+c_1)+\mu_2(1+c_2))V_m(\vec{x}) \\ &\quad + \min_{\vec{s}\in S}\Big\{\sum_{i=1,2}\mathbf{1}_{(x_i>0)}\mu_i s_i\left(V_m(\vec{x}-\vec{e}_i)-V_m(\vec{x})\right)\Big\}, \qquad (8.6) \end{aligned}$$

with S the capacity region, and $\vec{e}_i$ the i-th unit vector. Choosing $\tilde{C}(\vec{x}) = d_1x_1 + d_2x_2$, we obtain $V_{m+1}(\vec{x}) = \min_{\pi\in\bar{\Pi}} \mathbb{E}(d_1N_1^\pi(m+1) + d_2N_2^\pi(m+1)|\vec{N}(0) = \vec{x})$. When instead $\tilde{C}(\vec{x}) = \mathbf{1}_{(x_1+x_2>y)}$, we obtain $V_{m+1}(\vec{x}) = \min_{\pi\in\bar{\Pi}} \mathbb{P}(N_1^\pi(m+1) + N_2^\pi(m+1) > y|\vec{N}(0) = \vec{x})$. If for all m (and for all $y \geq 0$) we can choose the same minimizing action in (8.6) (the optimal action may depend on the state $\vec{x}$), then the corresponding stationary policy minimizes the mean holding cost $\mathbb{E}(d_1N_1(t) + d_2N_2(t))$ (when $\tilde{C}(\vec{x}) = d_1x_1 + d_2x_2$) or stochastically minimizes the total number of users (when $\tilde{C}(\vec{x}) = \mathbf{1}_{(x_1+x_2>y)}$), respectively, at every instant in time.

In the next two lemmas we establish convenient properties of $V_m(\cdot)$, without specifying the function $\tilde{C}(\cdot)$.

Lemma 8.2.1. *If $\tilde{C}(\cdot)$ is non-decreasing in x_1 and x_2, then $V_m(\cdot)$ is non-decreasing in x_1 and x_2 for all m.*

Proof: The statement follows directly from the definition of $V_m(\cdot)$. □

The set S is convex, hence for non-decreasing cost functions the minimizing action in (8.6) will be one of the extreme points of S. From Lemma 8.2.1 it follows that $\sum_{i=1,2}\mathbf{1}_{(x_i>0)}\mu_i s_i\left(V_m(\vec{x}-\vec{e}_i)-V_m(\vec{x})\right) \leq 0$, hence the minimizing action in (8.6) will not be $(0,0) \in S$. This implies that we can rewrite the function $V_{m+1}(\cdot)$ as follows:

$$\begin{aligned} &V_{m+1}(\vec{x}) = \lambda_1 V_m(\vec{x}+\vec{e}_1) + \lambda_2 V_m(\vec{x}+\vec{e}_2) \\ &+ \min\Big(\mu_1 V_m((x_1-1)^+, x_2) + (\mu_2+\mu_1c_1+\mu_2c_2)V_m(\vec{x}), \\ &\qquad \mu_2 V_m(x_1,(x_2-1)^+) + (\mu_1+\mu_1c_1+\mu_2c_2)V_m(\vec{x}), \\ &\qquad \mu_1c_1V_m((x_1-1)^+,x_2) + \mu_2c_2V_m(x_1,(x_2-1)^+) + (\mu_1+\mu_2)V_m(\vec{x})\Big). \quad (8.7) \end{aligned}$$

The next lemma shows that under certain conditions on the cost function, the minimizing action in (8.7) will be to always serve classes 1 and 2 in parallel, whenever possible, independent of the remaining time horizon. The proof uses Lemma 8.2.1 and may be found in Appendix 8.A.

Lemma 8.2.2. *Assume $c_1 + c_2 \geq 1$ and that $\tilde{C}(\cdot)$ is non-decreasing in x_1 and x_2. If $Z = \tilde{C}$ satisfies*

$$\begin{aligned}(\mu_1 + \mu_2)Z(\vec{x}) + \mu_1 c_1 Z(\vec{x} - \vec{e}_1) + \mu_2 c_2 Z(\vec{x} - \vec{e}_2)\\ \leq \min(\mu_1 Z(\vec{x} - \vec{e}_1) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2)Z(\vec{x}),\\ \mu_2 Z(\vec{x} - \vec{e}_2) + (\mu_1 + \mu_1 c_1 + \mu_2 c_2)Z(\vec{x})),\end{aligned} \tag{8.8}$$

for all $x_1, x_2 > 0$, then the same is true for $Z = V_m$, for all $m \geq 0$.

From the lemma above, we obtain that a policy that always serves both classes in parallel whenever possible, minimizes the mean holding cost, and stochastically minimizes the total number of users.

Proposition 8.2.3. *Assume $c_1 + c_2 \geq 1$. Let $\pi^* \in \bar{\Pi}$ be a policy that serves both classes in parallel whenever possible and let $\pi \in \bar{\Pi}$. Assume $\vec{N}^{\pi^*}(0) = \vec{N}^{\pi}(0)$. If $(d_2\mu_2 \leq) d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$, then*

$$\mathbb{E}(d_1 N_1^{\pi^*}(t) + d_2 N_2^{\pi^*}(t)) \leq \mathbb{E}(d_1 N_1^{\pi}(t) + d_2 N_2^{\pi}(t)), \quad \textit{for all } t \geq 0.$$

If in addition $d_1 = d_2$, then

$$N_1^{\pi^*}(t) + N_2^{\pi^*}(t) \leq_{st} N_1^{\pi}(t) + N_2^{\pi}(t), \quad \textit{for all } t \geq 0.$$

Proof: If $(d_2\mu_2 \leq) d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$, then the non-decreasing cost function $\tilde{C}(x_1, x_2) = d_1 x_1 + d_2 x_2$ satisfies (8.8). Lemma 8.2.2 implies that serving both classes in parallel (whenever possible) is always the minimizing action in (8.7) and hence the corresponding policy minimizes $\mathbb{E}(d_1 N_1(t) + d_2 N_2(t))$, at any time $t \geq 0$.

Now assume $d_1 = d_2$. In that case, the non-decreasing cost function $\tilde{C}(x_1, x_2) = \mathbf{1}_{(x_1 + x_2 > y)}$ satisfies (8.8). Lemma 8.2.2 implies that serving both classes in parallel (whenever possible) is always the minimizing action in (8.7) and hence the corresponding policy stochastically minimizes $\mathbb{P}(N_1(t) + N_2(t) > y)$, at any time $t \geq 0$, and for all $y \geq 0$. □

8.2.2 General structure of an average-cost optimal policy

Section 8.2.1 treats optimal policies restricted to the case $d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$. In this section we explore the general structure of an optimal policy.

When $d_1\mu_1 > d_2\mu_2$, maximizing the weighted user departure rate would imply that an optimal policy will never serve class 2 individually when class 1 is also present. At the same time, serving class 2 individually does not give the highest possible total service capacity either, since $c_1 + c_2 > 1$. Therefore, it is natural that this action should not be chosen by an optimal policy. This fact is proved in Proposition 8.2.5. First we state a lemma that in fact holds for generally distributed inter-arrival times and service requirements and, in particular, holds irrespective of the values for μ_1 and μ_2. The proof may be found in Appendix 8.B.

Lemma 8.2.4. *(This lemma holds for generally distributed inter-arrival times and service requirements.) Assume $c_1 + c_2 > 1$. Let $\tilde{\pi}$ be a policy that sometimes does* *serve class 2 individually while there are class-1 users present. Define policy π to be the policy that uses the same allocation as $\tilde{\pi}$ when possible, except when policy $\tilde{\pi}$ serves class 2 individually. In that case policy π serves classes 1 and 2 in parallel (if possible).*

Consider the same realizations of the arrival processes and service requirements. Then the following sample-path inequalities hold for all $t \geq 0$:

$$S_1^{\pi}(t) \geq S_1^{\tilde{\pi}}(t), \tag{8.9}$$

$$S_1^{\pi}(t) + S_2^{\pi}(t) \geq S_1^{\tilde{\pi}}(t) + S_2^{\tilde{\pi}}(t), \tag{8.10}$$

$$(1-c_2)S_1^{\pi}(t) + c_1 S_2^{\pi}(t) \geq (1-c_2)S_1^{\tilde{\pi}}(t) + c_1 S_2^{\tilde{\pi}}(t). \tag{8.11}$$

Proposition 8.2.5. *Assume $d_1\mu_1 \geq d_2\mu_2$ and $c_1 + c_2 > 1$. For any policy $\tilde{\pi}$ that serves class 2 individually when there is work of class 1 present, there exists a modified policy π that never serves class 2 individually when class 1 is present and that does not worse than $\tilde{\pi}$, i.e.,*

$$\mathbb{E}(d_1 N_1^{\pi}(t) + d_2 N_2^{\pi}(t)) \leq \mathbb{E}(d_1 N_1^{\tilde{\pi}}(t) + d_2 N_2^{\tilde{\pi}}(t)), \quad \textit{for all } t \geq 0.$$

Proof: Let $\tilde{\pi}$ be a policy that sometimes does serve class 2 individually while there are class-1 users present. Define policy π as in Lemma 8.2.4 and hence the sample-path inequalities (8.9) and (8.10) hold. Multiplying (8.9) by $d_1\mu_1 - d_2\mu_2 \geq 0$ and (8.10) by $d_2\mu_2$ and adding the two inequalities gives that $d_1\mu_1 S_1^{\pi}(t) + d_2\mu_2 S_2^{\pi}(t) \geq d_1\mu_1 S_1^{\tilde{\pi}}(t) + d_2\mu_2 S_2^{\tilde{\pi}}(t)$ and hence by (8.1) we obtain

$$d_1\mu_1 W_1^{\pi}(t) + d_2\mu_2 W_2^{\pi}(t) \leq d_1\mu_1 W_1^{\tilde{\pi}}(t) + d_2\mu_2 W_2^{\tilde{\pi}}(t), \quad \text{for all } t \geq 0. \tag{8.12}$$

Since we assumed exponentially distributed service requirements and we consider only non-anticipating policies, we have $\mathbb{E}(W_i^{\pi}(t)) = \frac{1}{\mu_i}\mathbb{E}(N_i^{\pi}(t))$. By taking expectations on both sides in (8.12), we obtain $\mathbb{E}(d_1 N_1^{\pi}(t) + d_2 N_2^{\pi}(t)) \leq \mathbb{E}(d_1 N_1^{\tilde{\pi}}(t) + d_2 N_2^{\tilde{\pi}}(t))$. Hence policy π is not worse than $\tilde{\pi}$ and policy π never serves class 2 individually when there is work of class 1 present. □

In Section 8.2.1 we explicitly found an optimal policy when $d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$. Hence, the remaining interesting case is when $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$. We are interested in a policy that minimizes $\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}(\int_0^T (d_1 N_1^{\pi}(t) + d_2 N_2^{\pi}(t)))\mathrm{d}t$ over all policies $\pi \in \bar{\Pi}$. According to [123, Corollary 20] such an average-cost optimal policy exists.

When $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$, there is a tradeoff when users of both classes are present. On one hand serving class 1 individually maximizes the weighted user departure rate since $d_1\mu_1 > d_1\mu_1 c_1 + c_2\mu_2 c_2$. However, serving classes 1 and 2 simultaneously maximizes the speed at which the total workload in the system decreases. When seeking an average-optimal policy, by Proposition 8.2.5 we only need to consider policies that never serve class-2 users individually when there are also class-1 users present. The decision between whether to serve class 1 individually or classes 1 and 2 jointly depends on the number of class-1 and class-2 users present in the system. Intuitively, one may expect that the optimal policy can be characterized by a switching curve, i.e., there exists a switching curve $h(\cdot)$ such that if $N_2(t) \geq h(N_1(t))$, then it is optimal to serve classes 1 and 2 in parallel, and otherwise it is optimal to serve class 1 individually. For a model with slightly different behavior near the boundaries, the authors in [17] state that value iteration techniques can be used to prove that an average-cost optimal policy is indeed characterized by such a switching curve (a proof will be included in a forthcoming paper by the authors of [17]). We expect that for our model, the existence of a switching curve can be proved using similar steps as in the proof of Proposition 4.3.9. However, value iteration will not provide us with any information concerning the shape of the curve. Therefore, in the remainder of the chapter we seek policies that are close to optimal by investigating two limiting regimes. In Section 8.3 this is done for a fluid-scaled system, and asymptotically fluid-optimal policies are derived, which are characterized by a switching curve. Optimality results for the heavy-traffic regime are reviewed in Section 8.6.

8.3 Fluid analysis

In this section we consider the stochastic queue length processes under a fluid scaling and investigate close to optimal policies for the unsolved case $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$. In order to do so, it will be convenient to first study the related deterministic fluid control model. This will be done in Section 8.3.1. For this relatively simple model we derive optimal controls, which are described by switching curves. Using these curves we derive in Section 8.3.2 switching curves that provide asymptotically fluid-optimal policies in the stochastic model.

8.3.1 Optimal fluid control

In this section we focus on the deterministic fluid control model, which arises from the original stochastic model by only taking into account the mean drifts. A fluid process is a solution $n(t) = (n_1(t), n_2(t))$ of the following equations:

$$n_i(t) = n_i + \lambda_i t - U_i(t)\mu_i - U_c(t)\mu_i c_i, \ i = 1, 2, \tag{8.13}$$

$$n_i(t) \geq 0, \ i = 1, 2. \tag{8.14}$$

Here $n = (n_1, n_2) \in \mathbb{R}_+^2$ and $U_j(t) = \int_0^t u_j(v)\mathrm{d}v$, $j = 1, 2, c$, such that for all $v \geq 0$,

$$u_1(v) + u_2(v) + u_c(v) \leq 1, \tag{8.15}$$

$$u_j(v) \geq 0, \ j = 1, 2, c, \tag{8.16}$$

and the functions $u_j(\cdot)$ are measurable, $j = 1, 2, c$. The subscript c refers to "combined service", i.e., serving both classes in parallel. Note that $U_j(\cdot)$ is Lipschitz continuous with constant less than or equal to 1. Hence, it is absolutely continuous which implies that it is differentiable almost everywhere [112]. Then, $n_i(\cdot)$ is differentiable almost everywhere as well, and at regular points (a regular point is a value of t at which $n_i(t)$ is differentiable) we have

$$\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = \lambda_i - u_i(t)\mu_i - u_c(t)\mu_i c_i,\ i = 1, 2. \tag{8.17}$$

Under the stability conditions, the fluid model can be drained in finite time, as is stated in the following lemma.

Lemma 8.3.1. *If (8.4) and (8.5) are satisfied, then the policy that serves classes 1 and 2 in parallel whenever possible, drains the fluid model in finite time and keeps the system empty from that moment on.*

Proof: We consider the workload fluid processes $w_i(t) := \frac{n_i(t)}{\mu_i}$, $i = 1, 2$. From (8.17) we have $\frac{\mathrm{d}w_i(t)}{\mathrm{d}t} = \rho_i - u_i(t) - u_c(t)c_i$, $i = 1, 2$, at regular points. Focus on the policy that serves classes 1 and 2 in parallel whenever possible. Assume $w_1(t), w_2(t) > 0$. Then $u_c(t) = 1$. By (8.4), there is a class i with $\frac{\rho_i}{c_i} < 1$. Hence, $\frac{\mathrm{d}w_i(t)}{\mathrm{d}t} = \rho_i - c_i < 0$ and class i will eventually be drained to zero. When at that time the workload in class j ($j \neq i$) is strictly positive (while $w_i(t) = 0$), we have $u_c(t) = \frac{\rho_i}{c_i}$ and $u_j(t) = 1 - \frac{\rho_i}{c_i}$. From (8.5) this gives $\frac{\mathrm{d}w_i(t)}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}w_j(t)}{\mathrm{d}t} = \rho_j - 1 + \frac{\rho_i}{c_i} - \frac{\rho_i}{c_i}c_j = \rho_j + \frac{\rho_i}{c_i}(1 - c_j) - 1 < 0$. Hence, class j must eventually become empty as well. □

A policy π for the fluid control model is described by the control functions $u_1^\pi(t)$, $u_2^\pi(t)$ and $u_c^\pi(t)$ (we also write $U_j^\pi(t) = \int_0^t u_j^\pi(v)\mathrm{d}v$). A corresponding trajectory is denoted by $n^\pi(t)$. We are interested in finding an (average-cost) optimal fluid control that minimizes

$$\int_0^\infty (d_1 n_1^\pi(t) + d_2 n_2^\pi(t))\mathrm{d}t, \quad \text{with } (n^\pi(t), u^\pi(t)) \text{ satisfying (8.13)–(8.16).} \tag{8.18}$$

(Different from Chapter 5, we will omit the term "average-cost" since we do not consider other optimality criteria.) We denote an optimal control by $u_j^*(t), j = 1, 2, c$, and a corresponding optimal trajectory by $n_i^*(t)$, $i = 1, 2$. Before proceeding to find $n^*(t)$ and $u^*(t)$, we first prove in the next lemma that an optimal pair $(n^*(t),\ u^*(t))$ exists. In addition, the lemma states that if $u^*(t)$ is an optimal control for the infinite-horizon problem, then it is also optimal for the finite-horizon problem whenever the horizon is large enough. This property will be useful to prove convergence of the stochastic model in Section 8.3.2. The proof of the lemma goes along similar lines as the proof of Lemma 5.2.4 and may be found in Appendix 8.C.

Lemma 8.3.2. *If (8.4) and (8.5) are satisfied, then there exists a control $u^*(t)$ and a corresponding trajectory $n^*(t)$ that solves the minimization problem (8.18).*

In addition, there exists a function $H : \mathbb{R} \to \mathbb{R}$ such that,

$$\begin{aligned}\min_{n(t)\ s.t.\ (8.13)-(8.16)} \int_0^D (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t &= \int_0^D (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t \\ &= \int_0^\infty (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t,\end{aligned}$$

for all $D \geq H(d_1 n_1 + d_2 n_2)$, with n the initial state.

For the stochastic model we know that when $d_1\mu_1 \geq d_2\mu_2$, it is never optimal to serve class 2 exclusively when also work of class 1 is present, see Proposition 8.2.5. In the fluid control model this is true as well, as is stated in the lemma below. The proof may be found in Appendix 8.D.

Lemma 8.3.3. *Assume (8.4) and (8.5) are satisfied, $d_1\mu_1 \geq d_2\mu_2$, and $c_1+c_2 > 1$. Then, for any policy $\tilde{\pi}$ that allows $u_2^{\tilde{\pi}}(t) > 0$ when $n_1^{\tilde{\pi}}(t) > 0$, there exists a modified policy π, with $u_2^{\pi}(t) = 0$ whenever $n_1^{\pi}(t) > 0$, that does not do worse than $\tilde{\pi}$, i.e., $d_1 n_1^{\pi}(t) + d_2 n_2^{\pi}(t) \leq d_1 n_1^{\tilde{\pi}}(t) + d_2 n_2^{\tilde{\pi}}(t)$, for all $t \geq 0$.*

In case $d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$, the control that serves both classes in parallel whenever possible is optimal, i.e., $u_c^*(t) = 1$ when $n_1(t), n_2(t) > 0$, and $u_c^*(t) = \min(\frac{\rho_j}{c_j}, 1)$, $u_i^*(t) = 1 - u_c^*(t)$ when $n_j(t) = 0$ and $n_i(t) > 0$, for $i \neq j$, $i, j = 1, 2$. This follows from the fact that the above-described policy minimizes the time to empty the system, while at the same time, it maximizes the weighted departure rate at any moment in time. We do not include a formal proof of this fact, since the main objective of this section is to investigate close-to-optimal policies for parameter choices that did not allow us to exactly determine the optimal policy for the stochastic model. (Proposition 8.2.3 discusses an optimal policy for the stochastic model when $d_1\mu_1 \leq d_1\mu_1 c_1 + d_2\mu_2 c_2$.)

In the remainder of this section we concentrate on the case $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$, for which the following lemma enables us to prove that the optimal policy in the fluid control model is characterized by a switching curve.

Lemma 8.3.4. *Assume (8.4) and (8.5) are satisfied, $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$, and $c_1 + c_2 > 1$. Consider a trajectory starting in $\tilde{n} \in \{n \in \mathbb{R}_+^2 : n_1 > 0, n_2 \geq 0\}$ with the following properties: (i) first class 1 is served exclusively during a contiguous period, and then (ii) we switch to serving both classes simultaneously during another contiguous period. Let $\hat{n}$ be the end point of this trajectory.*

Among all feasible trajectories that move from $\tilde{n}$ to $\hat{n}$ without coinciding with the $n_1 = 0$ axis, the unique path that minimizes $d_1 n_1(t) + d_2 n_2(t)$ at all times (until reaching $\hat{n}$), is exactly the trajectory described above.

Proof: Since we consider only trajectories from $\tilde{n}$ to $\hat{n}$ that do not coincide with the $n_1 = 0$ axis, by Lemma 8.3.3 we know that the best path does not spend any time serving class 2 individually. Denote by U_1 (U_c) the cumulative amount of time spent serving class 1 individually (classes 1 and 2 in parallel). The net change in the amount of fluid in the two classes can be written as

$$\begin{aligned} \hat{n}_1 - \tilde{n}_1 &= (\lambda_1 - \mu_1)U_1 + (\lambda_1 - c_1\mu_1)U_c, \\ \hat{n}_2 - \tilde{n}_2 &= \lambda_2 U_1 + (\lambda_2 - c_2\mu_2)U_c. \end{aligned}$$

Under the necessary stability conditions (8.4) and (8.5) this has a unique solution for U_1 and U_c. Hence, all trajectories spend the same cumulative amount of time serving both classes in parallel as well as serving class 1 individually.

The rate at which the cost decreases when $n_1(t) > 0$ is given by $\frac{\mathrm{d}(d_1 n_1(t) + d_2 n_2(t))}{\mathrm{d}t} = d_1\lambda_1 + d_2\lambda_2 - u_1(t) d_1\mu_1 - u_c(t)(d_1\mu_1 c_1 + d_2\mu_2 c_2)$. Since $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$,

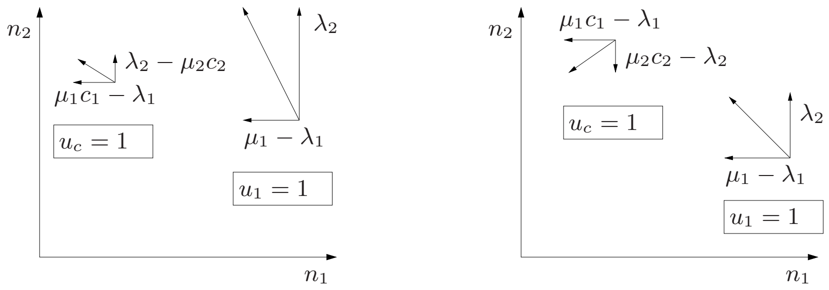


Figure 8.1: Drift vectors for $\rho_1 < c_1$ and $\rho_2 > c_2$ (left), and $\rho_1 < c_1$ and $\rho_2 < c_2$ (right), respectively.

first serving only class 1 initially maximizes the rate at which $d_1 n_1(t) + d_2 n_2(t)$ decreases. Hence, this minimizes $d_1 n_1(t) + d_2 n_2(t)$ at all times (until reaching $\hat{n}$). □

For the fluid control model we can now determine the optimal control. We make a distinction between whether $\rho_1 < c_1$ or $\rho_1 \geq c_1$. Note that, cf. Bellman's principle of optimality, we only need to consider policies that base their actions on the current state $n(t)$, because of the infinite horizon and the fact that the parameters do not depend on the current time t.

Case $\rho_1 < c_1$

When $\rho_1 < c_1$, a sufficient condition for the system to drain in finite time is $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$ (see Lemma 8.3.1). Depending on ρ_2 and c_2, the drifts are as in Figure 8.1. In Proposition 8.3.5 we describe the optimal fluid control, which is characterized by a linear switching curve. In Figure 8.2 the optimal trajectory is shown. In order to state the proposition it is convenient to define α as

$$\max\Big(0, \quad \frac{c_2 - \rho_2}{c_1 - \rho_1} + \frac{c_1}{c_1 + c_2 - 1} \cdot \frac{1 - \rho_2 - \frac{\rho_1}{c_1}(1 - c_2)}{c_1 - \rho_1} \cdot \frac{d_1\mu_1 - d_1\mu_1 c_1 - d_2\mu_2 c_2}{d_2\mu_2}\Big). \tag{8.19}$$

Note that under the conditions of Proposition 8.3.5, it holds that $\alpha > \frac{c_2 - \rho_2}{c_1 - \rho_1}$.

