Fluid Limit Approximations of Stochastic Networks

ISBN: 978-90-6464-775-8

Cover design by Maria Remerova. The cover visualises the general idea of fluid limits: they provide a "helicopter view" of the system trajectories which strips away unessential details and allows to see some key features.

Print: GVO Drukkers & Vormgevers B.V. | Ponsen & Looijen, Ede

VRIJE UNIVERSITEIT

Fluid Limit Approximations of Stochastic Networks

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op dinsdag 20 mei 2014 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Maria Remerova

geboren te Chita, Rusland

promotoren: prof.dr. A.P. Zwart
prof.dr. S.G. Foss

To my parents, Vladimir Frolkov and Svetlana Frolkova

Acknowledgements

This thesis would not have been possible without the help and encouragement of many people. I take this opportunity to express my gratitude to them.

First and foremost, I want to thank my advisors Bert Zwart and Sergey Foss, who have always so generously and enthusiastically shared their knowledge with me and kept me motivated. Bert and Sergey Georgievich, I am deeply grateful for your devoted guidance, your patience with my style of working, and your genuine caring attitude. Sergey Georgievich, thank you for converting me from a careless writer to a self-critical one. Bert and Maria, you helped me a great deal to settle in this country, I really appreciate it. Also thank you for the advice to try teaching at the TU/e which I am very much enjoying!

The results described in this dissertation were obtained in collaboration with Bert, Sergey and Josh Reed. Josh, it was truly a pleasure working with you. Thank you for coming all the way from the US to serve on my defence committee. I also thank the other committee members Rob van den Berg, Sem Borst, Michel Mandjes and Floske Spieksma for the careful reading of my manuscript and valuable comments, and for cooperation in setting the defence date.

I gratefully acknowledge the financial support for my Ph.D. project provided by The Netherlands Organization for Scientific Research (NWO). I thank my institute, the CWI Amsterdam, for making my relocation to the Netherlands a fuss-free and joyful experience, and for the many opportunities I have had there as a Ph.D. student: from various career trainings to language courses to ping-pong tables.

I thank the Dutch Network on the Mathematics of Operations Research (LNMB) for the excellent courses in queueing theory and OR. In particular, I enjoyed the convex analysis course by Erik Balder, and I apply the techniques I learned there in my dissertation.

A lot of excitement in my Ph.D. life has been due to conferences and working visits since I love travelling so much. I will cherish the memories of the cozy Cambridge and the holy Jerusalem and many others.

I am happy to have met many new friends during these years: Sihan, Demeter, Kerem, Jan-Pieter, Joost, Arnoud, Natalia (the CWI crowd so far), Melania (VU), Lera (RUG), Tanya (UvA), the list being far from complete. To you, guys, and to my very special old friend Lena (Twente), thank you (check all that apply ☺) for our collaborative achievements in the LNMB courses and research, for the whole lot of your help on all sorts of

occasions including production of this thesis, for the fun we had together, and for you being there to lend me your ear and give me good advice.

To Mitya, my husband, from the bottom of my heart, thank you for being with me all the way, for loving me and never reproaching me. I wish I had half of your kindness.

Finally, I want to thank my parents and brother. Although me moving so far away to do my studies was not easy for them, they chose to trust and support my aspirations. Mum, Dad, Yegor, thank you so much! I am where I am now because of you, and every compliment and congratulation I get on my work is equally yours! I hope this book will be some joy to you.

Masha
March 2014

Contents

Chapter 1

Introduction

1.1 Stochastic networks

This thesis belongs to queueing theory — a domain of mathematics that specialises in modeling real-life service systems such as mobile, computer and manufacturing networks, and pursues the goal of explaining, predicting and controlling the behavior of these systems. Modeling and analysis tools of queuing theory come mainly from probability theory and operations research.

In a service system, some participants (*servers*) deliver a certain type of service to others (*customers*) according to some rule (*service discipline*). Depending on a particular application, these notions can read differently. For example, in a computer network, servers are links, customers are packets of data, and the service discipline is called a transmission protocol. There are simple models that involve a single waiting line, or a *queue*. More complicated models that consist of a number of queues interacting through the service discipline and/or customer *routing* are called *stochastic networks*, where "stochastic" refers to the random components of the model.

A probabilistic approach to modeling of real-life service systems is very natural. Indeed, when we deal with human behavior, there is always room for uncertainty. For example, call centers do not know when exactly in the course of the day they will receive calls and how long answering them will take. This can be modeled as a random *arrival process* of customers to the system and random *service times*. Assumptions on their distributions can be made on the basis of empirical data. The service discipline and the behavior of servers can involve randomness as well. For example, in a wireless network, access points can be viewed as servers in the sense that they provide connection to the Internet for their clients. A new client connects at random to one of the access points with the strongest signal, and when this access point goes down (the client has no control over this, i.e. it happens at a random epoch from the client's point of view), he has to pick a new connection.

There are different ways to characterize the behavior of a stochastic network depending on the purpose of the analysis. Sometimes average characteristics are sufficient, but

here we are interested in a more detailed description — in how the state of the system evolves over time. The state of the network could be simply the population size, or it can include additional variables such as residual service times of the customers present. Generally, a *Markovian* description of the system state is preferable, i.e. a sufficiently rich description enabling one to make predictions based solely on the current system state, without use of any earlier history. The theory of Markov processes is well developed and suggests powerful analysis tools (see Norris [81], Meyn and Tweedie [78], Nummelin [82], Brémaud [24], Ethier and Kurtz [40]). From such processes, one expects more regularity. In particular, they often allow approximations by solutions to differential equations (note that solutions to differential equations can be restored from the latest value available just as Markov processes). Typically, the state of the network exhibits jumps: for example, when a customer arrives to or departs from the system. Such jumps are conventionally assumed to be right-continuous with finite left limits, so that the processes associated with the network are (random) elements in a special functional space called the *Skorokhod space*. We discuss the connection between Markov processes and differential equations in more detail in Section 1.2, and the Skorokhod space — in Section 1.5.

For fundamental models and results in queueing theory, we refer to the textbooks by Asmussen [5], Adan and Resing [2], Baccelli and Brémaud [7], Borovkov [14], Cohen [30], Khinchin [61], Takács [102].

In the present work, we especially focus on *impatience* of customers, which means that they may leave the system before their service has been completed. In different contexts there are different reasons for abandonments: in streaming media, to provide an acceptable quality of service, the data has to be transmitted within a given time interval, see [85]; in health care, patients may die while waiting for an organ transplant, see [101]. In general, impatience of customers comes naturally in *overloaded* systems as a reaction to long waiting times. The overload regime is another focus of this thesis. For a survey on overloaded and critically loaded queueing systems with impatient customers, we refer to Ward [111].

1.2 Fluid limits

It is usually the case in queueing theory that elegant exact analysis and tractable results are only possible for models of a relatively simple design and/or under restrictive (e.g. exponential) stochastic assumptions. Very often one faces a dilemma: on one hand, simple models and simple stochastic assumptions are not really practically relevant; but on the other hand, generalizations complicate the analysis significantly. In such situations, considering approximations could help, in particular fluid limit approximations.

Strictly speaking, by a fluid limit approximation of a stochastic network, we mean a fluid limit of the stochastic process keeping track of the state of the network.

Fluid limits of a stochastic process arise as a result of scaling of this process in a certain fashion, which is called a *fluid scaling*, or a *law-of-large-numbers scaling*. A (vague) definition of a fluid scaling could read as: in a fluid scaled process, jumps whose size is of

order $1/r$ occur at a rate of order r. Then the scaling parameter r is taken to ∞, and the limit processes are called *fluid limits*.

To illustrate the idea, we give two basic examples, where the definition of fluid scaling given above is justified by the compression of space and speeding up time by a factor $r > 0$.

Example 1.1 (**FLLN**, the functional law of large numbers)**.** Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with a finite mean $\mathbb{E}X_1 = a$. Consider the summation process and its fluid scaled version:

$$S(t) := \sum_{i=1}^{\lfloor t \rfloor} X_i, \quad \overline{S}^r(t) := S(rt)/r, \quad t \in \mathbb{R}_+.$$

By Chen and Yao [26], the linear function

$$\overline{S}(t) := at, \quad t \in \mathbb{R}_+,$$

is the fluid limit of the summation process $S(\cdot)$ in the sense that

$$\overline{S}^r(\cdot) \to \overline{S}(\cdot) \quad \text{a.s.,} \quad \text{u.o.c.} \quad \text{as } r \to \infty,$$

where the abbreviation "u.o.c." stands for uniform convergence on compact sets.

Example 1.2 ($GI/G/1$ **queue).** Consider a single server queue where interarrival times of customers are i.i.d. with mean $1/\lambda$, and their service times are i.i.d. with mean $1/\mu$, independent from the arrival process. Customers are served in the order of arrival. Denote by $Q(l,t)$ the number of customers at the queue at time $t \in \mathbb{R}_+$ given that at $t = 0$ there are l customers. In this example, we take the initial condition as the scaling parameter and consider the fluid scaled processes

$$\overline{Q}^l(t) := Q(l, lt)/l, \quad t \in \mathbb{R}_+. \tag{1.1}$$

By Chen and Yao [26], the a.s.-u.o.c. limit of the processes (1.1) is given by

$$\overline{Q}(t) := \max\{1 + (\lambda - \mu)t, 0\}, \quad t \in \mathbb{R}_+. \tag{1.2}$$

Remark 1.1. The notation $GI/G/1$ was proposed by Kendall [59], where the first factor GI indicates that interarrival times are generally distributed and independent, the second factor G stands for generally distributed service times, and the last factor 1 means that there is a single server. If the interarrival or service time distribution is exponential, the corresponding factor should be replaced by M (stands for Markovian). We work with $M/G/1$ queues in Chapter 4.

In the above examples, the fluid limit allows to see the system from afar, providing a caricature of the original stochastic process. The general philosophy of rescaling processes is to strip away unessential details to be able to see some key features. Sometimes, to obtain non-trivial and insightful fluid limits, one has to do more work than just zooming out. Instead of a fixed model, one might need to consider a sequence of models with,

for example, more and more patient customers (see Chapters 2 and 5), or where, instead of speeding up time, one should let the system capacity and the arrival rates grow large (see Chapter 3). There are also *heavy-traffic* fluid scalings that assume the load intensity to approach the value critical for the system stability (see Whitt [115]).

The convergence to fluid limits does not have to be in the strong a.s. sense. More often, when the fluid scaling involves a sequence of systems, weak convergence results are proven. In Section 1.5, we discuss how they can be proven.

We now give a simple example where a fluid scaling different from the space-time scaling is required.

Example 1.3 ($M/M/\infty$ **queue).** Consider a system with infinitely many servers, Poisson arrivals of customers at rate λ, and their service times being distributed exponentially with parameter μ, mutually independent and independent from the arrival process. Service of a customer starts immediately upon his arrival at any server that is not occupied.

If we scale this system in space and time by the initial state like in Example 1.2, the point-wise a.s. limit of the processes (1.1) will be the indicator function

$$\mathbb{I}\{t = 0\}, \quad t \in \mathbb{R}_+, \tag{1.3}$$

which is not right-continuous. So fluid limits in the Skorokhod topology do not exist. In addition, the limit (1.3) does not provide much insight.

Now, instead of zooming out, consider a sequence of $M/M/\infty$ queues such that, in queue l, the initial condition is l and the arrival rate is $l\lambda$. The mean service time is $1/\mu$ in all queues. Denote the queue length process of queue l by $Q_*^l(\cdot)$. By Robert [95], the scaled processes

$$\overline{Q}_*^l(\cdot) := Q_*^l(\cdot)/l$$

converge in distribution, u.o.c. to the fluid limit

$$\overline{Q}_*(t) := \lambda/\mu + (1 - \lambda/\mu)e^{-\mu t}, \quad t \in \mathbb{R}_+. \tag{1.4}$$

Remark 1.2. Again, the notation $M/M/\infty$ is due to Kendall [59]. To accommodate interarrival times that are generally (but identically) distributed and independent, the first factor M should be replaced by GI, and if the arrival process is general (no i.i.d. assumption) — by G. If service times are i.i.d. and generally distributed, the second factor M should be changed into G. In Chapter 3, we work with $M/G/\infty$ and $G/G/\infty$ queues.

It may seem that, since fluid limits average stochastic processes over long time intervals or over large populations of customers, they should be deterministic. This is indeed the case for the majority of stochastic models arising in queueing, but not for all of them.

Example 1.4. Consider a queue that is served by two servers. At time $t = 0$, there are at least two customers in the system, and there are more arriving according to a Poisson process of rate λ. If a customer is served by server i, his service time has an exponential

distribution with parameter μ_i, $i = 1, 2$. All service times are mutually independent and do not depend on the arrival process. The service discipline is the following. At $t = 0$, each server is serving a customer. Further, upon completing a service, the server picks the next customer from the queue, but if the queue is empty at that moment, the server abandons the system forever. We assume that, when the queue is served by both servers, it is stable, and when one of the servers leaves, the queue blows up, i.e.

$$\mu_1, \mu_2 < \lambda < \mu_1 + \mu_2.$$

We scale the queue length process in the same way as in Example 1.2. In this case, the fluid scaled processes (1.1) converge a.s., u.o.c. to the random fluid limit $\overline{Q}(\cdot)$ whose trajectories are depicted in Figure 1.1. There are three possible scenarios:

$$\begin{aligned} A_{1,2} &= \{\text{both servers leave}\}, \\ A_i &= \{\text{only server } i \text{ leaves}\}, \quad i = 1, 2, \end{aligned}$$

that occur with probabilities

$$\begin{aligned} \mathbb{P}\{A_{1,2}\} &= \frac{2\mu_1\mu_2}{\lambda(\mu_1 + \mu_2)}, \\ \mathbb{P}\{A_i\} &= \frac{\mu_i(\lambda - \mu_{3-i})}{\lambda(\mu_1 + \mu_2)}, \quad i = 1, 2. \end{aligned}$$

The second segment of the fluid limit has slope λ on the event $A_{1,2}$, and slope $\lambda - \mu_{3-i}$ on the event A_i, $i = 1, 2$.

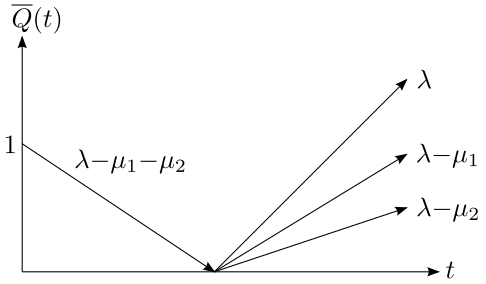

Figure 1.1: Random fluid limit in Example 1.4

The last example might seem artificial, but it is in fact a part of a more general model introduced by Kovalevski et al. [62]. In [62], the servers do not leave forever but switch to another queue and then switch back after long time intervals, which can be viewed as processor sharing. Chapter 4 studies another queueing system with a random fluid limit — a polling model that arises naturally in many applications. There are only a few examples of this phenomenon in the literature, but they prompt the following conclusion (first noticed by Robert [95]): under passage to the fluid dynamics, randomness gets preserved at those system states where transition rates are discontinuous.

The ideas of rescaling processes originated from hydrodynamic studies of many-particle

systems and gained steady popularity in queueing theory in the 90's. There are two basic applications of fluid limits in analysis of stochastic networks, which we discuss in the next two sections.

1.3 Motivation 1: fluid limits for stability

With stability of a Markov process, we mean that the time the process needs to come back (or close) to where it started has a finite mean. The formal term is "positive recurrence" if the state space is countable, and "positive Harris recurrence" in case of an uncountable state space.

As fluid limits emerged in queueing theory, they were mainly used as a tool for investigating stability of stochastic networks, and turned out to be a very efficient and universal tool. In this section, we briefly review how the idea developed, and discuss in what sense fluid limits should be (un)stable to guarantee (in)stability of the underlying system. For a comprehensive and rigorous treatment of the topic, we refer to Bramson [23].

Development of the idea The natural (and necessary) stability conditions for a stochastic network with a conservative service discipline (one that does not allow idling of servers if the system is non-empty) are that the intensity of traffic offered to each server is less than its processing capacity. The necessity of these conditions was proven, for example, by Rybko [98] under the assumption that arrivals are Poisson and service requirements are of phase-type. For a long time, these conditions were hypothetically considered to be sufficient as well. Kelly [55] and Massey [73] found wide classes of stochastic networks that have product-form stationary distributions, and hence satisfy the hypothesis. The first counterexamples to the hypothesis appeared in the literature in the early 1990's and were, in fact, deterministic, see [65] and [25]. The first stochastic counterexample was provided by Rybko and Stolyar [99]: they studied a two-station-two-customer-class network that is stable under the natural stability condition if the service discipline is FCFS, and transient if certain priority disciplines are used. Later, Bramson [19] came up with examples of two-station stochastic networks for which the natural conditions fail to guarantee stability even if the discipline is FCFS. In fact, the FCFS discipline it the very cause of instability in his examples, where customers re-enter the same station multiple times, and so the queue at this station is building up while the other station is idling.

Coming back to the work of Rybko and Stolyar [99], another main innovation of that paper was the way of proving stability. Instead of the stochastic network itself, they first looked at its formal deterministic analogue which they defined by a set of differential equations imitating the transition rates in the original system. They showed that the natural conditions imply, in a sense, stability of the deterministic analogue, and then they translated the proof onto the original stochastic model. So the role of the simpler deterministic model was to provide insight about the more involved stochastic model.

It was also mentioned that the original model, when properly scaled, might approach its deterministic analogue, but it was not proven.

In the subsequent works by Dai [31] and Stolyar [100], the approach of [99] was extended to multiclass-multistation stochastic networks with a wide range of service disciplines allowed, and it was also simplified. Dai [31] and Stolyar [100] did not carry out the parallel analysis of the stochastic network and its deterministic analogue. Instead, they proved that the deterministic model is the fluid limit of the stochastic model. From this convergence and additional integrability assumptions, they concluded that the stability of the fluid limit implies the stability of the original system. In such a form, the approach became popular in the literature.

Stability of fluid limits Traditionally, for investigating stability, the norm of the system initial state is taken as the fluid scaling parameter, like in Example 1.2. We will assume such a scaling throughout the rest of this section. Also, for simplicity, we will assume that the state space is $\mathbb{R}_+$. Consequently, all fluid limits considered below start from 1.

By Dai [31] and Stolyar [100], the stability of fluid limits was defined as follows:

(D.1) there exists a finite constant time T such that, for any fluid limit $\overline{X}(\cdot)$, we have $\overline{X}(T) = 0$ a.s. for all $t \geq T$,

or equivalently, although formally weaker,

(D.2) there exists an $\varepsilon \in (0,1)$ and, for any fluid limit $\overline{X}(\cdot)$, a finite constant time $T_{\overline{X}}$ such that $\overline{X}(T_{\overline{X}}) \leq \varepsilon$ a.s.

In fact, in both [31] and [100], and in all papers mentioned so far in this section, fluid limits are deterministic processes. In this case, by [100], (D.1) and (D.2) are equivalent to:

(D.3) for any fluid limit $\overline{X}(\cdot)$, we have $\inf_{t\in\mathbb{R}_+} \overline{X}(t) < 1$.

However, for random fluid limits, the conditions (D.1) and (D.2) are too restrictive: they are sufficient but far from necessary for stability of the underlying Markov process. In this situation, Kumar and Meyn [64] proved that the following milder condition is enough:

(R.1) fluid limits are uniformly L_p-stable, $p > 0$, i.e. $\sup_{\overline{X}(\cdot)} \mathbb{E}(\overline{X}(t))^p \to 0$ as $t \to \infty$, where the supremum is taken over all fluid limits.

Foss and Kovalevskii [42] also worked in this direction and proposed another notion of fluid limit stability (implying stability of the underlying Markov process) which is more general than (D.1) and (D.2) and which operates with stopping times. Their later work with Topchii [62] uses this notion as well:

(R.2) for all fluid limits $\overline{X}(\cdot)$, there exist stopping times $\tau_{\overline{X}}$ which are uniformly integrable and such that, for some $\varepsilon \in (0,1)$ and for all $\overline{X}(\cdot)$, we have $\overline{X}(\tau_{\overline{X}}) \leq \varepsilon$ a.s.,

or equivalently,

(R.3) for all fluid limits $\overline{X}(\cdot)$, there exist stopping times $\widetilde{\tau}_{\overline{X}}$ which are uniformly integrable and such that, for some $\varepsilon \in (0,1)$ and for all $\overline{X}(\cdot)$, we have $\mathbb{E}|\overline{X}(\widetilde{\tau}_{\overline{X}})| \leq \varepsilon$.

Preceding the works mentioned above, Malyshev *et al.* (see [71] and the references therein) had investigated stability of random walks on the integer lattice $\mathbb{Z}_+^l$. They operated with *second vector fields* rather than fluid limits (the two approaches being closely related though). In essence, the results of Malyshev *et al.* accommodate both deterministic and random (branching) fluid limits.

Instability via fluid limits A number of papers established instability of queueing networks via fluid limit approximations. Meyn [77] and Dai [32] proved that if all fluid limits are unstable (in different senses), then the underlying Markov process is transient. Note that to prove stability, one typically has to show that all fluid limits are stable, hence there should be at least one unstable fluid limit to make the original system transient. So the condition of [77] and [32] seems far from necessary, and the following result by Puhalskii and Rybko [86] looks more natural: a Markov process is transient if there exists a (what they call) "essential" unstable fluid limit, while some of the other fluid limits may be stable.

1.4 Motivation 2: fluid limits as approximations

As was discussed in the previous section, one can investigate stability of stochastic networks via their fluid limit approximations. Now, directly as approximations, fluid limits work best for systems that blow up (because of overload or because their capacities grow large). The reason is that, for such systems, fluid limits are typically bounded away from zero. Returning to Example 1.3 ($M/M/\infty$ queue), the fluid limit (1.4) is non-zero everywhere, and so

$$Q_*^l(\cdot) \sim l\overline{Q}_*(\cdot)$$

is a reasonable estimate of the queue length process for l big enough. In Example 1.2, the fluid limit (1.2) stays away from zero if we assume overload ($\lambda > \mu$), and for large l and t of order l, the approximation

$$Q(l,t) \sim l\,\overline{Q}(t/l) = l + (\lambda - \mu)t$$

makes sense. On the other hand, if $\lambda < \mu$, the fluid limit (1.2) gets absorbed at zero after a finite time and this guarantees stability of the queue, but the candidate approximation $l\,\overline{Q}(t/l)$ becomes in this case an indeterminate form $0 \times \infty$ and is of no use. In other words, when the fluid limit turns zero, too much information is stripped away to be able to estimate the value of the original process.