Proposition 8.3.5. *Let $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 < c_1$ and $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$. An optimal control $u^*(t)$ in the fluid control model is*

- *$u_1^*(t) = 1$, if $n_2(t) < \alpha\frac{\mu_2}{\mu_1} n_1(t)$.*
- *$u_c^*(t) = 1$, if $n_2(t) \geq \alpha\frac{\mu_2}{\mu_1} n_1(t)$ and $n_1(t) > 0$.*
- *$u_c^*(t) = \frac{\rho_1}{c_1}$ and $u_2^*(t) = 1 - \frac{\rho_1}{c_1}$, if $n_1(t) = 0$.*

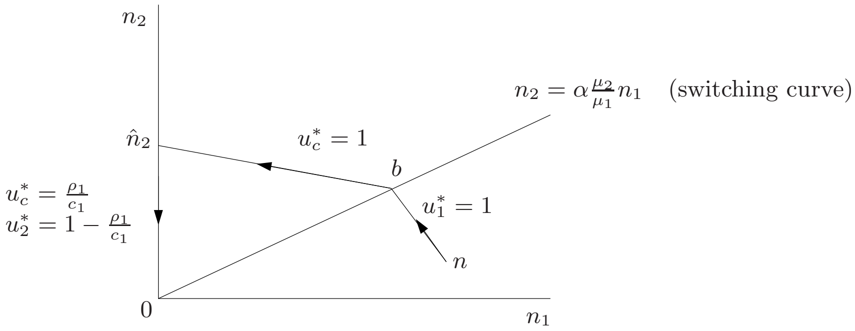

n_2
$n_2 = \alpha \frac{\mu_2}{\mu_1} n_1$ (switching curve)
$\hat{n}_2$
$u_c^* = 1$
b
$u_c^* = \frac{\rho_1}{c_1}$
$u_2^* = 1 - \frac{\rho_1}{c_1}$
$u_1^* = 1$
n
0
n_1

Figure 8.2: Optimal trajectory of the fluid control model when $\rho_1 < c_1$.

Proof: If $n_1(t) > 0$, when seeking an optimal control, by Lemma 8.3.3 we only need to consider controls with $u_2(t) = 0$ and $u_1(t) + u_c(t) = 1$. Hence, from $\frac{\mathrm{d}n_1(t)}{\mathrm{d}t} = \lambda_1 - u_1(t)\mu_1 - u_c(t)\mu_1 c_1$, and the fact that $\rho_1 < c_1 < 1$, class 1 remains empty once it hits zero. So $\frac{\mathrm{d}n_1(t)}{\mathrm{d}t} = 0$, or equivalently, $\rho_1 - u_1(t) - u_c(t)c_1 = 0$ when $n_1(t) = 0$.

We can now determine the optimal allocation for points with $n_1(t) = 0$. Class 1 is kept empty, hence an optimal fluid control will maximize the departure rate of class 2. We should therefore maximize $u_2(t)\mu_2 + u_c(t)\mu_2 c_2$ given that $\rho_1 - u_1(t) - u_c(t)c_1 = 0$, $u_1(t) + u_2(t) + u_c(t) = 1$ and $u_j(t) \geq 0$. Solving this we obtain

$$u_c^*(t) = \frac{\rho_1}{c_1},\ u_1^*(t) = 0 \text{ and } u_2^*(t) = 1 - \frac{\rho_1}{c_1},$$

when $n_1(t) = 0$.

Now assume we start at time $t = 0$ in $n(0) = n = (n_1, n_2)$ with $n_1 > 0$ and $n_2 \geq 0$. At some point an optimal trajectory will hit the vertical axis for the first time. This point will be denoted by $\hat{n} = (0, \hat{n}_2)$, see Figure 8.2. Note that the path from n to $\hat{n}$ that first serves class 1 individually and at some point switches to serving both classes in parallel, is always feasible (see the drift vectors in Figure 8.1). Hence, by Lemma 8.3.4 this path is also the optimal path from n to $\hat{n}$. The turning point where the switch occurs is denoted by $b = (b_1, b_2)$, see again Figure 8.2. We can calculate the costs corresponding to a certain turning point b. Let $T(x, y)$ be the time it takes to go from point x to y in the plane. We have $T(n, b) = \frac{n_1 - b_1}{\mu_1 - \lambda_1}$, $T(b, \hat{n}) = \frac{b_1}{\mu_1 c_1 - \lambda_1}$, and

$$T(\hat{n}, 0) = \frac{\hat{n}_2}{u_2\mu_2 + u_c\mu_2 c_2 - \lambda_2} = \frac{\hat{n}_2}{\mu_2 - \mu_2\frac{\rho_1}{c_1}(1 - c_2) - \lambda_2},$$

with $\hat{n}_2 = b_2 + T(b, \hat{n})(\lambda_2 - \mu_2 c_2)$ and $b_2 = n_2 + T(n, b)\lambda_2$. Let $K_n(b_1) = \int_0^\infty (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t$ be the cost of the fluid trajectory going from n to the origin when the turning point is $b = (b_1, b_2)$. Note that $b_2 = n_2 + \frac{n_1 - b_1}{\mu_1 - \lambda_1}\lambda_2$, hence b_2 is uniquely

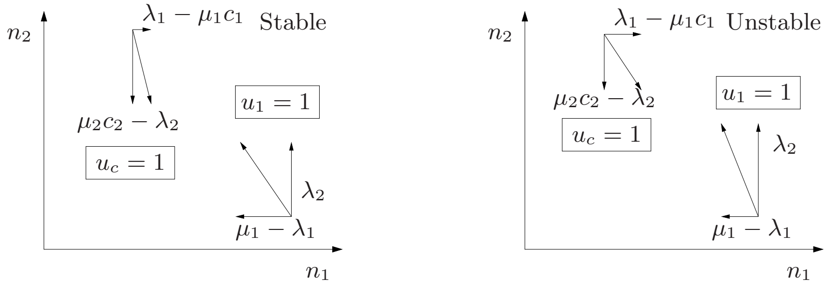


Figure 8.3: Drift vectors for $\rho_1 \geq c_1$ and $\rho_2 < c_2$. Left figure: $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$ and hence there are policies that give a stable system. Right figure: $\rho_1 > 1 - \frac{\rho_2}{c_2}(1-c_1)$ and hence the system is unstable.

determined by b_1 and n. We have

$$\begin{aligned} K_n(b_1) = T(n,b)\Big(\frac{d_1(n_1+b_1)}{2} + \frac{d_2(n_2+b_2)}{2}\Big) + T(b,\hat{n})\Big(\frac{d_1 b_1}{2} + \frac{d_2(b_2+\hat{n}_2)}{2}\Big) \\ +T(\hat{n},0)\frac{d_2\hat{n}_2}{2}. \qquad (8.20) \end{aligned}$$

It can be checked that the function $K_n(b_1)$ is a quadratic function in b_1 and when minimizing the cost in (8.20), the optimal turning point b lies on the line $b_2 = \alpha\frac{\mu_2}{\mu_1}b_1$. Hence, if $n_2(t) < \alpha\frac{\mu_2}{\mu_1}n_1(t)$, then $u_1^*(t) = 1$, and if $n_2(t) \geq \alpha\frac{\mu_2}{\mu_1}n_1(t)$ and $n_1(t) > 0$, then $u_c^*(t) = 1$. This completes the characterization of an optimal control. □

Case $\rho_1 \geq c_1$

When $\rho_1 \geq c_1$, the stability condition is $\rho_2 < c_2$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$ (see (8.4) and (8.5)). Hence $\frac{\rho_2}{1-\rho_1} \leq \frac{c_2-\rho_2}{\rho_1-c_1}$ and the drifts are as in the left picture in Figure 8.3. When $\rho_1 \geq 1 - \frac{\rho_2}{c_2}(1-c_1)$, the system is unstable which corresponds to the picture on the right in Figure 8.3. The optimal fluid control is described in the next proposition, and in Figure 8.4 the optimal trajectory is shown.

Proposition 8.3.6. *Let $d_1\mu_1 > d_1\mu_1c_1 + d_2\mu_2c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 \geq c_1$, $\rho_2 < c_2$, and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$. An optimal policy in the fluid control model is to give priority to class 1, i.e.,*

- *$u_1^*(t) = 1$ if $n_1(t) > 0$.*
- *$u_c^*(t) = \frac{1-\rho_1}{1-c_1}$ and $u_1^*(t) = \frac{\rho_1-c_1}{1-c_1}$ if $n_1(t) = 0$.*

The proof of Proposition 8.3.6 below does not give much insight into the result. Therefore, we first provide some intuition for the fact that the control u^* as defined

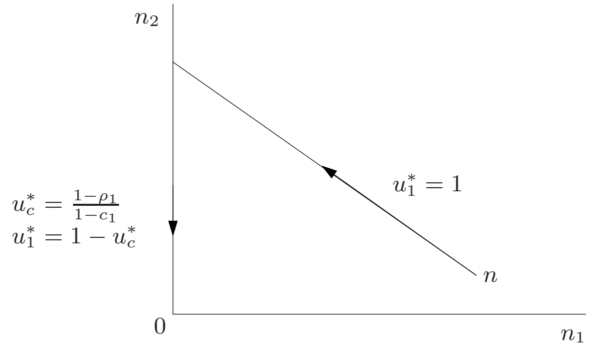


Figure 8.4: Optimal trajectory of the fluid control model when $\rho_1 \geq c_1$.

above, is optimal when $\rho_1 > c_1$: Using Lemma 8.3.4 it can be argued that as long as $n_1(t) > 0$, an optimal action is $u_1^*(t) = 1$. Hence, once $n_1(t) = 0$, this optimal control will keep class 1 empty ($\rho_1 < 1$). An optimal fluid control will now choose allocations $u_j^*(t)$ such that the departure rate for class 2, $u_2(t)\mu_2 + u_c(t)\mu_2 c_2$, is maximized subject to $u_1(t) + u_c(t)c_1 = \rho_1$, $u_1(t) + u_2(t) + u_c(t) = 1$ and $u_j(t) \geq 0$. The unique solution to this is $u_2^*(t) = 0$, $u_1^*(t) = \frac{\rho_1 - c_1}{1-c_1}$ and $u_c^*(t) = \frac{1-\rho_1}{1-c_1}$ when $n_1(t) = 0$.

Proof of Proposition 8.3.6: Consider the control $u^*(t)$ as defined in Proposition 8.3.6. The corresponding trajectory is denoted by $n^*(t)$. Its cost-to-go function is defined as $K_{(t,n)} := \int_t^\infty (d_1 n_1^*(s) + d_2 n_2^*(s) | n(t) = n)\mathrm{d}s = \int_0^\infty (d_1 n_1^*(s) + d_2 n_2^*(s) | n(0) = n)\mathrm{d}s$, for $n = (n_1, n_2) \in \mathbb{R}_+^2$. Hence, we can drop the dependence on t, and write K_n for the cost-to-go starting in state n. A sufficient condition for optimality of $u^*(t)$ is that its cost-to-go function K_n satisfies the "Hamilton-Jacobi-Bellman" partial differential equation:

$$0 = \min_{u \text{ s.t.}(8.14)-(8.16)} \Big(\frac{\partial K_n}{\partial n_1} \cdot (\lambda_1 - \mu_1(u_1 + c_1 u_c)) + \frac{\partial K_n}{\partial n_2} \cdot (\lambda_2 - \mu_2(u_2 + c_2 u_c)) + d_1 n_1 + d_2 n_2 \Big), \qquad (8.21)$$

for all $n_1, n_2 \geq 0$, and that the control $u^*(t)$ is a corresponding minimizing action, [39, Section 5.5]. In the remainder of the proof we show that this is indeed satisfied.

The cost-to-go function is easily derived. Let $\hat{n} = (0, \hat{n}_2)$ denote the point where the trajectory $n^*(t)$ hits the vertical axis, see Figure 8.4. Hence, $\hat{n}_2 = n_2 + n_1 \frac{\lambda_2}{\mu_1 - \lambda_1}$ and also $\hat{n}_2 = T(\hat{n}, 0) \cdot (\mu_2 c_2 \frac{1-\rho_1}{1-c_1} - \lambda_2) = T(\hat{n}, 0) \cdot \mu_2 \frac{c_2}{1-c_1} \cdot (1 - \rho_1 - \frac{\rho_2}{c_2}(1 - c_1))$, with $T(\hat{n}, 0)$ the time it takes to move from $\hat{n}$ to 0. So

$$\begin{aligned} K_n &= T(n,\hat{n})\Big(\frac{d_1 n_1}{2} + \frac{d_2(n_2+\hat{n})}{2}\Big) + T(\hat{n},0)\frac{d_2 \hat{n}_2}{2} \\ &= \frac{n_1}{\mu_1-\lambda_1}\Big(\frac{d_1 n_1}{2} + \frac{d_2(2n_2 + n_1\frac{\lambda_2}{\mu_1-\lambda_1})}{2}\Big) + \frac{d_2(n_2 + n_1\frac{\lambda_2}{\mu_1-\lambda_1})^2}{2\mu_2\frac{c_2}{1-c_1}\cdot(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))}. \end{aligned}$$

In the Hamilton-Jacobi-Bellman equation we are interested in the function

$$\begin{aligned} &\frac{\partial K_n}{\partial n_1}\cdot(\lambda_1-\mu_1(u_1+c_1u_c)) + \frac{\partial K_n}{\partial n_2}\cdot(\lambda_2-\mu_2(u_2+c_2u_c)) + d_1n_1 + d_2n_2 \\ &= (\lambda_1-\mu_1(u_1+c_1u_c))\cdot\Big(\frac{d_1n_1}{\mu_1-\lambda_1} + \frac{d_2}{\mu_1-\lambda_1}(n_1\frac{\lambda_2}{\mu_1-\lambda_1}+n_2) \\ &\quad + \frac{d_2\frac{\lambda_2}{\mu_1-\lambda_1}}{\mu_2\frac{c_2}{1-c_1}\cdot(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))}(n_1\frac{\lambda_2}{\mu_1-\lambda_1}+n_2)\Big) + (\lambda_2-\mu_2(u_2+c_2u_c))\cdot \\ &\quad \Big(\frac{d_2n_1}{\mu_1-\lambda_1} + \frac{d_2}{\mu_2\frac{c_2}{1-c_1}\cdot(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))}(n_1\frac{\lambda_2}{\mu_1-\lambda_1}+n_2)\Big) \\ &\quad + d_1n_1 + d_2n_2 \\ &= \frac{1}{\mu_1-\lambda_1}\cdot\Big(\rho_1(a_{11}n_1+a_{12}n_2) + \rho_2(a_{21}n_1+a_{22}n_2) \\ &\qquad\qquad -n_1(u_1a_{11}+u_c(c_1a_{11}+c_2a_{21})+u_2a_{21}) \\ &\qquad\qquad -n_2(u_1a_{12}+u_c(c_1a_{12}+c_2a_{22})+u_2a_{22})\Big) \\ &\quad + d_1n_1 + d_2n_2, \end{aligned} \tag{8.22}$$

where

$$\begin{aligned} a_{11} &= d_1\mu_1 + d_2\frac{\lambda_2}{1-\rho_1}(1+\frac{\frac{\rho_2}{c_2}(1-c_1)}{(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))}) \\ &= d_1\mu_1 + d_2\mu_2\frac{\rho_2}{1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)}, \\ a_{12} &= d_2\cdot\left(\mu_1 + \mu_1\frac{\frac{\rho_2}{c_2}(1-c_1)}{(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))}\right) = d_2\mu_1\frac{1-\rho_1}{1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)}, \\ a_{21} &= d_2\cdot\left(\mu_2 + \mu_2\frac{\frac{\rho_2}{c_2}(1-c_1)}{1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)}\right) = d_2\mu_2\frac{1-\rho_1}{1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)}, \\ a_{22} &= d_2\frac{\mu_1-\lambda_1}{\frac{c_2}{1-c_1}(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1))} = d_2\mu_1\frac{(1-\rho_1)\frac{1-c_1}{c_2}}{1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)}. \end{aligned}$$

Elementary calculation shows that for $u_1 = 1$, and $u_2 = u_c = 0$, equation (8.22) is equal to zero. In addition, under the conditions as stated in Proposition 8.3.6, it holds that $a_{11} > c_1a_{11} + c_2a_{21} > a_{21}$ and $a_{12} = c_1a_{12} + c_2a_{22} > a_{22}$. Hence, when

$n_1(t) > 0$, the minimizing action is $u_1(t) = 1, u_2(t) = 0, u_c(t) = 0$, which is indeed prescribed by the control strategy u^*.

When $n_1 = 0$, equation (8.22) is equal to

$$n_2 + \frac{1}{\mu_1 - \lambda_1}\Big(n_2(\rho_1 a_{12} + \rho_2 a_{22}) - n_2(u_1 a_{12} + u_c(c_1 a_{12} + c_2 a_{22}) + u_2 a_{22})\Big). \quad (8.23)$$

Again simple calculations show that this is equal to 0 for all u with $u_1 + u_c = 1$ and $u_2 = 0$. Besides $u_1 + u_2 + u_c \leq 1$, we have the restriction $u_1 + c_1 u_c \leq \rho_1$ (because $n_1 = 0$). Since $a_{12} = c_1 a_{12} + c_2 a_{22} > a_{22}$, any control with $u_1 + u_c = 1$ and $u_2 = 0$ such that $u_1 + c_1 u_c \leq \rho_1$, will minimize (8.23). The control $u_1^*(t) = \frac{\rho_1 - c_1}{1 - c_1}$, $u_c^*(t) = \frac{1-\rho_1}{1-c_1}$ and $u_2^*(t) = 0$ is therefore indeed a minimizing action. □

8.3.2 Asymptotically fluid-optimal policies for $\rho_1 \neq c_1$

In this section we discuss the theoretical foundations that justify the use of the optimal control in the fluid model as proxies for the optimal policies in the stochastic model. In particular, we prove that under a fluid scaling, the stochastic processes of the numbers of users under certain switching-curve policies, converge to the optimal fluid trajectory $n^*(t)$ as determined in Section 8.3.1. Using the latter, we then show that these switching-curve policies are asymptotically fluid-optimal in the stochastic model.

On a common probability space we construct a sequence of processes depending on the initial state. To be precise, for a given policy π we let $N_i^{\pi,r}(t)$ denote the number of class-i users at time t when the initial state equals $N_i^r(0) = rn_i$, $i = 1, 2$, with $r \in \mathbb{N}$. All processes $N^{\pi,r}(t)$ share the same sequences of arrivals and service requirements. For a given policy π, denote by $T_I^{\pi,r}(t)$ the cumulative amount of time during the interval $(0, t]$ that neither class is served, by $T_i^{\pi,r}(t)$ the cumulative amount of time that was spent serving class i individually, $i = 1, 2$, and by $T_c^{\pi,r}(t)$ the cumulative amount of time that was spent serving classes 1 and 2 in parallel. Then, $T_I^{\pi,r}(t) + T_1^{\pi,r}(t) + T_2^{\pi,r}(t) + T_c^{\pi,r}(t) = t$, and

$$N_i^{\pi,r}(t) = rn_i + E_i(t) - F_i(T_i^{\pi,r}(t)) - F_{c,i}(T_c^{\pi,r}(t)), \quad i = 1, 2, \quad (8.24)$$

with $E_i(t)$ a Poisson process with rate λ_i, $F_i(\cdot)$ a Poisson process with rate μ_i and $F_{c,i}(\cdot)$ a Poisson process with rate $c_i\mu_i$, [48].

We will be interested in the processes under the fluid scaling, i.e., both time and space are scaled linearly by the parameter r:

$$\overline{N}_i^{\pi,r}(t) := \frac{N_i^{\pi,r}(rt)}{r} \quad \text{and} \quad \overline{T}_j^{\pi,r}(t) := \frac{T_j^{\pi,r}(rt)}{r}.$$

Limit points for $\overline{N}_i^{\pi,r}(t)$ and $\overline{T}_j^{\pi,r}(t)$ are described in the next lemma.

Lemma 8.3.7. *For almost all sample paths ω there exists a subsequence r_k such that*

$$\lim_{k\to\infty} \overline{N}_i^{\pi,r_k}(t) = \overline{N}_i^{\pi}(t), \quad i = 1,2, \ u.o.c.,$$
$$\lim_{k\to\infty} \overline{T}_j^{\pi,r_k}(t) = \overline{T}_j^{\pi}(t), \quad j = I,1,2,c, \quad u.o.c.$$

Furthermore, $(\overline{N}^{\pi}, \overline{T}^{\pi})$ satisfies for $i = 1,2,\ j = I,1,2,c$,

$$\overline{N}_i^{\pi}(t) = n_i + \lambda_i t - \mu_i \overline{T}_i^{\pi}(t) - \mu_i c_i \overline{T}_c^{\pi}(t), \tag{8.25}$$

$\overline{N}_i(t) \geq 0$, $\overline{T}_j^{\pi}(0) = 0$, $\overline{T}_I^{\pi}(t) + \overline{T}_1^{\pi}(t) + \overline{T}_2^{\pi}(t) + \overline{T}_c^{\pi}(t) = t$, and $\overline{T}_j^{\pi}(t)$ are non-decreasing and Lipschitz continuous functions.

The notation u.o.c. stands for uniform convergence on compact sets. We call the processes $\overline{T}_j^{\pi}(t), j = I,1,2,c$, and $\overline{N}_i^{\pi}(t), i = 1,2$ (as obtained in Lemma 8.3.7) fluid limits for initial fluid level n and policy π.

Proof of Lemma 8.3.7: Making use of (8.24) and the fact that $\overline{T}_j^{\pi,r}(t), j = 1,2,c$, is Lipschitz continuous with a constant less than or equal to 1, the proof follows similarly as that of [44, Theorem 4.1]. Note that the Poisson assumptions are in fact not needed for the result of this lemma to hold. □

Similar to Chapter 5, we take as cost in the stochastic model $\mathbb{E}\Big(\int_0^D (d_1 N_1^{\pi,r}(t) + d_2 N_2^{\pi,r}(t))\mathrm{d}t\Big)$, with $D > 0$. As $r \to \infty$, this will tend to infinity. In order to obtain a non-trivial limit we divide the cost by r^2 and consider a horizon that grows linearly in r. So we are interested in

$$\mathbb{E}\Big(\int_0^{r\cdot D} \frac{d_1 N_1^{\pi,r}(t) + d_2 N_2^{\pi,r}(t)}{r^2}\mathrm{d}t\Big) = \mathbb{E}\Big(\int_0^D (d_1 \overline{N}_1^{\pi,r}(t) + d_2 \overline{N}_2^{\pi,r}(t))\mathrm{d}t\Big).$$

We have the following lower bound on the scaled cost.

Lemma 8.3.8. *For any policy π we have*

$$\begin{aligned}\liminf_{r\to\infty} \mathbb{E}\Big(\int_0^D (d_1 \overline{N}_1^{\pi,r}(t) + d_2 \overline{N}_2^{\pi,r}(t))\mathrm{d}t\Big) \quad &\geq \quad \int_0^D (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t \\ &= \quad \int_0^\infty (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t,\end{aligned}$$

whenever $D \geq H(d_1 n_1 + d_2 n_2)$. Here $n^*(t)$ represents an optimal solution of (8.18) for initial state n and $H(\cdot)$ is as defined in Lemma 8.3.2.

Proof: Using Lemmas 8.3.2 and 8.3.7, the proof follows exactly the same steps as the proof of Lemma 5.2.8 and will therefore not be included here. □

As described in Section 1.6.3, a policy is asymptotically fluid-optimal when the lower bound is obtained. Hence, policy π^* is asymptotically fluid-optimal when

$$\lim_{r\to\infty} \mathbb{E}\Big(\int_0^D (d_1 \overline{N}_1^{\pi^*,r}(t) + d_2 \overline{N}_2^{\pi^*,r}(t))\mathrm{d}t\Big) = \int_0^\infty (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t,$$

with $D \geq H(d_1 n_1 + d_2 n_2)$ and $n^*(t)$ an optimal solution of (8.18) for initial state n. In the remainder of this section we characterize these policies.

Case $\rho_1 < c_1$

We first consider the case $\rho_1 < c_1$. In Proposition 8.3.5 we found that the optimal switching curve for the fluid control problem was given by $h(x_1) = \alpha\frac{\mu_2}{\mu_1}x_1$, with α as defined in (8.19). In the following lemma we show that under this switching curve, the fluid-scaled processes of the original stochastic model have a unique limit, which is described by the optimal trajectory of the fluid control model. The proof may be found in Appendix 8.E.

Lemma 8.3.9. *Assume $c_1 + c_2 > 1$, $\rho_1 < c_1$ and $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$. Denote by π^* the policy with switching curve $h(x_1) = \alpha\frac{\mu_2}{\mu_1}x_1$, with α as defined in (8.19). The functions $\overline{T}_j^{\pi^*}(t)$ are differentiable almost everywhere, and for each regular point t it holds that*

$$\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t)}{\mathrm{d}t} = 1, \ \textit{if} \ \overline{N}_2^{\pi^*}(t) < \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(t), \tag{8.26}$$

$$\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t} = 1, \ \textit{if} \ \overline{N}_2^{\pi^*}(t) \geq \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(t) \ \textit{and} \ \overline{N}_1^{\pi^*}(t) > 0, \tag{8.27}$$

$$\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t} = \frac{\rho_1}{c_1} \ \textit{and} \ \frac{\mathrm{d}\overline{T}_2^{\pi^*}(t)}{\mathrm{d}t} = 1 - \frac{\rho_1}{c_1}, \quad \textit{if} \ \overline{N}_1^{\pi^*}(t) = 0 \ \textit{and} \ \overline{N}_2^{\pi^*}(t) > 0, \tag{8.28}$$

and $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_2^{\pi^*}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_I^{\pi^*}(t)}{\mathrm{d}t} = 1.$

In particular, $\overline{N}^{\pi^*}(t)$ is uniquely determined by

$$\overline{N}^{\pi^*}(t) = n^*(t), \tag{8.29}$$

with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$ as defined in Proposition 8.3.5.

From Lemma 8.3.9 we obtain that the linear switching curve provides a policy that is asymptotically fluid-optimal for the original stochastic model.

Proposition 8.3.10. *Let $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$ and $c_1 + c_2 > 1$. If $\rho_1 < c_1$ and $\rho_2 < 1 - \frac{\rho_1}{c_1}(1 - c_2)$, then the policy π^* with switching curve $h(x_1) = \alpha\frac{\mu_2}{\mu_1}x_1$ is asymptotically fluid-optimal, with α as defined in (8.19).*

Proof: For a given sample path ω, let r_k be a subsequence such that

$$\liminf_{r\to\infty}\int_0^D (d_1\overline{N}_1^{\pi^*,r}(t)+d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t=\lim_{k\to\infty}\int_0^D (d_1\overline{N}_1^{\pi^*,r_k}(t)+d_2\overline{N}_2^{\pi^*,r_k}(t))\mathrm{d}t.$$

From Lemma 8.3.7 it follows that for almost all ω there exists a subsequence r_{k_l} of r_k such that $\lim_{l\to\infty}\overline{N}^{\pi^*,r_{k_l}}(t)=\overline{N}^{\pi^*}(t)$, u.o.c.. Since every fluid limit $\overline{N}^{\pi^*}(t)$ coincides with the optimal fluid control solution $n^*(t)$ (see equation (8.29)) we obtain $\lim_{l\to\infty}\overline{N}_i^{\pi^*,r_{k_l}}(t)=n_i^*(t)$, $i=1,2$. Since the functions $\overline{N}_i^{\pi^*,r_{k_l}}(t)$, $i=1,2$, converge uniformly on the set $[0,D]$, we can interchange the limit and the integral, so that

$$\begin{aligned}&\liminf_{r\to\infty}\int_0^D (d_1\overline{N}_1^{\pi^*,r}(t)+d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t\\&=\lim_{l\to\infty}\int_0^D (d_1\overline{N}_1^{\pi^*,r_{k_l}}(t)+d_2\overline{N}_2^{\pi^*,r_{k_l}}(t))\mathrm{d}t=\int_0^D(d_1n_1^*(t)+d_2n_2^*(t))\mathrm{d}t.\end{aligned}$$

The same holds for the lim sup and we can conclude that for almost all ω,

$$\lim_{r\to\infty}\int_0^D (d_1\overline{N}_1^{\pi^*,r}(t)+d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t=\int_0^D(d_1n_1^*(t)+d_2n_2^*(t))\mathrm{d}t. \tag{8.30}$$

We also have that $\int_0^D(d_1\overline{N}_1^{\pi^*,r}(t)+d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t$ is uniformly integrable. This follows from the same argument as in the proof of [44, Lemma 4.5] (see the proof of Proposition 5.2.10 for more details). We obtain

$$\begin{aligned}&\limsup_{r\to\infty}\mathbb{E}\Big(\int_0^D (d_1\overline{N}_1^{\pi^*,r}(t)+d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t\Big)\\&=\lim_{m\to\infty}\mathbb{E}\Big(\int_0^D (d_1\overline{N}_1^{\pi^*,r_m}(t)+d_2\overline{N}_2^{\pi^*,r_m}(t))\mathrm{d}t\Big)\\&=\mathbb{E}\Big(\int_0^D(d_1n_1^*(t)+d_2n_2^*(t))\mathrm{d}t\Big)=\int_0^D(d_1n_1^*(t)+d_2n_2^*(t))\mathrm{d}t,\end{aligned} \tag{8.31}$$

where in the second step we used (8.30) and uniform integrability to interchange the limit and expectation (see [27, Theorem 3.5]).

Equation (8.31) holds in particular for $D>H(d_1n_1+d_2n_2)$. Together with Lemma 8.3.8 we can conclude that π^* is asymptotically fluid-optimal. □

Case $\rho_1>c_1$

In Proposition 8.3.6 we found that for the fluid control problem it is optimal to give class 1 priority whenever present when $\rho_1\geq c_1$. A straightforward translation of this policy to the original stochastic model would be to give preemptive priority to class-1 users. However, the stability conditions under this policy are $\rho_1+\rho_2<1$, which are more stringent than the necessary stability conditions as given in (8.4) and (8.5). Hence, a more precise interpretation of the optimal fluid control is needed to avoid an unstable system.

Note that the optimal policy in the fluid control model can keep the system stable under (8.4) and (8.5), since on the vertical axis the fluid model partly serves class 1 individually and partly serves both classes in parallel. This suggests that for the stochastic model we should as well serve classes 1 and 2 in parallel when the process moves close to the vertical axis. So there is a switching curve in the original model that lies close to the vertical axis such that it is non-observable in the fluid limit. However, under the optimal fluid control, class-2 users are never served individually ($u_2^*(t) = 0$). Hence, in the stochastic model, the switching curve should be such that, after fluid scaling, the time that the stochastic process spends on the vertical axis tends to zero. This indicates that the curve should not be too close to the vertical axis.

In the next proposition we state that the above is achieved by a policy with a switching curve of the shape $h(x_1) = \mathrm{e}^{x_1/\gamma}$ (for γ large enough), i.e., such a policy is asymptotically fluid-optimal. We make no claim for small γ. In Section 8.4 reasonable estimates for γ are obtained. Note that the case $\rho_1 = c_1$ is excluded from the proposition. This setting will be discussed in Section 8.5.

Proposition 8.3.11. *Let $d_1\mu_1 > d_1\mu_1 c_1 + d_2\mu_2 c_2$ and $c_1 + c_2 > 1$. Assume $\rho_1 > c_1$, $\rho_2 < c_2$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1 - c_1)$. The policy $\pi^*(\gamma)$ with switching curve $h(x_1) = \mathrm{e}^{x_1/\gamma}$ is asymptotically fluid-optimal for $\gamma > 0$ large enough.*
In addition, the fluid limit is uniquely determined by $\overline{N}^{\pi^*(\gamma)}(t) = n^*(t)$, with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$ as defined in Proposition 8.3.6.

Proof: Let $\overline{N}_i^{\pi^*(\gamma)}(t)$, $i = 1, 2$, $\overline{T}_j^{\pi^*(\gamma)}(t)$, $j = I, 1, 2, c$, be a fluid limit of policy $\pi^*(\gamma)$. The function $\overline{T}_j^{\pi^*(\gamma)}(\cdot)$ is absolutely continuous. Using the same techniques as in [53, Section 7] it follows that for each regular point t, the derivatives satisfy:

$$\frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = 1, \quad \text{if } \overline{N}_1^{\pi^*(\gamma)}(t) > 0, \tag{8.32}$$

$$\frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = \frac{\rho_1 - c_1}{1 - c_1} \text{ and } \frac{\mathrm{d}\overline{T}_c^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = \frac{1 - \rho_1}{1 - c_1}, \quad \text{if } \overline{N}_1^{\pi^*(\gamma)}(t) = 0, \tag{8.33}$$

for γ large enough. In fact, this can be checked using the following correspondence between the processes ξ_t, x_t and $T(t)$ in [53], and our equivalents: $\xi_t^3 = N_1^{\pi^*(\gamma)}(t)$, $\xi_t^1 = N_2^{\pi^*(\gamma)}(t)$, $x_t^3 = \overline{N}_1^{\pi^*(\gamma)}(t)$, $x_t^1 = \overline{N}_2^{\pi^*(\gamma)}(t)$, $T_{01}(t) = \overline{T}_1^{\pi^*(\gamma)}(t)$ and $T_{11}(t) = \overline{T}_c^{\pi^*(\gamma)}(t)$, and mapping our parameters $c_1, c_2, \mu_1, \mu_2, \lambda_1$ and λ_2, such that the drifts in the interior of Figure 4 in [53] correspond to the drifts in our Figure 8.3. Note that the drifts on the boundaries cannot be matched, but this does not influence the fluid analysis. From (A.12) in [53] it follows that for γ large enough, $\frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = 1$ if $\overline{N}_1^{\pi^*(\gamma)}(t) > 0$, hence we obtain (8.32). In addition, for γ large enough it follows from (A.13) in [53] that

$$\frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\overline{T}_c^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = 1, \quad \text{if } \overline{N}_1^{\pi^*(\gamma)}(t) = 0. \tag{8.34}$$

Since $\frac{1}{\mu_1}\frac{\mathrm{d}\overline{N}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = \rho_1 - \frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} - c_1\frac{\mathrm{d}\overline{T}_c^{\pi^*(\gamma)}(t)}{\mathrm{d}t}$ and $\rho_1 < 1 = \frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t}$ when $\overline{N}_1^{\pi^*(\gamma)}(t) > 0$, class 1 remains empty once it hits zero. Hence, if $\overline{N}_1^{\pi^*(\gamma)}(t) = 0$, then $\frac{1}{\mu_1}\frac{\mathrm{d}\overline{N}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = 0$, i.e., $\frac{\mathrm{d}\overline{T}_1^{\pi^*(\gamma)}(t)}{\mathrm{d}t} + c_1\frac{\mathrm{d}\overline{T}_c^{\pi^*(\gamma)}(t)}{\mathrm{d}t} = \rho_1$, which, together with (8.34), implies (8.33).

From (8.25), (8.32) and (8.33) it follows that $\overline{N}_i^{\pi^*(\gamma)}(t)$ is uniquely determined. Using the correspondence $u_j^*(t) = \frac{\mathrm{d}\overline{T}_j^{\pi^*(\gamma)}(t)}{\mathrm{d}t}$, $j = 1, 2, c$, it follows from Proposition 8.3.6 that $\overline{N}^{\pi^*(\gamma)}(t) = n^*(t)$, with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$ as defined in Proposition 8.3.6. The remainder of the proof is similar to the proof of Proposition 8.3.10. □

8.4 Discussion for the case $\rho_1 > c_1$

When $\rho_1 < c_1$, an asymptotically fluid-optimal policy can be characterized by a linear switching curve for which the slope has been exactly determined. When $\rho_1 > c_1$, Proposition 8.3.11 proves that an exponential switching curve $h(x_1) = \mathrm{e}^{x_1/\gamma}$ is asymptotically fluid-optimal for any $\gamma > 0$ that is *large enough*. However, it is not straightforward to determine a good value for γ. The purpose of the remainder of this section is to determine a reasonable *rule of thumb*.

An asymptotically fluid-optimal policy π^* satisfies

$$\begin{aligned}\mathbb{E}\Big(\int_0^{r\cdot D}(d_1N_1^{\pi^*,r}(t) + d_2N_2^{\pi^*,r}(t))\mathrm{d}t\Big) &= r^2\cdot\mathbb{E}\Big(\int_0^{D}(d_1\overline{N}_1^{\pi^*,r}(t) + d_2\overline{N}_2^{\pi^*,r}(t))\mathrm{d}t\Big)\\ &= r^2\int_0^\infty(d_1n_1^*(t) + d_2n_2^*(t))\mathrm{d}t + \mathrm{o}(r^2).\end{aligned}$$

Hence, one way to determine a reasonable value for γ is by choosing that value for γ that minimizes the next order term, $\mathrm{o}(r^2)$. For the discrete-time version of our model it is possible to find an estimate of this term, using the techniques of [54].

Consider a discrete-time system with Bernoulli arrivals. In an interval of length Δ, a class-i user arrives with probability $\lambda_i\Delta$, and it leaves the system with probability $\mu_is_i\Delta$, $s \in S$ (with S the capacity region as defined in Section 8.1). We are interested in policies with exponential switching curves with parameter γ, and denote the state in the interval k by $N_i^\gamma(k)$, $i = 1, 2$. Hence, $s(k) = (1, 0)$ if $N_2^\gamma(k) < \mathrm{e}^{N_1^\gamma(k)/\gamma}$, and $s(k) = (c_1, c_2)$ if $N_2^\gamma(k) \geq \mathrm{e}^{N_1^\gamma(k)/\gamma}$ and $N_1^\gamma(k) > 0$. When $\Delta \to 0$, this approximates the continuous-time system with Poisson arrivals and exponentially distributed service requirements. (The user departure rate in the discrete model is μ_is_i, which is equal to the user departure rate in the stochastic model.)

Following the reasoning in [54] we consider different realizations of the queue length process, indexed by a superscript $r \in \mathbb{N}$. We take as initial point $n^r = (\gamma \ln[rn_2], [rn_2])$ and as time horizon $r \cdot D$ for some fixed D with $n_2 > D$. We then write $\mathbb{E}(\sum_{k=0}^{r \cdot D} N_1^{\gamma,r}(k) + N_2^{\gamma,r}(k)) = \sum_{k=1}^{4} V_k^{\gamma}(n^r)$ with

$$
\begin{aligned}
V_1^{\gamma}(n^r) &= d_1 \sum_{k=0}^{r \cdot D} \mathbb{E}(N_1^{\gamma,r}(k)), \\
V_2^{\gamma}(n^r) &= d_2 \sum_{k=0}^{r \cdot D} (n_2^r + k(\lambda_2 - \frac{1-\rho_1}{1-c_1}\mu_2 c_2)), \\
V_3^{\gamma}(n^r) &= d_2 \sum_{k=0}^{r \cdot D} \frac{\mu_2 c_2}{\mu_1(1-c_1)}(n_1^r - \mathbb{E}(N_1^{\gamma,r}(k))), \\
V_4^{\gamma}(n^r) &= d_2 \sum_{k=0}^{r \cdot D} \mu_2 \frac{c_1 + c_2 - 1}{1-c_1} \mathbb{E}(v_k^{\gamma,r}),
\end{aligned}
$$

where $v_k^{\gamma,r} = \sum_{m=0}^{k-1} \mathbf{1}_{(N_1^{\gamma,r}(m)=0)}$ is the number of times the process serves class 2 individually. The asymptotic behavior of $v_k^{\gamma,r}$, and hence of $V_4^{\gamma}(n^r)$, involves studying $\mathbb{P}(N_1^r(rt) = 0)$ as $r \to \infty$. Since $c_1 < \rho_1 < 1$ and $\rho_1 < 1 - \frac{\rho_2}{c_2}(1-c_1)$, the drift of the process (after fluid scaling) is away from the vertical axis and towards the switching curve. Therefore, we can use the large-deviation results in [54] in order to obtain the asymptotic behavior of $V_4^{\gamma}(n^r)$. In particular, it can be shown that

$$
\begin{aligned}
V_1^{\gamma}(n^r) &= d_1 D r \gamma \ln(r) + \mathrm{O}(r), \\
V_2^{\gamma}(n^r) &= r^2 \int_0^{\infty} (d_1 n_1^*(t) + d_2 n_2^*(t)) \mathrm{d}t + \mathrm{O}(r), \\
V_3^{\gamma}(n^r) &= \mathrm{O}(r), \\
V_4^{\gamma}(n^r) &= d_2 \mu_2 \frac{c_1 + c_2 - 1}{1-c_1} \cdot r^{2-\beta(\Delta)\gamma + \mathrm{o}(1)},
\end{aligned}
$$

as $r \to \infty$, with $\beta(\Delta) = \ln(\frac{\rho_1}{c_1}\frac{1-\mu_1 c_1 \Delta}{1-\lambda_1 \Delta})$, and $n^*(t)$ the optimal fluid trajectory corresponding to initial state $(0, n_2)$. Obviously, $V_2^{\gamma}(n^r)$ represents the first-order term, which coincides with what we expected. As said before, we are interested in the order of the next term. This term is of the order $r^{2-\beta(\Delta)\gamma}$ when $\gamma < 1/\beta(\Delta)$, and of the order $r\ln(r)$ when $\gamma \geq 1/\beta(\Delta)$. This indicates that setting $\gamma \geq 1/\beta(\Delta)$ gives good second-order asymptotics. In our numerical experiments in Section 8.7.1 (for the continuous-time setting), this is indeed observed for large loads, as the performance severely degrades for small values of γ (see Figure 8.12). For large values of γ, performance is also suboptimal: When $\gamma > 1/\beta(\Delta)$, the second-order term is given by $D\gamma r \ln r$, so that it is not attractive to choose γ too large either. However, performance turns out to be less sensitive to small changes in γ for large values of γ, see also our numerical experiments in Section 8.7.1.

Letting $\Delta \to 0$, we have $\lim_{\Delta \to 0} \beta(\Delta) = \ln(\frac{\rho_1}{c_1})$. Choosing as estimate $\gamma =$

$1/\ln(\frac{\rho_1}{c_1})$ in the continuous-time system proves to be a reasonable *rule of thumb* in all our experiments, see Section 8.7.1.

8.5 Discussion for the case $\rho_1 = c_1$

In Section 8.3.2 we obtained asymptotically fluid-optimal policies when $\rho_1 \neq c_1$. In this section we discuss the unsolved case, $\rho_1 = c_1$. The optimal fluid control is to serve class 1 individually whenever possible, i.e., the switching curve lies on the vertical axis. As mentioned before, this policy can be unnecessarily unstable. Hence, it is expected that a stochastically optimal policy will have a switching curve, which is non-observable in the fluid limit.

By value iteration we computed numerically an average-cost optimal policy in the case $\rho_1 = c_1$. (In Section 8.7 we will compute such policies for scenarios with $\rho_1 \neq c_1$.) We found that an optimal policy is characterized by a switching curve that resembles a quadratic function, see Figure 8.5. In this section we explain this quadratic shape.

We first describe the stochastic behavior below and above the switching curve. In states below the switching curve, the drift is towards this curve. Let us now concentrate on the free process that corresponds to the behavior in states above the switching curve ($N_2(t) \geq h(N_1(t))$ and $N_1(t) > 0$), i.e., both classes are served in parallel. As in Section 5.3.2, the following properties of the free process can be derived: Since $\rho_1 = c_1$, the fluid-scaled number of class-1 users has zero drift and the diffusion-scaled number of class-1 users converges to a Brownian motion. In addition, the fluid-scaled number of class-2 users has a negative drift $\lambda_2 - \mu_2 c_2$.

The optimal fluid control indicates that a switching curve should be such that it is close to the vertical axis. Letting the switching curve be too close, however, poses the risk that the process spends too much time on the vertical axis. The

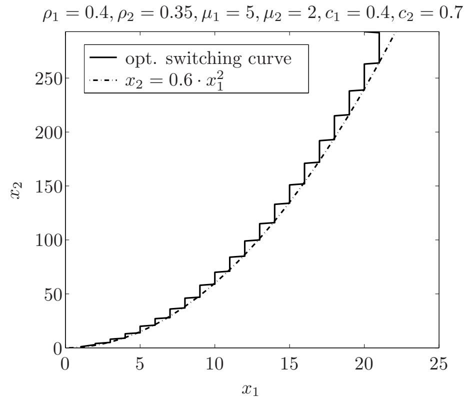


Figure 8.5: Optimal switching curve and a quadratic approximation when $\rho_1 = c_1$.

latter should be avoided, since the optimal fluid control indicates that, under a fluid scaling, no time should be spent on serving class 2 individually. As described above, the fluctuations in the number of class 1 users in linear time $\mathrm{O}(r)$ are of the order $\mathrm{O}(\sqrt{r})$, while the number of class-2 users decreases linearly in time. This indicates that a switching curve of the shape $h(x_1) = kx_1^2$, $k > 0$, strikes the right balance between these two goals. For comparison: a linear switching curve would be impossible to reach, therefore the policy would not profit from serving class 1 individually. On the other hand, under an exponential switching curve it is too easy to move to the vertical axis, thus risking to be a considerable amount of time on the vertical axis.

8.6 Optimality in heavy traffic

One of the goals of this chapter is to describe policies that approximate the optimal policy rather well (in cases where the optimal policy could not be determined explicitly). In Section 8.3 we did so by considering a simpler (fluid) model that only took into account the mean drifts. We proved that certain policies are asymptotically fluid-optimal, and therefore are potentially close to optimal in the original stochastic model as well. In this section we discuss another approach to obtain policies that are in some sense approximately optimal: We review optimality results available in the literature for a heavy-traffic regime. These results can be used as approximations for the original system when the load is rather high, however, there is no guarantee for the performance of these policies in moderately-loaded systems. Therefore, in Section 8.7 we numerically compare (under moderate load conditions) the performance of the policies that are optimal in heavy traffic with our asymptotically fluid-optimal policies. Note that both of these policies are motivated by a certain asymptotic regime, and beforehand it is unclear how well they perform outside these regimes.

The maximum stability conditions of the parallel two-server model are given in (8.4) and (8.5). Equivalently, we may say that the system can be kept stable when the vector $\vec{\lambda} = (\lambda_1, \lambda_2)$ lies in the interior of the stability set as depicted in

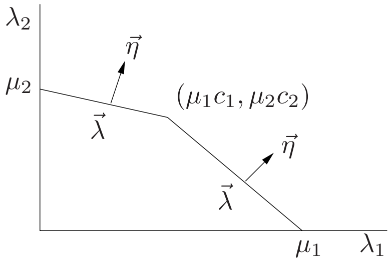


Figure 8.6: Stability set.

Figure 8.6. The system is said to be in heavy traffic when the vector $\vec{\lambda}$ lies on the northeast boundary of the stability set. As mentioned in Section 1.5.3, policies that are in some sense asymptotically optimal in a heavy-traffic setting have been investigated in the case of complete resource pooling. The complete resource pooling condition is satisfied if the outer normal, $\vec{\eta}$, to the stability set at that $\vec{\lambda}$ is unique up to scaling and all its coordinates are strictly positive. Hence, $\lambda_i > 0$ for $i = 1, 2$ and $\vec{\lambda} \neq (\mu_1 c_1, \mu_2 c_2)$. More precisely, the parameters of a heavily-loaded system under the resource pooling condition correspond to one of the following two regions:

- Region A: $\lambda_2 = \mu_2 - \frac{\lambda_1}{\mu_1 c_1}\mu_2(1 - c_2)$, $\lambda_2 > \mu_2 c_2$, and $\mu_1 c_1 > \lambda_1 > 0$, i.e., $\rho_2 = 1 - \frac{\rho_1}{c_1}(1 - c_2)$, $\rho_2 > c_2$, and $c_1 > \rho_1 > 0$. The outer normal vector to a point in this region is $\vec{\eta} = (\mu_2(1 - c_2), \mu_1 c_1)$.
- Region B: $\lambda_1 = \mu_1 - \frac{\lambda_2}{\mu_2 c_2}\mu_1(1 - c_1)$, $\lambda_1 > \mu_1 c_1$, and $\mu_2 c_2 > \lambda_2 > 0$, i.e., $\rho_1 = 1 - \frac{\rho_2}{c_2}(1 - c_1)$, $\rho_1 > c_1$, and $c_2 > \rho_2 > 0$. The outer normal vector to a point in this region is $\vec{\eta} = (\mu_2 c_2, \mu_1(1 - c_1))$.