This thesis focuses on the approximation aspect of fluid limits, and hence considers densely populated systems.

Trajectory-wise approximations Typically fluid limits are characterised as solutions to (systems of) integral/differential equations that arise as the limits under fluid scaling of dynamic equations satisfied by the original stochastic processes. In Example 1.2 ($GI/GI/1$ queue), formula (1.2) of the fluid limit emerges from the following basic equation for the queue length process: for $t \in \mathbb{R}_+$,

$$Q(l,t) = l + \max\Big\{n \geq 0\colon \textstyle\sum_{i=1}^{n} \tau_i \leq t\Big\} - \max\Big\{n \geq 0\colon \textstyle\sum_{i=1}^{n} \sigma_i \leq \int_0^t \mathbb{I}\{Q(l,s) > 0\}\, ds\Big\},$$

where $\{\tau_i\}_{i\in\mathbb{N}}$ are the i.i.d. inter-arrival times with mean $1/\lambda$, and $\{\sigma_i\}_{i\in\mathbb{N}}$ are the i.i.d. service times with mean $1/\mu$.

Since passing to the fluid dynamics averages the original stochastic process in a way, an initial guess on the fluid limit equations can be made based on the following heuristics. Suppose that the underlying process is I-dimensional. If the drift (the product of the average jump size and the total jump rate) of the original stochastic process at state $\mathbf{x} = (x_1, \ldots, x_I)$ is $\mathbf{b}(\mathbf{x}) = (b_1, \ldots, b_I)(\mathbf{x})$, then it is natural to expect that the fluid limit equation will look as follows:

$$\mathbf{x}'(\cdot) = \mathbf{b}(\mathbf{x}(\cdot)). \tag{1.5}$$

Addressing Example 1.2 ($GI/GI/1$ queue) again, the growth rate of the queue length due to arrivals is always λ. If we additionally assume overload, the queue never empties, and its decay rate due to departures is always μ. Hence, the fluid limit should satisfy

$$\overline{Q}\,'(\cdot) = \lambda - \mu,$$

which is in accordance with (1.2).

These heuristics can be made rigorous in most cases, making the fluid limits solutions to deterministic differential equations. This explains the popularity of fluid limit approximations, since the analysis of differential equations is often more tractable, and also more efficient from a computational point of view.

Philosophically, even if the fluid limit is random, it is still a simpler process than the original one and would be welcomed as an approximation. But in practice, there are only few examples when a precise characterisation of a random fluid limit is obtained; in Chapter 4 we present one of them.

Uniqueness of fluid limits The question of uniqueness of a fluid limit can be viewed as a matter of definition. One can "force" a fluid limit to be unique by imposing more requirements on it. The additional requirements should of course have prototypes that hold for the original process. So one should balance between the work it costs to derive the fluid limit properties from their stochastic analogs and the actual purposes of the analysis. To be used directly as approximations, fluid limits require the most precise description possible, in particular uniqueness is desired. At the same time, in heavy-traffic approximations, one does not care about uniqueness of fluid limits, but rather

about their uniform convergence to the set of *invariant solutions* of the fluid limit equations, see [21, 117].

Fixed-point approximations for stationary distributions Stationary measures of stochastic networks (distribution, queue length, blocking probabilities, *etc.*) are among the most often used performance measures. Sometimes, even if formulas defining those measures are known, the actual values might be problematic to compute in practice: due to high dimensionality or because the available formulas are implicit and difficult to invert. In this case, taking limits of the available expressions under proper fluid scalings might simplify them and eliminate the computational problems. For example, in Kelly [56], loss networks are studied, which were, in fact, one of the first applications of fluid limits as approximations. Although explicit expressions for blocking probabilities in [56] are known, they involve summations over a large number of states, which grows rapidly with the network capacity. A large-capacity fluid scaling provides a simple product-form approximation, also known as the *Erlang fixed point*.

Computationally impractical formulas for stationary distributions is not the worst-case scenario, however. For complicated networks (in terms of policies or too general stochastic assumptions), expressions for the stationary distribution are difficult to derive in the first place. In such situations fluid limits might still be of help. Intuitively, under passing to the fluid dynamics, the stationary distribution of a stochastic process should turn into an invariant solution, or a *fixed point*, of the fluid limit equations. For now we assume that the fixed point is unique. This intuition is sometimes wrong, but even in case it is correct, it is usually difficult to make it rigorous.

Typically, the proof of this type of results consists of two steps, both being highly nontrivial in general:

- asymptotic stability of the fixed-point, i.e. convergence of solutions of the fluid limit equations to the fixed point in a long time run;
- interchange of limits (the limit under the fluid scaling and the long-time limit).

For the first step of this scheme, the method of Lyapunov functions is considered rather common. The following proposition (see e.g. Hartman [50]) applies it to the general fluid limit equation (1.5).

Proposition 1.1 (**Asymptotic stability via Lyapunov functions).** *Suppose that the function* $\mathbf{b}(\mathbf{x})$ *is defined in an open set* $S \subseteq \mathbb{R}^K$ *and is continuous there, and that for any* $\mathbf{x}(0) \in S$, *there exists a unique solution* $\mathbf{x}(\cdot)$ *to equation* (1.5). *Suppose also that there exists a function* $L(\mathbf{x})$ *that is continuously differentiable in* S *and such that*

- $L(\mathbf{x})$ *is non-negative in* S, $L(\mathbf{x}) \to \infty$ *as* $\|\mathbf{x}\| \to \infty$;
- *the derivative of* $L(\mathbf{x})$ *with respect to* (1.5), *defined by*

$$dL(\mathbf{x})/dt := \sum_{i=1}^{I} b_i(\mathbf{x})\, \partial L/\partial x_i(\mathbf{x}),$$

is non-positive in S.

If there exists a unique $\mathbf{x}^*$ *such that* $dL(\mathbf{x}^*)/dt = 0$*, then it is the unique fixed point of* (1.5) *and any other solution* $\mathbf{x}(t)$ *of* (1.5) *converges to* $\mathbf{x}^*$ *as* $t \to \infty$*.*

In Chapter 5 we discuss how difficult it can be to find a Lyapunov function. In Chapter 3, we develop a new method of proving asymptotic stability of the fixed point that derives recursive asymptotic bounds for fluid limits.

As for the second step — the interchange of limits — one possible strategy is to prove the fluid limit result for the system running in the stationary regime. Then, on the one hand, the fluid limit is a stationary process whose distribution at each time instant is the limit under the fluid scaling of the stationary distribution of the original model. On the other hand, by the first step, the fluid limit converges to the fixed point over time. This is the approach of Chapter 3 of this thesis and [53].

As was mentioned before, fluid limit approximations of stationary distributions do not always work. The fluid limit equations might have multiple fixed points, and it might be unclear how the latter are related to the original stochastic model. Examples of multiple fixed-points are quite rare though; we give one in Chapter 3. Another thing that can go wrong is the interchange of limits; Kang and Ramanan [53] construct an example where it does not hold.

Precision estimate For the FLLN (Example 1.1), which is the fluid limit of a random walk, it is well-known how to estimate the deviation between the pre-limit and the limit. Namely, the functional central limit theorem (FCLT) states that, if the random walk steps X_i have a positive and finite variance σ^2, then we have the weak convergence

$$\sqrt{r}(\overline{S}^r - \overline{S})(\cdot) \Rightarrow \sigma W(\cdot),$$

where $W(\cdot)$ a standard Brownian motion.

There is also a functional law of iterated logarithm (FLIL) — a corollary of the FCLT that estimates the maximum error of the FLLN in a finite interval:

$$\sup_{t\in[0,T]} |S(t) - \overline{S}(t)| = O(\sqrt{T \log\log T}) \quad \text{a.s.}$$

Similar precision estimates can be derived for other processes as well. The limits in FCLT type of results are called *diffusion limits*. Typically, they are represented by Brownian motions; also Ornstein-Uhlenbeck processes are quite common. To derive diffusion limits, one assumes the stochastic primitives forming the process (i.e. the inter-arrival and service times, etc.) to have finite second moments.

If the stochastic primitives have moments of a higher order, even further refinements are possible, i.e. such that take into account the fluid limit approximation of the process and the diffusion approximation of the error of the fluid approximation. For light-tailed distributions, an exponential rate of convergence to the fluid limit can be shown.

For some basic queueing models, different types of error estimates in fluid limit approximations, and also approximations beyond diffusion limits, are given in Chen and

Yao [26]. See [90, 88] for diffusion approximations of Jackson networks and bandwidth-sharing networks, respectively; heavy-traffic diffusion approximations of various queueing models are developed in [114, 116, 83, 84, 67] and [112, 113, 87].

1.5 Methods of proving convergence to fluid limits

Although it is often easy to guess differential/integral equations satisfied by the fluid limits, the justification of this guess can be challenging. In this section we outline the three traditional techniques of proving convergence of Markov processes to their fluid limits: the **C**-tightness criterion (applied in Chapters 2, 3 and 5), the martingale representation (applied in Chapter 2) and convergence of generators. These standard methods are rather universal, but by no means exhaustive. There are specific models that assume specific tools, see e.g. Chapter 4, where the analysis is based on an embedded branching process.

To begin with, we discuss the "working" function space in queueing theory (and in this thesis as well).

Skorokhod space As we already mentioned, time evolution of a stochastic network is considered to be a right-continuous process with well-defined left limits (the French abbreviation *càdlàg* is also used). This assumption comes naturally: suppose, for example, that there is a discontinuity at time epoch t due to a customer arrival. It means that the new customer is present in the system starting from the moment t, and immediately prior to t there has been one customer less.

In what follows, we assume that the state space S of the stochastic network is endowed with metric $r\colon S \times S \to \mathbb{R}_+$, and denote the space of càdlàg functions $f\colon \mathbb{R}_+ \to S$ by

$$\mathbf{D}(\mathbb{R}_+, S).$$

The space $\mathbf{D}(\mathbb{R}_+, S)$ is called Skorokhod when endowed with the Skorokhod J_1 metric. We will not give its cumbersome definition since we are not using it anywhere in the thesis, but we will discuss the main idea. In this metric, two càdlàg functions are close if they have jumps at close time epochs and the magnitudes of the corresponding jumps are close: $f_n(\cdot) \to f(\cdot)$ as $n \to \infty$ if, for any $T > 0$, there exists a sequence of strictly increasing bijections $g_n\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\sup_{t\in[0,T]} |g_n(t) - t| \to 0 \quad \text{and} \quad \sup_{t\in[0,T]} r(f_n(g_n(t)), f(t)) = 0. \tag{1.6}$$

The topology induced on $\mathbf{D}(\mathbb{R}_+, S)$ by the Skorokhod J_1 metric is much richer than the topology of uniform convergence on compact sets, since the latter requires the corresponding jumps occur exactly at the same times. For example, the convergence

$$f_n(\cdot) := (1 + 1/n)\, \mathbb{I}\{\cdot \in [0, 1 + 1/n)\} \to \mathbb{I}\{\cdot \in [0, 1)\} =: f(\cdot)$$

takes place in the Skorokhod metric but does not hold uniformly. Indeed, take the time changes $g_n(t) := (1 + 1/n)\, t$, $t \in \mathbb{R}_+$. Then

$$\sup_{t\in[0,T]} |g_n(t) - t| = T/n \quad \text{and} \quad \sup_{t\in[0,T]} r(f_n(g_n(t)), f(t)) = 1/n,$$

implying that the definition (1.6) holds; but, for a $T > 1$ and all n big enough,

$$\sup_{t\in[0,T]} |f_n(t) - f(t)| = 1 + 1/n \not\to 0.$$

However, the following equivalence holds, which we often exploit in the subsequent chapters of the thesis.

Proposition 1.2. *If the limit is continuous, then the convergence takes place in both the Skorokhod J_1 and uniform metrics.*

For more background on Skorokhod spaces, we refer to Billingsley [9], Ethier and Kurtz [40] and Whitt [115]. Now we proceed with three traditional methods of proving convergence to fluid limits.

C-tightness criterion For establishing existence and continuity of weak fluid limits (these two properties combined are referred to as **C**-*tightness* of the family of the fluid scaled processes), the conditions of *compact containment* and *oscillation control* are quite popular. They read as (1.7) and (1.8) in the following proposition, see e.g. Ethier and Kurtz [40].

Proposition 1.3 (C-tightness). *Let the metric space (S, r) be complete and separable, and let $\{X_n(\cdot)\}_{n\in\mathbb{N}}$ be a sequence of stochastic processes with sample paths in $\mathbf{D}(\mathbb{R}_+, S)$. Then $\{X_n(\cdot)\}_{n\in\mathbb{N}}$ is relatively compact in $\mathbf{D}(\mathbb{R}_+, S)$ in the sense of convergence in distribution and the limit processes are a.s. continuous if the following two conditions hold:*

- *for any rational $t \geq 0$ and any $\varepsilon > 0$, there exists a compact set $K_{\varepsilon,t} \subset S$ such that*

$$\underline{\lim}_{n\to\infty} \mathbb{P}\{X_n(t) \in K_{\varepsilon,t}\} \geq 1 - \varepsilon; \tag{1.7}$$

- *for any $T > 0$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that*

$$\underline{\lim}_{n\to\infty} \mathbb{P}\{\omega(X_n, T, \delta) \leq \varepsilon\} \geq 1 - \varepsilon, \tag{1.8}$$

where, for a function $x(\cdot) \in \mathbf{D}(\mathbb{R}_+, S)$,

$$\omega(x, T, \delta) := \sup\{r(x(t), x(s))\colon s, t \in [0, T], |s - t| < \delta\}$$

is its modulus of continuity.

After the existence of fluid limits is established, one needs to derive equations characterising them from the fluid scaled equations for the original stochastic process. In

order to do that, it might be necessary to switch from the weak convergence to a.s.-convergence (by Skorokhod's representation theorem) and elaborate on trajectory-wise convergence in the particular state space S, like in [47, 121] and Chapter 3. Sometimes it is enough to have the weak convergence and combine it with laws of large numbers and the *continuous mapping theorem* (weak convergence being preserved by mappings that are continuous at the limit), like in Chapters 2 and 5.

Martingale representation Often rewriting of the dynamics equation for the original stochastic process in a certain "smart" way allows to see clearly that, under the fluid scaling, it differs from the guessed fluid limit equation by a negligible term. To illustrate the idea, we will assume the state space to be countable, for example $\mathbb{Z}_+^I$.

In what follows, bold notations are for state vectors. Denote by $\mathbf{X}(\cdot)$ the stochastic process under consideration and by $q(\mathbf{x}, \mathbf{y})$ its jump rate from state $\mathbf{x}$ to state $\mathbf{y}$. Define also $\boldsymbol{\beta}(\mathbf{x}) = \sum_{\mathbf{y}\neq\mathbf{x}}(\mathbf{y}-\mathbf{x})q(\mathbf{x},\mathbf{y})$. By Darling and Norris [33], in the representation

$$\mathbf{X}(t) = \mathbf{X}(0) + \int_0^t \boldsymbol{\beta}(\mathbf{X}(s))ds + \mathbf{M}(t), \tag{1.9}$$

the process $\mathbf{M}(\cdot)$ is a zero mean martingale. By suitable martingale inequalities and laws of large numbers, one can show that, under the fluid scaling, $\mathbf{M}(\cdot)$ vanishes and $\boldsymbol{\beta}(\cdot)$ converges to the drift $\mathbf{b}(\cdot)$. Hence (1.9) converges to the integral version of the fluid limit equation (1.5), that is

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{b}(\mathbf{x}(s))ds.$$

The final step of the proof could be a general result like [90, Lemma 1] that claims that convergence of integral equations guarantees existence of a solution to the limit equation and convergence of the pre-limit solutions to the limit solution. This approach is used, for example in [36]. More examples where such a martingale representation is efficient can be found in Darling and Norris [33] and Robert [95]. In Ethier and Kurtz [40], the method is developed for uncountable state spaces.

Convergence of infinitesimal generators The local dynamics of Markov processes is characterised by their infinitesimal generators, which are operators mapping functions on the state space S into other functions. Namely, for a time-homogeneous Markov process $X(\cdot)$, its generator A is (formally) given by

$$Af(x) = \lim_{t\to 0} \frac{\mathbb{E}\{f(X(t))|X(0)=x\} - f(x)}{t}, \quad x \in S.$$

In order for Markov processes to converge in distribution, their generators should converge on a class $\mathcal{F}$ of *test functions*, which we do not specify here; i.e. (with obvious notation)

$$X_n(\cdot) \Rightarrow X(\cdot) \quad \text{if} \quad \sup_{x\in S} |A_n f(x) - Af(x)| \to 0, \quad f \in \mathcal{F}. \tag{1.10}$$

The drawback of this approach is that it is restricted to the cases when there exist explicit formulas for the generators. On the other hand, if the exact formulas are known, the proof of the RHS of (1.10) is usually straightforward (often via Taylor expansion). For more details on the theory behind the method, we refer to Ethier and Kurtz [40]; for examples of its application, see [37, 76, 89].

1.6 Overview of the thesis

Here we provide a summary of the subsequent chapters of this thesis. They all develop fluid limit approximations, but the models they consider are very different and thus the analysis techniques vary greatly from one chapter to another. Each chapter corresponds to a paper.

- Chapter 2 is based on [43] M. Frolkova, S. Foss, and B. Zwart. Fluid limits for an ALOHA-type model with impatient customers. *Queueing Systems*, 72:69–101, 2012.
- Chapter 3 is based on [92] M. Remerova, J. Reed, and B. Zwart. Fluid limits for bandwidth-sharing networks with rate constraints. Accepted for publication in *Mathematics of Operations Research*, 2013.
- Chapter 4 is based on [93] M. Remerova, S. Foss, and B. Zwart. Random fluid limit of an overloaded polling model. *Advances in Applied Probability*, 46:76–101, 2014.
- Chapter 5 is based on [91] M. Remerova and B. Zwart. Fluid limits of a PS-queue with multistage service. In preparation, 2013.

Chapter 2: An ALOHA-type model with impatient customers ALOHA protocols are random multiple-access protocols. They are designed for networks with star configurations where multiple client nodes talk to the hub node at the same frequency. Consequently, if there are two or more client nodes talking simultaneously, they are all in conflict, preventing each other from being heard by the hub. The common idea of ALOHA protocols is "try to send your data and, if your message collides with another transmission, try resending later".

We study a generalisation of the conventional centralised time-slotted ALOHA model where impatience of users is allowed, which we assume to be caused by the overload regime. We apply to the (multidimensional) population process a time-space fluid scaling that lets users become more and more patient. Our first result is a description of fluid limits as solutions to a system of deterministic differential equations. The most challenging part of the proof of this result is to eliminate problems at zero. They arise because of the centralised protocol, which assumes that each time slot each user tries to transmit with probability one over the total population. The second main result of Chapter 2 is convergence of fluid limits over time to the unique fixed point. We prove it by means of a Lyapunov function.

Chapter 3: Bandwidth-sharing networks with rate constraints In a bandwidth-sharing network, elastic flows compete for service on several links. Link capacities are redistributed among the flows as their population changes, and bandwidth allocations are chosen in such a way that the network utility is always maximised. This setting was introduced by Massoulié & Roberts [97, 75], and nowadays is considered classical.

In Chapter 3, we modify the classical bandwidth-sharing setting by imposing constraints on processing rates of individual flows. We study the behavior of the model under the large capacity scaling, and with that we mean that the rate constraints, flow sizes and their patience times remain of a fixed order while the network capacity and arrival rates grow large. Note that this scenario is standard in practice. Under general stochastic assumptions, we characterise fluid limits of a process that contains full information about the system state, including residual flow sizes and their residual patience times. In particular, we extend the fluid limit result of Reed and Zwart [88] for Markovian stochastic assumptions. We also prove a new type of result for bandwidth-sharing networks: convergence of the network stationary distribution to the fixed point of the fluid limit equations under the fluid scaling (we need stricter stochastic assumptions here). Moreover, we show that, in many cases, the fixed point solves a strictly concave optimization problem, and thus can be computed in a polynomial time, which is a surprisingly efficient way to approximate such a complicated stochastic model.

Chapter 4: Random fluid limit of an overloaded polling model For many basic queueing systems, fluid limits are deterministic functions. In Chapter 4, we study a cyclic polling model under conditions that lead to a random fluid limit. These conditions are zero initial state and overload. We allow a wide class of service disciplines, which we call "multigated" and which assume a random number of iterations of conventional gated service. Exhaustive policy is in this class as well. Such disciplines ensure that the system population evolves as a multitype branching process. This provides us with our main tool — the Kesten-Stigum theorem that characterises long-time behavior of supercritical (or "overloaded") branching processes. The scaling regime we apply in this chapter is simply zooming out, and the fluid limit we obtain has a rather interesting structure. Firstly, the fluid limit oscillates frequently often in the neighborhood of zero. Secondly, all its trajectories can be mapped by a linear time-space scaling into the same deterministic function. An additional contribution of this chapter is that we develop a method of proving finiteness of moments of the busy period in an $M/G/1$ queue. It is inspired by a technical moment condition of the Kesten-Stigum theorem.

Chapter 5: A processor-sharing queue with multistage service This chapter considers a Markovian processor-sharing queue where service of each customer consists of several stages with independent service requirements. As in Chapter 2, the system is overloaded and customers are impatient. We develop fluid limit approximations of the per-stage population process and characterise them as solutions to a system of deterministic differential equations. We also establish relations between our fluid limits and measure-valued fluid limits in Gromoll et al. [47] for a processor-sharing queue with single-stage service. This allows us to prove that all fluid limits stabilise to the unique

fixed point over time. Additionally, we discuss Lyapunov functions for processor sharing. Existence of Lyapunov functions for models that combine impatience and routing (note that multistage service corresponds to tandem routing) is an open problem. We suggest partial solutions.

1.7 Notation

Here we list the notations common for all of the subsequent chapters.

To define x as equal to y, we write $x := y$ or $y =: x$. We abbreviate the left-hand side and right-hand side of an equation as "LHS" and "RHS", respectively.

The standard sets are: the natural numbers $\mathbb{N} := \{1, 2, \ldots\}$, integers $\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$ and non-negative integers $\mathbb{Z}_+ = \{0\} \cup \mathbb{N}$, the real line $\mathbb{R} := (-\infty, \infty)$ and non-negative half-line $\mathbb{R}_+ := [0, \infty)$.

By e we denote the base of the natural logarithm.

The following operations are defined on $x, y \in \mathbb{R}$:

$$\begin{aligned} &x \vee y := \max\{x, y\}, \quad x \wedge y := \min\{x, y\}, \\ &x^+ := x \vee 0, \quad x^- := (-x) \vee 0, \\ &\lfloor x \rfloor := \max\{n \in \mathbb{Z} \colon n \leq x\}. \end{aligned}$$

The upper limit and lower limit are $\overline{\lim}$ and $\underline{\lim}$, respectively.

All vector notations are boldface. Unless stated otherwise, the coordinates of an I-dimensional vector are denoted by the same symbol (regular instead of bold) with subscripts $1, \ldots, I$ added. Overlining, tildes, sub- and superscripts of vectors remain in their coordinates as well. For example: $\overline{\mathbf{Q}}^r(t) = (\overline{Q}_1^r, \ldots, \overline{Q}_I^r)(t)$, $\boldsymbol{\zeta}^* = (\zeta_1^*, \ldots, \zeta_I^*)$, $\mathbf{L}_i = L_{i,1}, \ldots, L_{i,I}$.

By $\mathbf{0}$ we denote the vector whose coordinates are all zeros, and by $\mathbf{1}$ the vector with all coordinates equal 1. In $\mathbb{R}^I$, we work with two norms: the supremum norm $\|\mathbf{x}\| := \max_{1 \leq i \leq I} |x_i|$, and the L_1-norm $\|\mathbf{x}\|_1 := \sum_{i=1}^I |x_i|$. The vector inequalities hold coordinate-wise. The coordinate-wise product of vectors of the same dimensionality I is denoted by

$$\mathbf{x} \times \mathbf{y} := (x_1 y_1, \ldots, x_I, y_I).$$

For metric spaces S_1 and S_2, $\mathbf{C}(S_1, S_2)$ stands for the space of continuous functions $f \colon S_1 \to S_2$. For a metric space S, $\mathbf{D}(\mathbb{R}_+, S)$ stands for the space of functions $f \colon \mathbb{R}_+ \to S$ that are right-continuous with left limits. We endow $\mathbf{D}(\mathbb{R}_+, S)$ with the Skorokhod J_1-topology. For a function $f(\cdot)$ defined on (a subset of) $\mathbb{R}$, $f'(t)$ denotes its derivative at t.

The complement of an event E is denoted by $\overline{E}$, and its indicator function by $\mathbb{I}\{E\}$. The indicator function $\mathbb{I}_A(\cdot)$ of an arbitrary set A is defined by $\mathbb{I}_A(x) := \mathbb{I}\{x \in A\}$.

The signs $\Rightarrow$, $\stackrel{\mathrm{d}}{=}$ and $\leq_{\mathrm{st}}$, $\geq_{\mathrm{st}}$ stand for weak convergence, equality in distribution and stochastic order, respectively. Recall that, for real-valued r.v.'s X and Y, $X \leq_{\mathrm{st}} Y$ if, for all $t \in \mathbb{R}$, one has $\mathbb{P}\{X > t\} \leq \mathbb{P}\{Y > t\}$.

Chapter 2

An ALOHA-type Model with Impatient Customers

2.1 Introduction

ALOHA-type algorithms are intended to govern star networks in which multiple client machines send data packets to the hub machine at the same frequency. Thus, collisions of packets being transmitted simultaneously are possible (clients know nothing about each other's intentions to transmit data and can not prevent collisions). Such algorithms assume the following acknowledgment mechanism. If data has been received correctly at the hub, which is possible only if no collisions occurred during its transmission, then a short acknowledgment packet is sent to the client. If a client has not received an acknowledgment after a short wait time, then it retransmits the packet after waiting a randomly selected time interval with distribution specified by the ALOHA protocol that governs the network.

The pioneering ALOHA computer network, also known as the ALOHAnet, was developed at the university of Hawaii under the leadership of Norman Abramson (see [1], where Abramson first proposed the ALOHA multi-access algorithm). The goal was to use low-cost commercial radio equipment to connect users on Oahu and the other Hawaiian islands with the central computer on the main Oahu campus. The ALOHAnet became operational in 1971, providing the first demonstration of a wireless data network. Nowadays the ALOHA random access techniques are widely used in WiFi, mobile telephone networks and satellite communications.

To give an example, we describe the conventional centralised time-slotted ALOHA model. Here "slotted time" means that users enter the system and abandon it, initiate and finish transmissions at times $n = 1, 2, \ldots$. The arrival process forms an i.i.d. sequence $\{A(n)\}_{n\in\mathbb{N}}$; all service times are assumed to equal 1. "Centralised model" means that the total number of users in the system is always known. Let $Q(n)$ denote the total number of users at time n. For any n, at the beginning of the n-th time slot, which is the time interval $[n, n+1)$, each of the $Q(n)$ customers present in the system starts his transmission with probability $p(n)$ (and does not with probability $1 - p(n)$) in-

dependently of the others. If two or more users attempt transmissions simultaneously, then the transmissions collide, and hence fail, causing the users to remain in the system in order to retransmit later. After a successful transmission the user leaves immediately. Note that, for any time slot, given there are m customers each starting his transmission with probability p, the probability of a successful transmission equals $mp(1-p)^{m-1}$ and is maximised by $p = 1/m$. So we assume that $p(n) = 1/Q(n)$. In such a setting, the population process $\{Q(n)\}_{n\in\mathbb{Z}_+}$ forms a Markov chain that is positive recurrent if $\mathbb{E}A(1) < e^{-1}$ (the system is stable) and transient if $\mathbb{E}A(1) > e^{-1}$ (the system is unstable). Stability conditions for other ALOHA-type models can be found in [13, 39, 48, 79].

In this chapter, we extend the framework described above allowing impatience of users. A user might abandon at the end of each time slot with a probability that is fixed with respect to time, independently of his previous history and decisions of other users. To distinguish between different levels of patience, we allow multiple classes of users. Within a class, all users have the same abandonment probability. We assume that impatience of users is caused by the overload regime $\mathbb{E}A(1) > e^{-1}$.

The results of the chapter concern fluid limits for the multiclass population process, where the fluid scaling combines zooming-out with letting users become more and more patient. For any initial state, the fluid limit is unique and solves a system of deterministic differential equations. We also show that the fluid limit equations have a unique fixed point. We suggest a quadratic Lyapunov function to prove that fluid limits with different initial states stabilise to the fixed point with time.

One of the possible generalisations of the model treated here is to allow interference of transmissions only if the distance between the corresponding client machines is small, and this is a subject of our future research. Such an extension of the network topology was proposed by Bordenave et al. in [12]. They do not take impatience into account and develop fluid limits to determine whether the stochastic model is stable or not.

The chapter is organised as follows. In Section 2.2, we present a detailed description of the stochastic model. In Section 2.3, we introduce and analyse a *fluid model*, which is a system of deterministic differential equations that are analogous to dynamic equations for the stochastic model. In Section 2.4, we specify the fluid scaling and state our main result — convergence of the fluid scaled population processes to solutions of the fluid model. The subsequent sections contain the proofs of the results stated in Sections 2.3 and 2.4. Namely, in Section 2.5, we establish existence and uniqueness of fluid model solutions. In Section 2.6, we show that the fluid model has a unique and asymptotically stable fixed point. In Section 2.7, we prove the result of Section 2.4, and in Section 2.8 some auxiliary statements.

2.2 Stochastic model

This section contains a detailed description of the stochastic model under study. In particular, it derives the dynamic equation for the system workload. All stochastic primitives introduced here are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

Stochastic assumptions and the protocol We consider an ALOHA-type service system with impatient customers. The system includes the waiting room where customers arrive to, and the server. Time is slotted, i.e. arrivals and abandonments may occur only at time instants $n = 1, 2, \ldots$. Time slot n is the time interval $[n, n+1)$. We assume that there are $I < \infty$ classes of customers.

The arrival process is denoted by $\{\mathbf{A}(n)\}_{n\in\mathbb{N}}$, where $\mathbf{A}(n) = (A_1(n), \ldots, A_I(n))$ and $A_i(n)$ is the number of class i customers arriving at time n. The coordinates of the vectors $\mathbf{A}(n)$ are allowed to be dependent, and the vectors $\mathbf{A}(n)$ themselves are i.i.d. copies of a random vector $\mathbf{A} = (A_1, \ldots, A_I)$. We assume that

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_I) := \mathbb{E}\mathbf{A} \in (0, \infty)^I.$$

Let $Q_i(n)$ be the number of class i customers present in the system at time n. As can be seen from the further description, $Q_i(n)$ coincides with the workload at time n due to class i customers. Hence, $\{\mathbf{Q}(n) = (Q_1(n), \ldots, Q_I(n))\}_{n\in\mathbb{Z}_+}$ denotes both the population process and the workload process.

Each customer brings a packet that takes exactly a single time slot to be transmitted to the server. He also sets a deadline for transmission: once the deadline expires, the customer leaves the system even if his packet has not been transmitted yet. In this case, we say that the customer has abandoned the system due to impatience. Patience times of class i customers have a geometrical distribution with parameter p_i and take values greater than or equal to 1. Introduce also the vector

$$\mathbf{p} = (p_1, \ldots, p_I)$$

of impatience parameters. Patience times of different customers are mutually independent.

We now describe how a transmission occurs. At the beginning of time slot n, each customer, independently of the others, starts transmission with probability $1/\|\mathbf{Q}(n)\|_1$ (and does not with probability $1 - 1/\|\mathbf{Q}(n)\|_1$). If there is one customer transmitting, then the transmission is going to be successful. Otherwise a collision happens. At the end of the time slot, customers learn the result. If a customer has succeeded to send his packet, he immediately leaves the system. If a customer has failed his transmission or chose not to transmit during this time slot, then, given he is from class i, with probability p_i he leaves due to impatience, and with probability $1 - p_i$ he stays in the system to try to (re)transmit his packet later.

Throughout the chapter we assume the following.

Assumption 2.1. *The input process is non-trivial:* $\mathbb{P}\{\|\mathbf{A}\|_1 \geq 2\} > 0$.

Assumption 2.2. *The mean amount of work arriving to the system per time slot exceeds the stability threshold for the corresponding model with no impatience of customers:* $\|\boldsymbol{\lambda}\|_1 > e^{-1}$.

Remark 2.1. The results of the chapter can be generalised to the case when patience times are not simply geometrical random variables but finite mixtures of those. It suf-

fices, for all $i = 1, \ldots, I$, to split customer class i into k_i new classes, where k_i is the number of mixture components in the patience time distribution for class i customers.

Population dynamics The sequence $\{\mathbf{Q}(n)\}_{n\in\mathbb{Z}_+}$ forms a time-homogeneous Markov Chain. Its dynamics can be described as follows: given a history $\{\mathbf{Q}(m)\}_{0\le m\le n}$ up to time n with $\mathbf{Q}(n) = \mathbf{x}$, we have, for $i = 1, \ldots, I$,

$$Q_i(n+1) \stackrel{\mathrm{d}}{=} x_i + A_i - D_i^{\mathrm{t}}(\mathbf{x}) - D_i^{\mathrm{a}}(\mathbf{x}), \tag{2.1}$$

where

- $D_i^{\mathrm{t}}(\mathbf{x})$ represents the number of class i customers who are present in the system at time n but will leave at the end of time slot n because of a successful transmission:

  $$D_i^{\mathrm{t}}(\mathbf{x}) = \mathbb{I}\left\{B_i\left(\mathbf{x}\right) = 1\right\} \prod\nolimits_{j\neq i} \mathbb{I}\left\{B_j\left(\mathbf{x}\right) = 0\right\},$$

  with $B_i(\mathbf{x})$ having the binomial distribution $B\left(x_i, 1/\|\mathbf{x}\|_1\right)$ if $\mathbf{x} \neq \mathbf{0}$ and $B_i(\mathbf{0}) = 0$;

- $D_i^{\mathrm{a}}(\mathbf{x})$ represents the number of class i customers who are present in the system at time n but will abandon at the end of time slot n due to impatience rather than a successful transmission: given $\mathbf{x} - \mathbf{D}^{\mathrm{t}}(\mathbf{x}) = \mathbf{y}$,

  $$D_i^{\mathrm{a}}(\mathbf{x}) = \widetilde{B}_i(\mathbf{y})$$

  with $\widetilde{B}_i(\mathbf{y})$ having the binomial distribution $B\left(y_i, p_i\right)$;

- the random elements $\mathbf{A}$ and $B_i\left(\mathbf{x}\right)$, $\widetilde{B}_j\left(\mathbf{y}\right)$, $i, j \in \{1, \ldots, I\}$, are mutually independent.

Remark 2.2. The number $D_i^{\mathrm{t}}(\mathbf{x})$ of successful transmissions by class i customers at a time slot may take only values 0 and 1. Moreover, $\|\mathbf{D}^{\mathrm{t}}(\mathbf{x})\|_1 \le 1$.

2.3 Fluid model

In the present section, we define a deterministic analogue of the stochastic model described in the previous section. As time and space are appropriately normalised, we expect that the difference equation (2.1) can be approximated by a differential equation where the rate of increase is due to arrival rates, and the rate of decrease due to service completions and abandonments. In the single class case, one may expect such a differential equation to look like (we omit the class index) $z'(t) = \lambda - e^{-1} - pz(t)$, since the throughput of ALOHA is e^{-1}. In the multiclass case, this naturally extends to $z_i'(t) = \lambda_i - e^{-1} z_i(t)/\|\mathbf{z}(t)\| - p_i z_i(t)$, $i = 1, \ldots, I$. This will be made rigorous in Section 2.4. We now proceed more formally.

Definition 2.1. Denote by $\mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ the class of continuous functions $\mathbf{z}\colon \mathbb{R}_+ \to \mathbb{R}_+^I$ such that $\mathbf{z} \neq \mathbf{0}$ for all $t \neq 0$.

Definition 2.2. For a $\mathbf{z}^0 \in \mathbb{R}_+^I$, a solution to the integral equation

$$\mathbf{z}(t) = \mathbf{z}^0 + t\,\boldsymbol{\lambda} - e^{-1}\int_0^t \mathbf{m}(\mathbf{z}(s))\,ds - \mathbf{p}\times\int_0^t \mathbf{z}(s)\,ds, \quad t \in \mathbb{R}_+, \tag{2.2}$$

that belongs to $\mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ is called a *fluid model solution (FMS) with initial state* $\mathbf{z}^0$. The function $\mathbf{m}\colon \mathbb{R}_+^I \to \mathbb{R}_+^I$ is given by

$$\mathbf{m}(\mathbf{x}) = \begin{cases} \mathbf{x}/\|\mathbf{x}\|_1 & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1 & \text{if } \mathbf{x} = \mathbf{0}. \end{cases}$$

Remark 2.3. For a function $\mathbf{z}(\cdot) \in \mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$, equation (2.2) is equivalent to the Cauchy problem

$$\begin{aligned} &\mathbf{z}'(t) = \boldsymbol{\lambda} - \mathbf{p}\times\mathbf{z}(t) - e^{-1}\mathbf{m}(\mathbf{z}(t)), \quad t > 0, \\ &\mathbf{z}(0) = \mathbf{z}^0. \end{aligned} \tag{2.3}$$

Remark 2.4. Although $\mathbf{m}(\mathbf{0})$ does not appear in (2.3), it needs to be defined for further use in Section 2.7. We assign to $\mathbf{m}(\mathbf{0})$ the value of $\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$, which is the limit of $\mathbf{m}(\cdot)$ along FMS's trajectories as they approach $\mathbf{0}$. Indeed, the only point where a fluid model solution can take the value of $\mathbf{0}$ is $t = 0$. Let $\mathbf{z}(\cdot)$ be an FMS starting from $\mathbf{z}(0) = \mathbf{0}$. For the moment suppose that $\mathbf{z}(\cdot)$ is continuously differentiable at $t = 0$. Then (2.3) and Taylor's expansion give, for small $t \in (0, \infty)$,

$$\mathbf{z}'(t) = \boldsymbol{\lambda} - \mathbf{p}\times\mathbf{z}(t) - e^{-1}\frac{t\mathbf{z}'(0) + \mathbf{o}(t)}{\sum_{i=1}^I z_i'(0)t + o(t)}, \tag{2.4}$$

where $o(t) \in \mathbb{R}$, $\mathbf{o}(t) \in \mathbb{R}^I$, and $o(t)/t \to 0$, $\mathbf{o}(t)/t \to \mathbf{0}$ as $t \to 0$. The continuity of $\mathbf{z}'(\cdot)$ at $t = 0$ and Assumption 2.2 guarantee that $\sum_{i=1}^I z_i'(0) > 0$, so we pass to the limit as $t \to 0$ on both sides of (2.4) and get $\mathbf{z}'(0) = \boldsymbol{\lambda} - e^{-1}\mathbf{z}'(0)/\left(\sum_{i=1}^I z_i'(0)\right)$. The last equation has a unique solution $\mathbf{z}'(0) = \left(1 - e^{-1}/\|\boldsymbol{\lambda}\|_1\right)\boldsymbol{\lambda}$, which implies existence of the limit $\lim_{t\to 0}\mathbf{m}(\mathbf{z}(t)) = \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$. Later on (see Section 2.5, Property 2.2), we prove that, for any function $\mathbf{z}\colon \mathbb{R}_+ \to \mathbb{R}_+^I$ that is continuous with $\|\mathbf{z}(t)\| > 0$ for all $t \neq 0$ and that solves (2.3), there exists the derivative $\mathbf{z}'(0) = \left(1 - e^{-1}/\|\boldsymbol{\lambda}\|_1\right)\boldsymbol{\lambda}$, and hence exists the limit $\lim_{t\to 0}\mathbf{m}(\mathbf{z}(t)) = \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$.

In the remainder of the section, we discuss properties of FMS's.

Existence and uniqueness of FMS's If the initial state is non-zero, existence and uniqueness of an FMS follow by the classical results from the theory of ordinary differential equations. Otherwise the proof is rather involved. The outline follows below; see Section 2.5 for the full proof.

Theorem 2.1. *For any initial state, a fluid model solution exists and is unique.*

One-dimensional case. Equation (2.2) turns into $z(t) = z^0 + \left(\lambda - e^{-1}\right)t - p\int_0^t z(s)ds$, and

its unique solution is given by $z(t) = z^0 e^{-p\,t} + \left((\lambda - e^{-1})/p\right)\left(1 - e^{-p\,t}\right)$.

Multidimensional case, non-zero initial state. Uniqueness follows easily by the Gronwall inequality (see for example Hartman [50]).

Proposition 2.1 (Gronwall inequality). *Suppose that functions $u(\cdot)$ and $v(\cdot)$ are non-negative and continuous in $[a,b]$, and that there exists a constant $C \geq 0$ such that $v(t) \leq C + \int_a^t v(s)u(s)ds$, $a \leq t \leq b$. Then $v(t) \leq C \exp \int_a^t u(s)ds$, $a \leq t \leq b$. In particular, if $C = 0$, then $v(\cdot) \equiv 0$ in $[a,b]$.*

Since the first order partial derivatives of the function $\mathbf{m}(\cdot)$ are bounded on all sets

$$\mathbb{R}^I_\delta := \left\{\mathbf{x} \in \mathbb{R}^I_+ \colon \|\mathbf{x}\| \geq \delta\right\}, \quad \delta > 0, \tag{2.5}$$

this function is Lipschitz continuous on all such sets. Let $c(\delta)$ be a Lipschitz constant for $\mathbf{m}(\cdot)$ on $\mathbb{R}^I_\delta$ with respect to the supremum norm, i.e. $\|\mathbf{m}(\mathbf{x}) - \mathbf{m}(\mathbf{x})\| \leq c(\delta)\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^I_\delta$. Suppose that $\mathbf{z}(\cdot)$ and $\widetilde{\mathbf{z}}(\cdot)$ are two FMS's with the same non-zero initial state. They are continuous and non-zero at every point, and hence, for each $t \in (0,\infty)$, there exists a $\delta(T) > 0$ such that $\mathbf{z}(t)$, $\widetilde{\mathbf{z}}(t) \in \mathbb{R}^I_{\delta(T)}$, $0 \leq t \leq T$. We put $\triangle\mathbf{z}(\cdot) = \mathbf{z}(\cdot) - \widetilde{\mathbf{z}}(\cdot)$ and $\triangle\mathbf{m}(\mathbf{z}) = \mathbf{m}(\mathbf{z}) - \mathbf{m}(\widetilde{\mathbf{z}})$. Then

$$-\triangle\mathbf{z}(t) = e^{-1}\int_0^t \triangle\mathbf{m}(\mathbf{z}(s))\,ds + \mathbf{p} \times \int_0^t \triangle\mathbf{z}(s)\,ds,$$

and, for $t \in [0,T]$, we have

$$\sup_{s\in[0,t]} \|\triangle\mathbf{z}(s)\| \leq \left(e^{-1}c(\delta(T)) + \|\mathbf{p}\|\right)\int_0^t \sup_{u\in[0,s]} \|\triangle\mathbf{z}(u)\|\,ds.$$

By the last inequality and the Gronwall inequality, we have $\sup_{s\in[0,t]} \|\triangle\mathbf{z}(s)\| \leq 0$, $0 \leq t \leq T$. Since T is arbitrary, $\mathbf{z}(\cdot)$ and $\widetilde{\mathbf{z}}(\cdot)$ coincide on $\mathbb{R}_+$.

Existence of an FMS with a non-zero initial state can be shown by applying the Peano existence theorem (the proof is postponed to Section 2.5).

Multidimensional case, zero initial state. The technique used in the previous case does not work here because it is based on the continuity properties of the function $\mathbf{m}(\cdot)$ that fail as $\mathbf{0}$ comes into play. So a different approach is required. We introduce a family of integral equations depending on parameters $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}^I_+$ that includes (for $(\varepsilon, \boldsymbol{\pi}) = (0, \mathbf{p})$) an equation equivalent to the Cauchy problem (2.3) with $\mathbf{z}^0 = \mathbf{0}$. We show that each equation of this family has a solution. If $\varepsilon > 0$, then uniqueness of an $(\varepsilon, \boldsymbol{\pi})$-solution is straightforward to show. In order to prove uniqueness of a $(0, \boldsymbol{\pi})$-solution, we derive a proper estimate for it via solutions with other parameters. The whole idea of this proof is adopted from Borst et al. [15].

Invariant FMS Recall that a constant solution to a system of differential/integral equations is called an *invariant solution*, or a *fixed point*. Now we characterise the (unique) invariant solution of the fluid model equation (2.3).