In Section 8.6.1 we briefly state the results from [19, 20]. There, the authors prove that threshold-based policies asymptotically minimize the (scaled) holding cost in heavy traffic. In Section 8.6.2 we recall the definition of Max-Weight policies and describe the results concerning their behavior in heavy traffic as obtained in [89, 132]. The model studied in [19, 20, 89, 132] is in fact a slight variation of the model we consider in this chapter. First, a server cannot have two or more users of the same class in service. (Note that this restriction has no impact in the case of exponential service requirements.) In addition, once a server starts serving a user, this user has to obtain its full service from this server. Finally, their model has slightly different behavior near the boundaries: when $N_i(t) = 1$, their model can have a departure rate of at most $\mu_i c_i$ for class i, since a single user cannot be served simultaneously by the two servers. In the model we consider, we can have a departure rate of μ_i.

8.6.1 Threshold policies

In [19, 20], the parallel-server model is investigated with an arbitrary number of servers and classes, and i.i.d. inter-arrival times and service requirements. In this section we collect results specific for the parallel two-server model.

In [19, 20], the authors consider a sequence of parameters indexed by r, μ_i^r and λ_i^r ($\rho_i^r = \frac{\lambda_i^r}{\mu_i^r}$), with $\lambda_i^r \to \lambda_i, \mu_i^r \to \mu_i$ such that $\lambda_1, \lambda_2, \mu_1$ and μ_2 correspond either to Region A or Region B. An additional condition involves the rate at which the system converges:

$$\lim_{r\to\infty} \sqrt{r}\mu_i^r(\rho_i^r - \rho_i) = \theta_i, \quad \text{with} \ \ \theta_i \in \mathbb{R}, \ i = 1, 2.$$

Let $N_i^{\pi,r}(t)$ be the number of class-i users in the r-th system under policy π, and let $\hat{N}_i^{\pi,r}(t) = \frac{N_i^{\pi,r}(rt)}{\sqrt{r}}$ be the diffusion-scaled number of class-i users. It is assumed that the system is initially empty. Define $\hat{J}^r(\pi) := \mathbb{E}(\int_0^\infty \mathrm{e}^{-\xi t}(\hat{N}_1^{\pi,r}(t) + \hat{N}_2^{\pi,r}(t))\mathrm{d}t)$

where $\xi > 0$ is a constant. In [19, 20], a sequence of policies $\tilde{\pi}^r$ is called *asymptotically optimal in heavy traffic* when $\lim_{r\to\infty} \hat{J}^r(\tilde{\pi}^r) \leq \liminf_{r\to\infty} \hat{J}^r(\pi^r)$ for any sequence of policies π^r.

In case $d_1\mu_1 \leq d_1\mu_1c_1 + d_2\mu_2c_2$, they prove that an asymptotically optimal policy in heavy traffic is to serve both classes in parallel whenever possible. For $d_1\mu_1 > d_1\mu_1c_1 + d_2\mu_2c_2$ the following result holds:

Proposition 8.6.1 ([20])**.** *Assume $d_1\mu_1 > d_1\mu_1c_1 + d_2\mu_2c_2$ and $c_1 + c_2 > 1$, and consider a heavy-traffic setting with complete resource pooling.*

- *If (ρ_1, ρ_2) corresponds to Region A, then the policy that serves classes 1 and 2 in parallel whenever possible, is an asymptotically optimal policy in heavy traffic.*
- *If (ρ_1, ρ_2) corresponds to Region B, then the sequence of threshold policies that serves class 1 individually when $N_1(t) > h \cdot \ln(\sqrt{r})$ (with $h > 0$ large enough), and that otherwise serves classes 1 and 2 in parallel, is asymptotically optimal in heavy traffic.*

Denote by $Th(r)$ the minimum value for the threshold Th such that the r-th system is stable under the threshold policy that serves class 1 individually when $N_1(t) > Th$ and serves classes 1 and 2 in parallel otherwise. In [137] it is shown that any threshold Th with $Th > Th(r)$, makes the r-th system stable. In addition, $Th(r)/\ln(\sqrt{r}) \to \hat{h}$ for some constant $\hat{h} > 0$. This shows that the threshold $h\cdot\ln(\sqrt{r})$ in the above proposition is of a minimum order.

In Section 8.7.1 we will evaluate the performance of threshold-based policies in the moderately-loaded case, and compare it with the optimal policy found numerically, and with the asymptotically fluid-optimal policies.

8.6.2 Max-Weight policies

The Max-Weight policies are defined in Section 1.5.2. In this section we summarize the heavy-traffic results on Max-Weight policies as described in [89, 132]. We like to emphasize that an important property of the Max-Weight policies is that they maintain a stable system under the maximum stability conditions [132].

The authors of [89] consider a parallel-server model with K classes of users and L servers. They assume i.i.d. inter-arrival times and service requirements and consider a sequence of systems indexed by r, λ_i^r, with $\lambda_i^r \to \lambda_i$, while keeping the parameters of the service requirements fixed. The parameters λ_i, $i = 1, \ldots, K$, are such that the system is in heavy traffic and the complete resource pooling condition is satisfied. In addition, $\lim_{r\to\infty} \sqrt{r}(\lambda_i^r - \lambda_i) = \theta_i$, with $\theta_i \in \mathbb{R}$, $i = 1, \ldots, K$. The initial state converges under the diffusion scaling such that $\lim_{r\to\infty} \frac{N_i^r(0)}{\sqrt{r}} = m_i$ with $(\gamma_1 m_1^\beta, \ldots, \gamma_K m_K^\beta)$ proportional to $\vec{\eta}$. As before, $\vec{\eta}$ is the outer normal vector to the stability set at $\vec{\lambda}$.

The next proposition states the results for the Max-Weight policy, which is denoted by MW. In particular, for a heavy traffic setting it states that the Max-

Weight policy minimizes (under diffusion scaling) both the cost, $\sum_i C_i(N_i(t)) = \sum_i \gamma_i \cdot (N_i(t))^{\beta+1}$, and the "virtual" workload, $\sum_i \eta_i N_i(t)$, at all times.

Proposition 8.6.2 ([89])**.** *Consider a heavy-traffic setting with complete resource pooling. For any policy $\pi \in \bar{\Pi}$ it holds that*

$$\lim_{r\to\infty} \sum_i \gamma_i \cdot (\hat{N}_i^{MW,r}(t))^{\beta+1} \leq \liminf_{r\to\infty} \sum_i \gamma_i \cdot (\hat{N}_i^{\pi,r}(t))^{\beta+1},$$

and

$$\lim_{r\to\infty} \sum_i \eta_i \hat{N}_i^{MW,r}(t) \leq \liminf_{r\to\infty} \sum_i \eta_i \hat{N}_i^{\pi,r}(t).$$

for all time t. In addition, the vector $\vec{\eta}$ is proportional to the vector

$$\lim_{r\to\infty} (\gamma_1 \cdot (\hat{N}_1^{MW,r}(t))^{\beta}, \ldots, \gamma_K \cdot (\hat{N}_K^{MW,r}(t))^{\beta}). \tag{8.35}$$

The result that the vector in (8.35) is proportional to $\vec{\eta}$, is referred to as a state-space collapse, since the dimension of the K-dimensional process decreases to one. A similar result was obtained in Chapter 2 for DPS-based policies.

Note that the Max-Weight policy does not minimize the holding cost $d_1\hat{N}_1(t) + d_2\hat{N}_2(t)$, since β must be strictly positive. However, the Max-Weight policy can be used to come close to this setting, for example, by setting $\beta > 0$ very small and $\gamma_i = d_i, i = 1, \ldots, K$. An alternative option is by making use of the fact that the Max-Weight policy does minimize the virtual workload $\sum_i \eta_i \hat{N}_i(t)$. Hence, when trying to minimize the holding cost among the Max-Weight policies, it is best to set the parameters (γ_i's and β) such that $\hat{N}_k^{MW}(t)$ is as large as possible, where k is such that $\eta_k/d_k \geq \eta_i/d_i$ for all $i \neq k$. For this reason, in [89] it is suggested that in heavy traffic a good choice for the parameters is $\beta = 1$, $\gamma_i = 1$, $i \neq k$ and $\gamma_k = \epsilon_k$, with $\epsilon_k > 0$ small, since the state space collapse result implies that then $\hat{N}_k^{MW}(t)$ will become relatively large compared to $\hat{N}_i^{MW}(t), i \neq k$.

For the parallel two-server model as considered in this chapter, the Max-Weight policy is as follows:

- Serve class 1 individually when $N_2(t) < \left(\frac{\gamma_1(1-c_1)\mu_1}{\gamma_2 c_2 \mu_2}\right)^{\frac{1}{\beta}} N_1(t)$.
- Serve classes 1 and 2 in parallel when $\left(\frac{\gamma_1(1-c_1)\mu_1}{\gamma_2 c_2 \mu_2}\right)^{\frac{1}{\beta}} N_1(t) \leq N_2(t) < \left(\frac{\gamma_1 c_1 \mu_1}{\gamma_2(1-c_2)\mu_2}\right)^{\frac{1}{\beta}} N_1(t)$.
- Serve class 2 individually when $\left(\frac{\gamma_1 c_1 \mu_1}{\gamma_2(1-c_2)\mu_2}\right)^{\frac{1}{\beta}} N_1(t) \leq N_2(t)$.

Hence the Max-Weight policy has two linear switching curves. In Figure 8.7 these switching curves are plotted. Note that in heavy traffic, the state space collapses to the line $x_2 = \left(\frac{c_1\mu_1\gamma_1}{(1-c_2)\mu_2\gamma_2}\right)^{\frac{1}{\beta}} x_1$ in the case of Region A and to the line $x_2 = \left(\frac{(1-c_1)\mu_1\gamma_1}{c_2\mu_2\gamma_2}\right)^{\frac{1}{\beta}} x_1$ in the case of Region B.

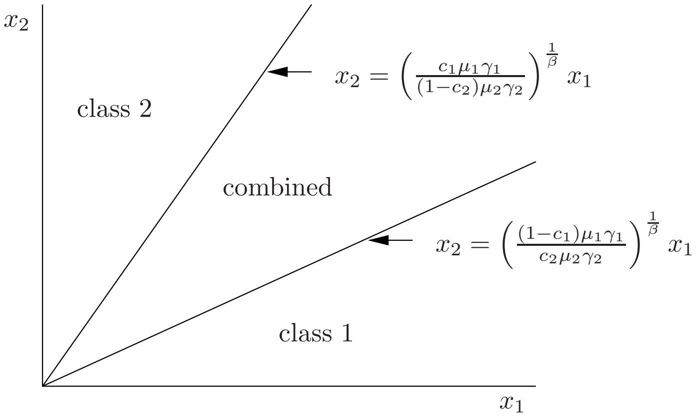

x_2
class 2
$x_2 = \left(\frac{c_1\mu_1\gamma_1}{(1-c_2)\mu_2\gamma_2}\right)^{\frac{1}{\beta}} x_1$
combined
$x_2 = \left(\frac{(1-c_1)\mu_1\gamma_1}{c_2\mu_2\gamma_2}\right)^{\frac{1}{\beta}} x_1$
class 1
x_1

Figure 8.7: The Max-Weight policy.

In Section 8.7.2 we will investigate the performance of the Max-Weight policies in the moderately-loaded case, and compare it with the optimal policy found numerically, and with the asymptotically fluid-optimal policies.

8.7 Numerical evaluation

The average-cost optimal policy for the original stochastic model can be computed numerically by value iteration after truncating the state space. Figures 8.8 and 8.9 illustrate for various scenarios that the optimal policy is characterized by a switching curve. We note that finding these optimal curves numerically was extremely time-consuming. Figure 8.8 considers the setting $\rho_1 < c_1$. We observe that the optimal switching curve is linear and coincides exactly with the asymptotically fluid-optimal switching curve $h(x_1) = \alpha\frac{\mu_2}{\mu_1}x_1$ from Proposition 8.3.10. Figure 8.9 corresponds to a scenario with $\rho_1 > c_1$ and illustrates that the optimal strategy resembles an exponentially-shaped curve, which agrees with Proposition 8.3.11. An optimal switching curve for a setting with $\rho_1 = c_1$ can be found in Figure 8.5. In that case, the curve coincided with a quadratic function. In the remainder of this section we will assess the gains that can be achieved by choosing the best switching-curve policies.

8.7.1 Switching-curve policies

We have conducted a large set of simulation experiments to assess the effectiveness of different switching-curve policies. We simulate in the order of 10^6 busy periods and are interested in the mean total number of users, i.e., $d_1 = d_2 = 1$.

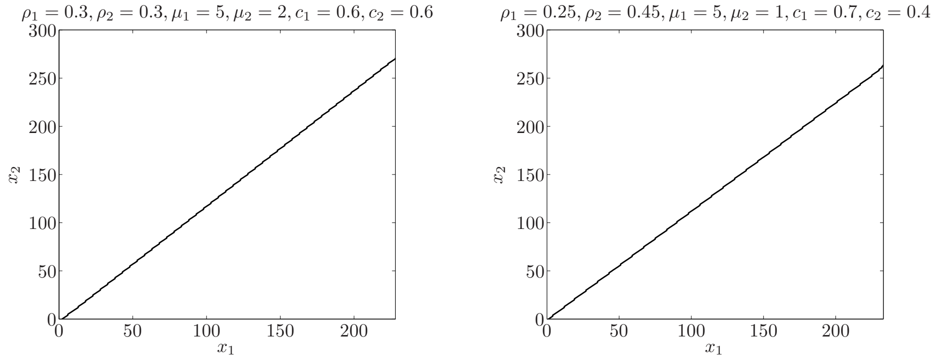


Figure 8.8: Optimal switching curve when $\rho_1 < c_1$, $\rho_2 < c_2$ (left), and $\rho_1 < c_1$ and $\rho_2 > c_2$ (right).

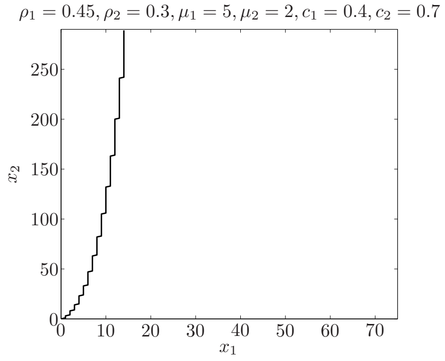


Figure 8.9: Optimal switching curve when $\rho_1 > c_1$ and $\rho_2 < c_2$.

Case $\rho_1 < c_1$

When $\rho_1 < c_1$, an asymptotically fluid-optimal policy is described by the linear switching curve as stated in Proposition 8.3.10. In Figure 8.10 we focus on this case and plot the total mean number of users under policies with a linear switching curve $h(x_1) = kx_1$, $k \geq 0$ (obtained by simulation). On the horizontal axis we vary the value of k. Note that $k = 0$ corresponds to always serving both classes in parallel. When the slope grows large ($k \to \infty$), the policy gives higher priority to serving class 1 exclusively (whenever present). Note that strict priority for class 1 leads to instability if $\rho_1 + \rho_2 > 1$, which can be the case even if the stability conditions (8.4) and (8.5) are met. The two graphs on the left in Figure 8.10 correspond to a moderately-loaded system. There we also plot the optimal policy found numerically by value iteration. We observe that when the parameter k is chosen well, the linear switching-curve policy coincides with the optimal policy. The two graphs on the

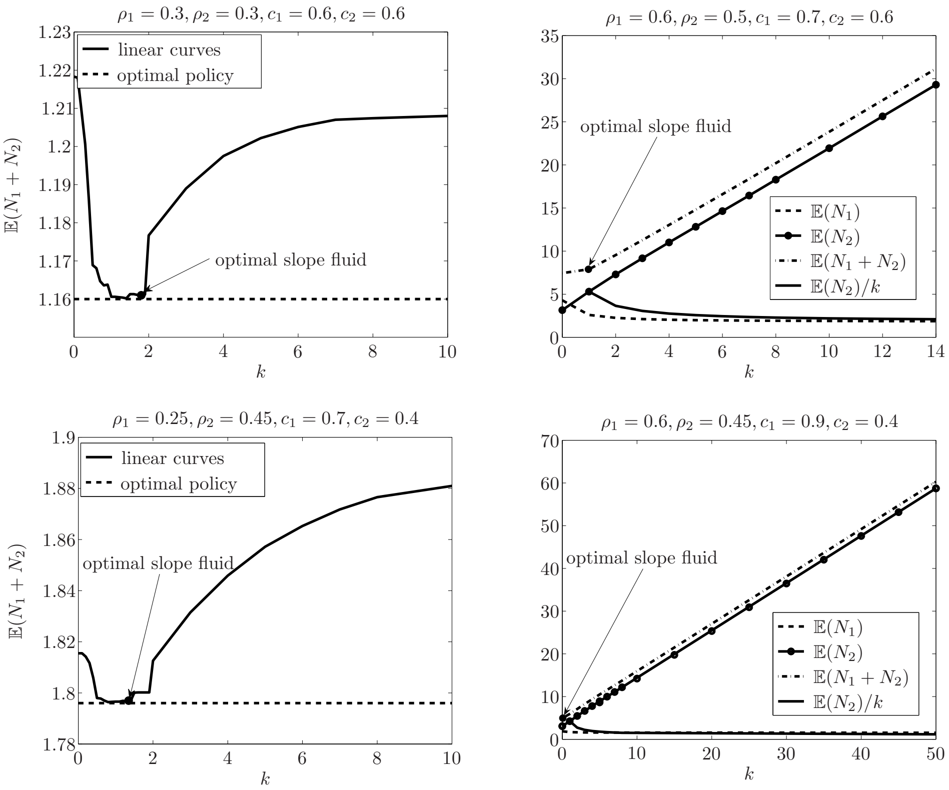


Figure 8.10: Total mean number of users for policies with a linear switching curve ($\mu_1 = 10, \mu_2 = 1$). The marker indicates the optimal slope for the fluid approximation. The two graphs on the top row correspond to cases with $\rho_1 < c_1$ and $\rho_2 < c_2$. The lower graphs have $\rho_1 < c_1$ and $\rho_2 > c_2$.

right in Figure 8.10 represent a heavily-loaded system. We did not determine the optimal policy for this parameter setting, since this is extremely time-consuming. Choosing k very large implies that the mean number of users will be large (since $\rho_1 + \rho_2 > 1$). It seems that a good choice for heavily-loaded systems is $k = 0$, i.e., always serve both classes in parallel. In a heavy-traffic setting with $\rho_1 < c_1$ (and necessarily $\rho_2 > c_2$ while $\rho_2 + \frac{\rho_1}{c_1}(1 - c_2) \to 1$) we see that the policy that always serves both classes in parallel is also the asymptotically optimal policy as found by both the fluid analysis (since then $\alpha = 0$, see (8.19), so the optimal slope in the fluid model is equal to 0) and the heavy-traffic analysis.

In Figure 8.11 we repeated the experiment for different parameter choices to illustrate that in the case of a moderately-loaded system the relative differences in performance between the optimal linear policy (obtained numerically by value iteration) and the strategy that maximizes the service capacity at all times (slope $k = 0$) can be quite significant.

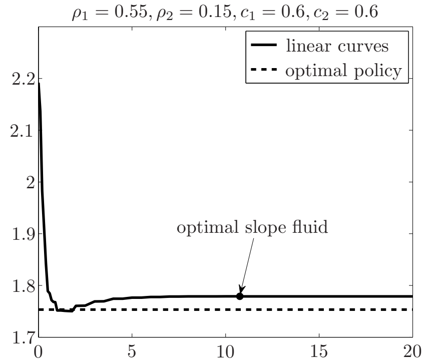


Figure 8.11: Total mean number of users for policies with a linear switching curve ($\mu_1 = 10, \mu_2 = 1$). The marker indicates the optimal slope for the fluid approximation.

An important observation in Figures 8.10 and 8.11 is that the slope as defined in (8.19) corresponding to the asymptotically fluid-optimal policy (denoted in the figures by "optimal slope fluid") is always close to optimal and performs very well.

In the two graphs on the right in Figure 8.10, we observe that the total mean number of users grows linearly in k as $k \to \infty$. In the remark below we give some intuition for this effect.

Remark 8.7.1. Consider the policy with a linear switching curve $h(x_1) = kx_1$. If k tends to ∞, then the system dynamics tend to a priority queue where class 1 is given preemptive priority. When $\rho_1 + \rho_2 < 1$, this policy is stable, and we indeed observe in the two graphs on the left in Figure 8.10 and in Figure 8.11 that the mean number of users will converge to a constant. However, when $\rho_1 + \rho_2 > 1$, this policy is not stable, and $\mathbb{E}(N_1 + N_2)$ will grow infinitely large as $k \to \infty$. The two graphs on the right in Figure 8.10 suggest that the mean number of users grows linearly in k as $k \to \infty$. This can be intuitively understood as follows.

Conditioned on $jk \leq N_2(t) < (j+1)k$, class 1 has as departure rate $\mu_1 c_1$ if $N_1(t) \leq j$, and μ_1 otherwise. For a given j, let $\pi(j)$ denote the equilibrium distribution for the process with departure rates as described above. Hence, $\pi_i(j) = \pi_0(j)\big(\frac{\rho_1}{c_1}\big)^i$ if $i \leq j$ and $\pi_i(j) = \pi_0(j)\big(\frac{\rho_1}{c_1}\big)^j \rho_1^{i-j}$ if $i > j$. If d is large, we assume that class 1 reaches equilibrium during the time that $jk \leq N_2(t) < (j+1)k$. Then the mean departure rate for class 2 is $\mu_2(j) := \mu_2\pi_0(j) + \mu_2 c_2 \sum_{i=1}^{j} \pi_i(j)$ (when $jk \leq N_2(t) < (j+1)k$), since both classes are served in parallel whenever $N_2(t) \geq kN_1(t)$. It can be checked that $\mu_2(j)$ is increasing in j, hence there exists a j^* such that $\mu_2(j^*-1) < \lambda_2 \leq \mu_2(j^*)$ (for convenience we set $\mu_2(-1) = 0$). Note that $j^* > 0$, unless $\rho_1 + \rho_2 < 1$. Hence, if $jk \leq N_2(t) < (j+1)k$ with $j < j^*$, then the mean drift in class 2 is positive, and the probability that the increase in $N_2(t)$ is O(k) tends to 1 as $k \to \infty$. If $jk \leq N_2(t) < (j+1)k$ with $j \geq j^*$, then the mean drift in class 2 is negative. Hence, the probability that the decrement of $N_2(t)$ is of order $\mathrm{O}(k)$ tends to 1 as $k \to \infty$. It is therefore plausible that the process $N_2(t)/k$ will most of the time be around the level j^*.

If the region $(j^*+1)k \leq N_2(t)$ is not reached (which is not a strong assumption, since this region will be rarely visited as $k \to \infty$), then the number of class-1 users can be upper bounded by the number of class-1 users in a system with departure rates $\mu_1 c_1$ if $N_1(t) \leq j^*$ and μ_1 otherwise. Since j^* does not depend on k, the upper bound for the number of class-1 users does not scale with k.

For the parameters used in the graph on the top right in Figure 8.10, the j^* is equal to 2. We observe in the figure that $\mathbb{E}(N_2)/k$ indeed converges to $j^* = 2$ and that $\mathbb{E}(N_1)$ does not scale with k. For the parameters that belong to the graph on the bottom right in Figure 8.10, the j^* is equal to 1. In that case too, we observe in the figure that $\mathbb{E}(N_2)/k$ indeed converges to $j^* = 1$ and that $\mathbb{E}(N_1)$ does not scale with k.

Case $\rho_1 > c_1$

When $\rho_1 > c_1$, an asymptotically fluid-optimal policy is described by an exponential switching curve as stated in Proposition 8.3.11. In Figure 8.12 we consider several parameter settings with $\rho_1 > c_1$, and plot the total mean number of users under policies with switching curves of the shape $h(x_1) = \mathrm{e}^{x_1/\gamma}$ (obtained by simulation). On the horizontal axis we vary the value of γ. Note that when γ grows large, this approaches the policy that always serves both classes in parallel. We observed that the best choice for the parameter γ, delivers virtually the same performance as the optimal policy (found numerically by value iteration). The large-deviation analysis further suggests that $\gamma = \frac{1}{\ln(\rho_1/c_1)}$ is a safe choice, see Section 8.4 (denoted in the figures by "rule of thumb"). In the three graphs on the left column in Figure 8.12 this corresponds to $\gamma = 8.5$. We observe that in fact the better choices for the parameter γ are smaller than 8.5. Still, the large-deviation result gives a safe estimate (the policy is stable) with better performance than the capacity-maximizing strategy (serving both classes in parallel whenever possible, i.e., $\gamma \to \infty$). In the three graphs on the right column in Figure 8.12, the rule of thumb is equal to $\gamma = 2.5$. In this case, the rule of thumb is very close to the optimal policy. In general, in all our tests we observed that the rule of thumb for γ proves to be a reasonable choice.

Recall that when $\rho_1 > c_1$, a threshold policy is asymptotically optimal in a heavily-loaded system. That is, both classes should be served in parallel whenever the number of class-1 users is below or equal to some threshold $Th \geq 0$. When the threshold grows large, this coincides with the policy that always serves both classes in parallel. In Figure 8.12 we consider a moderately-loaded system. We vary the value of the threshold Th, and plot the total mean number of users (obtained by simulation). For certain small values of the threshold, the threshold policy performs rather well. However, when the threshold is chosen too small, the performance of the system can degrade considerably and become unstable. In fact, for a system with large loads ($\rho_1 + \rho_2 > 1$), the policy with $Th = 0$ is unstable. In the two

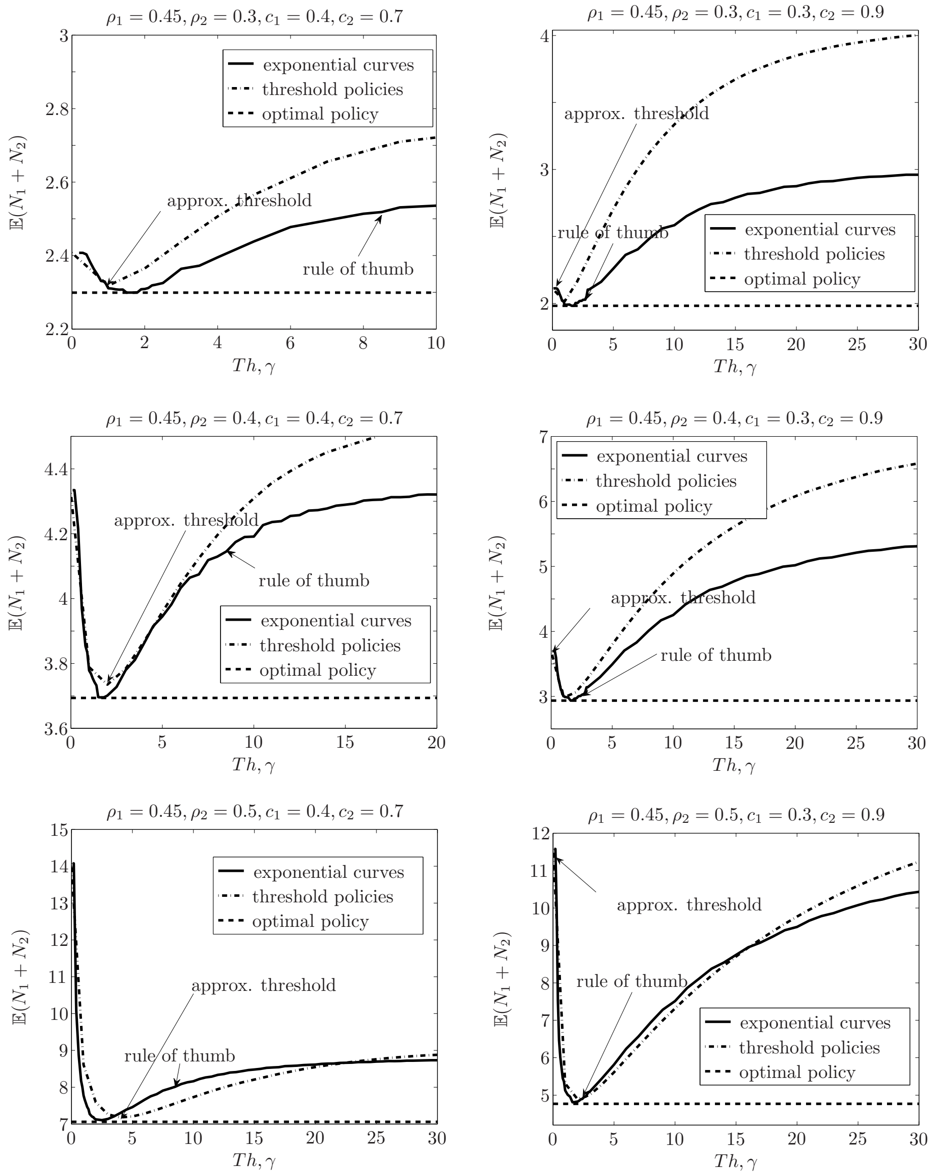


Figure 8.12: Total mean number of users when $\rho_1 > c_1$ and $\rho_2 < c_2$ for policies with exponential switching curves (as a function of γ), and for threshold policies (as a function of Th). In the graphs on the left column we chose $\mu_1 = 5, \mu_2 = 2$, and in the graphs on the right column we chose $\mu_1 = 10, \mu_2 = 1$.

graphs on the last row in Figure 8.12 (where still $\rho_1 + \rho_2 < 1$) we already see that the total number of users doubles when the threshold is set equal to 0. In [129] the authors propose a method to obtain estimates for the value of the threshold. For the settings in Figure 8.12 we have calculated the estimates for the threshold using their method (denoted in the figures by "approx. threshold"). We see that in the figures on the left column, the approximation for the threshold matches exactly with the best threshold value. However, in the figures on the right column, the approximation of the threshold is too small, which results in severe performance degradation in case of high loads ($\rho_1 = 0.45, \rho_2 = 0.5$).

In general, for the case $\rho_1 > c_1$ our fluid-based method (rule of thumb) proved to be a rather safe option, while threshold policies (using the approximation in [129]) may sometimes perform better, but can also be far from optimal. Although this is supported by a rather extensive set of experiments, it remains as a challenge to provide a theoretical basis for the robustness of fluid-based policies.

8.7.2 Max-Weight policies

As stated in Section 8.6.2, Max-Weight policies can be close to optimal in a heavy-traffic setting. In this section we investigate the performance of the Max-Weight policies in a *moderately-loaded* system and compare this to the performance of the asymptotically fluid-optimal policies as found in this chapter. We compare the total mean number of users and distinguish between whether $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$ or $\mu_1 c_1 + \mu_2 c_2 < \mu_1$. We will observe that in the second case the fluid-based policies can outperform the Max-Weight policies, and that the parameter choices for the Max-Weight policies as suggested by the heavy-traffic results are not necessarily a good choice in a moderately-loaded system.

Case $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$

From Section 8.2.1 we know that when $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$, the policy which serves classes 1 and 2 in parallel whenever possible, stochastically minimizes the total mean number of users present in the system. Note that when $\mu_1 c_1 + \mu_2 c_2 \geq \mu_1$, the Max-Weight policy with $\gamma_1 = \gamma_2 = 1$ and β close to zero, will almost always serve both classes in parallel. Numerically we observed that the Max-Weight policy (with $\gamma_1 = \gamma_2 = 1$ and β close to zero) turns out to be very effective and nearly matches the performance of the optimal policy. For this reason, we have not included any graphs for this case.

Case $\mu_1 c_1 + \mu_2 c_2 < \mu_1$

When $\mu_1 c_1 + \mu_2 c_2 < \mu_1$, the asymptotically fluid-optimal policy we proposed is described by a switching curve $h(x_1)$ (either linear, quadratic or exponential), where class 1 is served in states below the switching curve, and classes 1 and 2 are served in parallel in states above the switching curve. We compare these policies with Max-Weight policies. We choose the parameters as described in Section 8.6.2. So we take $\gamma_1 = \gamma_2 = 1$ and $\beta = 10^{-4}$. When $\mu_1 > \mu_2$ and $\mu_1 c_1 + \mu_2 c_2 < \mu_1$, we

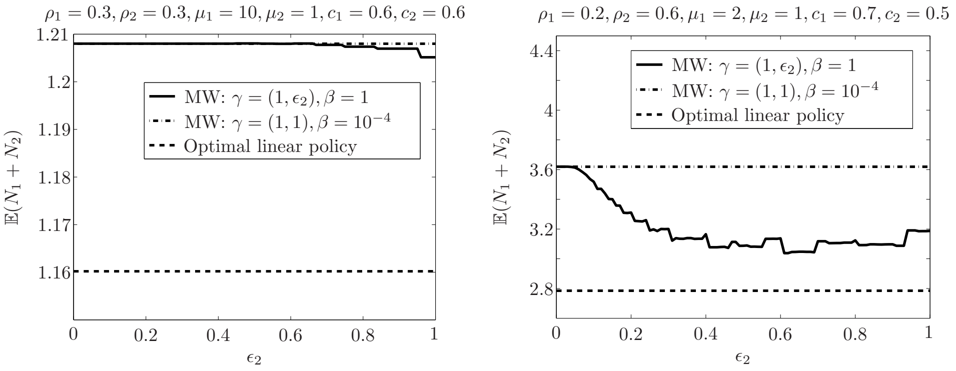


Figure 8.13: Total mean number of users under Max-Weight policies and under the optimal linear switching curve, with $\mu_1 c_1 + \mu_2 c_2 < \mu_1$ and $\rho_1 < c_1$.

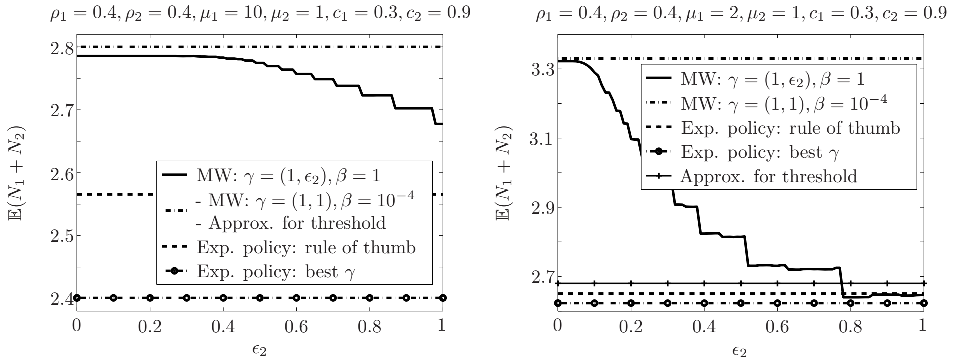


Figure 8.14: Total mean number of users under Max-Weight policies and under exponential switching curves, with $\mu_1 c_1 + \mu_2 c_2 < \mu_1$ and $\rho_1 > c_1$.

have $\eta_1 < \eta_2$, both in Region A and in Region B of Figure 8.6. Hence, we will also consider Max-Weight policies with $\gamma_1 = 1$, $\gamma_2 = \epsilon_2, \epsilon_2 > 0$, and $\beta = 1$.

In Figures 8.13 and 8.14, we compare (by simulation) the performance of the Max-Weight policies (referred to as MW) with the performance of the best linear or exponential switching-curve policies. On the horizontal axis we vary ϵ_2 and on the vertical axis we plot the total mean number of users under the various policies. First of all, we note that in both Figures 8.13 and 8.14, the Max-Weight policy with $\beta = 10^{-4}$ and $\gamma_i = 1$, $i = 1, 2$, performs rather poorly. This is not surprising, since if $\mu_1 c_1 + \mu_2 c_2 < \mu_1$, then the Max-Weight policy (with $\beta = 10^{-4}$ and $\gamma_1 = \gamma_2 = 1$) will almost always serve class 1 individually, which is far from optimal.

For the parameters as in Figure 8.13 (left), the fluid approximation suggests that if $N_2(t) \leq 1.8N_1(t)$, then it is good to serve class 1, and otherwise to serve both classes in parallel. Numerically, we found that this is also the best linear policy for the stochastic model. The Max-Weight policy (with $\beta = 1$ and $\gamma = (1, \epsilon_2)$) will serve class 1 most of the time, since that is the prescribed action in states such that $N_2(t) \leq 6\frac{2}{3\epsilon_2}N_1(t)$. From the figure, we see that this is only 5% worse than the optimal linear policy. For the parameters as in Figure 8.13 (right), the fluid approximation serves always classes 1 and 2 in parallel. Numerically, we found that this is also the best linear policy for the stochastic model. The Max-Weight policy (with $\beta = 1$ and $\gamma = (1, \epsilon_2)$) however, serves class 1 individually as soon as $N_2(t) \leq \frac{12}{10\epsilon_2}N_1(t)$. These states will be visited more often when $\epsilon_2 \downarrow 0$. In the figure, the performance degrades from 15% worse ($\epsilon_2 = 1$), to 30% worse ($\epsilon_2 \downarrow 0$), compared with the optimal linear policy.

In Figure 8.14, the parameters are such that an exponential switching curve is asymptotically fluid-optimal. We plot the performance of both the best exponential switching curve (determined numerically), and of the exponential switching curve where γ is set according to the rule of thumb, i.e., $\gamma = \frac{1}{\ln(\rho_1/c_1)} = 3.48$. For $\mu_1 = 10$, the Max-Weight policy (with $\beta = 1$ and $\gamma = (1, \epsilon_2)$) is about 15% worse compared with the best exponential policy. For $\mu_1 = 2$, it is close to optimal when $\epsilon_2 = 1$, but the performance degrades when $\epsilon_2 \downarrow 0$. Observe that in both cases the policy with an exponential switching curve where γ is chosen according to the rule of thumb, performs rather well. We have also calculated the performance of the threshold policy as suggested in [129]. For the setting of Figure 8.14 (left) it suggests a threshold equal to 0, in which case there are approximately 2.8 users in the system. Hence, the proposed policy does not give good performance. For the setting of Figure 8.14 (right) it suggests a threshold equal to 1, in which case there are approximately 2.68 users in the system. This is rather close to optimal.

8.8 Concluding remarks

We have studied optimal policies for a parallel two-server model where the highest service capacity is achieved when serving both classes in parallel. Through a fluid limit analysis we determined the shape of close-to-optimal policies, which can be characterized either by linear, quadratic or exponential switching curves. The results yield directly usable estimates for efficient policies in the stochastic setting, comparing favorably with threshold-based policies and Max-Weight policies for moderately-loaded regimes

Several extensions to this work are of interest. For example, it is interesting to investigate how our results change if the capacity is also favorably affected by the numbers of users within each class. For example, in wireless networks the aggregate transmission rate increases with the number of users, due to opportunistic scheduling, which exploits multiuser diversity [83]. An intermediate step that is of interest on its own would be to consider our current model with several possible service capacity vectors when serving classes in parallel. For example, if in addition to the service capacities c_1 and c_2 we can choose $\tilde{c}_1$ and $\tilde{c}_2$ that are not in the convex hull depicted in Figure 1.6. A third direction of interest is to study our model with more than two classes. This could also serve as an intermediate step towards handling multiuser diversity gains as mentioned above, which is presumably more difficult to handle.

Appendix

8.A Proof of Lemma 8.2.2

The proof is by induction on the time index m. For $m = 0$ the statement holds. In order to apply induction, assume it holds for $Z = V_m$. We will show that it holds for $Z = V_{m+1}$ as well. Define

$$\begin{aligned}\tilde{V}_{m+1}(\vec{x}) \quad &:= \quad V_{m+1}(\vec{x}) - \lambda_1 V_m(\vec{x} + \vec{e}_1) - \lambda_2 V_m(\vec{x} + \vec{e}_2) \\ &= \quad \min\Big(\mu_1 V_m((x_1-1)^+, x_2) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2) V_m(\vec{x}), \\ &\qquad \mu_2 V_m(x_1, (x_2-1)^+) + (\mu_1 + \mu_1 c_1 + \mu_2 c_2) V_m(\vec{x}), \\ &\qquad \mu_1 c_1 V_m((x_1-1)^+, x_2) + \mu_2 c_2 V_m(x_1, (x_2-1)^+) + (\mu_1 + \mu_2) V_m(\vec{x})\Big).\end{aligned}$$

By assumption, $\lambda_1 V_m(\vec{x} + \vec{e}_1) + \lambda_2 V_m(\vec{x} + \vec{e}_2)$ satisfies the inequality. In order to prove that $V_{m+1}(\cdot)$ does as well, it is sufficient to show that $\tilde{V}_{m+1}(\cdot)$ does. This will be done in the remainder of the proof. We will show that

$$\begin{aligned}&(\mu_1 + \mu_2)\tilde{V}_{m+1}(\vec{x}) + \mu_1 c_1 \tilde{V}_{m+1}(\vec{x} - \vec{e}_1) + \mu_2 c_2 \tilde{V}_{m+1}(\vec{x} - \vec{e}_2) \\ &\leq \mu_1 \tilde{V}_{m+1}(\vec{x}) + \mu_2 \tilde{V}_{m+1}(\vec{x} - \vec{e}_2) + (\mu_1 c_1 + \mu_2 c_2)\tilde{V}_{m+1}(\vec{x})\end{aligned} \tag{8.36}$$

is indeed satisfied for all $x_1, x_2 > 0$. The proof of

$$\begin{aligned}&(\mu_1 + \mu_2)\tilde{V}_{m+1}(\vec{x}) + \mu_1 c_1 \tilde{V}_{m+1}(\vec{x} - \vec{e}_1) + \mu_2 c_2 \tilde{V}_{m+1}(\vec{x} - \vec{e}_2) \\ &\leq \mu_1 \tilde{V}_{m+1}(\vec{x} - \vec{e}_1) + \mu_2 \tilde{V}_{m+1}(\vec{x}) + (\mu_1 c_1 + \mu_2 c_2)\tilde{V}_{m+1}(\vec{x})\end{aligned}$$

follows exactly the same steps, but with the role of class 1 and class 2 interchanged.

First assume $x_1 > 0$ and $x_2 = 1$. By definition of $\tilde{V}_{m+1}(\cdot)$ we can write

$$\begin{aligned}&\mu_2 \tilde{V}_{m+1}(x_1, 1) + \mu_1 c_1 \tilde{V}_{m+1}(x_1 - 1, 1) + \mu_2 c_2 \tilde{V}_{m+1}(x_1, 0) \\ &\leq \mu_2[\mu_2 V_m(x_1, 0) + (\mu_1 c_1 + \mu_2 c_2 + \mu_1) V_m(x_1, 1)] \\ &\quad + \mu_1 c_1[\mu_2 V_m(x_1 - 1, 0) + (\mu_1 c_1 + \mu_2 c_2 + \mu_1) V_m(x_1 - 1, 1)] \\ &\quad + \mu_2 c_2[\mu_1 V_m(x_1 - 1, 0) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2) V_m(x_1, 0)].\end{aligned} \tag{8.37}$$

Rearranging terms in (8.37), gives

$$\begin{aligned}&\mu_1[\mu_2 V_m(x_1, 1) + \mu_1 c_1 V_m(x_1 - 1, 1) + \mu_2 c_2 V_m(x_1, 0)] \\ &+ (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_m(x_1, 1) + \mu_1 c_1 V_m(x_1 - 1, 1) + \mu_2 c_2 V_m(x_1, 0)] \\ &+ \mu_2[\mu_2 V_m(x_1, 0) + \mu_2 c_2 V_m(x_1, 0)] \\ &+ \mu_1 \mu_2[(c_1 + c_2) V_m(x_1 - 1, 0) - c_2 V_m(x_1, 0)].\end{aligned} \tag{8.38}$$

Since $V_m(\cdot)$ is increasing in x_1 (see Lemma 8.2.1), $c_1 + c_2 \geq 1$, and since (8.8) holds by induction for $V_m(\cdot)$, the expression in (8.38) is less than or equal to

$$\begin{aligned}
&\mu_1[\mu_2 V_m(x_1,0) + (\mu_1 c_1 + \mu_2 c_2)V_m(x_1,1)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_m(x_1,1) + \mu_1 c_1 V_m(x_1-1,1) + \mu_2 c_2 V_m(x_1,0)]\\
&\quad + \mu_2[\mu_2 V_m(x_1,0) + \mu_2 c_2 V_m(x_1,0)]\\
&\quad + \mu_1\mu_2[(c_1-1)V_m(x_1,0) + V_m(x_1-1,0)]\\
&= \mu_2[\mu_1 V_m(x_1-1,0) + (\mu_2 + \mu_1 c_1 + \mu_2 c_2)V_m(x_1,0)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[(\mu_1+\mu_2)V_m(x_1,1) + \mu_1 c_1 V_m(x_1-1,1) + \mu_2 c_2 V_m(x_1,0)],
\end{aligned} \tag{8.39}$$

where in the last step we rearranged the terms. Since (8.8) holds by induction for $V_m(\cdot)$, the expression in (8.39) is equal to $\mu_2 \tilde{V}_{m+1}(x_1,0) + (\mu_1 c_1 + \mu_2 c_2)\tilde{V}_{m+1}(x_1,1)$. Hence, (8.36) is proved.

Now assume $x_1 > 0$ and $x_2 > 1$. By definition of $\tilde{V}_{m+1}(\cdot)$ we can write

$$\begin{aligned}
&\mu_2 \tilde{V}_{m+1}(\vec{x}) + \mu_1 c_1 \tilde{V}_{m+1}(\vec{x}-\vec{e}_1) + \mu_2 c_2 \tilde{V}_{m+1}(\vec{x}-\vec{e}_2)\\
&\leq \mu_2[\mu_2 V_m(\vec{x}-\vec{e}_2) + (\mu_1 + \mu_1 c_1 + \mu_2 c_2)V_m(\vec{x})]\\
&\quad + \mu_1 c_1[\mu_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 V_m(\vec{x}-\vec{e}_1-\vec{e}_2) + (\mu_1 c_1 + \mu_2 c_2)V_m(\vec{x}-\vec{e}_1)]\\
&\quad + \mu_2 c_2[\mu_1 V_m(\vec{x}-\vec{e}_2) + \mu_2 V_m(\vec{x}-2\vec{e}_2) + (\mu_1 c_1 + \mu_2 c_2)V_m(\vec{x}-\vec{e}_2)].
\end{aligned} \tag{8.40}$$

Rearranging terms in (8.40), shows that it equals

$$\begin{aligned}
&\mu_1[\mu_2 V_m(\vec{x}) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 c_2 V_m(\vec{x}-\vec{e}_2)]\\
&\quad + \mu_2[\mu_2 V_m(\vec{x}-\vec{e}_2) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2) + \mu_2 c_2 V_m(\vec{x}-2\vec{e}_2)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_m(\vec{x}) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 c_2 V_m(\vec{x}-\vec{e}_2)].
\end{aligned} \tag{8.41}$$

Since (8.8) holds by induction for $V_m(\cdot)$, the expression in (8.41) is less than or equal to

$$\begin{aligned}
&\mu_1[\mu_2 V_m(\vec{x}-\vec{e}_2) + \mu_1 c_1 V_m(\vec{x}) + \mu_2 c_2 V_m(\vec{x})]\\
&+ \mu_2[\mu_2 V_m(\vec{x}-\vec{e}_2) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2) + \mu_2 c_2 V_m(\vec{x}-2\vec{e}_2)]\\
&+ (\mu_1 c_1 + \mu_2 c_2)[\mu_2 V_m(\vec{x}) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 c_2 V_m(\vec{x}-\vec{e}_2)]\\
&= \mu_2[(\mu_1+\mu_2)V_m(\vec{x}-\vec{e}_2) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1-\vec{e}_2) + \mu_2 c_2 V_m(\vec{x}-2\vec{e}_2)]\\
&\quad + (\mu_1 c_1 + \mu_2 c_2)[(\mu_1+\mu_2)V_m(\vec{x}) + \mu_1 c_1 V_m(\vec{x}-\vec{e}_1) + \mu_2 c_2 V_m(\vec{x}-\vec{e}_2)],
\end{aligned} \tag{8.42}$$

where in the last step we rearranged the terms. Since (8.8) holds by induction for $V_m(\cdot)$, the expression in (8.42) is equal to $\mu_2 \tilde{V}_{m+1}(\vec{x}-\vec{e}_2) + (\mu_1 c_1 + \mu_2 c_2)\tilde{V}_{m+1}(\vec{x})$. Hence, (8.36) is proved. □

8.B Proof of Lemma 8.2.4

We use t^+ to denote any element in an interval $(t, t+\delta]$, for a sufficiently small $\delta > 0$. Throughout the proof we use that

$$W_i(t) > 0 \text{ implies } W_i(t^+) > 0, \text{ and that } W_i(t) = 0 \text{ implies } W_i(t^+) = 0. \tag{8.43}$$

This follows since the workload process $W_i(t)$, $i = 1, 2$, is right-continuous and increases only with an arrival.

Note that $S_i(t)$, $i = 1, 2$, is continuous. In order to show relation (8.9), we therefore consider the first time instant t such that (8.9) holds with equality and is violated immediately after time t. So $S_1^{\pi}(t) = S_1^{\tilde{\pi}}(t)$, and by (8.1) also $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t)$, while $s_1^{\pi}(t^+) < s_1^{\tilde{\pi}}(t^+)$, so that $S_1^{\pi}(t^+) < S_1^{\tilde{\pi}}(t^+)$. Since $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t)$, by (8.43) and by construction of policy π we obtain that $s_1^{\pi}(t^+) \geq s_1^{\tilde{\pi}}(t^+)$. This gives contradiction and hence (8.9) holds for all $t \geq 0$.