Theorem 2.2. *Suppose Assumption 2.2 holds. Then there exists a unique invariant solution of the fluid model equation* (2.3)*, which is given by*

$$z_i^* = \frac{\lambda_i}{x + p_i}, \quad i = 1, \ldots, I, \tag{2.6}$$

where x solves

$$\sum_{i=1}^{I} \frac{p_i \lambda_i}{x + p_i} = \|\boldsymbol{\lambda}\|_1 - e^{-1}; \tag{2.7}$$

and any fluid model solution $\mathbf{z}(t)$ *converges to* $\mathbf{z}^*$ *as* $t \to \infty$*.*

Theorem 2.2 asserts global asymptotic stability of the invariant FMS $\mathbf{z}^*$; we prove it by means of a Lyapunov function in Section 2.6.

2.4 Fluid limit theorem

In this section, we characterise the asymptotic behaviour under a fluid scaling of the population process of the stochastic model introduced in Section 2.2, justifying the heuristics given in the previous section. First we specify the fluid scaling. Let r be a positive number. Consider the stochastic model from Section 2.2 with the impatience parameters p_i replaced by p_i/r, $i = 1, \ldots, I$. Denote the population process of the r^{th} model by $\mathbf{Q}^r(\cdot)$, and scale it by r both in space and time,

$$\overline{\mathbf{Q}}^r(t) := \mathbf{Q}^r\left(\lfloor rt \rfloor\right)/r, \quad t \in \mathbb{R}_+. \tag{2.8}$$

The fluid-scaled population processes (2.8) take values in the Skorokhod space $\mathbf{D}(\mathbb{R}_+, \mathbb{R}^I)$. We refer to weak limits of the processes (2.8) along subsequences $r \to \infty$ as *fluid limits*.

Now we formulate the main theorem of this chapter and highlight the basic steps of the proof; the detailed proof will follow in Section 2.7.

Theorem 2.3. *Suppose Assumptions 2.1 and 2.2 hold and that* $\overline{\mathbf{Q}}^r(0) \Rightarrow \mathbf{z}^0$ *as* $r \to \infty$*, where* $\mathbf{z}^0$ *is a random vector. Then the fluid limit exists and coincides a.s. with the unique FMS with initial state* $\mathbf{z}^0$*. In particular, if the limit initial state* $\mathbf{z}^0$ *is deterministic, the corresponding fluid limit is a deterministic function.*

The first and the most difficult step of the proof of Theorem 2.3 shows that the fluid-scaled population is bounded away from zero outside $t = 0$. Together with a martingale representation, this allows us to prove that the fluid-scaled population satisfies an integral equation that differs from the fluid model equation (2.2) by negligible terms (for r large enough). Then we establish $\mathbf{C}$-tightness of the family of the fluid-scaled processes by applying Proposition 1.3 and show that fluid limits a.s. satisfy the fluid model equation (2.2).

2.5 Proof of Theorem 2.1

We split the proof into two parts, for a non-zero and zero initial state.

2.5.1 Non-zero initial state

Here we have to prove the existence result only, see Section 2.3 for the proof of uniqueness. First we derive bounds for an FMS using the following lemma.

Lemma 2.1. *Let S be either a finite interval $[0,T]$ or the half-line $\mathbb{R}_+$, and let a real-valued function $x(t)$ be continuous in S and differentiable in $S \setminus \{0\}$. Suppose that a constant C is such that $x(t) \geq C$ for $t \in S \setminus \{0\}$ implies $x'(t) \leq 0$. Then* $\sup_{t\in S} x(t) \leq \max\{x(0), C\}$.

Proof. Let $\varepsilon > 0$. Suppose that $x(t) \in (\max\{x(0), C\}, \max\{x(0), C\} + \varepsilon]$. Then, starting from time t, the function $x(t)$ is decreasing at least until it reaches level C. So $\sup_{t\in S} x(t) \leq \max\{x(0), C\} + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, $\sup_{t\in S} x(t) \leq \max\{x(0), C\}$. □

Bounds for an FMS Let S be either a finite interval $[0,T]$ or the half-line $\mathbb{R}_+$. Suppose that a function $\mathbf{z}(\cdot)$ is continuous, non-negative and does not hit zero in S, and that it solves the fluid model equation (2.3) in S. Then $\|\mathbf{z}(t)\|_1 = \sum_{i=1}^I z_i(t)$ and the derivative $\|\mathbf{z}(t)\|_1'$ exists for all $t \in S$. Summing up the coordinates of equation (2.3), we get

$$\|\boldsymbol{\lambda}\|_1 - e^{-1} - p^*\|\mathbf{z}(t)\|_1 \leq \|\mathbf{z}(t)\|_1' \leq \|\boldsymbol{\lambda}\|_1 - e^{-1} - p_*\|\mathbf{z}(t)\|_1, \quad t \in S,$$

where $p_* = \min_{1\leq i\leq I} p_i$ and $p^* = \max_{1\leq i\leq I} p_i$. Then Lemma 2.1 applied to $x(\cdot) = \|\mathbf{z}(\cdot)\|_1'$ and $C = (\|\boldsymbol{\lambda}\|_1 - e^{-1})/p_*$, and $x(\cdot) = -\|\mathbf{z}(\cdot)\|_1'$ and $C = (\|\boldsymbol{\lambda}\|_1 - e^{-1})/p^*$, implies that

$$\sup_{t\in S} \|\mathbf{z}(t)\|_1 \leq \max\left\{\|\mathbf{z}(0)\|_1, \frac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{p_*}\right\} =: u\left(\mathbf{z}(0)\right), \tag{2.9}$$

$$\inf_{t\in S} \|\mathbf{z}(t)\|_1 \geq \min\left\{\|\mathbf{z}(0)\|_1, \frac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{p^*}\right\} =: l\left(\mathbf{z}(0)\right). \tag{2.10}$$

By Assumption 2.2, we have $u(\mathbf{z}(0)) > 0$ for any non-negative $\mathbf{z}(0)$, and $l(\mathbf{z}(0)) > 0$ for any non-negative and non-zero $\mathbf{z}(0)$.

Equation (2.3) and inequality (2.9) imply that $z_i'(t) \leq \lambda_i - e^{-1}u^{-1}(\mathbf{z}(0))z_i(t) - p_i z_i(t)$, $i = 1, \ldots, I$. Then, by Lemma 2.1, we get

$$\sup_{t\in S} \mathbf{z}_i(t) \leq \max\left\{z_i(0), \frac{\lambda_i}{e^{-1}u^{-1}\left(\mathbf{z}(0)\right) + p_i}\right\} =: u_i\left(\mathbf{z}(0)\right). \tag{2.11}$$

Similarly, if $\mathbf{z}(0) \neq \mathbf{0}$, then inequality (2.10) and Lemma 2.1 yield, for $i = 1, \ldots, I$,

$$\inf_{t \in S} \mathbf{z}_i(t) \geq \min \left\{ z_i(0), \frac{\lambda_i}{e^{-1} l^{-1}(\mathbf{z}(0)) + p_i} \right\} =: l_i(\mathbf{z}(0)). \tag{2.12}$$

Remark 2.5. If $\mathbf{z}(0) \neq \mathbf{0}$, then the bound (2.12) and the fact that $z_i'(0) = \lambda_i > 0$ for $z_i(0) = 0$ imply,

$$\text{for all } \delta > 0, \quad \inf_{t \in S, t \geq \delta} \min_{1 \leq i \leq I} z_i(t) > 0.$$

Existence of an FMS with a non-zero initial state The key tool used in this proof is the Peano existence theorem (see e.g. Hartman [50]).

Proposition 2.2 (Peano). *Suppose that a function* $\mathbf{f}\colon \mathbb{R} \times \mathbb{R}^I \to \mathbb{R}^I$ *is continuous in the rectangle* $B = \left\{(t, \mathbf{x})\colon t_0 \leq t \leq t_0 + a, \max_{1 \leq i \leq I} |x_i - x_i^0| \leq b\right\}$. *Let* $M \geq \sup_{(t,\mathbf{x}) \in B} \|\mathbf{f}(t, \mathbf{x})\|$ *and* $\alpha = \min\{a, b/M\}$. *Then the Cauchy problem*

$$\begin{aligned} \mathbf{x}'(t) &= \mathbf{f}\big(t, \mathbf{x}(t)\big), \\ \mathbf{x}(t_0) &= \mathbf{x}^0 \end{aligned}$$

has a solution in the interval $[t_0, t_0 + \alpha]$ *such that* $\mathbf{x}(t) \in B$.

First note that it suffices to show existence of a non-negative solution to (2.3). By Remark 2.5, it will not hit zero at $t \in (0, \infty)$, and hence will be an FMS.

Further note that it suffices to consider only initial states with all coordinates positive. Indeed, if $\mathbf{z}(0)$ is non-zero, there exists a rectangle $B = \{\max_{1 \leq i \leq I} |z_i - z_i(0)| \leq b\}$ that does not contain zero, and, consequently, the mapping $\mathbf{f}(\cdot) = e^{-1}\mathbf{m}(\cdot) + \mathbf{p}$ is continuous in B. Let $M = \sup_{\mathbf{z} \in B} \|\mathbf{f}(\mathbf{z})\|$ and $\alpha = b/M$. By the Peano theorem, there exists a solution to (2.3) in the interval $[0, \alpha]$. If $z_i(0) > 0$, then, by continuity of $\mathbf{z}(\cdot)$, there exists a $t_i \leq \alpha$ such that $z_i(t) \geq z_i(0)/2$, $t \leq t_i$. If $z_i(0) = 0$, then $z_i'(0) = \lambda_i > 0$, and there exists a $t_i \leq \alpha$ such that $z_i(t) = \lambda_i t(1 + o(1)) \geq \lambda_i t/2$, $t \leq t_i$. Therefore, with $\beta = \min_{1 \leq i \leq I} t_i$, we have $\inf_{t \leq \beta} \|\mathbf{z}(t)\|_1 > 0$ and $z_i(\beta) > 0$ for all i.

Suppose now that $z_i(0) > 0$ for all i. We show that there exists a constant $\alpha^* > 0$ such that any non-negative solution $\mathbf{z}^{(T)}(\cdot)$ to (2.3) that is defined in an interval $[0, T]$ can be continued onto $[0, T + \alpha^*]$ remaining non-negative (α^* does not depend on T and $\mathbf{z}^{(T)}(\cdot)$). This will complete the proof. Define the rectangle

$$B^* = \left\{0 < l_i(\mathbf{z}(0))/2 \leq z_i \leq u_i(\mathbf{z}(0)) + l_i(\mathbf{z}(0))/2, \quad 1 \leq i \leq I\right\}$$

and the constants

$$\begin{aligned} M^* &= \sup_{\mathbf{z} \in B^*} \|\mathbf{f}(\mathbf{z})\|, \\ b^* &= \min_{1 \leq i \leq I} l_i(\mathbf{z}(0))/2, \\ \alpha^* &= b^*/M^*. \end{aligned}$$

Consider $T = 0$. Let $B_0 = \{\max_{1 \leq i \leq I} |z_i - z_i(0)| \leq b^*\}$, then $M^* \geq \sup_{\mathbf{z} \in B_0} \|\mathbf{f}(\mathbf{z})\|$ be-

cause $B_0 \subseteq B^*$. By the Peano theorem, there exists a solution to (2.3) in the interval $[0, \alpha^*]$, and it is non-negative because $B_0 \subseteq \mathbb{R}_+^I$. Consider $T > 0$ and a non-negative solution $\mathbf{z}^{(T)}(\cdot)$ to (2.3) defined in $[0, T]$. By the bounds (2.11) and (2.12), we have $l_i(\mathbf{z}(0)) \leq z_i^{(T)}(T) \leq u_i(\mathbf{z}(0))$ for all i. Let $B_T = \{\max_{1 \leq i \leq I} |z_i - z_i^{(T)}(T)| \leq b^*\}$, then $M^* \geq \sup_{\mathbf{z} \in B_T} \|\mathbf{f}(\mathbf{z})\|$ because $B_T \subseteq B^*$. By the Peano theorem, in the interval $[0, \alpha^*]$, there exists a solution $\mathbf{x}^{(T)}(\cdot)$ to the Cauchy problem

$$\begin{aligned} \mathbf{x}'(t) &= \boldsymbol{\lambda} - \mathbf{p} \times \mathbf{x}(t) - e^{-1}\mathbf{m}(\mathbf{x}(t)), \\ \mathbf{x}(0) &= \mathbf{z}^{(T)}(T), \end{aligned}$$

and it is non-negative because $B_T \subseteq \mathbb{R}_+^I$. Then

$$\mathbf{z}^{(T+\alpha^*)}(t) := \begin{cases} \mathbf{z}^{(T)}(t), & t \in [0, T], \\ \mathbf{x}^{(T)}(t - T), & t \in [T, T + \alpha^*] \end{cases}$$

is a non-negative solution to (2.3) in $[0, T + \alpha^*]$.

2.5.2 Zero initial state

This proof is based on the ideas of Borst et al. [15]. We introduce a family of auxiliary integral equations parametrised by $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+^I$. By further Lemma 2.2, the equation with parameters $(0, \mathbf{p})$ is equivalent to the fluid model equation with zero initial condition. By Property 2.1, each auxiliary equation has a solution. By Property 2.3, for any $\boldsymbol{\pi} \in \mathbb{R}_+^I$, a solution to the equation with parameters $(0, \boldsymbol{\pi})$ is unique.

Lemma 2.2. ***(Equivalent description of the fluid model)*** *For any initial state $\mathbf{z}^0$, the set of fluid model solutions coincides with the set of functions $\mathbf{z}(\cdot) \in \mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ that solve the following system of integral equations: for $i = 1, \ldots, I$, $t \in \mathbb{R}_+$,*

$$\begin{aligned} z_i(t) = z_i^0 \exp\left(-p_i t - \int_0^t \frac{e^{-1}}{\|\mathbf{z}(s)\|_1} \, ds \right) \\ + \lambda_i \int_0^t \exp\left(-p_i(t - s) - \int_s^t \frac{e^{-1}}{\|\mathbf{z}(x)\|_1} \, dx \right) ds. \end{aligned} \tag{2.13}$$

Proof. As we differentiate equations (2.13), the fluid model equation (2.3) follows. We now show that (2.3) implies (2.13). Let $\mathbf{z}(\cdot)$ be an FMS with initial state $\mathbf{z}^0$ and consider the following Cauchy problem with respect to $\mathbf{u}(\cdot)$: for $i = 1, \ldots, I$,

$$\begin{aligned} u_i'(t) &= \lambda_i - \left(p_i + \frac{e^{-1}}{\|\mathbf{z}(t)\|_1} \right) u_i(t), \quad t > 0, \\ u_i(0) &= z_i^0. \end{aligned} \tag{2.14}$$

If (2.14) has a continuous solution, it must be unique. Indeed, suppose that $\mathbf{u}(\cdot)$, $\widetilde{\mathbf{u}}(\cdot)$

are two continuous solutions to (2.14). Let $\mathbf{Q}(\cdot) = \mathbf{u}(\cdot) - \widetilde{\mathbf{u}}(\cdot)$. Then, for $i = 1, \dots, I$,

$$\begin{aligned} w_i'(t) &= -\left(p_i + \frac{e^{-1}}{\|\mathbf{z}(t)\|_1}\right) w_i(t), \quad t > 0, \\ w_i(0) &= 0. \end{aligned}$$

Lemma 2.1 applied to $x(\cdot) = w_i(\cdot)$ and $C = 0$, and $x(\cdot) = -w_i(\cdot)$ and $C = 0$, $i = 1, \dots, I$, implies that $\mathbf{Q}(\cdot) \equiv \mathbf{0}$.

Finally, $\mathbf{z}(\cdot)$ is a solution to (2.14) by (2.3). Differentiating of the RHS of (2.13) implies that it is a solution to (2.14), too. Since (2.14) has a unique continuous solution, $\mathbf{z}(\cdot)$ coincides with the RHS of (2.13). □

Auxiliary fluid model solutions and their properties For each $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+^I$ we introduce the operator $\mathbf{F}^{(\varepsilon,\boldsymbol{\pi})} \colon \mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I) \to \mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ defined by: for $t \in \mathbb{R}_+$, $i = 1, \dots, I$,

$$F_i^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{u})(t) = \varepsilon + \lambda_i \int_0^t \exp\left(-\pi_i(t-s) - \int_s^t \frac{e^{-1}}{\|\mathbf{u}(x)\|_1}\, dx\right) ds.$$

Definition 2.3. Let $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+^I$. A function from $\mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ solving the integral equation

$$\mathbf{z}(t) = \mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{z})(t), \quad t \in \mathbb{R}_+, \tag{2.15}$$

is called an $(\varepsilon, \boldsymbol{\pi})$-*fluid model solution* (for short, we write "$(\varepsilon, \boldsymbol{\pi})$-*FMS* ").

Further we establish a number of properties of the auxiliary fluid model solutions defined above, including the existence and uniqueness of an $(\varepsilon, \boldsymbol{\pi})$-FMS for any $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+^I$.

For $\boldsymbol{\pi} \in \mathbb{R}_+^I$, put

$$\begin{aligned} \pi^{\mathrm{u}} &= \max_{1 \le i \le I} \pi_i, & \pi^{\mathrm{l}} &= \min_{1 \le i \le I} \pi_i, \\ \boldsymbol{\pi}^{\mathrm{u}} &= (\pi^{\mathrm{u}}, \dots, \pi^{\mathrm{u}}), & \boldsymbol{\pi}^{\mathrm{l}} &= (\pi^{\mathrm{l}}, \dots, \pi^{\mathrm{l}}). \end{aligned}$$

Property 2.1. *In what follows,* $\mathbf{z}^{(\varepsilon,\boldsymbol{\pi})}(\cdot)$ *denotes an* $(\varepsilon, \boldsymbol{\pi})$*-FMS.*

(i) *For any* $(\varepsilon, \boldsymbol{\pi}) \in \mathbb{R}_+ \times \mathbb{R}_+^I$, *there exists an* $(\varepsilon, \boldsymbol{\pi})$*-FMS.*

(ii) *If* $\varepsilon > 0$, *then an* $(\varepsilon, \boldsymbol{\pi})$*-FMS is unique.*

(iii) *If* $\pi_1 = \ldots = \pi_I$, *then a* $(0, \boldsymbol{\pi})$*-FMS is unique and given by*

$$z_i^{(0,\boldsymbol{\pi})}(t) = \begin{cases} \dfrac{\lambda_i}{\|\boldsymbol{\lambda}\|_1}\left(\|\boldsymbol{\lambda}\|_1 - e^{-1}\right) t & \text{if } \pi_1 = 0, \\ \dfrac{\lambda_i}{\|\boldsymbol{\lambda}\|_1} \dfrac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{\pi_1}\left(1 - e^{-\pi_1 t}\right) & \text{if } \pi_1 > 0, \end{cases} \tag{2.16}$$

(iv) *If* $\varepsilon > \delta \geq 0$, *then* $\mathbf{z}^{(\varepsilon,\boldsymbol{\pi})}(t) \geq \mathbf{z}^{(\delta,\boldsymbol{\pi})}(t)$, $t \in \mathbb{R}_+$.

(v) *A* $(0,\boldsymbol{\pi})$*-FMS admits the bounds* $\mathbf{z}^{(0,\boldsymbol{\pi}^{\mathrm{u}})}(t) \leq \mathbf{z}^{(0,\boldsymbol{\pi})}(t) \leq \mathbf{z}^{(0,\boldsymbol{\pi}^{\mathrm{l}})}(t)$, $t \in \mathbb{R}_+$.

Proof. (i) *for* $\varepsilon > 0$. Put $\mathbf{z}^0(\cdot) \equiv \boldsymbol{\varepsilon} := (\varepsilon, \ldots, \varepsilon)$, and, for $n \geq 0$, $\mathbf{z}^{n+1}(\cdot) = \mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{z}^n)(\cdot)$. Then $\mathbf{z}^1(t) \geq \boldsymbol{\varepsilon} = \mathbf{z}^0(t)$, $t \in \mathbb{R}_+$. The operator $\mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}$ is monotone, that is, $\mathbf{u}(t) \geq \mathbf{v}(t)$ for all $t \in \mathbb{R}_+$ implies $\mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{u})(t) \geq \mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{v})(t)$ for all $t \in \mathbb{R}_+$. Also $\mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{u})(t) \leq \boldsymbol{\varepsilon} + t\boldsymbol{\lambda}$ for all $\mathbf{u}(\cdot) \in \mathcal{G}(\mathbb{R}_+, \mathbb{R}_+^I)$ and all $t \in \mathbb{R}_+$. Hence, for each $t \in \mathbb{R}_+$, the sequence $\{\mathbf{z}^n(t)\}_{n\in\mathbb{Z}_+}$ is non-decreasing and bounded from above, and there exists the pointwise limit of $\mathbf{z}^n(\cdot)$ as $n \to \infty$; denote it by $\mathbf{z}(\cdot)$. Now we show that $\mathbf{z}(\cdot)$ is an $(\varepsilon,\boldsymbol{\pi})$-FMS. Take an arbitrary $t \in \mathbb{R}_+$. For all $n \in \mathbb{Z}_+$, we have $\mathbf{z}^n(\cdot) \geq \boldsymbol{\varepsilon}$, which implies $1/\|\mathbf{z}^n(\cdot)\|_1 \leq 1/(K\varepsilon)$. Then the dominated convergence theorem guarantees that $\int_s^t e^{-1}/\|\mathbf{z}^n(x)\|_1 dx \to \int_s^t e^{-1}/\|\mathbf{z}(x)\|_1 dx$ as $n \to \infty$, $s \in [0,t]$. Since $\exp\bigl(-\pi_i(t-s) - \int_s^t e^{-1}/\|\mathbf{z}^n(x)\|_1 dx\bigr) \leq 1$, $s \in [0,t]$, by the previous argument and the dominated convergence theorem, we obtain $\mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{z}^n)(t) \to \mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{z})(t)$ as $n \to \infty$. So, indeed, $\mathbf{z}(\cdot)$ satisfies equation (2.15) for all $t \in \mathbb{R}_+$.

(ii) Let $\mathbf{z}(\cdot)$ and $\widetilde{\mathbf{z}}(\cdot)$ be two $(\varepsilon,\boldsymbol{\pi})$-FMS's. Take an arbitrary $t \in (0,\infty)$. Since $\pi_i(t-s) + \int_s^t e^{-1}/\|\mathbf{z}(x)\|_1 dx \leq (\|\boldsymbol{\pi}\| + (eK\varepsilon)^{-1})T$, $0 \leq s \leq t \leq T$, and the same is true for $\widetilde{\mathbf{z}}(\cdot)$, by Lipschitz continuity of $\exp(\cdot)$ on compact sets, there exists a constant $\alpha(T)$ such that, for $t \leq T$,

$$\begin{aligned}\|\mathbf{z}(t) - \widetilde{\mathbf{z}}(t)\| &\leq \alpha(T)e^{-1}\int_0^t\int_s^t \Bigl|1/\|\mathbf{z}(x)\|_1 - 1/\|\widetilde{\mathbf{z}}(x)\|_1\Bigr| dx\, ds \\ &\leq \alpha(T)Te^{-1}T\int_0^t \Bigl|1/\|\mathbf{z}(x)\|_1 - 1/\|\widetilde{\mathbf{z}}(x)\|_1\Bigr| dx.\end{aligned}$$

Then, by Lipschitz continuity of $1/\|\cdot\|_1$ on the set $\mathbb{R}_\varepsilon^I$ (defined by (2.5)) and by the Gronwall inequality, $\mathbf{z}(\cdot)$ and $\widetilde{\mathbf{z}}(\cdot)$ must coincide in all finite intervals $[0,T]$, and hence they coincide on $\mathbb{R}_+$.

(iii) Due to Lemma 2.2, $(0,\boldsymbol{\pi})$-FMS's are defined by the Cauchy problem

$$\begin{aligned}&\mathbf{z}'(t) = \boldsymbol{\lambda} - \pi_1\mathbf{z}(t) - e^{-1}\mathbf{z}(t)/\|\mathbf{z}(t)\|_1, \quad t > 0,\\ &\mathbf{z}(0) = \mathbf{0}.\end{aligned}$$

Summing up its coordinates, we get the Cauchy problem

$$\begin{aligned}&\|\mathbf{z}(t)\|_1' = (\|\boldsymbol{\lambda}\|_1 - e^{-1}) - \pi_1\|\mathbf{z}(t)\|_1, \quad t > 0,\\ &\|\mathbf{z}(0)\|_1 = 0,\end{aligned}$$

which admits a unique solution

$$\|\mathbf{z}(t)\|_1 = \begin{cases} \bigl(\|\boldsymbol{\lambda}\|_1 - e^{-1}\bigr)\, t & \text{if } \pi_1 = 0,\\ (\|\boldsymbol{\lambda}\|_1 - e^{-1})(1 - e^{-\pi_1 t})/\pi_1 & \text{if } \pi_1 > 0.\end{cases}$$

In the case of $\varepsilon = 0$ and $\pi_1 = \ldots = \pi_I$, equation (2.15) implies that $\mathbf{z}(\cdot)/\|\mathbf{z}(\cdot)\|_1 \equiv$

$\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$. Then the unique $(0,\boldsymbol{\pi})$-FMS is given by (2.16).

(i) *for* $\varepsilon = 0$, *and* (v) In order to prove the existence, consider the sequence $\mathbf{z}^0(\cdot) := \mathbf{z}^{(0,\boldsymbol{\pi}^{\mathrm{u}})}(\cdot)$, $\mathbf{z}^{n+1}(\cdot) := \mathbf{F}^{(0,\boldsymbol{\pi})}(\mathbf{z}^n)(\cdot)$, $n \in \mathbb{Z}_+$. By the reasoning analogous to that in the case of $\varepsilon > 0$, the point-wise limit of this sequence is a $(0,\boldsymbol{\pi})$-FMS. Further consider the sequence $\mathbf{z}^0(\cdot) := \mathbf{z}^{(0,\boldsymbol{\pi})}(\cdot)$, $\mathbf{z}^{n+1}(\cdot) := \mathbf{F}^{(0,\boldsymbol{\pi}^{\mathrm{u}})}(\mathbf{z}^n)(\cdot)$, $n \in \mathbb{Z}_+$. It is non-increasing in n, and its point-wise limit is a $(0,\boldsymbol{\pi}^{\mathrm{u}})$-FMS. Then $\mathbf{z}^{(0,\boldsymbol{\pi}^{\mathrm{u}})}(t) \le \mathbf{z}^0(t) = \mathbf{z}^{(0,\boldsymbol{\pi})}(t)$ for all $t \in \mathbb{R}_+$. Similarly, the second bound holds.

(iv) Consider the sequence $\mathbf{z}^0(\cdot) := \mathbf{z}^{(\delta,\boldsymbol{\pi})}(\cdot)$, $\mathbf{z}^{n+1}(\cdot) := \mathbf{F}^{(\varepsilon,\boldsymbol{\pi})}(\mathbf{z}^n)(\cdot)$, $n \in \mathbb{R}_+$. It is non-decreasing in n, and its point-wise limit is the $(\varepsilon,\boldsymbol{\pi})$-FMS. Then $\mathbf{z}^{(\varepsilon,\boldsymbol{\pi})}(t) \ge \mathbf{z}^0(t) = \mathbf{z}^{(\delta,\boldsymbol{\pi})}(t)$ for all $t \in \mathbb{R}_+$. □

We proceed with properties of $(0,\boldsymbol{\pi})$-FMS's at $t = 0$ (*cf.* Remark 2.4).

Property 2.2. *For any* $(0,\boldsymbol{\pi})$*-FMS* $\mathbf{z}^{(0,\boldsymbol{\pi})}(\cdot)$*, its right derivative at* $t = 0$ *is well defined and* $(\mathbf{z}^{(0,\boldsymbol{\pi})})'(0) = (1 - e^{-1}/\|\boldsymbol{\lambda}\|_1)\boldsymbol{\lambda}$*. Also the limit* $\lim_{t\to 0} \mathbf{z}^{(0,\boldsymbol{\pi})}(t)/\|\mathbf{z}^{(0,\boldsymbol{\pi})}(t)\|_1 = \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$ *exists.*

Proof. Here are three possibilities: either $\pi^{\mathrm{u}} \ge \pi^{\mathrm{l}} > 0$ or $\pi^{\mathrm{u}} > \pi^{\mathrm{l}} = 0$, or $\pi^{\mathrm{u}} = \pi^{\mathrm{l}} = 0$. We prove the property in the first case, the other two cases can be treated similarly. By Property 2.1, (iii) and (v), for $i = 1,\dots,I$, $t \in \mathbb{R}_+$,

$$\frac{\lambda_i}{\|\boldsymbol{\lambda}\|_1}\frac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{\pi^{\mathrm{u}}}(1 - e^{-\pi^{\mathrm{u}}t}) \le z_i^{(0,\boldsymbol{\pi})}(t) \le \frac{\lambda_i}{\|\boldsymbol{\lambda}\|_1}\frac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{\pi^{\mathrm{l}}}(1 - e^{-\pi^{\mathrm{l}}t}).$$

Then, for any sequence $t_n \to 0$, $n \to \infty$,

$$\begin{aligned}\overline{\lim}_{n\to\infty}\frac{z_i^{(0,\boldsymbol{\pi})}(t_n)}{t_n} &\le \frac{\lambda_i}{\|\boldsymbol{\lambda}\|_1}\left(\|\boldsymbol{\lambda}\|_1 - e^{-1}\right)\lim_{n\to\infty}\frac{1 - e^{-\pi^{\mathrm{l}}t_n}}{\pi^{\mathrm{l}}t_n}\\ &= \frac{\lambda_i}{\|\boldsymbol{\lambda}\|_1}\left(\|\boldsymbol{\lambda}\|_1 - e^{-1}\right).\end{aligned}$$

Similarly, $\underline{\lim}_{n\to\infty} z_i^{(0,\boldsymbol{\pi})}(t_n)/t_n \ge (\lambda_i/\|\boldsymbol{\lambda}\|_1)\left(\|\boldsymbol{\lambda}\|_1 - e^{-1}\right)$. Hence, the derivative exists. The second result follows from Taylor's expansion, as was discussed in Remark 2.4. □

Finally, we show uniqueness of a $(0,\boldsymbol{\pi})$-FMS by estimating it via the auxiliary FMS's with other parameters.

Property 2.3. *Fix a* $\boldsymbol{\pi} \in \mathbb{R}_+^I$*. For short,* $\mathbf{z}^\varepsilon$ *denotes an* $(\varepsilon,\boldsymbol{\pi})$*-FMS.*

(i) *For* $\varepsilon > 0$ *and the function* $\varphi_\varepsilon(t) := \int_0^t Ke^{-1}/\|\mathbf{z}^\varepsilon(s)\|_1 ds$*,*

$$\|\mathbf{z}^\varepsilon(t) - \mathbf{z}^0(t)\|_1 \le \varepsilon\left(K + \|\boldsymbol{\pi}\|_1 + \varphi_\varepsilon(t)\right), \quad t \in \mathbb{R}_+. \tag{2.17}$$

(ii) *If* $\varepsilon > 0$, $\varepsilon \to 0$*, then* $\varepsilon\varphi_\varepsilon(t) \to 0$ *for any* $t \in (0,\infty)$.

(iii) *A $(0, \boldsymbol{\pi})$-FMS is unique.*

Proof. (i) Let $\varepsilon \geq 0$. By differentiating equation (2.15), we get, for $i = 1, \ldots, I$,

$$\begin{aligned}(z_i^\varepsilon)'(t) &= \lambda_i - \left(\pi_i + \frac{e^{-1}}{\|\mathbf{z}^\varepsilon(t)\|_1}\right)(z_i^\varepsilon(t) - \varepsilon), \quad t > 0,\\ z_i^\varepsilon(0) &= \varepsilon.\end{aligned}$$

Then integrating over $[0, t]$ yields

$$z_i^\varepsilon(t) - \varepsilon = \lambda_i t - \int_0^t \left(\pi_i + \frac{e^{-1}}{\|\mathbf{z}^\varepsilon(s)\|_1}\right)(z_i^\varepsilon(s) - \varepsilon)\, ds, \quad t \in \mathbb{R}_+,$$

which, after taking the sum in all coordinates, can be rewritten as

$$\begin{aligned}&\textstyle\sum_{i=1}^I z_i^\varepsilon(t) + \sum_{i=1}^I \int_0^t \pi_i z_i^\varepsilon(s) ds\\ &= (\|\boldsymbol{\lambda}\|_1 - e^{-1})t + \varepsilon K + \varepsilon\|\boldsymbol{\pi}\|_1 t + \varepsilon\varphi_\varepsilon(t), \quad t \in \mathbb{R}_+.\end{aligned}$$

The last equation implies that, for $\varepsilon > 0$,

$$\begin{aligned}&\textstyle\sum_{i=1}^I \left(z_i^\varepsilon(t) - z_i^0(t)\right) + \sum_{i=1}^I \int_0^t \pi_i \left(z_i^\varepsilon(s) - z_i^0(s)\right) ds\\ &= \varepsilon K + \varepsilon\|\boldsymbol{\pi}\|_1 t + \varepsilon\varphi_\varepsilon(t), \quad t \in \mathbb{R}_+.\end{aligned} \tag{2.18}$$

Due to Property 2.1, (v), for $\varepsilon > 0$, both sums in the LHS of (2.18) have non-negative summands. By omitting the second sum, we obtain the bound (2.17).

(ii) Suppose that $\pi^{\mathrm{u}} > 0$ (the other case can be treated similarly). Property 2.1, (iv) and (v), together with $\|\mathbf{z}^\varepsilon(\cdot)\|_1 \geq K\varepsilon$, implies that, for $\varepsilon > 0$,

$$\varphi_\varepsilon(t) = \int_0^t \frac{Ke^{-1}}{\|\mathbf{z}^\varepsilon(s)\|_1}\, ds \leq \int_0^t \frac{Ke^{-1}}{\max\{K\varepsilon, \|\mathbf{z}^{(0,\pi^{\mathrm{u}})}(s)\|_1\}}\, ds.$$

By (2.16), $\|\mathbf{z}^{(0,\pi^{\mathrm{u}})}(s)\|_1 \leq I\varepsilon$ if and only if

$$s \leq f(\varepsilon) := \frac{1}{\pi^{\mathrm{u}}} \ln \frac{\|\boldsymbol{\lambda}\|_1 - e^{-1}}{\|\boldsymbol{\lambda}\|_1 - e^{-1} - K\varepsilon\pi^{\mathrm{u}}}.$$

We have $f(\varepsilon) > 0$ for ε small enough and $f(\varepsilon) \to 0$, $\varepsilon \to 0$. Put $\beta = Ke^{-1}/(\|\boldsymbol{\lambda}\|_1 - e^{-1})$,

then

$$\begin{aligned}\varepsilon\varphi_\varepsilon t &= e^{-1}f(\varepsilon) + \varepsilon\beta\pi^{\mathrm{u}}\int_{f(\varepsilon)}^{t}\frac{1}{1-e^{-\pi^{\mathrm{u}}s}}\,ds\\ &= e^{-1}f(\varepsilon) + \varepsilon\beta\pi^{\mathrm{u}}\left(t - f(\varepsilon) + \frac{1}{\pi^{\mathrm{u}}}\ln\left(1-e^{-\pi^{\mathrm{u}}t}\right)\right)\\ &\quad - \left(\frac{\beta\varepsilon}{f(\varepsilon)}\right)\left(f(\varepsilon)\ln\left(1-e^{-\pi^{\mathrm{u}}f(\varepsilon)}\right)\right).\end{aligned}$$

In the very RHS of the last equation, convergence of the first two summands to 0 as $\varepsilon \to 0$ is clear. The first multiplier of the last summand tends to a finite constant, and the second multiplier tends to 0.

(iii) Suppose that $\mathbf{z}^0(\cdot)$ and $\widetilde{\mathbf{z}}^0(\cdot)$ are two $(0,\boldsymbol{\pi})$-FMS's. For any $t \in (0,\infty)$, by (i) and (ii), $\|\mathbf{z}^0(t) - \widetilde{\mathbf{z}}^0(t)\|_1 \le \|\mathbf{z}^0(t) - \mathbf{z}^\varepsilon(t)\|_1 + \|\mathbf{z}^\varepsilon(t) - \widetilde{\mathbf{z}}^0(t)\|_1 \to 0$ as $\varepsilon \to 0$. Hence, $\mathbf{z}^0(\cdot)$ and $\widetilde{\mathbf{z}}^0(\cdot)$ must coincide in $\mathbb{R}_+$. □

2.6 Proof of Theorem 2.2

Existence and uniqueness The function $f(x) = \sum_{i=1}^{I} p_i\lambda_i/(x+p_i)$ is continuous and strictly decreasing in $(0,\infty)$, and takes all values between $\|\boldsymbol{\lambda}\|_1$ and 0 as x goes along $(0,\infty)$. Then, by Assumption 2.2, there exists a unique $\mathbf{z}^*$ satisfying (2.6)-(2.7), and all its coordinates are positive. Invariant FMS's are defined by the equation

$$\boldsymbol{\lambda} - \mathbf{p}\times\mathbf{z}^* - e^{-1}\mathbf{z}^*/\left\|\mathbf{z}^*\right\|_1 = 0. \tag{2.19}$$

In order to prove the first part of the theorem, we have to check that (2.6)-(2.7) is a solution to (2.19), and that, if there is a solution to (2.19), then it is necessarily (2.6)-(2.7).

By plugging (2.6)-(2.7) into (2.19) multiplied by $\|\mathbf{z}^*\|_1$, we obtain, for $i = 1,\dots,I$,

$$\begin{aligned}&\left(\lambda_i - \frac{p_i\lambda_i}{x+p_i}\right)\sum_{j=1}^{I}\frac{\lambda_j}{x+p_j} - \frac{e^{-1}\lambda_i}{x+p_i}\\ &= \frac{\lambda_i x}{x+p_i}\sum_{j=1}^{I}\frac{\lambda_j}{x+p_j} - \frac{e^{-1}\lambda_i}{x+p_i}\\ &= \frac{\lambda_i}{x+p_i}\left(\sum_{j=1}^{I}\frac{\lambda_j x}{x+p_j} - e^{-1}\right)\\ &= \frac{\lambda_i}{x+p_i}\left(\sum_{j=1}^{I}\lambda_j - \sum_{j=1}^{I}\frac{\lambda_j p_j}{x+p_j} - e^{-1}\right) = 0.\end{aligned}$$

So, indeed, (2.6)-(2.7) is an invariant FMS.

Suppose now that $\mathbf{z}^*$ is a solution to (2.19). By solving coordinate i of (2.19) with respect

to z_i^*, we get

$$z_i^* = \frac{\lambda_i}{p_i + e^{-1}/\|\mathbf{z}^*\|_1}, \quad i = 1, \dots, I.$$

Plug the last relation into the sum of the coordinates of (2.19), then

$$\sum_{i=1}^I \frac{p_i \lambda_i}{p_i + e^{-1}/\|\mathbf{z}^*\|_1} = \|\boldsymbol{\lambda}\|_1 - e^{-1},$$

Hence, $\mathbf{z}^*$ satisfies (2.6) with $x = e^{-1}/\|\mathbf{z}^*\|_1$.

Stability By Remark 2.5 and Property 2.1, (iii) and (v), any FMS at any time $t > 0$ has all coordinates positive. Then it suffices to show convergence to the invariant point for FMS's that start in the interior of $\mathbb{R}_+^I$. This, in turn, follows from Proposition 1.1 with the open set $E = (0, \infty)^I$, equation (2.3) and the Lyapunov function

$$L(\mathbf{z}) = \sum_{i=1}^I \frac{y_i^2}{z_i^*}, \quad \text{where } y_i := z_i - z_i^*.$$

By plugging y_i's into (2.3), we get, for all i,

$$z_i' = -p_i y_i + e^{-1} \frac{z_i^*}{\|\mathbf{z}^*\|_1} - e^{-1} \frac{y_i + z_i^*}{\|\mathbf{y} + \mathbf{z}^*\|_1}.$$

Then,

$$L'(\mathbf{z}) = \sum_{i=1}^I \frac{2 y_i z_i'}{z_i^*} = -\sum_{i=1}^I \frac{2 p_i y_i^2}{z_i^*} - \frac{2e^{-1}}{\|\mathbf{y} + \mathbf{z}^*\|_1 \|\mathbf{z}^*\|_1} \Sigma(\mathbf{y}),$$

where

$$\Sigma(\mathbf{y}) := \sum_{i=1}^I \frac{y_i^2}{z_i^*/\|\mathbf{z}^*\|_1} - \left(\sum_{i=1}^I y_i\right)^2.$$

We have to check that $L'(\mathbf{z})$ is non-positive on $(0, \infty)^I$, or equivalently, that $\Sigma(\mathbf{y})$ is non-negative in $\mathbb{R}^I$. The latter indeed holds by convexity of the quadratic function:

$$\begin{aligned}\left(\sum_{i=1}^I y_i\right)^2 &\le \left(\sum_{i=1}^I \frac{z_i^*}{\|\mathbf{z}^*\|_1} \frac{y_i}{z_i^*/\|\mathbf{z}^*\|_1}\right)^2 \\ &\le \sum_{i=1}^I \frac{z_i^*}{\|\mathbf{z}^*\|_1} \left(\frac{y_i}{z_i^*/\|\mathbf{z}^*\|_1}\right)^2 \le \sum_{i=1}^I \frac{y_i^2}{z_i^*/\|\mathbf{z}^*\|_1}.\end{aligned}$$

2.7 Proof of Theorem 2.3

The proof is organised as follows. Section 2.7.1 contains a representation of the population process before and after the fluid scaling. In Section 2.7.2, we formulate two auxiliary results (Lemmas 2.3 and 2.4), and then show that the family of the fluid-scaled

processes is **C**-tight and that fluid limits are a.s. FMS's. Proofs of Lemmas 2.3 and 2.4 are given in Sections 2.7.3 and 2.7.4, respectively.

We assume that all stochastic models of this section are defined on the same probabilty space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

2.7.1 A representation of the population process

Throughout the proof, unless otherwise stated, we use the following representation of the population dynamics:

$$\mathbf{Q}^r(n+1) = \mathbf{Q}^r(n) + \mathbf{A}(n+1) - \mathbf{D}^{\mathrm{t}}\left(n, \mathbf{Q}^r(n)\right) - \mathbf{D}^{r,\mathrm{a}}\left(n, \mathbf{Q}^r(n)\right), \tag{2.20}$$

where, for $\mathbf{x} \in \mathbb{Z}_+^I$ and $i = 1, \ldots, I$,

$$D_i^{\mathrm{t}}(n, \mathbf{x}) = \mathbb{I}\left\{\sum_{j=1}^{i-1} p_j(\mathbf{x}) \le U(n) < \sum_{j=1}^{i} p_j(\mathbf{x})\right\}$$

$$p_i(\mathbf{x}) = \begin{cases} B\left(x_i, 1/\|\mathbf{x}\|_1\right)(\{1\}) \prod_{j \neq i} B\left(x_j, 1/\|\mathbf{x}\|_1\right)(\{0\}), & \mathbf{x} \neq \mathbf{0}, \\ 0, & \mathbf{x} = \mathbf{0}, \end{cases}$$

$$D_i^{r,\mathrm{t}}(n, \mathbf{x}) = \sum_{j=1}^{x_i - D_i^{\mathrm{t}}(n,\mathbf{x})} \xi_i^r(n, j),$$

and

- $\{U(n)\}_{n \in \mathbb{Z}_+}$ is an i.i.d. sequence, and $U(0)$ is distributed uniformly over $[0, 1]$,
- $B(x_i, 1/\|\mathbf{x}\|_1)$ is the binomial distribution with parameters x_i and $1/\|\mathbf{x}\|_1$,
- $\{\xi_i^r(n, j)\}_{j \in \mathbb{N}}$, $n \in \mathbb{Z}_+$, $i = 1, \ldots, I$, are independent i.i.d. sequences of Bernoulli r.v.'s, and $\mathbb{P}\{\xi_i^r(n, 1) = 1\} = p_i/r = 1 - \mathbb{P}\{\xi_i^r(n, 1) = 0\}$,
- the sequences $\{\mathbf{A}(n)\}_{n \in \mathbb{N}}$, $\{U(n)\}_{n \in \mathbb{Z}_+}$ and $\{\xi_i^r(n, j)\}_{j \in \mathbb{N}}$, $n \in \mathbb{Z}_+$, $i = 1, \ldots, I$, are mutually independent and also do not depend on the initial condition $\mathbf{Q}^r(0)$.