Let time t be the first time instant such that either (8.10) or (8.11) holds with equality and is violated immediately after time t. We will show that such a t does not exist. The remainder of the proof consists of two parts, depending on whether equation (8.10) or equation (8.11) is the first to be violated.

Part I: Assume (8.10) is the first equation that fails to hold, i.e., $S_1^{\pi}(t) + S_2^{\pi}(t) = S_1^{\tilde{\pi}}(t) + S_2^{\tilde{\pi}}(t)$, and by (8.1) also $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$, while $s_1^{\pi}(t^+) + s_2^{\pi}(t^+) < s_1^{\tilde{\pi}}(t^+) + s_2^{\tilde{\pi}}(t^+)$, so that $S_1^{\pi}(t^+) + S_2^{\pi}(t^+) < S_1^{\tilde{\pi}}(t^+) + S_2^{\tilde{\pi}}(t^+)$. We will show that

$$W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t) \quad \text{implies} \quad W_i^{\pi}(t) = W_i^{\tilde{\pi}}(t), \quad i = 1, 2. \tag{8.44}$$

By (8.43) and by construction of policy, $W_i^{\pi}(t) = W_i^{\tilde{\pi}}(t)$, $i = 1, 2$, implies that $s_1^{\pi}(t^+) + s_2^{\pi}(t^+) \geq s_1^{\tilde{\pi}}(t^+) + s_2^{\tilde{\pi}}(t^+)$, and hence we reach a contradiction. So let us prove (8.44).

- We first assume that there is an interval $[u, t)$ in which policy $\tilde{\pi}$ has more work in the system compared to policy π, i.e., $W_1^{\pi}(v) + W_2^{\pi}(v) < W_1^{\tilde{\pi}}(v) + W_2^{\tilde{\pi}}(v)$ for all $v \in [u, t)$, and at time t, $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$. We can choose this interval such that $\tilde{\pi}$ has made up for the lost capacity in one of the three ways described below. Define $M_c^{\tilde{\pi}}(u, t)$ as the cumulative amount of time that both classes are served in parallel under policy $\tilde{\pi}$ in the time interval $[u, t)$.

  (i) During the interval $[u, t)$ policy $\tilde{\pi}$ has work in the system, while policy π has an empty system.

  (ii) In the interval $[u, t)$ we have $M_c^{\tilde{\pi}}(u, t) > 0$, while policy π serves class 1 with service capacity 1. Hence $W_2^{\pi}(v) = 0$ and $W_1^{\pi}(v) > 0$, for all $v \in [u, t)$.

  (iii) In the interval $[u, t)$ we have $M_c^{\tilde{\pi}}(u, t) > 0$, while policy π serves class 2 with service capacity 1. Hence $W_1^{\pi}(v) = 0$ and $W_2^{\pi}(v) > 0$, for all $v \in [u, t)$.

  Note that the three cases are mutually exclusive. We will show that (8.44) holds for (i), (ii) and (iii). Although not mentioned explicitly, in all three cases we use that a possible arrival at time t alters the workload in both systems in the same way. Let t^- denote any element in an interval $[t - \delta, t)$ with $\delta > 0$ sufficiently small.

In case (i) we have $W_i^{\pi}(t^-) = 0$ for $i = 1, 2$. Since at time t it holds that $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$, we obtain that $W_i^{\tilde{\pi}}(t^-) = 0,\ i = 1, 2$. Hence, we have $W_i^{\pi}(t) = W_i^{\tilde{\pi}}(t),\ i = 1, 2$.

In case (ii) we have that $W_2^{\pi}(t^-) = 0$, hence $W_2^{\pi}(t) \leq W_2^{\tilde{\pi}}(t)$. From $W_1^{\pi}(t) + W_2^{\pi}(t) = W_1^{\tilde{\pi}}(t) + W_2^{\tilde{\pi}}(t)$ and $W_1^{\pi}(t) \leq W_1^{\tilde{\pi}}(t)$ (follows from (8.1) and (8.9)), we obtain $W_2^{\pi}(t) \geq W_2^{\tilde{\pi}}(t)$. Hence, $W_i^{\pi}(t) = W_i^{\tilde{\pi}}(t)$, $i = 1, 2$.

In case (iii) we have

$$M_c^{\tilde{\pi}}(u,t)(c_1 + c_2 - 1) = W_1^{\tilde{\pi}}(u) + W_2^{\tilde{\pi}}(u) - W_2^{\pi}(u), \tag{8.45}$$

since the total amount of additional capacity that policy $\tilde{\pi}$ gets compared to policy π in the interval $[u, t)$ (left-hand side in (8.45)), is equal to the difference in the total workload at time u (right-hand side in (8.45)). Since $W_1^{\pi}(u) = 0$, from (8.1) and (8.11) we obtain that $c_1 W_2^{\pi}(u) = (1 - c_2)W_1^{\pi}(u) + c_1 W_2^{\pi}(u) \leq (1 - c_2)W_1^{\tilde{\pi}}(u) + c_1 W_2^{\tilde{\pi}}(u)$. Rewriting this gives $W_1^{\tilde{\pi}}(u) \leq \frac{c_1}{c_1+c_2-1}(W_1^{\tilde{\pi}}(u) + W_2^{\tilde{\pi}}(u) - W_2^{\pi}(u)) = c_1 M_c^{\tilde{\pi}}(u,t)$. Note that $S_1^{\tilde{\pi}}(t) - S_1^{\tilde{\pi}}(u) \geq c_1 M_c^{\tilde{\pi}}(u,t)$ and $A_1(u, t^-) = 0$ (since $W_1^{\pi}(v) = 0$ for all $v \in [u, t)$). Together this gives $W_1^{\tilde{\pi}}(t^-) = W_1^{\tilde{\pi}}(u) + A_1(u,t^-) - (S_1^{\tilde{\pi}}(t) - S_1^{\tilde{\pi}}(u)) \leq 0$. Since we also know that $W_1^{\pi}(t^-) = 0$, it follows that $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t)$, and hence $W_2^{\pi}(t) = W_2^{\tilde{\pi}}(t)$.

- Now consider the case when there is an interval $[w, t]$ such that $W_1^{\pi}(v) + W_2^{\pi}(v) = W_1^{\tilde{\pi}}(v) + W_2^{\tilde{\pi}}(v)$ for all $v \in [w, t]$ and $W_1^{\pi}(w^-) + W_2^{\pi}(w^-) < W_1^{\tilde{\pi}}(w^-) + W_2^{\tilde{\pi}}(w^-)$. From the previous item, we obtain that $W_i^{\pi}(w) = W_i^{\tilde{\pi}}(w)$, $i = 1, 2$. Together with the fact that in the interval $[w, t]$ the total workload is equal under both policies, and by construction of policy π, it follows that $\tilde{\pi}$ did not serve class 2 individually while π serves both classes in parallel. Hence, $W_i^{\pi}(v) = W_i^{\tilde{\pi}}(v)$ for all $v \in [w, t]$, $i = 1, 2$.

Part II: Assume (8.11) is the first equation that fails to hold, i.e., $(1 - c_2)S_1^{\pi}(t) + c_1 S_2^{\pi}(t) = (1 - c_2)S_1^{\tilde{\pi}}(t) + c_1 S_2^{\tilde{\pi}}(t)$, and by (8.1) also $(1 - c_2)W_1^{\pi}(t) + c_1 W_2^{\pi}(t) = (1-c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$, while $(1-c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) < (1-c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+)$; so that $(1-c_2)S_1^{\pi}(t^+) + c_1 S_2^{\pi}(t^+) < (1-c_2)S_1^{\tilde{\pi}}(t^+) + c_1 S_2^{\tilde{\pi}}(t^+)$. With slight abuse of notation, let $f_1(t^+), f_2(t^+), f_c(t^+), f_I(t^+)$ be the coefficients that define the capacity vector in the capacity region S under policy $\tilde{\pi}$ at time t^+, i.e., $(s_1^{\tilde{\pi}}(t^+), s_2^{\tilde{\pi}}(t^+)) = f_1(t^+)\cdot(1,0) + f_2(t^+)\cdot(0,1) + f_c(t^+)\cdot(c_1,c_2) + f_I(t^+)\cdot(0,0)$. Note that $1 = f_1(t^+) + f_2(t^+) + f_c(t^+) + f_I(t^+)$. We have the following possibilities:

- If $W_1^{\pi}(t) > 0$ and $W_2^{\pi}(t) > 0$, then by (8.43) and by definition of policy π we have $(s_1^{\pi}(t^+), s_2^{\pi}(t^+)) = f_1(t^+)\cdot(1,0) + (f_c(t^+) + f_2(t^+))\cdot(c_1,c_2) + f_I(t^+)\cdot(0,0)$, hence $(1 - c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) = (1 - c_2)(f_1(t^+) + c_1(f_c(t^+) + f_2(t^+))) + c_1 c_2(f_c(t^+) + f_2(t^+)) = (1 - c_2)(f_1(t^+) + c_1 f_c(t^+)) + c_1(f_2(t^+) + c_2 f_c(t^+)) = (1 - c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+)$.

- If $W_1^{\pi}(t) = 0$ and $W_2^{\pi}(t) > 0$, then, by definition, policy π serves class 2 individually for a fraction of time $1 - f_I(t^+)$ and otherwise idles. So $(1 - c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) = c_1(1 - f_I(t^+))$. Since $c_1 + c_2 > 1$, we have that

$c_1(1-f_I(t^+)) \geq (1-c_2)f_1(t^+) + c_1(f_c(t^+) + f_2(t^+)) = (1-c_2)(f_1(t^+) + c_1 f_c(t^+)) + c_1(f_2(t^+) + c_2 f_c(t^+)) = (1-c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+)$.

- If $W_1^{\pi}(t) > 0$ and $W_2^{\pi}(t) = 0$, then we have $(1-c_2)W_1^{\pi}(t) = (1-c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$ and $W_1^{\pi}(t) \leq W_1^{\tilde{\pi}}(t)$ (by (8.9)). Hence $W_1^{\pi}(t) = W_1^{\tilde{\pi}}(t)$ and $0 = W_2^{\pi}(t) = W_2^{\tilde{\pi}}(t)$. By (8.43) we obtain $f_2(t^+) = 0$, so by definition of policy π, $s_i^{\pi}(t^+) = s_i^{\tilde{\pi}}(t^+)$, $i = 1, 2$.

- If $W_1^{\pi}(t) + W_2^{\pi}(t) = 0$, then $0 = (1-c_2)W_1^{\tilde{\pi}}(t) + c_1 W_2^{\tilde{\pi}}(t)$. By (8.43) we have $W_i^{\pi}(t^+) = W_i^{\tilde{\pi}}(t^+) = 0$, and hence $(1-c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) = (1-c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+) = 0$.

For all the four possibilities we reached a contradiction with $(1-c_2)s_1^{\pi}(t^+) + c_1 s_2^{\pi}(t^+) < (1-c_2)s_1^{\tilde{\pi}}(t^+) + c_1 s_2^{\tilde{\pi}}(t^+)$ and this concludes the proof. □

8.C Proof of Lemma 8.3.2

By the Filippov-Cesari theorem [122, Chapter 2.8], there exists an optimal control $u^{*D}(t)$ and a corresponding optimal trajectory $n^{*D}(t)$ for the problem $\min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^D (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t$, for any $D \geq 0$.

For the moment, assume there exists a function $H(\cdot)$ such that

$$n_1^{*D}(t) + n_2^{*D}(t) = 0, \quad \text{for all } t \geq H(d_1 n_1 + d_2 n_2), \text{ with } n = (n_1, n_2) \text{ the initial state.} \tag{8.46}$$

The proof of (8.46) will be given later on. From (8.46) we obtain

$$\begin{aligned}
&\min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^{\infty} (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t \\
&\geq \min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^{D} (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t \\
&= \int_0^{D} (d_1 n_1^{*D}(t) + d_2 n_2^{*D}(t))\mathrm{d}t = \int_0^{\infty} (d_1 n_1^{*D}(t) + d_2 n_2^{*D}(t))\mathrm{d}t \\
&\geq \min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^{\infty} (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t,
\end{aligned} \tag{8.47}$$

for all $D \geq H(d_1 n_1 + d_2 n_2)$. Hence, $(u^{*D}(t), n^{*D}(t))$ is an optimal solution of (8.18). In particular, this implies the existence result for the minimization problem (8.18). In addition, from (8.47) we obtain that for any optimal trajectory $n^*(t)$ of (8.18), it holds that

$$\begin{aligned}
&\min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^\infty (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t \\
&= \int_0^\infty (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t \geq \int_0^D (d_1 n_1^*(t) + d_2 n_2^*(t))\mathrm{d}t \\
&\geq \min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^D (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t \\
&= \min_{n(t) \text{ s.t. } (8.13)-(8.16)} \int_0^\infty (d_1 n_1(t) + d_2 n_2(t))\mathrm{d}t,
\end{aligned}$$

for all $D \geq H(d_1 n_1 + d_2 n_2)$. This proves the lemma under the condition that there indeed exists a function $H(\cdot)$ satisfying (8.46). The latter will be shown in the remainder of the proof. We use similar arguments as in [88, Proposition 6.1].

Denote by π^p the policy that always serves classes 1 and 2 in parallel whenever possible. Let $n^p(t)$ be the trajectory that corresponds to policy π^p. Under the stability conditions we know that $n^p(t)$ hits zero after a finite time and then remains empty, see Lemma 8.3.1. Denote by $T^p(\tilde{n}, n')$ the time it takes for policy π^p to move from $\tilde{n}$ to n'. Then, the depletion time, $T^p(\tilde{n}, 0)$, can be written as follows

$$T^p(\tilde{n}, 0) = T^p(\tilde{n}, \text{axes}) + \frac{y_1(\tilde{n})}{\mu_1(1 - \frac{\rho_2}{c_2}(1-c_1) - \rho_1)} + \frac{y_2(\tilde{n})}{\mu_2(1 - \frac{\rho_1}{c_1}(1-c_2) - \rho_2)}, \quad (8.48)$$

where $T^p(\tilde{n}, \text{axes}) = \min\left(\frac{\tilde{n}_1}{(\mu_1 c_1 - \lambda_1)^+}, \frac{\tilde{n}_2}{(\mu_2 c_2 - \lambda_2)^+}\right)$ is the time until the trajectory hits either one of the axes, and $y(\tilde{n})$ represents the point where the trajectory hits the axis when started in $\tilde{n}$. Note that $y_1(\tilde{n}) = \tilde{n}_1 - T^p(\tilde{n}, \text{axes}) \cdot \mu_1(c_1 - \rho_1)$ and $y_2(\tilde{n}) = \tilde{n}_2 - T^p(\tilde{n}, \text{axes}) \cdot \mu_2(c_2 - \rho_2)$. Hence, the depletion time scales as follows: $T^p(a \cdot \tilde{n}, 0) = a \cdot T^p(\tilde{n}, 0)$, $a \geq 0$.

Let $0 < \zeta < 1$ be fixed, and $x > 0$. We now have the following upper bound for all initial states n with $d_1 n_1 + d_2 n_2 = x$:

$$\begin{aligned}
&\int_0^D (d_1 n_1^{*D}(t) + d_2 n_2^{*D}(t))\mathrm{d}t \leq \int_0^D (d_1 n_1^p(t) + d_2 n_2^p(t))\mathrm{d}t \\
&\leq \sup_{0 \leq t \leq D} \{d_1 n_1^p(t) + d_2 n_2^p(t)\} \cdot T^p(n, 0) \leq x \cdot \zeta \cdot (1-\zeta) \cdot H(x). \quad (8.49)
\end{aligned}$$

Here the function $H(x)$ is defined as

$$H(x) := \frac{\beta}{\zeta \cdot (1-\zeta)} \cdot \sup_{l: d_1 l_1 + d_2 l_2 = x} \{T^p(l, 0)\},$$

with the constant

$$\beta := 1 + \max(0, \frac{d_1(\lambda_1 - \mu_1 c_1) + d_2(\lambda_2 - \mu_2 c_2)}{\mu_1 c_1 - \lambda_1}, \frac{d_1(\lambda_1 - \mu_1 c_1) + d_2(\lambda_2 - \mu_2 c_2)}{\mu_2 c_2 - \lambda_2}),$$

so that for all initial states n with $d_1n_1+d_2n_2=x$ it holds that $\sup_{0\leq t\leq D}\{d_1n_1^p(t)+d_2n_2^p(t)\}=\max\left(x,\ x+T^p(n,\text{axes})\cdot(d_1(\lambda_1-\mu_1c_1)+d_2(\lambda_2-\mu_2c_2))\right)\leq\beta\cdot x$.

From (8.48) it easily follows that $T^p(l,0)$ is continuous in l. This implies that $\sup_{l:d_1l_1+d_2l_2=x}T^p(l,0)<\infty$ and in particular $H(x)<\infty$ for all $x>0$. Assume $D\geq H(x)$ (in particular, $D\geq(1-\zeta)\cdot H(x)$). From (8.49) we obtain that

$$\tau(x):=\arg\min_{t\geq 0}\{d_1n_1^{*D}(t)+d_2n_2^{*D}(t)\leq x\cdot\zeta\}\leq(1-\zeta)\cdot H(x),\tag{8.50}$$

for all initial states n with $d_1n_1+d_2n_2=x$.

From continuity of $n^{*D}(t)$ it follows that $d_1n_1^{*D}(\tau(x))+d_2n_2^{*D}(\tau(x))=x\cdot\zeta$. Hence, if $n^{*D}(0)=(n_1,n_2)$, then $n^{*D}\left(\sum_{m=1}^{\infty}\tau((d_1n_1+d_2n_2)\zeta^{m-1})\right)=(0,0)$. Note that $H(a\cdot x)=a\cdot H(x)$, $a\geq 0$. Together with (8.50) it follows that $\sum_{m=1}^{\infty}\tau((d_1n_1+d_2n_2)\zeta^{m-1})\leq\sum_{m=1}^{\infty}\zeta^{m-1}(1-\zeta)\cdot H(d_1n_1+d_2n_2)=H(d_1n_1+d_2n_2)<\infty$. Hence, relation (8.46) holds. □

8.D Proof of Lemma 8.3.3

We construct policy π below. Note that $u_2^{\pi}(t)=0$ when $n_1^{\pi}(t)>0$.

- If $n_1^{\pi}(t)>0$ and $n_2^{\pi}(t)>0$, then $u_c^{\pi}(t)=u_2^{\tilde\pi}(t)+u_c^{\tilde\pi}(t)$, $u_1^{\pi}(t)=u_1^{\tilde\pi}(t)$ and $u_2^{\pi}(t)=0$.
- If $n_1^{\pi}(t)=0$ and $n_2^{\pi}(t)>0$, then $u_c^{\pi}(t)=\min\left(u_2^{\tilde\pi}(t)+u_c^{\tilde\pi}(t),\frac{\rho_1}{c_1}\right)$, $u_1^{\pi}(t)=\min\left(u_1^{\tilde\pi}(t),\rho_1-c_1u_c^{\pi}(t)\right)$ and $u_2^{\pi}(t)=u_c^{\tilde\pi}(t)+u_1^{\tilde\pi}(t)+u_2^{\tilde\pi}(t)-u_c^{\pi}(t)-u_1^{\pi}(t)$.
- If $n_1^{\pi}(t)>0$ and $n_2^{\pi}(t)=0$, then $u_c^{\pi}(t)=\min\left(u_2^{\tilde\pi}(t)+u_c^{\tilde\pi}(t),\frac{\rho_2}{c_2}\right)$, $u_1^{\pi}(t)=u_c^{\tilde\pi}(t)+u_1^{\tilde\pi}(t)+u_2^{\tilde\pi}(t)-u_c^{\pi}(t)$ and $u_2^{\pi}(t)=0$.
- If $n_1^{\pi}(t)=0$ and $n_2^{\pi}(t)=0$, then take $u^{\pi}(t)$ such that $\rho_i=u_i^{\pi}(t)+c_iu_c^{\pi}(t)$, $i=1,2$.

Once $n_1^{\pi}(t)+n_2^{\pi}(t)=0$, policy π will keep the system empty from that moment on (this is possible since the stability conditions are satisfied). Therefore, we will focus on states with $n_1^{\pi}(t)+n_2^{\pi}(t)>0$.

For policies π and $\tilde\pi$, we will prove the following inequalities:

$$U_1^{\pi}(t)+c_1U_c^{\pi}(t)\geq U_1^{\tilde\pi}(t)+c_1U_c^{\tilde\pi}(t),\tag{8.51}$$

$$U_1^{\pi}(t)+U_2^{\pi}(t)+(c_1+c_2)U_c^{\pi}(t)\geq U_1^{\tilde\pi}(t)+U_2^{\tilde\pi}(t)+(c_1+c_2)U_c^{\tilde\pi}(t),\tag{8.52}$$

$$(1-c_2)U_1^{\pi}(t)+c_1(U_2^{\pi}(t)+U_c^{\pi}(t))\geq(1-c_2)U_1^{\tilde\pi}(t)+c_1(U_2^{\tilde\pi}(t)+U_c^{\tilde\pi}(t)).\tag{8.53}$$

They are similar to the inequalities of the stochastic model (8.9)–(8.11) when setting $S_i(t)=U_i(t)+c_iU_c(t)$. When multiplying (8.51) by $d_1\mu_1-d_2\mu_2\geq 0$ and (8.52) by $d_2\mu_2$ and adding the two inequalities, we obtain $d_1\mu_1U_1^{\pi}(t)+d_2\mu_2U_2^{\pi}(t)+(d_1\mu_1c_1+d_2\mu_2c_2)U_c^{\pi}(t)\geq d_1\mu_1U_1^{\tilde\pi}(t)+d_2\mu_2U_2^{\tilde\pi}(t)+(d_1\mu_1c_1+d_2\mu_2c_2)U_c^{\tilde\pi}(t)$. By (8.13) we get $d_1n_1^{\pi}(t)+d_2n_2^{\pi}(t)\leq d_1n_1^{\tilde\pi}(t)+d_2n_2^{\tilde\pi}(t)$ for all $t\geq 0$, which was to be proved.

The remainder of the appendix is devoted to the proof of inequalities (8.51)–(8.53). Throughout the proof, we consider the workload fluid processes $w_i(\cdot) = n_i(t)/\mu_i$, $i = 1, 2$.

Note that $U_j(t)$, $j = 1, 2, c$, is continuous. In order to show (8.51), we therefore consider the first time instant t such that (8.51) holds with equality and is violated immediately after time t. So $U_1^{\pi}(t) + c_1 U_c^{\pi}(t) = U_1^{\tilde{\pi}}(t) + c_1 U_c^{\tilde{\pi}}(t)$, and by (8.13) also $n_1^{\pi}(t) = n_1^{\tilde{\pi}}(t)$, while $u_1^{\pi}(t^+) + c_1 u_c^{\pi}(t^+) < u_1^{\tilde{\pi}}(t^+) + c_1 u_c^{\tilde{\pi}}(t^+)$, so $n_1^{\pi}(t^+) > n_1^{\tilde{\pi}}(t^+)$. Since $n_1^{\pi}(t^+) > 0$, by construction of policy π we obtain $u_1^{\pi}(t) + c_1 u_c^{\pi}(t) \geq u_1^{\tilde{\pi}}(t) + c_1 u_c^{\tilde{\pi}}(t)$, which gives contradiction. Hence (8.51) holds for all $t \geq 0$.

Let time t be the first time instant that either (8.52) or (8.53) holds with equality and is violated immediately after time t. The remainder of the proof consists of two parts, depending on whether equation (8.52) or (8.53) is the first to be violated.