For short, we put

$$h(x) = \begin{cases} 0, & x = 0, \\ (1 - 1/x)^{x-1}, & x \ge 1. \end{cases}$$

Then, in particular,

$$\begin{aligned} p_i(\mathbf{x}) &= h\big(\|\mathbf{x}\|_1\big) m_i\big(\mathbf{x}\big), \\ \mathbb{E}\Big[\mathbf{D}^{\mathrm{t}}\left(i, \mathbf{Q}^r(i)\right) \Big| \mathbf{Q}^r(i)\Big] &= h\left(\|\mathbf{Q}^r(i)\|_1\right) \mathbf{m}\left(\mathbf{Q}^r(i)\right), \\ \mathbb{E}\Big[\mathbf{D}^{r,\mathrm{a}}\left(i, \mathbf{Q}^r(i)\right) \Big| \mathbf{Q}^r(i)\Big] &= \frac{\mathbf{p}}{r} \times \Big(\mathbf{Q}^r(i) - h\left(\|\mathbf{Q}^r(i)\|_1\right) \mathbf{m}\left(\mathbf{Q}^r(i)\right)\Big). \end{aligned}$$

We now transform the workload dynamics into an integral equation that, as we show later, differs from the fluid model equation (2.2) by the terms that vanish as $r \to \infty$. For

any $n \in \mathbb{Z}_+$, we have

$$\begin{aligned}\mathbf{Q}^r(n) &= \mathbf{Q}^r(0) + \sum_{i=1}^{n} \mathbf{A}(i) - \sum_{i=0}^{n-1} \mathbf{D}^{\mathrm{t}}\left(i, \mathbf{Q}^r(i)\right) - \sum_{i=0}^{n-1} \mathbf{D}^{r,\mathrm{a}}\left(i, \mathbf{Q}^r(i)\right) \\ &= \mathbf{Q}^r(0) + n\boldsymbol{\lambda} - \sum_{i=0}^{n-1} h\left(\|\mathbf{Q}^r(i)\|_1\right) \mathbf{m}\left(\mathbf{Q}^r(i)\right) \\ &\quad - \frac{\mathbf{p}}{r} \times \sum_{i=0}^{n-1} \left(\mathbf{Q}^r(i) - h\left(\|\mathbf{Q}^r(i)\|_1\right) \mathbf{m}\left(\mathbf{Q}^r(i)\right)\right) + \mathbf{M}^r(n), \end{aligned} \tag{2.21}$$

where the sequence $\{\mathbf{M}^r(n)\}_{n\in\mathbb{Z}_+}$ forms a zero-mean martingale since

$$\begin{aligned}\mathbf{M}^r(n) = &\sum_{i=1}^{n} \left(\mathbf{A}(i) - \mathbb{E}\mathbf{A}(i)\right) \\ &- \sum_{i=0}^{n-1} \left(\mathbf{D}^{\mathrm{t}}\left(i, \mathbf{Q}^r(i)\right) - \mathbb{E}\left[\mathbf{D}^{\mathrm{t}}\left(i, \mathbf{Q}^r(i)\right) \middle| \mathbf{Q}^r(i)\right]\right) \\ &- \sum_{i=0}^{n-1} \left(\mathbf{D}^{r,\mathrm{a}}\left(i, \mathbf{Q}^r(i)\right) - \mathbb{E}\left[\mathbf{D}^{r,\mathrm{a}}\left(i, \mathbf{Q}^r(i)\right) \middle| \mathbf{Q}^r(i)\right]\right). \end{aligned}$$

Introduce the fluid-scaled version of the martingale $\{\mathbf{M}^r(n)\}_{n\in\mathbb{Z}_+}$ analogous to that of the workload process:

$$\overline{\mathbf{Q}}^r(t) = \mathbf{Q}^r\left(\lfloor rt \rfloor\right)/r, \quad \overline{\mathbf{M}}^r(t) = \mathbf{M}^r\left(\lfloor rt \rfloor\right)/r.$$

Then equation (2.21) turns into the integral equation

$$\begin{aligned}\overline{\mathbf{Q}}^r(t) = &\overline{\mathbf{Q}}^r(0) + \frac{\lfloor rt \rfloor}{r} \boldsymbol{\lambda} \\ &- \left(\mathbf{1} - \frac{\mathbf{p}}{r}\right) \times \int_0^{\lfloor rt \rfloor / r} h\left(r\|\overline{\mathbf{Q}}^r(s)\|_1\right) \mathbf{m}\left(\overline{\mathbf{Q}}^r(s)\right) ds \\ &- \mathbf{p} \times \int_0^{\lfloor rt \rfloor / r} \overline{\mathbf{Q}}^r(s)\, ds + \overline{\mathbf{M}}^r(t), \end{aligned} \tag{2.22}$$

where

$$\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^I.$$

Finally, we rewrite equation (2.22) as

$$\overline{\mathbf{Q}}^r(t) = \overline{\mathbf{Q}}^r(0) + t\boldsymbol{\lambda} - e^{-1} \int_0^t \mathbf{m}\left(\overline{\mathbf{Q}}^r(s)\right) ds - \mathbf{p} \times \int_0^t \overline{\mathbf{Q}}^r(s) ds + \mathbf{G}^r(t), \tag{2.23}$$

where

$$\begin{aligned}\mathbf{G}^r(t) &= \overline{\mathbf{M}}^r(t) + \mathbf{G}^{r,1}(t) + \mathbf{G}^{r,2}(t) + \mathbf{G}^{r,3}(t), \\ \mathbf{G}^{r,1}(t) &= \left(\lfloor rt \rfloor / r - t\right)\boldsymbol{\lambda}, \\ \mathbf{G}^{r,2}(t) &= e^{-1} \int_0^t \mathbf{m}\left(\overline{\mathbf{Q}}^r(s)\right) ds \\ &\quad - \left(\mathbf{1} - \frac{\mathbf{p}}{r}\right) \times \int_0^{\lfloor rt \rfloor / r} h\left(r\|\overline{\mathbf{Q}}^r(s)\|_1\right) \mathbf{m}\left(\overline{\mathbf{Q}}^r(s)\right) ds, \end{aligned}$$

$$\mathbf{G}^{r,3}(t) = \mathbf{p} \times \int_{\lfloor rt \rfloor / r}^{t} \overline{\mathbf{Q}}^{r}(s)\, ds.$$

2.7.2 C-tightness and limiting equations

We first discuss convergence $\mathbf{G}^r(\cdot) \Rightarrow \mathbf{0}$ as $r \to \infty$ in $D(\mathbb{R}_+, \mathbb{R}^I)$. By the FLLN and Proposition 1.2, $\mathbf{G}^{r,1}(\cdot) \Rightarrow \mathbf{0}$ as $r \to \infty$. Weak convergence to zero of the three other summands in $\mathbf{G}^r(\cdot)$ follows from Lemmas 2.3 and 2.4.

Lemma 2.3. *Let Assumptions 2.1 and 2.2 hold. Then*

(i) *for any $\delta > 0$ and $\varepsilon > 0$, there exists a $\gamma = \gamma(\delta, \varepsilon) > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}\left\{\varphi^r(\gamma r) \le \delta r\right\} \ge 1 - \varepsilon,$$

where

$$\varphi^r(\gamma r) := \inf\{n \ge 0 : \|\mathbf{Q}^r(n)\|_1 \ge \gamma r\},$$

(ii) *for any $\varepsilon > 0$ and $\Delta > \delta > 0$, there exists a $C = C(\varepsilon, \delta, \Delta) > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}\{\inf_{\delta \le t \le \Delta} \|\overline{\mathbf{Q}}^r(t)\|_1 \ge C\} \ge 1 - \varepsilon, \tag{2.24}$$

(iii) $\mathbf{G}^{r,2}(\cdot) \Rightarrow \mathbf{0}$ *in* $D(\mathbb{R}_+, \mathbb{R}^I)$ *as* $r \to \infty$.

Lemma 2.4. *Suppose* $\overline{\mathbf{Q}}^r(0) \Rightarrow \mathbf{z}^0$ *as* $r \to \infty$. *Then* $\mathbf{G}^{r,3}(\cdot) \Rightarrow \mathbf{0}$ *and* $\overline{\mathbf{M}}^r(\cdot) \Rightarrow \mathbf{0}$ *in* $D(\mathbb{R}_+, \mathbb{R}^I)$.

Now we are in a position to prove the theorem. The proof consists of two steps. First, we establish **C**-tightness of a family of the scaled processes $\overline{\mathbf{Q}}^r(\cdot)$ such that $\overline{\mathbf{Q}}^r(0) \Rightarrow \mathbf{z}^0$ as $r \to \infty$. Second, we show that all weak limits of such a family a.s. coincide with the FMS starting from $\mathbf{z}^0$.

C-tightness We prove the **C**-tightness by applying Proposition 1.3, i.e. we show that asymptotically the scaled processes $\overline{\mathbf{Q}}^r(\cdot)$ live on a compact set and have small oscillations.

The compact containment condition (1.7) follows easily by the upper bound

$$\overline{\mathbf{Q}}^r(t) \le \overline{\mathbf{Q}}^r(0) + \textstyle\sum_{n=1}^{\lfloor rt \rfloor} \mathbf{A}(n)/r \Rightarrow \mathbf{z}^0 + t\,\boldsymbol{\lambda}.$$

Now we establish the oscillation control (1.8). By (2.23), we have, for all s, t and i,

$$\overline{Q}_i^r(t) - \overline{Q}_i^r(s) = \lambda_i(t-s) - e^{-1}\int_s^t m_i(\overline{\mathbf{Q}}^r(x))\,dx - p_i\int_s^t \overline{Q}_i^r(x)\,dx + G_i^r(t) - G_i^r(s).$$

Since $m_i(\cdot) \le 1$, it follows for $s, t \in [0, T]$, $|s-t| < \delta$ that

$$|\overline{Q}_i^r(t) - \overline{Q}_i^r(s)| \le (\lambda_i + e^{-1})\delta + p_i\delta \sup_{0\le x\le T}\|\overline{Q}_i^r(x)\| + 2\sup_{0\le x\le T}\|G_i^r(x)\|,$$

where, again by (2.23),

$$\sup_{0\le x\le T}\|\overline{Q}_i^r(x)\| \le \|\overline{\mathbf{Q}}^r(0)\| + \|\boldsymbol{\lambda}\|T + \sup_{0\le x\le T}\|G_i^r(x)\|.$$

The last two bounds put together give the following upper bound on oscillations of the scaled process $\overline{\mathbf{Q}}^r(\cdot)$:

$$\begin{aligned}\omega(\overline{\mathbf{Q}}^r, T, \delta) &:= \sup\{\|\overline{\mathbf{Q}}^r(t) - \overline{\mathbf{Q}}^r(s)\| : s, t \in [0,T], |s-t| < \delta\}\\ &\le (\|\boldsymbol{\lambda}\| + e^{-1})\delta + \|\mathbf{p}\|\delta(\|\overline{\mathbf{Q}}^r(0)\| + \|\boldsymbol{\lambda}\|T)\\ &\quad + (2 + \|\mathbf{p}\|\delta)\sup_{0\le x\le T}\|\mathbf{G}^r(x)\|.\end{aligned}$$

By Lemmas 2.3 and 2.4, and Proposition 1.2, $\sup_{0\le x\le T}\|\mathbf{G}^r(x)\| \to 0$ as $r \to \infty$ for any $T > 0$; and also $\overline{\mathbf{Q}}^r(0) \to \mathbf{z}^0$. Then there exists a $\delta > 0$ such that the oscillation control condition (1.8) holds.

Fluid limits as FMS's Now we show that, if a sequence $\{\overline{\mathbf{Q}}^q(\cdot)\}_{q\to\infty}$ converges weakly in $\mathbf{D}(\mathbb{R}_+, \mathbb{R}^I)$, then its limit $\widetilde{\mathbf{Q}}(\cdot)$ a.s.

(i) is continuous,

(ii) does not vanish at $t \in (0, \infty)$,

(iii) satisfies the fluid model equation (2.2).

Then, by the uniqueness of FMS's, $\widetilde{\mathbf{Q}}(\cdot)$ a.s. coincides with the FMS starting from $\mathbf{z}^0$.

(i) Continuity of $\widetilde{\mathbf{Q}}(\cdot)$ follows by the $\mathbf{C}$-tightness.

(ii) By Lemma 2.3, for any $\varepsilon > 0$ and $n \ge 1$, there exists a constant $C(\varepsilon, 1/n, n) > 0$ such that

$$\underline{\lim}_{q\to\infty}\mathbb{P}\{\inf_{1/n\le t\le n}\|\overline{\mathbf{Q}}^q(t)\|_1 \ge C(\varepsilon, 1/n, n)\} \ge 1 - \varepsilon.$$

Since $\inf_{1/n\le t\le n}\|\mathbf{x}(t)\|$ is a continuous mapping in $D(\mathbb{R}_+, \mathbb{R}^I)$, one can choose a constant $\widetilde{C}(\varepsilon, 1/n, n) \in (0, C(\varepsilon, 1/n, n)]$ being a continuity point for the distribution of

$\inf_{1/n\le t\le n}\|\widetilde{\mathbf{Q}}(t)\|_1$. Then

$$\begin{aligned}1-\varepsilon &\le \lim_{q\to\infty}\mathbb{P}\{\inf_{1/n\le t\le n}\|\overline{\mathbf{Q}}^q(t)\|_1\ge \widetilde{C}(\varepsilon,1/n,n)\}\\ &= \mathbb{P}\{\inf_{1/n\le t\le n}\|\widetilde{\mathbf{Q}}(t)\|_1\ge \widetilde{C}(\varepsilon,1/n,n)\}\\ &\le \mathbb{P}\{\inf_{1/n\le t\le n}\|\widetilde{\mathbf{Q}}(t)\|_1>0\}.\end{aligned}$$

By taking the limit as $\varepsilon\to 0$ in the last inequality, we obtain $\mathbb{P}\{\inf_{1/n\le t\le n}\|\widetilde{\mathbf{Q}}(t)\|_1>0\}=1$ for any $n\ge 1$, which, in turn, implies that

$$\mathbb{P}\{\|\widetilde{\mathbf{Q}}(t)\|_1>0 \text{ for all } t\in(0,\infty)\}=1. \tag{2.25}$$

(iii) Fix a $t\in\mathbb{R}_+$. We introduce the mappings $\varphi_t^1,\varphi_t^2\colon D(\mathbb{R}_+,\mathbb{R}^I)\to\mathbb{R}^I$ defined by

$$\begin{aligned}\varphi_t^1(\mathbf{x}) &= \mathbf{x}(t)-\mathbf{x}(0)-t\boldsymbol{\lambda}+\mathbf{p}\times\int_0^t\mathbf{x}(s)\,ds,\\ \varphi_t^2(\mathbf{x}) &= e^{-1}\int_0^t\mathbf{m}(\mathbf{x}(s))\,ds.\end{aligned}$$

By Proposition 1.2, the mapping φ_t^1 is continuous at any continuous $\mathbf{x}(\cdot)$. By $m_i(\cdot)\le 1$, $i=1,\dots,I$, and the dominated convergence theorem, the mapping φ_t^2 is continuous at any continuous $\mathbf{x}(\cdot)$ that differs from zero everywhere except points forming a set of zero Lebesgue measure. Then, by the continuity of $\widetilde{\mathbf{Q}}(\cdot)$, (2.25) and the continuous mapping theorem, we have

$$\varphi_t^1(\overline{\mathbf{Q}}^q)+\varphi_t^2(\overline{\mathbf{Q}}^q)\Rightarrow\varphi_t^1(\widetilde{\mathbf{Q}})+\varphi_t^2(\widetilde{\mathbf{Q}})\quad\text{as } q\to\infty.$$

On the other hand, by (2.23),

$$\varphi_t^1(\overline{\mathbf{Q}}^q)+\varphi_t^2(\overline{\mathbf{Q}}^q)=\mathbf{G}^q(t)\Rightarrow\mathbf{0}\quad\text{as } q\to\infty.$$

Hence, for any $t\in\mathbb{R}_+$,

$$\mathbb{P}\{\varphi_t^1(\widetilde{\mathbf{Q}})+\varphi_t^2(\widetilde{\mathbf{Q}})=\mathbf{0}\}=\mathbb{P}\{\widetilde{\mathbf{Q}}(\cdot)\text{ satisfies (2.2) at } t\}=1.$$

Let Ω_t denote either of the events (they coincide) in the last equality. Then, again due to the continuity of $\widetilde{\mathbf{Q}}(\cdot)$ and (2.25),

$$\mathbb{P}\{\widetilde{\mathbf{Q}}(\cdot)\text{ satisfies (2.2) in }\mathbb{R}_+\}=\mathbb{P}\{\textstyle\bigcap\Omega_t\text{ over all rational } t\in\mathbb{R}_+\}=1.$$

2.7.3 Proof of Lemma 2.3

We split the proof into four parts. In the first two parts, we show that Assumptions 2.1 and 2.2 imply (i), and that (i) implies (ii), both for the single-class case. In the third part, we show that the total population $\|\mathbf{Q}(\cdot)\|_1$ of a multiclass model is bounded from below by that of a single-class model with suitable parameters. Then (i) and (ii) hold for the

multiclass case, too. In the last part, we show that (ii) implies (iii).

Assumptions 2.1 and 2.2 imply (i)**, single class case** For every γ from an interval $(0, \gamma^*]$, we construct a Markov chain (see $\{V_\gamma(n)\}_{n\in\mathbb{Z}_+}$ below) that, for all r large enough, is a lower bound for the population process until the latter first hits the set $[\gamma r, \infty)$. Then we choose a γ so as to have (i) with $\{V_\gamma(n)\}_{n\in\mathbb{Z}_+}$ in place of $\{Q^r(n)\}_{n\in\mathbb{Z}_+}$, and this completes the proof.

Throughout the proof, δ and ε are fixed.

Without loss of generality, we assume that, for all r, a.s. $Q^r(0) = 0$. Indeed, for all n and $x \geq y$,

$$\text{a.s.} \quad x - D^{\mathrm{t}}(n,x) - D^{r,\mathrm{a}}(n,x) \geq y - D^{\mathrm{t}}(n,y) - D^{r,\mathrm{a}}(n,y). \tag{2.26}$$

Property (2.26) says that the process $\{Q^r(n)\}_{n\in\mathbb{Z}_+}$ admits path-wise monotonicity: the bigger the initial value $Q^r(0)$ is, the bigger all the other values $Q^r(n)$, $n \geq 1$, are.

Further we make preparations needed to construct the lower-bound Markov chains. Let

$$h^* = e^{-1} + (\lambda - e^{-1})/2, \quad B(k, 1/k)(\{1\}) \leq h^* \text{ for } k \geq N. \tag{2.27}$$

Let $\{B(n)\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. r.v.'s with the binomial distribution $B(N, p)$.

We apply the following statement (see Section 2.8 for the proof) with $a = (\lambda - e^{-1})/4$.

Statement 2.1. *For any $a > 0$, there exists a $\gamma^* = \gamma^*(a)$ and a family of r.v.'s $\{\theta_\gamma\}_{0\leq\gamma\leq\gamma^*}$ with the following properties:*

(i) *the family $\{\theta_\gamma\}_{0\leq\gamma\leq\gamma^*}$ is uniformly integrable;*

(ii) *for any $\gamma \in [0, \gamma^*]$, $\mathbb{E}\theta_\gamma \leq a$;*

(iii) *$\theta_\gamma \Rightarrow \theta_0$ as $\gamma \to 0$;*

(iv) *for any $\gamma \in (0, \gamma^*]$, there exists an r_γ such that, for all $r \geq r_\gamma$, $\theta_\gamma \geq_{\mathrm{st}} B^r_\gamma$, where B^r_γ is a r.v. with the binomial distribution $B(\lfloor \gamma r \rfloor, p/r)$.*

For $\gamma \in (0, \gamma^*]$, let $\{\theta_\gamma(n)\}_{n\in\mathbb{N}}$ be an i.i.d. sequence with $\theta_\gamma(1) \stackrel{\mathrm{d}}{=} \theta_\gamma$, and assume that this sequence does not depend on $\{B(n)\}_{n\in\mathbb{N}}$.

Now we construct the lower-bound Markov chains. For r large enough, $Q^r(n) < N$ implies that

$$\begin{aligned} Q^r(n+1) - Q^r(n) &\stackrel{\text{a.s.}}{\geq} A(n+1) - 1 - \textstyle\sum_{i=1}^{N} \xi^r(n,i) \\ &\geq_{\mathrm{st}} A(1) - 1 - B(1), \end{aligned} \tag{2.28}$$

and $N \le Q^r(n) < \gamma r$ implies that

$$\begin{aligned} Q^r(n+1) - Q^r(n) &\overset{\text{a.s.}}{\ge} A(n+1) - \mathbb{I}\{U(n) \le h^*\} - \textstyle\sum_{i=1}^{\lfloor \gamma r \rfloor} \xi^r(n,i) \\ &\ge_{\text{st}} A(1) - \mathbb{I}\{U(0) \le h^*\} - \theta_\gamma. \end{aligned} \tag{2.29}$$

Introduce the two i.i.d. sequences:

$$x(n) = A(n) - 1 - B(n), \quad n \ge 1, \tag{2.30}$$

$$y_\gamma(n) = A(n) - \mathbb{I}\{U(n-1) \le h^*\} - \theta_\gamma(n), \quad n \ge 1, \tag{2.31}$$

and the two auxiliary Markov chains:

$$V_\gamma^r(0) = 0,$$

$$V_\gamma^r(n+1) = \begin{cases} (V_\gamma^r(n) + x(n+1))^+ & \text{if } V_\gamma^r(n) < N, \\ (V_\gamma^r(n) + y_\gamma(n+1))^+ & \text{if } N \le V_\gamma^r(n) < \gamma r, \\ V_\gamma^r(n) + A(n+1) - D^{\mathsf{t}}\left(n, V_\gamma^r(n)\right) & \\ -D^{r,\mathsf{a}}\left(n, V_\gamma^r(n)\right) & \text{if } V_\gamma^r(n) \ge \gamma r, \end{cases}$$

$$V_\gamma(0) = 0,$$

$$V_\gamma(n+1) = \begin{cases} (V_\gamma(n) + x(n+1))^+ & \text{if } V_\gamma(n) < N, \\ (V_\gamma(n) + y_\gamma(n+1))^+ & \text{if } V_\gamma(n) \ge N. \end{cases}$$

Put $\psi_\gamma^r(\gamma r)$ and $\psi_\gamma(\gamma r)$ to be the first hitting times of the set $[\gamma r, \infty)$ for the processes $\{V_\gamma^r(n)\}_{n\in\mathbb{Z}_+}$ and $\{V_\gamma(n)\}_{n\in\mathbb{Z}_+}$, respectively. Then $\psi_\gamma^r(\gamma r) = \psi_\gamma(\gamma r)$ for all r.