Part I: Assume (8.52) is the first equation that fails to hold, i.e., $U_1^{\pi}(t)+U_2^{\pi}(t)+(c_1+c_2)U_c^{\pi}(t) = U_1^{\tilde{\pi}}(t)+U_2^{\tilde{\pi}}(t)+(c_1+c_2)U_c^{\tilde{\pi}}(t)$, and by (8.13) also $w_1^{\pi}(t)+w_2^{\pi}(t) = w_1^{\tilde{\pi}}(t)+w_2^{\tilde{\pi}}(t)$, while $u_1^{\pi}(t^+) + u_2^{\pi}(t^+) + (c_1+c_2)u_c^{\pi}(t^+) < u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + (c_1+c_2)u_c^{\tilde{\pi}}(t^+)$. In what follows we use the following implication, which will be proved later on:

$$w_1^{\pi}(t) + w_2^{\pi}(t) = w_1^{\tilde{\pi}}(t) + w_2^{\tilde{\pi}}(t) \quad \text{implies} \quad w_i^{\pi}(t) = w_i^{\tilde{\pi}}(t), \quad i = 1, 2. \qquad (8.54)$$

We now distinguish between three cases: (i) If $w_1^{\pi}(t^+) > 0$ and $w_2^{\pi}(t^+) > 0$, then by construction of policy π, $u_c^{\pi}(t^+) \geq u_c^{\tilde{\pi}}(t^+)$. (ii) If $w_1^{\pi}(t^+) = 0$, then $0 = w_1^{\pi}(t)(= w_1^{\tilde{\pi}}(t))$, since $w_1(\cdot)$ is continuous. Policy π is able to keep class 1 empty at time t^+ while $\tilde{\pi}$ might not, so we have $\rho_1 = u_1^{\pi}(t^+) + c_1 u_c^{\pi}(t^+) \geq u_1^{\tilde{\pi}}(t^+) + c_1 u_c^{\tilde{\pi}}(t^+)$. In particular, $u_c^{\tilde{\pi}}(t^+) \leq \rho_1/c_1$, and by construction of policy π, this implies $u_c^{\pi}(t^+) \geq u_c^{\tilde{\pi}}(t^+)$. (iii) If $w_2^{\pi}(t^+) = 0$, then $0 = w_2^{\pi}(t)(= w_2^{\tilde{\pi}}(t))$, since $w_2(\cdot)$ is continuous. In a similar fashion as in the previous case, we obtain that $u_c^{\pi}(t^+) \geq u_c^{\tilde{\pi}}(t^+)$. Hence, in all cases it holds that $u_c^{\pi}(t^+) \geq u_c^{\tilde{\pi}}(t^+)$. Together with $c_1 + c_2 \geq 1$ and $u_1^{\pi}(t^+) + u_2^{\pi}(t^+) + u_c^{\pi}(t^+) = u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)$, we can conclude that $u_1^{\pi}(t^+) + u_2^{\pi}(t^+) + (c_1+c_2)u_c^{\pi}(t^+) \geq u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + (c_1+c_2)u_c^{\tilde{\pi}}(t^+)$, and we reach a contradiction. It now only remains to prove that the implication in (8.54) is satisfied. We distinguish between the following two cases:

- Assume there is an interval $[u, t)$ in which policy $\tilde{\pi}$ has more work in the system compared to policy π, i.e., $w_1^{\pi}(v) + w_2^{\pi}(v) < w_1^{\tilde{\pi}}(v) + w_2^{\tilde{\pi}}(v)$, for all $v \in [u, t)$. If the interval is such that $w_1^{\pi}(v) > 0$ and $w_2^{\pi}(v) > 0$, for all $v \in [u, t)$, then policy $\tilde{\pi}$ can never catch up with π (by construction of policy π). Hence, we can choose the interval $[u, t)$ such that:

  (i) For all $v \in [u, t)$, $w_2^{\pi}(v) = 0$ and $w_1^{\pi}(v) > 0$.

  (ii) For all $v \in [u, t)$, $w_1^{\pi}(v) = 0$ and $w_2^{\pi}(v) > 0$.

  Note that the two cases are mutually exclusive. We show that (8.54) holds in both cases.

  By continuity of $w_2^{\pi}(\cdot)$, in case (i) we have as well $w_2^{\pi}(t) = 0$. Hence, $w_1^{\pi}(t) = w_1^{\tilde{\pi}}(t) + w_2^{\tilde{\pi}}(t)$. By (8.13) and (8.51) we have $w_1^{\pi}(t) \leq w_1^{\tilde{\pi}}(t)$. Together this gives $w_2^{\tilde{\pi}}(t) = 0$ $(= w_2^{\pi}(t))$ and $w_1^{\tilde{\pi}}(t) = w_1^{\pi}(t)$. Hence, in case (i), (8.54) is proved.

Let $M_j^{\hat{\pi}}(u,t) = \int_u^t u_j^{\hat{\pi}}(s)\mathrm{d}s$ be the cumulative amount of time that activity j occurs under policy $\hat{\pi}$ in the time interval $[u,t)$. The total amount of additional capacity that policy $\tilde{\pi}$ gets compared with policy π in the interval $[u,t)$ is

$$\begin{aligned}&(c_1+c_2)M_c^{\tilde{\pi}}(u,t) + M_1^{\tilde{\pi}}(u,t) + M_2^{\tilde{\pi}}(u,t) - (c_1+c_2)M_c^{\pi}(u,t)\\ &- M_1^{\pi}(u,t) - M_2^{\pi}(u,t) = (c_1+c_2-1)(M_c^{\tilde{\pi}}(u,t) - M_c^{\pi}(u,t)),\end{aligned}$$

where we used that $M_c^{\tilde{\pi}}(u,t) + M_1^{\tilde{\pi}}(u,t) + M_2^{\tilde{\pi}}(u,t) = M_c^{\pi}(u,t) + M_1^{\pi}(u,t) + M_2^{\pi}(u,t)$. This is equal to the difference in the total workload at time u, so $(c_1+c_2-1)(M_c^{\tilde{\pi}}(u,t) - M_c^{\pi}(u,t)) = w_1^{\tilde{\pi}}(u) + w_2^{\tilde{\pi}}(u) - w_1^{\pi}(u) - w_2^{\pi}(u)$. In case (ii), $w_1^{\pi}(u) = 0$, hence we obtain from (8.13) and (8.53) that $c_1 w_2^{\pi}(u) = (1-c_2)w_1^{\pi}(u) + c_1 w_2^{\pi}(u) \leq (1-c_2)w_1^{\tilde{\pi}}(u) + c_1 w_2^{\tilde{\pi}}(u)$. Rewriting this gives

$$w_1^{\tilde{\pi}}(u) \leq \frac{c_1}{c_1+c_2-1}(w_1^{\tilde{\pi}}(u) + w_2^{\tilde{\pi}}(u) - w_2^{\pi}(u)) = c_1(M_c^{\tilde{\pi}}(u,t) - M_c^{\pi}(u,t)). \tag{8.55}$$

Note that $\rho_1(t-u) = c_1 M_c^{\pi}(u,t) + M_1^{\pi}(u,t)$ (since in case (ii) class 1 is kept empty under policy π), and $M_1^{\tilde{\pi}}(u,t) \geq M_1^{\pi}(u,t)$ (by definition of policy π). Together with (8.55) this gives

$$w_1^{\tilde{\pi}}(t) = w_1^{\tilde{\pi}}(u) + \rho_1(t-u) - c_1 M_c^{\tilde{\pi}}(u,t) - M_1^{\tilde{\pi}}(u,t) \leq 0.$$

By continuity of $w_1^{\pi}(\cdot)$, in case (ii) we have as well $w_1^{\pi}(t) = 0$. Hence it follows immediately from $w_1^{\pi}(t) + w_2^{\pi}(t) = w_1^{\tilde{\pi}}(t) + w_2^{\tilde{\pi}}(t)$ that $w_i^{\pi}(t) = w_i^{\tilde{\pi}}(t)$, $i = 1,2$.

- Now consider the case when there is an interval $[v,t]$ such that $w_1^{\pi}(u) + w_2^{\pi}(u) = w_1^{\tilde{\pi}}(u) + w_2^{\tilde{\pi}}(u)$ for all $u \in [v,t]$ and $w_1^{\pi}(v^-) + w_2^{\pi}(v^-) < w_1^{\tilde{\pi}}(v^-) + w_2^{\tilde{\pi}}(v^-)$. From the previous item, we obtain that $w_i^{\pi}(v) = w_i^{\tilde{\pi}}(v)$, $i = 1,2$. Together with the fact that in the interval $[v,t]$ the total workload is equal under both policies, and by construction of policy π, it follows that π does exactly the same as policy $\tilde{\pi}$. Hence, $w_i^{\pi}(u) = w_i^{\tilde{\pi}}(u)$ for all $u \in [v,t]$, $i = 1,2$.

Part II: Assume (8.53) is the first equation that fails to hold, i.e., $(1-c_2)U_1^{\pi}(t) + c_1(U_2^{\pi}(t) + U_c^{\pi}(t)) = (1-c_2)U_1^{\tilde{\pi}}(t) + c_1(U_2^{\tilde{\pi}}(t) + U_c^{\tilde{\pi}}(t))$, and by (8.13) also $(1-c_2)w_1^{\pi}(t) + c_1 w_2^{\pi}(t) = (1-c_2)w_1^{\tilde{\pi}}(t) + c_1 w_2^{\tilde{\pi}}(t)$, while $(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+)) < (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$. We have the following possibilities:

- If $w_1^{\pi}(t^+) > 0$ and $w_2^{\pi}(t^+) > 0$, then by definition of policy π we have $(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+)) = (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$.
- If $w_1^{\pi}(t^+) = 0$ and $w_2^{\pi}(t^+) > 0$, then we distinguish between the following three cases:

  (i) If $\rho_1 \leq c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$, then $u_1^{\pi}(t^+) = 0, u_2^{\pi}(t^+) = u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+) - \frac{\rho_1}{c_1}$ and $u_c^{\pi}(t^+) = \frac{\rho_1}{c_1}$. Since $c_1 + c_2 > 1$, we have

$$\begin{aligned}(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+)) &= c_1(u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))\\ &\geq (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)).\end{aligned}$$

(ii) If $c_1(u_2^{\tilde{\pi}}(t^+)+u_c^{\tilde{\pi}}(t^+)) < \rho_1 \leq u_1^{\tilde{\pi}}(t^+)+c_1(u_2^{\tilde{\pi}}(t^+)+u_c^{\tilde{\pi}}(t^+))$, then $u_1^{\pi}(t^+) = \rho_1 - c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)), u_2^{\pi}(t^+) = u_1^{\tilde{\pi}}(t^+) - \rho_1 + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$ and $u_c^{\pi}(t^+) = u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)$. Together with $c_1 + c_2 > 1$, we obtain

$$\begin{aligned}
&(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+))\\
&= (1-c_2)(\rho_1 - c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)))\\
&\quad + c_1(u_1^{\tilde{\pi}}(t^+) + u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+) - \rho_1 + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)))\\
&= (1-c_1-c_2)\rho_1 + c_1(c_1+c_2)(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)) + c_1 u_1^{\tilde{\pi}}(t^+)\\
&\geq (1-c_1-c_2)(u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))) + c_1(c_1+c_2)(u_2^{\tilde{\pi}}(t^+)\\
&\quad + u_c^{\tilde{\pi}}(t^+)) + c_1 u_1^{\tilde{\pi}}(t^+)\\
&= (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)).
\end{aligned}$$

(iii) If $u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)) < \rho_1$, then $u_1^{\pi}(t^+) = u_1^{\tilde{\pi}}(t^+), u_2^{\pi}(t^+) = 0$ and $u_c^{\pi}(t^+) = u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)$. So we have $(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+)) = (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$.

- If $w_1^{\pi}(t^+) > 0$ and $w_2^{\pi}(t^+) = 0$, then by continuity of $w_2^{\pi}(\cdot)$ we have $w_2^{\pi}(t) = 0$. Hence, $(1-c_2)w_1^{\pi}(t) = (1-c_2)w_1^{\tilde{\pi}}(t) + c_1 w_2^{\tilde{\pi}}(t)$. Since also $w_1^{\pi}(t) \leq w_1^{\tilde{\pi}}(t)$, this gives $w_1^{\pi}(t) = w_1^{\tilde{\pi}}(t)$ and $0 = w_2^{\pi}(t) = w_2^{\tilde{\pi}}(t)$. Note that when $w_2^{\tilde{\pi}}(t^+) = 0$, then $u_2^{\tilde{\pi}}(t^+)+c_2u_c^{\tilde{\pi}}(t^+) = \rho_2$. If instead $w_2^{\tilde{\pi}}(t^+) > 0$, then $u_2^{\tilde{\pi}}(t^+)+c_2u_c^{\tilde{\pi}}(t^+) < u_2^{\pi}(t^+)+c_2u_c^{\pi}(t^+) = \rho_2$ (the inequality follows from $0 = w_2^{\pi}(t) = w_2^{\tilde{\pi}}(t)$, and the fact that policy π is able to keep class 2 empty at time t^+, while policy $\tilde{\pi}$ is not). Hence, it holds that $u_2^{\tilde{\pi}}(t^+) + c_2u_c^{\tilde{\pi}}(t^+) \leq \rho_2$ (so also $u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+) \leq \frac{\rho_2}{c_2}$). By construction of policy π, this implies $u_c^{\pi}(t^+) = u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+)$, $u_1^{\pi}(t^+) = u_1^{\tilde{\pi}}(t^+)$ and $u_2^{\pi}(t^+) = 0$. Hence, $(1-c_2)u_1^{\pi}(t^+) + c_1(u_2^{\pi}(t^+) + u_c^{\pi}(t^+)) = (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$.

For all the three possibilities we reach a contradiction with $(1-c_2)u_1^{\pi}(t^+)+c_1(u_2^{\pi}(t^+)+u_c^{\pi}(t^+)) < (1-c_2)u_1^{\tilde{\pi}}(t^+) + c_1(u_2^{\tilde{\pi}}(t^+) + u_c^{\tilde{\pi}}(t^+))$ and this concludes the proof. □

8.E Proof of Lemma 8.3.9

Let $\overline{N}_i^{\pi^*}(t)$, $i = 1, 2$, $\overline{T}_j^{\pi^*}(t)$, $j = 1, 2, c, 0$, be a fluid limit of policy π^*. So the functions $\overline{N}_i^{\pi^*}(t)$, $i = 1, 2$, satisfy (8.25), and the functions $\overline{T}_j^{\pi^*}(\cdot)$, $j = I, 1, 2, c$, are absolutely continuous (follows from Lipschitz continuity), and hence are differentiable almost everywhere. Fix a sample path ω such that there is a subsequence r_k with $\lim_{r_k\to\infty} \overline{N}_i^{\pi^*,r_k}(t) = \overline{N}_i^{\pi^*}(t)$, $i = 1, 2$, u.o.c., and $\lim_{r_k\to\infty} \overline{T}_j^{\pi^*,r_k}(t) = \overline{T}_j^{\pi^*}(t)$, $j = 1, 2, c$, u.o.c.. Further, let $t > 0$ be a regular point of $\overline{T}_j^{\pi^*}(t)$ for all $j = I, 1, 2, c$.

First assume $\overline{N}_2^{\pi^*}(t) < \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(t)$. Then there is an $\epsilon > 0$ such that $\overline{N}_2^{\pi^*}(s) < \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(s)$ for $s \in [t-\epsilon, t+\epsilon]$. By the uniform convergence of $\overline{N}_i^{\pi^*,r_k}(\cdot)$ to $\overline{N}_i^{\pi^*}(\cdot)$, $i = 1, 2$, on $[t-\epsilon, t+\epsilon]$, we have $N_2^{\pi^*,r_k}(r_k s) < \alpha\frac{\mu_2}{\mu_1}N_1^{\pi^*,r_k}(r_k s)$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. Hence, under policy π^*, in the interval $[r_k(t-\epsilon), r_k(t+\epsilon)]$ class 1 is served and we obtain $\overline{T}_1^{\pi^*,r_k}(t+\epsilon) - \overline{T}_1^{\pi^*,r_k}(t-\epsilon) = 2\epsilon$. Letting $r_k \to \infty$ and $\epsilon \downarrow 0$ we obtain $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t)}{\mathrm{d}t} = 1$.

Now assume $\overline{N}_2^{\pi^*}(t) > \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(t)$ and $\overline{N}_1^{\pi^*}(t) > 0$. Then there is an ϵ such that $N_2^{\pi^*,r_k}(r_k s) > \alpha\frac{\mu_2}{\mu_1}N_1^{\pi^*,r_k}(r_k s)$ and $N_1^{\pi^*,r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. Under policy π^*, in this interval both classes are served in parallel, hence $\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t} = 1$.

Assume $\overline{N}_2^{\pi^*}(t) = \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(t)$ and $\overline{N}_1^{\pi^*}(t) > 0$. Then there is an ϵ such that $N_1^{\pi^*,r_k}(r_k s) > 0$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. In this interval, class 2 is never served individually, so $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s} + \frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s} = 1$, for any regular point $s \in [t-\epsilon, t+\epsilon]$. Together with (8.25), we obtain

$$\begin{aligned}
&\alpha\frac{\mu_2}{\mu_1}\frac{\mathrm{d}\overline{N}_1^{\pi^*}(s)}{\mathrm{d}s} - \frac{\mathrm{d}\overline{N}_2^{\pi^*}(s)}{\mathrm{d}s}\\
&= \mu_2\Big(-\alpha\cdot(-\rho_1 + \frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s} + c_1\frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s}) - \rho_2 + c_2\frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s}\Big)\\
&< \mu_2\Big(-\frac{c_2-\rho_2}{c_1-\rho_1}\cdot(-\rho_1 + \frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s} + c_1\frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s}) - \rho_2 + c_2\frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s}\Big)\\
&= \mu_2\Big(-\frac{c_2-\rho_2}{c_1-\rho_1}\cdot(c_1 - \rho_1 + \frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s}(1-c_1)) - \rho_2 + c_2 - c_2\frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s}\Big)\\
&= -\frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s}\cdot\frac{\mu_2}{c_1-\rho_1}\cdot\Big((c_2-\rho_2)(1-c_1) + c_2(c_1-\rho_1)\Big)\\
&= -\frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s}\cdot\frac{\mu_2 c_2}{c_1-\rho_1}\cdot\Big(1-\rho_1-\frac{\rho_2}{c_2}(1-c_1)\Big) \le 0, \qquad (8.56)
\end{aligned}$$

whenever $s \in [t-\epsilon, t+\epsilon]$ is a regular point. Here we used that $c_1 + c_2 > 1$, $\rho_1 < c_1 \le 1$, $\rho_2 < 1 - \frac{\rho_1}{c_1}(1-c_2)$, $\alpha > \frac{c_2-\rho_2}{c_1-\rho_1}$ and $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(s)}{\mathrm{d}s} + \frac{\mathrm{d}\overline{T}_c^{\pi^*}(s)}{\mathrm{d}s} = 1$. Equation (8.56) implies that if at a certain time $\overline{N}^{\pi^*}$ lies below the switching curve, then it moves towards the switching curve and if $\overline{N}^{\pi^*}$ lies on or above the switching curve, it will move away from (and above) the switching curve. Since at time t we are in a state on the switching curve, we have $\overline{N}_2^{\pi^*}(s) < \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(s)$ for $s \in [t-\epsilon, t)$ and $\overline{N}_2^{\pi^*}(s) > \alpha\frac{\mu_2}{\mu_1}\overline{N}_1^{\pi^*}(s)$ for $s \in (t, t+\epsilon]$. Note that $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t-)}{\mathrm{d}t} = 1$ and $\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t-)}{\mathrm{d}t} = 0$, while $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t+)}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t+)}{\mathrm{d}t} = 1$, so that the point t itself is not a regular point.

Finally assume $\overline{N}_1^{\pi^*}(t) = 0$ and $\overline{N}_2^{\pi^*}(t) > 0$. Then there is an $\epsilon > 0$ such that $\overline{N}_2^{\pi^*}(s) > \alpha\frac{\mu_1}{\mu_2}\overline{N}_1^{\pi^*}(s)$ for $s \in [t-\epsilon, t+\epsilon]$ and hence $N_2^{\pi^*,r_k}(r_k s) > \alpha\frac{\mu_1}{\mu_2}N_1^{\pi^*,r_k}(r_k s)$ for all r_k large enough and $s \in [t-\epsilon, t+\epsilon]$. In this interval class 1 is not served individually under policy π^*, hence $\frac{\mathrm{d}\overline{T}_1^{\pi^*}(t)}{\mathrm{d}t} = 0$. From (8.25) we then have

$$\frac{\mathrm{d}\overline{N}_1^{\pi^*}(t)}{\mathrm{d}t} = \lambda_1 - \mu_1 c_1 \frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t}. \tag{8.57}$$

Note that if $\overline{N}_1^{\pi^*}(t+\delta) > 0$, for all $0 < \delta < \Delta$, then $\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t+\delta)}{\mathrm{d}t} = 1$. Since $\rho_1 < c_1$, from (8.57) we see that class 1 will stay empty, and thus $\frac{\mathrm{d}\overline{T}_c^{\pi^*}(t)}{\mathrm{d}t} = \frac{\rho_1}{c_1}$. We conclude that (8.26)–(8.28) are satisfied for each fluid limit $\overline{T}^{\pi^*}(t)$.

From (8.25) and (8.26)–(8.28) it follows that $\overline{N}_i^{\pi^*}(t)$ is uniquely determined. Using the correspondence $u_j^*(t) = \frac{\mathrm{d}\overline{T}_j^{\pi^*}(t)}{\mathrm{d}t}$, $j = 1, 2, c$, with $u^*(t)$ as defined in Proposition 8.3.5, it follows that $\overline{N}^{\pi^*}(t) = n^*(t)$, with $n^*(t)$ the trajectory corresponding to the control $u^*(t)$.

Bibliography

[1] S. Aalto and U. Ayesta. SRPT applied to bandwidth-sharing networks. *Annals of Operations Research*, 170:3–19, 2009.

[2] S. Aalto, U. Ayesta, S.C. Borst, V. Misra, and R. Núñez-Queija. Beyond processor sharing. *Performance Evaluation Review*, 34(4):36–43, 2007.

[3] S. Aalto, U. Ayesta, and R. Righter. On the Gittins index in an M/G/1 queue. *Queueing Systems*, 2009. To appear.

[4] M. Abundo. Some conditional crossing results of Brownian motion over a piecewise-linear boundary. *Statistics & Probability Letters*, 58:131–145, 2002.

[5] E. Altman, K.E. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53:53–63, 2006.

[6] E. Altman, K.E. Avrachenkov, and R. Núñez-Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. *Advances in Applied Probability*, 36:839–853, 2004.

[7] E. Altman, T. Jimenez, and D. Kofman. DPS queues with stationary ergodic service times and the performance of TCP in overload. In *Proceedings of IEEE INFOCOM*, Hong Kong, 2004.

[8] M. Armony and N. Bambos. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems*, 44:209–252, 2003.

[9] S. Asmussen. *Applied Probability and Queues.* Springer-Verlag, New York, 2003.

[10] B. Ata and S. Kumar. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Annals of Applied Probability*, 15:331–391, 2005.

[11] B. Avi-Itzhak, H. Levy, and D. Raz. Quantifying fairness in queuing systems: Principles, approaches, and applicability. *Probability in the Engineering and Informational Sciences*, 22:495–517, 2008.

[12] K.E. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM*, Miami FL, USA, 2005.

[13] F. Avram. Optimal control of fluid limits of queueing networks and stochasticity corrections. *Mathematics of stochastic manufacturing systems. Lectures in Applied Mathematics*, 33:1–36, 1997.

[14] F. Avram, D. Bertsimas, and M. Ricard. Fluid models of sequencing problems in open queueing networks: An optimal control approach. *IMA Volumes in Mathematics and its Applications*, 71:199–234, 1995.

[15] N. Bäuerle. Asymptotic optimality of tracking policies in stochastic networks. *Annals of Applied Probability*, 10:1065–1083, 2000.

[16] N. Bäuerle. Optimal control of queueing networks: An approach via fluid models. *Advances in Applied Probability*, 34:313–328, 2002.

[17] M. Bayati, M. Sharma, and M.S. Squillante. Optimal scheduling in a multi-server stochastic network. *Performance Evaluation Review*, 34(3):45–47, 2006.

[18] L. Beghin and E. Orsingher. On the maximum of the generalized Brownian bridge. *Lithuanian Mathematical Journal*, 39:157–167, 1999.

[19] S.L. Bell and R.J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Annals of Applied Probability*, 11:608–649, 2001.

[20] S.L. Bell and R.J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electronic Journal of Probability*, 10:1044–1115, 2005.

[21] R.E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, 1957.

[22] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J.W. Roberts. Statistical bandwidth sharing: A study of congestion at flow level. In *Proceedings of ACM SIGCOMM*, pages 111–122, San Diego CA, USA, 2001.

[23] A. Ben Tahar and A. Jean-Marie. The fluid limit of the multiclass processor sharing queue. *INRIA Research Report*, RR-6867, 2009.