The processes $\{V_\gamma^r(n)\}_{n\in\mathbb{Z}_+}$ and $\{Q^r(n)\}_{n\in\mathbb{Z}_+}$ are related in the following way: $V_\gamma^r(n) = x$, $Q^r(n) = y$, where $x \le y$, implies $Q^r(n+1) \ge_{\text{st}} V_\gamma^r(n+1)$. Indeed, due to inequalities (2.26), (2.28) and (2.29),

$$\begin{aligned} Q^r(n+1) &= y + A(n+1) - D^{\mathsf{t}}(n,y) - D^{r,\mathsf{a}}(n,y) \\ &\overset{\text{a.s.}}{\ge} x + A(n+1) - D^{\mathsf{t}}(n,x) - D^{r,\mathsf{a}}(n,x) \\ &\begin{cases} \ge_{\text{st}} (x + x(n+1))^+ = V_\gamma^r(n+1), & \text{if } x < N, \\ \ge_{\text{st}} (x + y(n+1))^+ = V_\gamma^r(n+1), & \text{if } N \le x < \gamma r, \\ = V_\gamma^r(n+1), & \text{if } x \ge \gamma r. \end{cases} \end{aligned}$$

Then we get $\varphi^r(\gamma r) \ge_{\text{st}} \psi_\gamma^r(\gamma r) = \psi_\gamma(\gamma r)$ as a consequence of the following result (see Section 2.8 for the proof).

Statement 2.2. *Suppose $\{X(n)\}_{n\in\mathbb{Z}_+}$ and $\{Y(n)\}_{n\in\mathbb{Z}_+}$ are Markov chains with a common state space S, where S is a closed subset of $\mathbb{R}$, and deterministic initial states $X(0) \le Y(0)$. Suppose also that, for any $x \le y$ and any z,*

$$\mathbb{P}\{X(n+1) \ge z \big| X(n) = x\} \le \mathbb{P}\{Y(n+1) \ge z \big| Y(n) = y\}.$$

Then there exist Markov Chains $\{\widetilde{X}(n)\}_{n\in\mathbb{Z}_+}$ *and* $\{\widetilde{Y}(n)\}_{n\in\mathbb{Z}_+}$ *defined on a common probability space, distributed as* $\{X(n)\}_{n\in\mathbb{Z}_+}$ *and* $\{Y(n)\}_{n\in\mathbb{Z}_+}$*, respectively, and such that* $\widetilde{X}(n) \le \widetilde{Y}(n)$ *a.s. for all* n.

Now our goal is to choose a $\gamma \in (0, \gamma^*]$ so as to have

$$\underline{\lim}_{r\to\infty} \mathbb{P}\{\psi_\gamma(\gamma r) \le \delta r\} \ge 1 - \varepsilon. \tag{2.32}$$

To track the moments when the process $\{V_\gamma(n)\}_{n\in\mathbb{Z}_+}$ reaches level N from below and above, we recursively define the hitting times

$$\begin{aligned} &\tau_\gamma(0) = 0, && \tau_\gamma(i) = \inf\{n \ge \nu_\gamma(i-1)\colon V_\gamma(n) \ge N\}, \quad i \in \mathbb{N}, \\ &\nu_\gamma(0) = 0, && \nu_\gamma(i) = \inf\{n \ge \tau_\gamma(i)\colon V_\gamma(n) < N\}, \quad i \in \mathbb{N}. \end{aligned}$$

By convention, the infimum over the empty set is ∞. So, if either $\tau_\gamma(i) = \infty$, or $\nu_\gamma(i) = \infty$, then $\tau_\gamma(j) = \nu_\gamma(j) = \infty$ for all $j > i$.

Note that the r.v. $\tau(1) := \tau_\gamma(1)$ is a.s. finite and does not depend on γ because, for $n \le \tau(1)$, the process $\{V_\gamma(n)\}_{n\in\mathbb{Z}_+}$ is a reflected homogeneous random walk with i.i.d. increments $\{x(n)\}_{n\in\mathbb{N}}$, which are given by (2.30) and do not depend on γ. By Assumption 2.1, $\mathbb{P}\{x(1) > 0\} > 0$, then $\mathbb{E}\tau(1) < \infty$. Further, for any i, if $\nu_\gamma(i-1)$ is finite, then $\tau_\gamma(i)$ is finite, too, and the difference $\widetilde{\tau}_\gamma(i) := \tau_\gamma(i) - \nu_\gamma(i-1)$ is stochastically bounded from above by $\tau(1)$.

Let $q_\gamma(i) = \mathbb{P}\{\nu_\gamma(i) < \infty \big| \nu_\gamma(i-1) < \infty\}$. Then there exists a constant $\widetilde{q} < 1$ such that,

$$\text{for all } i \text{ and } \gamma \text{ small enough}, \quad q_\gamma(i) \le \widetilde{q}. \tag{2.33}$$

Indeed, for all $\gamma \in [0, \gamma^*]$, consider the random walks $Y_\gamma(n) := \sum_{i=1}^n y_\gamma(i)$ (here $y_0(n)$, $n \in \mathbb{N}$, are defined by (2.31) with $\gamma = 0$). By Statement 2.1, the family $\{y_\gamma(1)\}_{0\le\gamma\le\gamma^*}$ is uniformly integrable, and $y_\gamma(1) \Rightarrow y_0(1)$ as $\gamma \to 0$, which, together with $\mathbb{E}y_0(1) \ge (\lambda - e^{-1})/4 > 0$, implies that $\inf_{n\in\mathbb{Z}_+} Y_\gamma(n) \Rightarrow \inf_{n\in\mathbb{Z}_+} Y_0(n)$ as $\gamma \to 0$ (see Asmussen [5, Chapter X, Theorem 6.1]). Also $\mathbb{E}y_0(1) > 0$ implies that $\mathbb{P}\{\inf_{n\in\mathbb{Z}_+} Y_0(n) \ge 0\} =: p_0 > 0$ (see Asmussen [5, Chapter VIII, Theorem 2.4]). Then, for all i, $\mathbb{P}\{\nu_\gamma(i) = \infty \big| \nu_\gamma(i-1) < \infty\} \le \mathbb{P}\{\inf_{n\in\mathbb{Z}_+} Y_\gamma(n) \ge 0\} \to p_0 > 0$, and, for all i and γ small enough, $q_\gamma(i) \le 1 - p_0/2 =: \widetilde{q} < 1$.

Let $K_\gamma = \inf\{i \in \mathbb{N}\colon \nu_\gamma(i) = \infty\}$. By (2.33), the K_γ's are stochastically bounded from above by a geometric r.v. uniformly in γ small enough,

$$\mathbb{P}\{K_\gamma > i\} \le \widetilde{q}^{\,i}, \quad i \ge 0. \tag{2.34}$$

Further, for $i = 1, \dots, I_\gamma$, define the hitting times

$$\widetilde{\nu}_\gamma(i) = \inf\{n \in \mathbb{Z}_+\colon \textstyle\sum_{j=1}^n y_\gamma(\tau_\gamma(i) + j) \ge \gamma r\}.$$

Since $\mathbb{E}y_\gamma(1) > 0$, these r.v.'s are finite. We have $\psi_\gamma(\gamma r) \le \sum_{i=1}^{K_\gamma} (\widetilde{\tau}_\gamma(i) + \widetilde{\nu}_\gamma(i))$. Indeed,

if $\min\{i \in \mathbb{N}\colon \nu_\gamma(i) - \tau_\gamma(i) \ge \widetilde{\nu}_\gamma(i)\} = k$, then $k \le K_\gamma$ because $\nu_\gamma(K_\gamma) = \infty$, and

$$\begin{aligned}\psi_\gamma(\gamma r) &\le \tau_\gamma(k) + \widetilde{\nu}_\gamma(k)\\ &= (\tau_\gamma(1) - \nu_\gamma(0)) + (\nu_\gamma(1) - \tau_\gamma(1)) + \cdots + (\tau_\gamma(k) - \nu_\gamma(k)) + \widetilde{\nu}_\gamma(k)\\ &\le \widetilde{\tau}_\gamma(1) + \widetilde{\nu}_\gamma(1) + \cdots + \widetilde{\tau}_\gamma(k) + \widetilde{\nu}_\gamma(k)\\ &\le \widetilde{\tau}_\gamma(1) + \widetilde{\nu}_\gamma(1) + \cdots + \widetilde{\tau}_\gamma(K_\gamma) + \widetilde{\nu}_\gamma(K_\gamma).\end{aligned}$$

Now we are ready to complete the proof. By (2.34), there exist k_0 and $\widetilde{\gamma}$ such that $\mathbb{P}\{K_\gamma > k_0\} \le \varepsilon$ for all $\gamma \le \widetilde{\gamma}$. Put $\delta_0 = \delta/(2k_0)$ and $\gamma = \min\{\gamma^*, \widetilde{\gamma}, \delta_0(\lambda - e^{-1})/8\}$. Then

$$\begin{aligned}\mathbb{P}\{\psi_\gamma(\gamma r) > \delta r\} &\le \mathbb{P}\Big\{\textstyle\sum_{i=1}^{\infty}(\widetilde{\tau}_\gamma(i) + \widetilde{\nu}_\gamma(i))\mathbb{I}\{K_\gamma \ge i\} > \delta r\Big\}\\ &\le \mathbb{P}\Big\{\textstyle\sum_{i=1}^{k_0}(\widetilde{\tau}_\gamma(i) + \widetilde{\nu}_\gamma(i))\mathbb{I}\{K_\gamma \ge i\} > \delta r\Big\} + \varepsilon \qquad (2.35)\\ &\le \textstyle\sum_{i=1}^{k_0}(t_i^r + s_i^r) + \varepsilon,\end{aligned}$$

where

$$\begin{aligned}t_i^r &:= \mathbb{P}\{\widetilde{\tau}_\gamma(i)\mathbb{I}\{K_\gamma \ge i\} > \delta_0 R\},\\ s_i^r &:= \mathbb{P}\{\widetilde{\nu}_\gamma(i)\mathbb{I}\{K_\gamma \ge i\} > \delta_0 r\}.\end{aligned}$$

Since $\{K_\gamma \ge i\} \subseteq \{\nu(i-1) < \infty\}$ and $\tau(1)$ is a.s. finite,

$$\begin{aligned}t_i^r &\le \mathbb{P}\{\widetilde{\tau}_\gamma(i)\mathbb{I}\{\nu(i-1) < \infty\} > \delta_0 r\}\\ &= \mathbb{P}\{\widetilde{\tau}_\gamma(i) > \delta_0 r | \nu(i-1) < \infty\}\mathbb{P}\{\nu(i-1) < \infty\} \qquad (2.36)\\ &\le \mathbb{P}\{\tau(1) > \delta_0 R\} \to 0 \quad \text{as } r \to \infty.\end{aligned}$$

For $i = 1, \dots, K_\gamma$, we have $\{\sum_{j=1}^{\lfloor \delta_0 r\rfloor} y_\gamma(\tau_\gamma(i) + j) \ge \gamma r\} \subseteq \{\widetilde{\nu}_\gamma(i) \le \delta_0 r\}$. Then

$$\begin{aligned}s_i^r &= \mathbb{P}\{\widetilde{\nu}_\gamma(i) > \delta_0 r | K_\gamma \ge i\}\mathbb{P}\{K_\gamma \ge i\}\\ &\le \mathbb{P}\{\textstyle\sum_{j=1}^{\lfloor \delta_0 r\rfloor} y_\gamma(\tau_\gamma(i) + j) < \gamma r | K_\gamma \ge i\} \qquad (2.37)\\ &= \mathbb{P}\{Y_\gamma(\lfloor \delta_0 r\rfloor) < \gamma r\} \to 0 \quad \text{as } r \to \infty\end{aligned}$$

because a.s. $Y_\gamma(\lfloor \delta_0 r\rfloor)/r \to \delta_0 \mathbb{E} y_\gamma(1) \ge \delta_0(\lambda - e^{-1})/4 > \delta_0\gamma$.

Finally, (2.35)–(2.37) imply (2.32).

(i) **implies** (ii)**, single class case** By (i), we can choose a $\gamma > 0$ such that, for large r, the process $Q^r(\cdot)$ reaches level γr in time $\varphi^r(\gamma r) \le \delta r$ with high probability. Now we prove that, within the time horizon $[\varphi^r(\gamma r), \Delta r]$, there exists a minorant for $Q^r(\cdot)$ that, for large r, stays close to level γr with high probability. Then $Q^r(\cdot)$ stays higher than, for example, level $\gamma r/2$.

We now proceed more formally. Fix δ, Δ and ε. Take h^* and N the same as in (2.27). Take γ and r.v. θ_γ that satisfy (i) of Lemma 2.3, (ii) (with $a = (\lambda - e^{-1})/4$) of Statement 2.1 and (iv) of Statement 2.1. Let $\{\theta(n)\}_{n\in\mathbb{N}}$ be an i.i.d. sequence with $\theta(1) \stackrel{\mathrm{d}}{=} \theta_\gamma$, and assume that this sequence does not depend on $\{A(n)\}_{n\in\mathbb{N}}$ and $\{U(n)\}_{n\in\mathbb{Z}_+}$. Let $\{v^r(n)\}_{n\in\mathbb{N}}$ and $\{y(n)\}_{n\in\mathbb{N}}$ be i.i.d. sequences with $v^r(n) = A(n) - \mathbb{I}\{U(n-1) \le h^*\} - \sum_{i=1}^{\lfloor\gamma r\rfloor} \xi^r(n-1,i)$ and $y(n) = A(n) - \mathbb{I}\{U(n-1) \le h^*\} - \theta(n)$. Define the auxiliary processes

$$V^r(n) = \begin{cases} Q^r(n), & n < \varphi^r(\gamma r), \\ \lfloor\gamma r\rfloor, & n = \varphi^r(\gamma r), \\ \min\{\lfloor\gamma r\rfloor, V^r(n-1)+v^r(n)\}, & n > \varphi^r(\gamma r), \end{cases}$$

and

$$\begin{aligned} \widetilde{V}^r(0) &= 0, & \widetilde{V}^r(n) &= -(\widetilde{V}^r(n-1) + v^r(n))^-, \quad n \in \mathbb{N}, \\ Y(0) &= 0, & Y(n) &= -(Y(n-1) + y(n))^-, \quad n \in \mathbb{N}. \end{aligned}$$

The processes $Q^r(\cdot)$ and $V^r(\cdot)$ coincide within the time interval $[0, \varphi^r(\gamma r) - 1]$. Starting from time $\varphi^r(\gamma r)$, as long as $V^r(\cdot)$ stays above level N, it stays a minorant for $Q^r(\cdot)$. Indeed, for r large enough, given $N \le V^r(i) \le Q^r(i)$, $i = \varphi^r(\gamma r), \dots, n$, if $Q^r(n) \ge \lfloor\gamma r\rfloor$, then, by (2.26),

$$\begin{aligned} Q^r(n+1) &\stackrel{\text{a.s.}}{=} \lfloor\gamma r\rfloor + A(n+1) - D^{\mathrm{t}}(n, \lfloor\gamma r\rfloor) - D^{r,\mathrm{a}}(n, \lfloor\gamma r\rfloor) \\ &= \lfloor\gamma r\rfloor + v^r(n+1) \ge V^r(n+1), \end{aligned}$$

and, if $Q^r(n) < \lfloor\gamma r\rfloor$, then

$$\begin{aligned} Q^r(n+1) &\ge Q^r(n) + A(n+1) - \mathbb{I}\{U(n) \le h^*\} - \textstyle\sum_{i=1}^{\lfloor\gamma r\rfloor} \xi^r(n,i) \\ &\ge V^r(n) + v^r(n+1) \ge V^r(n+1). \end{aligned}$$

Further, by independence arguments, for r large enough, we have $y(1) \le_{\mathrm{st}} v^r(1)$, and $\min_{0\le i\le n} Y(i) \le_{\mathrm{st}} \min_{0\le i\le n} \widetilde{V}^r(i)$. Since $\mathbb{E}y(1) > 0$, $\min_{0\le i\le n} Y(i)/n \to 0$ a.s. Hence

$$\min_{0\le i\le \lfloor\Delta r\rfloor} \widetilde{V}^r(i)/r \Rightarrow 0 \quad \text{as } r \to \infty. \tag{2.38}$$

Now we are ready to complete the proof. Put $C = \gamma/2$ and define the events

$$\begin{aligned} E^r &= \{\min_{\lfloor\delta r\rfloor \le n \le \lfloor\Delta r\rfloor} Q^r(n) < rC\}, \\ A^r &= \{\varphi^r(\gamma r) \le \delta r\}, \\ B^r &= \{\min_{0\le i\le \lfloor\Delta r\rfloor} V^r(\varphi^r(\gamma r) + i) \ge 3\gamma r/4\}. \end{aligned}$$

Then $\mathbb{P}\{E^r\} \le \mathbb{P}\{E^r \cap A^r \cap B^r\} + \mathbb{P}\{\overline{A^r}\} + \mathbb{P}\{\overline{B^r}\}$, where

- $E^r \cap A^r \cap B^r \subseteq \{3\gamma r/4 \le \min_{\lfloor \delta R \rfloor \le n \le \lfloor \Delta r \rfloor} Q^r(n) < \gamma r/2\}$ with the RHS event being empty,
- $\overline{\lim}_{r\to\infty} \mathbb{P}\{\overline{A^r}\} \le \varepsilon$,
- by $\{V^r(\varphi^r(\gamma r)+n)\}_{n\in\mathbb{Z}_+} \stackrel{\mathrm{d}}{=} \{\widetilde{V}^r(n)+\lfloor \gamma r \rfloor\}_{n\in\mathbb{Z}_+}$ and (2.38), as $r\to\infty$,

$$\mathbb{P}\{\overline{B^r}\} = \mathbb{P}\{\min_{0\le n\le \lfloor \Delta r\rfloor} \widetilde{V}(n) < 3\gamma r/4 - \lfloor \gamma r \rfloor\} \to 0.$$

Hence, (ii) of Lemma (2.3) holds.

Single class bound for a multiclass class model Now we show that a model with multiple classes of customers can be coupled with a suitable single-class model in such a way that the population process of the single class model is majorised by the total population of the multiclass model within the whole time horizon $\mathbb{R}_+$. This, in particular, implies that statements (i) and (ii) of the lemma, proven in the single class case, are valid in the multiclass case, too.

For the multiclass model, we slightly modify the representation of the population process suggested in Section 2.7.1. We only change the terms that represent impatient abandonments. For $\mathbf{x} \in \mathbb{Z}_+^I$, let

$$D_i^{r,\mathrm{a}}(n,\mathbf{x}) = \sum_{j=x_1-D_1^{\mathrm{t}}(n,\mathbf{x})+\ldots+x_i-D_i^{\mathrm{t}}(n,\mathbf{x})}^{x_1-D_1^{\mathrm{t}}(n,\mathbf{x})+\ldots+x_{i-1}-D_{i-1}^{\mathrm{t}}(n,\mathbf{x})} \mathbb{I}\{U(n,j) \le p_i/r\},$$

where $\{U(n,i)\}_{i\in\mathbb{N}}$, $n \in \mathbb{Z}_+$, are mutually independent i.i.d. sequences of r.v.'s distributed uniformly over the interval $[0,1]$. We also assume that these sequences do not depend on the random elements $\mathbf{Q}^r(0)$, $\{\mathbf{A}(n)\}_{n\in\mathbb{N}}$ and $\{U(n)\}_{n\in\mathbb{Z}_+}$.

Consider a single-class model with

- initial condition $\widetilde{Q}^r(0) = \|\mathbf{Q}^r(0)\|_1$,
- arrival process $\widetilde{A}(n) = \|\mathbf{A}(n)\|_1$,
- reneging probability $\widetilde{p} = \max_{1\le i\le I} p_i$,

and define its dynamics as follows:

$$\widetilde{Q}^r(n+1) = \widetilde{Q}^r(n) + \widetilde{A}(n+1) - \widetilde{D}^{\mathrm{t}}\left(n, \widetilde{Q}^r(n)\right) - \widetilde{D}^{r,\mathrm{a}}\left(n, \widetilde{Q}^r(n)\right), \tag{2.39}$$

where, for $k \in \mathbb{Z}_+$,

$$\widetilde{D}^{\mathrm{t}}(n,k) = \mathbb{I}\left\{U(n) \le h(k)\right\},$$
$$\widetilde{D}^{r,\mathrm{a}}(n,k) = \sum_{i=1}^{k-\widetilde{D}^{\mathrm{t}}(n,k)} \mathbb{I}\left\{U(n,i) \le \widetilde{p}/r\right\},$$

and the r.v.'s $U(n)$, $U(n,i)$, $n \in \mathbb{Z}_+$, $i \in \mathbb{N}$, are those defining the multiclass model.

Then, in particular,

$$\begin{aligned} &\left\|\mathbf{D}^{\mathrm{t}}(n,\mathbf{x})\right\|_1 \overset{\text{a.s.}}{=} \mathbb{I}\{U(n) \le \textstyle\sum_{j=1}^{I} p_j(\mathbf{x})\} = \widetilde{D}^{\mathrm{t}}(n, \|\mathbf{x}\|_1), \\ &\mathbb{I}\{U(n,j) \le p_i/r\} \le \mathbb{I}\{U(n,j) \le \widetilde{p}/r\}. \end{aligned} \tag{2.40}$$

We show by induction that $\widetilde{Q}^r(\cdot)$ bounds $\|\mathbf{Q}^r(\cdot)\|_1$ from below. Let $N^r(n) = \|\mathbf{Q}^r(n)\|_1 - \widetilde{D}^{\mathrm{t}}\,(n, \|\mathbf{Q}^r(n)\|_1)$, and suppose that $\widetilde{Q}^r(n) \le \|\mathbf{Q}^r(n)\|_1$ a.s., then

$$\begin{aligned} \|\mathbf{Q}^r(n+1)\|_1 &\overset{\text{a.s.}}{\ge} \|\mathbf{Q}^r(n)\|_1 + A(n+1) - \widetilde{D}^{\mathrm{t}}\,(n, \|\mathbf{Q}^r(n)\|_1) \\ &\quad - \textstyle\sum_{i=1}^{N^r(n)} \mathbb{I}\{U(n,i) \le \widetilde{p}/R\} \\ &= \|\mathbf{Q}^r(n)\|_1 + \widetilde{A}(n+1) - \widetilde{D}^{\mathrm{t}}\,(n, \|\mathbf{Q}^r(n)\|_1) \\ &\quad - \widetilde{D}^{r,\mathrm{a}}\,(n, \|\mathbf{Q}^r(n)\|_1) \\ &\overset{\text{a.s.}}{\ge} \widetilde{Q}^r(n) + \widetilde{A}(n+1) - \widetilde{D}^{\mathrm{t}}(n, \widetilde{Q}^r(n)) - \widetilde{D}^{r,\mathrm{a}}(n, \widetilde{Q}^r(n)) \\ &= Q^r(n+1), \end{aligned}$$

where the first and last inequalities hold by (2.40) and (2.26) respectively, and the identity is due to representation (2.39).