[24] D. Bertsekas and R. Gallager. *Data Networks.* Prentice-Hall, Englewood Cliffs, 1987.

[25] S. Bhardwaj and R.J. Williams. Diffusion approximation for a heavily loaded multi-user wireless communication system with cooperation. *Queueing Systems*, 2009. DOI: 10.1007/s11134-009-9119-8.

[26] S. Bhardwaj, R.J. Williams, and A.S. Acampora. On the performance of a two user MIMO downlink system in heavy traffic. *IEEE Transactions on Information Theory*, 53:1851–1859, 2007.

[27] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1999.

[28] T. Bonald, S.C. Borst, N. Hegde, and A. Proutière. Wireless data performance in multi-cell scenarios. In *Proceedings of ACM SIGMETRICS/Performance*, pages 378–388, New York NY, USA, 2004.

[29] T. Bonald, S.C. Borst, and A. Proutière. Inter-cell coordination in wireless data networks. *European Transactions on Telecommunications*, 17:303–312, 2006.

[30] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. In *Proceedings of ACM SIGMETRICS/Performance*, pages 82–91, Boston MA, USA, 2001.

[31] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems*, 53:65–84, 2006.

[32] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Systems*, 44:69–100, 2003.

[33] T. Bonald and A. Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Systems*, 47:81–106, 2004.

[34] T. Bonald and A. Proutière. Flow-level stability of utility-based allocations for non-convex rate regions. In *Proceedings of CISS Conference on Information Sciences and Systems*, 2006.

[35] S.C. Borst, M. Mandjes, and M.J.G. van Uitert. Generalized processor sharing queues with heterogeneous traffic classes. *Advances in Applied Probability*, 35:806–845, 2003.

[36] S.C. Borst, R. Núñez-Queija, and A.P. Zwart. Sojourn time asymptotics in processor-sharing queues. *Queueing Systems*, 53:31–51, 2006.

[37] M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4:414–431, 1994.

[38] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17:237–238, 1985.

[39] L. Cesari. *Optimization – Theory and Applications.* Springer-Verlag, New York, 1983.

[40] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization.* Springer-Verlag, New York, 2001.

[41] J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12:245–284, 1979.

[42] J.W. Cohen and O.J. Boxma. *Boundary Value Problems in Queueing System Analysis*. North-Holland, Amsterdam, 1983.

[43] P.J. Courtois. *Decomposability: Queueing and Computer System Applications*. Academic Press, New York, 1977.

[44] J.G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.

[45] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a Fair Queueing Algorithm. In *Proceedings of ACM SIGCOMM*, pages 1–12, 1989.

[46] R. Egorova, S.C. Borst, and A.P. Zwart. Bandwidth-sharing networks in overload. *Performance Evaluation*, 64:978–993, 2007.

[47] M. El-Taha and S. Stidham. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishers, 1999.

[48] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[49] G. Fayolle, A. de la Fortelle, J.M. Lasgouttes, L. Massoulié, and J.W. Roberts. Best-effort networks: Modeling and performance analysis via large networks asymptotics. In *Proceedings of IEEE INFOCOM*, Anchorage AK, USA, 2001.

[50] G. Fayolle and R. Iasnogorodski. Two coupled processors: the reduction to a Riemann-Hilbert problem. *Zeitschrift für Wahrscheinlichkeitstheorie*, 47:325–351, 1979.

[51] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27:519–532, 1980.

[52] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, New York, 1971.

[53] A. Gajrat and A. Hordijk. Fluid approximation of a controlled multiclass tandem network. *Queueing Systems*, 35:349–380, 2000.

[54] A. Gajrat, A. Hordijk, and A. Ridder. Large-deviations analysis of the fluid approximation for a controllable tandem queue. *Annals of Applied Probability*, 13:1423–1448, 2003.

[55] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Annals of Applied Probability*, 16:56–90, 2006.

[56] E. Gelenbe and I. Mitrani. *Analysis and Synthesis of Computer Systems.* Academic Press, London, 1980.

[57] J.C. Gittins. *Multi-Armed Bandit Allocation Indices.* Wiley, Chichester, 1989.

[58] G. Grishechkin. On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Advances in Applied Probability*, 24:653–698, 1992.

[59] J. Hansen, C. Reynolds, and S. Zachary. Stability of processor sharing networks with simultaneous resource requirements. *Journal of Applied Probability*, 44:636–651, 2007.

[60] M. Harchol-Balter, K. Sigman, and A. Wierman. Asymptotic convergence of scheduling policies with respect to slowdown. *Performance Evaluation*, 49:241–256, 2002.

[61] J.M. Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies. *Annals of Applied Probability*, 8:822–848, 1998.

[62] J.M. Harrison and M.J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33:339–368, 1999.

[63] M. Haviv and J. van der Wal. Mean sojourn times for phase-type discriminatory processor sharing systems. *European Journal of Operational Research*, 189:375–386, 2008.

[64] O. Hernández-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria.* Springer-Verlag, New York, 1996.

[65] V. Jacobson. Congestion avoidance and control. In *Proceedings of ACM SIGCOMM*, pages 314–329, 1988.

[66] A. Jean-Marie and P. Robert. On the transient behavior of the processor-sharing queue. *Queueing Systems*, 17:129–136, 1994.

[67] W.N. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams. Fluid and Brownian approximations for an Internet congestion control model. In *Proceedings of IEEE CDC*, pages 3938–3943, 2004.

[68] W.N. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Annals of Applied Probability*, 2009. To appear.

[69] F.P. Kelly. *Stochastic Networks and Reversibility.* Wiley, Chichester, 1979.

[70] F.P. Kelly, L. Massoulié, and N.S. Walton. Resource pooling in congested networks: Proportional fairness and product form. *Queueing Systems*, 2009. To appear.

[71] F.P. Kelly, A. Maulloo, and D. Tan. Rate control in communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[72] F.P. Kelly and R.J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 14:1055–1083, 2004.

[73] D.G. Kendall. 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, 1953.

[74] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM*, Miami FL, USA, 2005.

[75] B. Kim and J. Kim. Comparison of DPS and PS systems according to DPS weights. *IEEE Communications Letters*, 10(7):558–560, 2006.

[76] J.F.C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B*, 24:383–392, 1962.

[77] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14:242–261, 1967.

[78] L. Kleinrock. *Queueing Systems, Vol. II: Computer Applications*. Wiley, New York, 1976.

[79] G.M. Koole. Monotonicity in Markov reward and decision chains: Theory and applications. *Foundations and Trends in Stochastic Systems*, 1:1–76, 2006.

[80] H.J. Kushner. *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*. Springer-Verlag, New York, 2001.

[81] P. Lieshout, S.C. Borst, and M. Mandjes. Heavy-traffic approximations for linear networks operating under alpha-fair bandwidth-sharing policies. In *Proceedings of ValueTools*, Pisa, Italy, 2006.

[82] J. Liu, A. Proutière, Y. Yi, M. Chiang, and V.H. Poor. Flow-level stability of data networks with non-convex and time-varying rate regions. In *Proceedings of ACM SIGMETRICS*, pages 239–250, San Diego CA, USA, 2007.

[83] X. Liu, E. Chong, and N. Shroff. A framework for opportunistic scheduling in wireless networks. *Computer Networks*, 41:451–474, 2003.

[84] Z. Liu, P. Nain, and D. Towsley. Sample path methods in the control of queues. *Queueing Systems*, 21:293–335, 1995.

[85] F.J. López and G. Sanz. Markovian couplings staying in arbitrary subsets of the state space. *Journal of Applied Probability*, 39:197–212, 2002.

[86] R.M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambridge Philos.*, 58:497–520, 1962.

[87] S.H. Lu and P.R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36:1406–1416, 1991.

[88] C. Maglaras. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Annals of Applied Probability*, 10:897–929, 2000.

[89] A. Mandelbaum and A.L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52:836–855, 2004.

[90] P. Marbach. Priority service and max-min fairness. In *Proceedings of IEEE INFOCOM*, New York NY, USA, 2002.

[91] A. Mas-Colell, M.D. Whinston, and J.R. Green. *Microeconomic Theory*. Oxford University Press, New York, 1995.

[92] W.A. Massey. Stochastic orderings for Markov processes on partially ordered spaces. *Mathematics of Operations Research*, 12:350–367, 1987.

[93] L. Massoulié. Structural properties of proportional fairness: Stability and insensitivity. *Annals of Applied Probability*, 17:809–839, 2007.

[94] L. Massoulié and J.W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201, 2000.

[95] S.P. Meyn. Stability and optimization of queueing networks and their fluid models. *Mathematics of Stochastic Manufacturing Systems. Lectures in Applied Mathematics*, 33:175–199, 1997.

[96] S.P. Meyn. Dynamic safety-stocks for asymptotic optimality in stochastic networks. *Queueing Systems*, 50:255–297, 2005.

[97] S.P. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, New York, 2008.

[98] S.P. Meyn. Stability and asymptotic optimality of generalized MaxWeight policies. *SIAM Journal on Control and Optimization*, 47:3259–3294, 2009.

[99] J.A. van Mieghem. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability*, 5:809–833, 1995.

[100] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8:556–567, 2000.

[101] A.M. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester, 2002.

[102] P. Nain and D. Towsley. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Transactions on Automatic Control*, 39:1853–1855, 1994.

[103] A. Novikov, V. Frishling, and N. Kordzakhia. Approximations of boundary crossing probabilities for a Brownian motion. *Journal of Applied Probability*, 36:1019–1030, 1999.

[104] R. Núñez-Queija. *Processor-Sharing Models for Integrated-Services Networks.* Ph.D. Thesis Eindhoven University of Technology, 2000.

[105] M. Nuyens and A. Wierman. The foreground-background queue: A survey. *Performance Evaluation*, 65:286–307, 2008.

[106] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching times and thresholds. *Performance Evaluation*, 61:347–369, 2004.

[107] T. Osogami, M. Harchol-Balter, A. Scheller-Wolf, and L. Zhang. Exploring threshold-based policies for load sharing. In *Forty-second Annual Allerton Conference on Communication, Control, and Computing*, pages 1012–1021, 2004.

[108] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling TCP Reno performance: A simple model and its empirical validation. *IEEE/ACM Transactions on Networking*, 8:133–145, 2000.

[109] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Networking*, 1:344–357, 1993.

[110] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York, 1994.

[111] B. Radunović and J.-Y. Le Boudec. Rate performance objectives of multihop wireless networks. *IEEE Transactions on Mobile Computing*, 3:334–349, 2004.

[112] M.M. Rao. *Measure Theory and Integration.* Wiley, New York, 1987.

[113] K.M. Rege and B. Sengupta. Queue length distribution for the discriminatory processor-sharing queue. *Operations Research*, 44:653–657, 1996.

[114] R. Righter and J.G. Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, 3:323–333, 1989.

[115] P. Robert. *Stochastic Networks and Queues.* Springer-Verlag, New York, 2003.

[116] K. Ross and N. Bambos. Geometry of packet switching: Maximal throughput cone scheduling algorithms. In *High-performance Packet Switching Architectures, eds. I. Elhanany and M. Hamdi.*, pages 81–99, Springer, London, 2007.

[117] S.M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.

[118] D. Rubenstein, J. Kurose, and D. Towsley. The impact of multicast layering on network fairness. *IEEE/ACM Transactions on Networking*, 10:169–182, 2002.

[119] M. Sakata, S. Noguchi, and J. Oizumi. Analysis of a processor-shared queueing model for time-sharing systems. In *Proc. of the 2nd Hawaii International Conference on System Sciences*, pages 625–628, 1969.

[120] L.E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:687–690, 1968.

[121] L.E. Schrage and L.W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.

[122] A. Seierstad and K. Sydsæter. *Optimal Control Theory with Economic Applications*. North-Holland, Amsterdam, 1987.

[123] L.I. Sennott. Average reward optimization theory for denumerable state spaces. In *Handbook of Markov Decision Processes, eds. E.A. Feinberg and A. Shwartz*, pages 153–172, Kluwer, 2002.

[124] L.I. Sennott. Value iteration in countable state average cost Markov decision processes with unbounded costs. *Annals of Operations Research*, 28:261–271, 2005.

[125] D. Shah and D. Wischik. The teleology of scheduling algorithms for switched networks under light load, critical load, and overload. 2009. Submitted.

[126] M. Shaked and J.G. Shanthikumar. *Stochastic Orders and their Applications*. Academic Press, San Diego, 1993.

[127] D.R. Smith. A new proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 26:197–199, 1978.

[128] M.S. Squillante. Stochastic analysis of multiserver systems. *Performance Evaluation Review*, 34(4):44–51, 2007.

[129] M.S. Squillante, C.H. Xia, D.D. Yao, and L. Zhang. Threshold-based priority policies for parallel-server systems with affinity scheduling. In *Proc. of the IEEE American Control Conference*, volume 4, pages 2992–2999, 2001.

[130] R. Srikant. *The Mathematics of Internet Congestion Control*. Birkhäuser, Boston, 2004.

[131] S. Stidham Jr. and R.R. Weber. A survey of Markov decision models for control of networks of queues. *Queueing Systems*, 13:291–314, 1993.

[132] A.L. Stolyar. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, 14:1–53, 2004.

[133] H. Takagi. *Queueing Analysis, Vol. I: Vacation and Priority Systems*. North-Holland, Amsterdam, 1991.

[134] B. Tan, L. Ying, and R. Srikant. Short-term fairness and long-term QoS. In *Proceedings of CISS Conference on Information Sciences and Systems*, Princeton University, 2008.

[135] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio network. *IEEE Transactions on Automatic Control*, 37:1936–1948, 1992.

[136] L. Tassiulas and S. Sarkar. Maxmin fair scheduling in wireless networks. In *Proceedings of IEEE INFOCOM*, New York NY, USA, 2002.

[137] T. Tezcan. Augmented fluid models and the stability of queueing systems. 2009. Under submission.

[138] H.C. Tijms. *A First Course in Stochastic Models*. Wiley, England, 2003.

[139] G. De Veciana, T.-L. Lee, and T. Konstantopoulos. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking*, 9:2–14, 2001.

[140] I.M. Verloop. *Efficient Flow Scheduling in Resource-Sharing Networks*. Master Thesis, Utrecht University, available at http://www.cwi.nl/∼maaike/articles/thesisVerloop.pdf, 2005.

[141] I.M. Verloop, U. Ayesta, and S.C. Borst. Comparison of bandwidth-sharing policies in a linear network. In *Proceedings of ValueTools*, Athens, Greece, 2008.

[142] I.M. Verloop, U. Ayesta, and S.C. Borst. Monotonicity properties for multiclass queueing systems. *Discrete Event Dynamic Systems*, 2009. DOI: 10.1007/s10626-009-0069-4.

[143] I.M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *CWI-report PNA-E0905*, 2009. Conditionally accepted for publication in *Operations Research*.

[144] I.M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of the M/PH/1 discriminatory processor sharing queue with phase-dependent weights. *Performance Evaluation Review*, 2009. To appear.

[145] I.M. Verloop and S.C. Borst. Heavy-traffic delay minimization in bandwidth-sharing networks. In *Proceedings of IEEE INFOCOM*, Anchorage AK, USA, 2007.

[146] I.M. Verloop, S.C. Borst, and R. Núñez-Queija. Stability of size-based scheduling disciplines in resource-sharing networks. *Performance Evaluation*, 62:247–262, 2005.

[147] I.M. Verloop, S.C. Borst, and R. Núñez-Queija. Delay optimization in bandwidth-sharing networks. In *Proceedings of CISS Conference on Information Sciences and Systems*, Princeton University, 2006.

[148] I.M. Verloop and R. Núñez-Queija. Efficient resource allocation in bandwidth-sharing networks. *Performance Evaluation Review*, 35(3):49–50, 2007.

[149] I.M. Verloop and R. Núñez-Queija. Asymptotically optimal parallel resource assignment with interference. *CWI-report PNA-E0805*, 2008. Conditionally accepted for publication in *Queueing Systems*.

[150] I.M. Verloop and R. Núñez-Queija. Assessing the efficiency of resource allocations in bandwidth-sharing networks. *Performance Evaluation*, 66:59–77, 2009.

[151] I.M. Verloop, R. Núñez-Queija, and S.C. Borst. Delay-optimal scheduling in bandwidth-sharing networks. *Performance Evaluation Review*, 34(1):365–366, 2006.

[152] H. Viswanathan and K. Kumaran. Rate scheduling in multiple antenna downlink wireless systems. *IEEE Transactions on Communications*, 53:645–655, 2005.

[153] N.S. Walton. Proportional fairness and its relationship with multi-class queueing networks. *Annals of Applied Probability*, 2009. To appear.

[154] G. Weiss. Optimal draining of fluid re-entrant lines: Some solved examples. *Royal Statistical Society Lecture Notes Series*, 4:19–34, 1996.

[155] A. Wierman. Fairness and classifications. *Performance Evaluation Review*, 34(4):4–12, 2007.

[156] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *Proceedings of ACM SIGMETRICS*, pages 238–249, San Diego CA, USA, 2003.

[157] S. Yang and G. De Veciana. Size-based adaptive bandwidth allocation: Optimizing the average QoS for elastic flows. In *Proceedings of IEEE INFOCOM*, pages 657–666, New York NY, USA, 2002.

[158] S. Yang and G. De Veciana. Enhancing both network and user performance for networks supporting best-effort traffic. *IEEE/ACM Transactions on Networking*, 12:349–360, 2004.

[159] H.-Q. Ye. Stability of data networks under an optimization-based bandwidth allocation. *IEEE Transactions on Automatic Control*, 48:1238–1242, 2003.

[160] H.-Q. Ye and D.D. Yao. Heavy-traffic optimality of a stochastic network under utility-maximizing resource allocation. *Operations Research*, 56:453–470, 2008.

[161] L. Ying, B. Tan, and R. Srikant. On the delay optimality of proportional fairness. In *Proceedings of Information Theory and Applications*, 2008.

Summary

Scheduling in stochastic resource-sharing systems

In this thesis we study queueing models that arise in the context of resource sharing in communication networks. We determine scheduling policies that (asymptotically) optimize the performance of the system, and evaluate policies that share the resources among the users in a fair manner.

Chapter 1 gives an overview of several concepts related to resource-sharing systems and introduces the queueing models and main techniques used throughout the thesis. We describe the work-conserving single-server system, the linear bandwidth-sharing network, and the parallel two-server model. In the latter two models the speed at which the system works depends on the scheduling decision taken and on the types of users presently in the system. These models can therefore be seen as extensions of the single-server model. The stochastic evolution of the numbers of users is determined by the scheduling policy, which specifies how the resources are shared among all users. We use sample-path techniques and stochastic dynamic programming tools in order to characterize policies that minimize the holding cost. Since the original stochastic model is not always tractable, we also investigate two limiting regimes. One of them is the heavy-traffic regime where we let the offered load approach the maximum capacity. In the other regime we consider the queue-length processes under the so-called fluid scaling.

In Chapter 2 we focus on the single-server system in heavy traffic and analyze a generalization of the DPS policy. More specifically, we consider phase-type distributed service requirements and allow customers to have different weights in various phases of their service. In our main result we establish a state-space collapse for the scaled steady-state queue length vector in heavy traffic. The result shows that in the limit, the queue length vector is the product of an exponentially distributed random variable and a deterministic vector. The proof consists in showing that the joint probability generating function of the queue lengths satisfies a partial differential equation that allows a closed-form solution after passing to the heavy-traffic limit. Our result has several interesting consequences for the standard DPS queue. We derive that, conditioned on the number of customers, the remaining service requirements of the various customers are independent and distributed according to

the forward recurrence times. In addition, we show that the scaled holding cost stochastically reduces as more preference is given according to the mean forward recurrence times of the service requirements.

In Chapters 3–7 we study the linear bandwidth-sharing network. This network provides a natural modeling framework for the dynamic flow-level interaction among data transfers in wired communication networks. It models the bandwidth sharing of data traffic that traverses multiple links and the cross traffic it meets on its route.

Size-based scheduling policies, such as SRPT and LAS, are popular mechanisms for reducing the number of users in single-server systems by favoring smaller requests over larger ones. In Chapter 3 we prove that straightforward extensions of such policies may cause instability effects in the linear network and will therefore certainly not yield optimal performance. For networks with sufficiently many nodes, instability phenomena may in fact arise at arbitrarily low traffic loads.

In Chapters 4–6 we turn to finding policies that minimize the holding cost in a linear network. In Chapter 4 we restrict the search to the class of non-anticipating policies and assume exponentially distributed service requirements. We show that simple priority rules are optimal for certain settings of the parameters of the service requirements. For the remaining parameter settings we prove that, in the case of a two-node linear network, an average-cost optimal policy is characterized by "switching curves", i.e., the policy dynamically switches between several priority rules. Since an exact characterization of these curves is not possible in general, we study in Chapter 5 the related fluid control problem. We show that the optimal fluid control can be explicitly described by linear switching curves. In most cases these curves provide asymptotically fluid-optimal policies in the original stochastic model as well. For some scenarios however, fluid-based switching curves may result in a policy that is unstable. In that case, the diffusion scaling is appropriate and efficient switching-curve policies have a square-root shape.

The class of weighted α-fair bandwidth-sharing policies is commonly accepted to model the dynamic flow-level bandwidth allocation as realized by packet-based protocols. Through numerical experiments we evaluate the performance of these policies by comparing them to asymptotically optimal policies. In Chapter 4 we find that the gap between α-fair policies and optimal policies is not that large provided the system load is moderate. In addition, the performance under α-fair policies is quite insensitive to α, as long as this value is not too small. In Chapter 5 we observe that *weighted* α-fair policies can approach the optimal performance when choosing the weights appropriately.

In Chapter 6 we consider a linear network with generally distributed service requirements and allow anticipating policies. We focus on policies that allocate the capacity across the classes such that stability of the system is guaranteed. Motivated by the size-based scheduling results for single-server systems, we then prioritize within a class the large requests over the small ones. These size-based scheduling policies are proven to be asymptotically optimal in a heavy-traffic setting for service requirements with bounded support. In addition, we show that these policies may outperform α-fair policies, which are non-anticipating, by an arbitrarily large factor when the load is sufficiently high.

In Chapter 7 we first focus on a multi-class queueing system with general inter-arrival times and service requirements, and give sufficient conditions in order to compare sample-path wise the workload and the number of users under different policies. This allows us to evaluate the performance of the system under various policies in terms of stability and the mean holding cost. We then apply this framework to the linear network under weighted α-fair policies. We obtain stability results and, in the case of exponentially distributed service requirements, establish monotonicity of the mean holding cost with respect to the fairness parameter α and the relative weights. In addition, we investigate the monotonicity properties in a heavy-traffic regime and perform numerical experiments. Furthermore, for a single-server system with two user classes we obtain that under DPS and GPS the mean holding cost is monotone with respect to their relative weights. This result is in line with the monotonicity result for DPS under a heavy-traffic setting as obtained in Chapter 2.

In Chapter 8 we focus on a parallel two-server model with two classes of users with exponentially distributed service requirements. The study of this model is motivated by scheduling questions in wireless cellular communication networks. It may model for example the power control of two interfering base stations. For certain choices of the parameters of the service requirements we give an exact characterization of an optimal policy. For the remaining cases we study the related deterministic fluid control model for which we show that the optimal control is described by a switching curve. Using similar techniques as in Chapter 5, we prove that policies characterized by either linear or exponential switching curves are asymptotically fluid-optimal in the original stochastic model. For a moderately-loaded system, we numerically compare these fluid-based policies with Max-Weight and threshold-based policies, which are known to be optimal in a heavy-traffic setting. We observe that the fluid-based and the threshold-based policies perform well, while significant performance gains can be achieved over Max-Weight policies.

About the author

Maaike (Ina Maria) Verloop was born in Amsterdam, The Netherlands, on August 31, 1982. She graduated from Grammar School (Veenlanden College, De Ronde Venen) in June 2000. In August 2005, she received her Master's degree in mathematics (cum laude) from the Universiteit Utrecht. Her Master thesis, entitled "Efficient flow scheduling in resource-sharing networks", was written during a 9 month internship at CWI (Centrum Wiskunde & Informatica), Amsterdam. She is recipient of the VVS thesis award 2006 for the best Master thesis in the field of Operations Research and Statistics in the Netherlands. In September 2005, she started as a Ph.D. student at CWI and the Technische Universiteit Eindhoven (TU/e), under the supervision of Sem Borst and Sindo Núñez Queija. In 2008, Maaike visited INRIA Paris–Rocquencourt for two months, hosted by Philippe Robert. Maaike defends her Ph.D. thesis at TU/e on November 26, 2009.