(ii) **implies** (iii) By Proposition 1.2, it is enough to show that, for any $\Delta > 0$ and all i,

$$\sup_{0\le s\le\Delta} \|G_i^{r,2}(s)\| \Rightarrow 0 \quad \text{as } r \to \infty. \tag{2.41}$$

Fix Δ and i. Recall that

$$G_i^{r,2}(t) = e^{-1}\int_0^t m_i^r(s)\,ds - \Big(1 - \frac{p_i}{r}\Big)\int_0^{\lfloor rt\rfloor/r} h^r(s) m_i^r(s)\,ds, \tag{2.42}$$

where $m_i^r(s) = m_i(r\overline{\mathbf{Q}}^r(s))$ and $h^r(s) = h(r\|\overline{\mathbf{Q}}^r(s)\|_1)$. First, we estimate the subtractor in (2.42). Since $r\|\overline{\mathbf{Q}}^r(\cdot)\|_1$ is integer-valued and non-negative, $h^r(\cdot) \le 1$. Also $m_i^r(\cdot) \le 1$. Then, for $t \in [0,\Delta]$, we have

$$\begin{aligned} &\left| \int_0^t h^r(s) m_i^r(s)\,ds - \Big(1 - \frac{p_i}{r}\Big)\int_0^{\lfloor rt\rfloor/r} h^r(s) m_i^r(s)\,ds \right| \\ &= \int_{\lfloor rt\rfloor/r}^t h^r(s) m_i^r(s)\,ds + \frac{p_i}{r}\int_0^{\lfloor rt\rfloor/r} h^r(s) m_i^r(s)\,ds \le \frac{1+p_i\Delta}{r}. \end{aligned}$$

Take $\delta < \Delta$, then

$$\begin{aligned} \sup_{0\le s\le\Delta} \|G_i^{r,2}(s)\| &\le \frac{1+p_i\Delta}{r} + \sup_{0\le t\le\Delta}\left| \int_0^t m_i^r(s)(e^{-1} - h^r(s))\,ds \right| \\ &\le \frac{1+p_i\Delta}{r} + (e^{-1}+1)\delta + \Delta \sup_{\delta\le s\le\Delta} |e^{-1} - h^r(s)|. \end{aligned} \tag{2.43}$$

Now we show that

$$x^r(\delta,\Delta) := \sup_{\delta\le s\le\Delta} \|e^{-1} - h^r(s)\| \Rightarrow 0 \quad \text{as } r\to\infty. \tag{2.44}$$

For any $\sigma>0$, $\varepsilon>0$ and $C(\delta,\Delta,\varepsilon)$ satisfying (2.24),

$$\begin{aligned}\{x^r(\delta,\Delta)\ge\sigma\} &\subseteq \{x^r(\delta,\Delta)\ge\sigma, \inf_{\delta\le s\le\Delta}\|\overline{\mathbf{Q}}^r(s)\|_1 \ge C(\delta,\Delta,\varepsilon)\}\\ &\quad\cup\{\inf_{\delta\le s\le\Delta}\|\overline{\mathbf{Q}}^r(s)\|_1 < C(\delta,\Delta,\varepsilon)\}\\ &\subseteq \{\sup_{s\ge R\,C(\delta,\Delta,\varepsilon)}|e^{-1}-h(s)|\ge\sigma\}\\ &\quad\cup\{\inf_{\delta\le s\le\Delta}\|\overline{\mathbf{Q}}^r(s)\|_1 < C(\delta,\Delta,\varepsilon)\}.\end{aligned}$$

Here the first event in the very RHS is empty for r large enough, and hence

$$\overline{\lim}_{r\to\infty}\mathbb{P}\{x^r(\delta,\Delta)\ge\sigma\} \le \overline{\lim}_{r\to\infty}\mathbb{P}\{\inf_{\delta\le s\le\Delta}\|\overline{\mathbf{Q}}^r(s)\|_1 < C(\delta,\Delta,\varepsilon)\} \le \varepsilon.$$

Since in the last inequality $\varepsilon>0$ is arbitrary, we have $\mathbb{P}\{x^r(\delta,\Delta)\ge\sigma\}\to 0$ as $r\to\infty$ for any $\sigma>0$, which gives (2.44).

Finally, (2.43) and (2.44) imply (2.41).

2.7.4 Proof of Lemma 2.4

We prove the result in the single-class case. The same proof is valid for each coordinate in the multiclass case.

Convergence of $G^{r,3}(\cdot)$ By Proposition 1.2, it suffices to show that, for any $T>0$, as $r\to\infty$,

$$\mu^r(T) := \sup_{0\le t\le T}\int_{\lfloor rt\rfloor/r}^{t}\overline{Q}^r(s)\,ds \Rightarrow 0. \tag{2.45}$$

Since $\overline{Q}^r(\cdot)$ is a constant function within the interval $[\lfloor rt\rfloor/r, t]$, we have

$$\begin{aligned}\mu^r(T) &= \sup_{0\le t\le T}\overline{Q}^r(t)(rt-\lfloor rt\rfloor)/r \le \sup_{0\le t\le T}\overline{Q}^r(t)/r\\ &\le Q^r(0)/r^2 + \textstyle\sum_{i=1}^{\lfloor Tr\rfloor}A(i)/r^2,\end{aligned}$$

which implies (2.45).

Convergence of $\overline{M}^r(\cdot)$ We represent the martingale $\{M^r(n)\}_{n\in\mathbb{Z}_+}$ as a sum of three other zero-mean martingales,

$$M^r(n) = M^{r,1}(n) - M^{r,2}(n) - M^{r,3}(n),$$

$$
\begin{aligned}
M^{r,1}(n) &= \textstyle\sum_{i=1}^{n} A(i) - n\lambda, \\
M^{r,2}(n) &= \textstyle\sum_{i=0}^{n-1} \sum_{m=1}^{Q^r(i)} \Big(\xi^r(i,m) - \dfrac{p}{r} \Big), \\
M^{r,3}(n) &= \textstyle\sum_{i=0}^{n-1} \Big(D^{\mathrm{t}}\left(i, Q^r(i)\right) - h\left(Q^r(i)\right) \Big) \\
&\quad + \textstyle\sum_{i=0}^{n-1} \sum_{m=Q^r(i)-D^{\mathrm{t}}(i,Q^r(i))+1}^{Q^r(i)} \Big(\xi^r(i,m) - \dfrac{p}{r} h\left(Q^r(i)\right) \Big).
\end{aligned}
$$

It suffices to show that, for any $T > 0$,

$$
\max_{1 \le n \le \lfloor rT \rfloor} |M^{r,j}(n)|/r \Rightarrow 0 \quad \text{as } r \to \infty, \quad j = 1, 2, 3, \tag{2.46}
$$

so we fix a $T > 0$ for the rest of the proof.

For $j = 1$, (2.46) follows from the FLLN.

For all r and n, we have

$$
|M^{r,2}(n+1) - M^{r,2}(n)| \le 4. \tag{2.47}
$$

Then

$$
\overline{M}^{r,2}(T) := |M^{r,2}(\lfloor rT \rfloor)|/r \Rightarrow 0 \quad \text{as } r \to \infty \tag{2.48}
$$

by the following result (see Andrews [4]).

Proposition 2.3. *Suppose that, for any n, $\{X^n(l)\}_{l \in \mathbb{N}}$ is a martingale difference and that the family $\{X^n(l)\}_{l,n \in \mathbb{N}}$ is uniformly integrable. Then*

$$
\frac{1}{n} \textstyle\sum_{l=1}^{n} X^n(l) \Rightarrow 0 \quad \textit{as } n \to \infty.
$$

By (2.47), (2.48) and Doob's martingale inequality, for any $\varepsilon, \sigma > 0$,

$$
\begin{aligned}
&\mathbb{P}\{\max_{1 \le n \le \lfloor rT \rfloor} |M^{r,2}(n)|/r > \varepsilon\} \le \varepsilon^{-1} \mathbb{E} \overline{M}^{r,2}(T) \\
&= \varepsilon^{-1} \mathbb{E}\left[\overline{M}^{r,2}(T) \mathbb{I}\{\overline{M}^{r,2}(T) > \sigma\} \right] + \varepsilon^{-1} \mathbb{E}\left[\overline{M}^{r,2}(T) \mathbb{I}\{\overline{M}^{r,2}(T) \le \sigma\} \right] \\
&\le \varepsilon^{-1} 4T\, \mathbb{P}\{\overline{M}^{r,2}(T) > \sigma\} + \varepsilon^{-1}\sigma.
\end{aligned}
$$

As we take $r \to \infty$ and then $\sigma \to 0$ in the last inequality, (2.46) with $j = 2$ follows.

Now we prove (2.46) for $j = 3$. The key tool of this proof is Markov's inequality. We have to show that, for any $\varepsilon > 0$, as $r \to \infty$,

$$
\mathbb{P}\{\max_{1 \le n \le \lfloor rT \rfloor} M^{r,3}(n)/r > \varepsilon\} \to 0, \tag{2.49a}
$$

$$
\mathbb{P}\{\min_{1 \le n \le \lfloor rT \rfloor} M^{r,3}(n)/r < -\varepsilon\} \to 0. \tag{2.49b}
$$

By Taylor's expansion, there exists an $\alpha^* > 0$ such that, for any $\alpha \in [0, \alpha^*]$, $\rho \in [0,1]$ and

r.v. $\xi(\rho)$ with $\mathbb{P}\{\xi(\rho)=1\}=\rho=1-\mathbb{P}\{\xi(\rho)=0\}$, we have

$$\mathbb{E}e^{\alpha(\xi(\rho)-\rho)}\le e^{\alpha^2\rho},\quad \mathbb{E}e^{\alpha(\rho-\xi(\rho))}\le e^{\alpha^2\rho}. \tag{2.50}$$

Since $Q^r(n)\le Q^r(0)+\sum_{i=1}^{\lfloor rT\rfloor}A(i), 0\le n\le\lfloor rT\rfloor$, and $Q^r(0)/r\Rightarrow z^0$, and $\sum_{i=1}^{\lfloor rT\rfloor}A(i)/r\to\lambda T$ a.s., for any $\delta>0$, there exists an $M(\delta)>0$ such that

$$\overline{\lim}_{r\to\infty}\mathbb{P}\{\max_{0\le n\le\lfloor rT\rfloor}Q^r(n)>M(\delta)r\}\le\delta. \tag{2.51}$$

Denote the event in (2.51) by $E^r(\delta)$ and put

$$\alpha(\delta)=\min\{\alpha^*,\varepsilon/(2M(\delta))\}. \tag{2.52}$$

We introduce the auxiliary martingale

$$\widetilde{M}^{r,\delta}(n)=\sum_{i=0}^{n-1}\sum_{m=1}^{\widetilde{Q}^{r,\delta}(i)}(\xi^r(i,m)-p/r),\quad n\in\mathbb{N},$$

where

$$\widetilde{Q}^{r,\delta}(i)=\max\{Q^r(i),M(\delta)r\}.$$

Note that, on $\overline{E^r(\delta)}$, we have $\widetilde{M}^{r,\delta}(n)=M^{r,3}(n), 1\le n\le\lfloor rT\rfloor$. Hence

$$\begin{aligned}&\mathbb{P}\{\max_{1\le n\le\lfloor rT\rfloor}M^{r,3}(n)/r>\varepsilon\}\\&\le\mathbb{P}\{\max_{1\le n\le\lfloor rT\rfloor}\widetilde{M}^{r,\delta}(n)>\varepsilon r\}+\mathbb{P}\{E^r(\delta)\}\\&\le\sum_{n=1}^{\lfloor rT\rfloor}\mathbb{P}\{\widetilde{M}^{r,\delta}(n)>\varepsilon r\}+\mathbb{P}\{E^N(\delta)\}.\end{aligned} \tag{2.53}$$

By Markov's inequality, (2.52) and (2.50), we have, for $1\le n\le\lfloor rT\rfloor$,

$$\begin{aligned}&\exp(\alpha(\delta)\varepsilon r)\mathbb{P}\{\widetilde{M}^{r,\delta}(n)>\varepsilon r\}\le\mathbb{E}\exp(\alpha(\delta)\widetilde{M}^{r,\delta}(n))\\&=\mathbb{E}\left[\mathbb{E}\left[\prod_{i=0}^{n-1}\prod_{m=1}^{\widetilde{Q}^{r,\delta}(i)}\exp\Big(\alpha(\delta)(\xi^r(i,m)-p/r)\Big)\Big|Q^r(0),\dots,Q^r(n-1)\right]\right]\\&\le\mathbb{E}\left[\exp\Big(\alpha^2(\delta)(p/r)M(\delta)rn\Big)\right]\le\exp\big(\alpha^2(\delta)M(\delta)r\big).\end{aligned} \tag{2.54}$$

By (2.51) and (2.52), bounds (2.53) and (2.54) imply that

$$\overline{\lim}_{r\to\infty}\mathbb{P}\{\max_{1\le n\le\lfloor rT\rfloor}M^{r,3}(n)/r>\varepsilon\}\le\delta,$$

where $\delta>0$ is arbitrary. Hence, convergence (2.49a) holds. Convergence (2.49b) can be treated similarly.

2.8 Proofs of auxiliary results

Proof of Statement 2.1. There exists an $r^* > 0$ such that, for all $n \in \mathbb{Z}_+$, $\gamma \in (0,1]$ and $r \geq r^*$,

$$e^n \mathbb{P}\{B^r_\gamma \geq n\} \leq \mathbb{E} \exp(B^r_\gamma) = (ep/r + 1 - p/r)^{\lfloor \gamma r \rfloor} \leq \exp((e-1)p) =: \mu.$$

Take N^* such that $\sum_{n>N^*} \mu n e^{-n} \leq a/2$, and $\gamma^* = \min\{1, a/(4p)\}$. Fix $\gamma \in (0, \gamma^*]$. Since the binomial distributions $B(\lfloor \gamma r \rfloor, p/r)$ converge weakly to the Poisson distribution $\mathrm{Pois}(\gamma p)$ as $r \to \infty$, there exists an $r_\gamma \geq r^*$ such that

$$\mathbb{P}\{B^r_\gamma = n\} \leq 2\mathrm{Pois}(\gamma r)(\{n\}), \quad r \geq r_\gamma,\ n = 0, \dots N^*.$$

For $\gamma \in (0, \gamma^*]$, put

$$\mathbb{P}\{\theta_\gamma \geq n\} = \mu e^{-n}, \quad n > N^*, \tag{2.55}$$

and

$$\mathbb{P}\{\theta_\gamma = n\} = \min\left\{2\mathrm{Pois}(\gamma r)(\{n\}), 1 - \mathbb{P}\{\theta_\gamma \geq n+1\}\right\}, \quad n = N^*, \dots, 0.$$

For $\gamma = 0$, put (2.55) and

$$\begin{aligned} \mathbb{P}\{\theta_0 \geq 0\} &= 1 - \mu e^{-N^*-1}, \\ \mathbb{P}\{\theta_0 = n\} &= 0, \quad n = 1, \dots, N_*. \end{aligned}$$

□

Proof of Statement 2.2. Define

$$\begin{aligned} P(x, \geq z) &= \mathbb{P}\left\{X(n+1) \geq z \middle| X(n) = x\right\}, \\ Q(y, \geq z) &= \mathbb{P}\left\{Y(n+1) \geq z \middle| Y(n) = y\right\}. \end{aligned}$$

Let $\{U(n); n \geq 0\}$ be an i.i.d. sequence with $U(0)$ distributed uniformly over $[0,1]$. Then put

$$\begin{aligned} \widetilde{X}(0) &= X(0), \\ \widetilde{X}(n+1) &= \sup\{z \in S\colon U(n) \geq 1 - P(\widetilde{X}(n), \geq z)\}, \\ \widetilde{Y}(0) &= Y(0), \\ \widetilde{Y}(n+1) &= \sup\{z \in S\colon U(n) \geq 1 - P(\widetilde{Y}(n), \geq z)\}. \end{aligned}$$

□

Chapter 3

Bandwidth-Sharing Networks with Rate Constraints

3.1 Introduction

Bandwidth-sharing policies as introduced by Massoulié and Roberts [97, 75] dynamically distribute network resources among a changing population of users. Processor sharing is an example of such a policy and assumes a single resource. Bandwidth-sharing networks are of great research and practical interest. Along with the basic application in telecommunications, e.g. Internet congestion control, they also have recently been suggested as a tool in analyzing problems in road traffic, see Kelly and Williams [58].

The main issues in bandwidth-sharing related research are stability conditions and performance evaluation. A variety of results regarding the first topic may be found in de Veciana et al. [34, 35], Bonald and Massoulié [10], Mo and Walrand [80], Massoulié [74], Bramson [22], Gromoll and Williams [45], and Chiang et al. [27]. As for the second topic, for special combinations of network topologies and bandwidth-sharing policies, the network stationary distribution may be shown to have a product form and be insensitive to the flow size distribution, see Bonald et al. [11]. In general however, approximation methods must be used, which is the subject matter of the present chapter. Fundamental papers on fluid limit approximations for bandwidth sharing-networks are Kelly and Williams [57] and Gromoll and Williams [46]. Some more results on fluid and diffusion approximations are to be found in Egorova et al. [38], Borst et al. [15], Kang et al. [54] and Ye and Yao [118, 119]. The latter works ignore the fact that generally in practice the maximum service rate of an individual user is constrained, as has been pointed out by Roberts [96].

To the best of our knowledge, Ayesta and Mandjes [6] were the first to deal with fluid and diffusion approximations of bandwidth-sharing networks with rate limitations. They consider two specific settings, first without rate constraints, and then they truncate the capacity constraints at the rate maxima. Reed and Zwart [88] develop a different approach in the context of general bandwidth-sharing networks. They incorporate the rate

constraints into the network utility maximization procedure that defines bandwidth allocations. Thus, users operating below the maximal rate are allowed to take up the bandwidth that is not used by other rate constrained users, and bandwidth allocations are Pareto optimal. An interesting feature of both [6] and [88] is the scaling regime. In contrast to the other papers on bandwidth-sharing mentioned above, which mostly focus on the large-time properties of networks with fixed-order parameters, [6] and [88] view networks on a fixed-time scale letting arrival rates and capacities grow large. This large capacity scaling reflects the fact that overall network capacity and individual user rate constraints may be of different orders of magnitude. For example, it is common that Internet providers set download speed limitations for individual users which are typically measured in megabits per second, while network capacities are measured in gigabits or terabits per second.

The framework of [88] is rather comprehensive. In particular, it allows abandonments of flows: each flow knows how long it can stay in the system and abandons as soon as its service is finished or its patience time expires, whichever happens earlier. This chapter builds upon [88] by relaxing its stochastic assumptions: we assume a general distribution for interarrival times and a general joint distribution for the size and patience time of a flow (in particular, the flow size and patience time are allowed to be dependent), while [88] assumes a Markovian setting with independent arrivals, flow sizes and patience times. We study the behavior of bandwidth-sharing networks in terms of measure-valued processes that are called state descriptors and that keep track of residual flow sizes and residual patience times. The first main result of the chapter is a fluid limit theorem (it generalizes the fluid limit result of [88] to non-Markovian stochastic assumptions). We show that the scaled state descriptors are tight with all weak limit points a.s. solving a system of deterministic integral equations. We provide a sufficient condition for these deterministic equations to have a unique solution. In terms of proof techniques, these results are closely related to previous work on bandwidth-sharing [46], processor-sharing with impatience [47], and bandwidth-sharing in overload conditions [15, 38]. The rate constraints play a crucial role in adopting the techniques of the cited papers. For example, [47] requires an additional assumption of overload to eliminate problems at zero. In our case however, due to the rate constraints, the network never empties, and the load conditions become irrelevant.

Our second main result, which is a new type of result for bandwidth-sharing networks, is convergence of the scaled network stationary distribution to the fixed point of the deterministic limit equations, provided the fixed point is unique. There is a similar result by Kang and Ramanan [53] for a call center model, but the techniques of [53] are different than ours. Applying the approach of Borst et al. [15], we prove that in many cases the fixed point can be found by solving an optimization problem with a strictly concave objective function and a polyhedral constraint set, and thus is unique and computable in polynomial time. We also construct an example with multiple fixed points, which is a feature that is distinctive from earlier cited works. For this group of results, we suggest novel proof ideas which we believe can be extended to models other than bandwidth-sharing. In particular, we derive equations for the lower and upper asymptotic bounds for fluid limits (see Theorem 3.4). For a wide class of networks, these equations can be solved, and then asymptotic stability of the fixed point can be shown. Another interesting idea is that, in the stationary regime, the properties of a

network depend on newly arriving flows only, since all initial flows are gone after some point (see Lemma 3.2). To guarantee existence of a unique stationary distribution, we assume Poisson arrivals. Poisson arrivals also imply $M/G/\infty$ bounds that we exploit heavily in the proofs.

The structure of the chapter is as follows. Section 3.2 describes the stochastic bandwidth-sharing model. Section 3.3 introduces deterministic integral equations mimicking the stochastic network, we call them the *fluid model*. Also Section 3.3 states sufficient conditions for a solution to the fluid model to be unique, and for a fixed solution to the fluid model to be unique and asymptotically stable. Sections 3.4 and 3.5 discuss convergence of the scaled state descriptor and its stationary distribution to the fluid model and its fixed point, respectively. Sections 3.6, 3.7 and 3.8 contain the proofs of the statements from Sections 3.3, 3.4 and 3.5. Section 3.9 proves auxiliary results. In the remainder of this section, we list the notations that are specific for the current chapter.

Notation First we describe the space of measures where the state descriptor of a bandwidth-sharing network takes values. For a measure ξ on $\mathbb{R}^2_+$ and a ξ-integrable functions $f\colon \mathbb{R}^2_+ \to \mathbb{R}$, define

$$\langle f, \xi \rangle := \int_{\mathbb{R}^2_+} f d\xi.$$

If $\boldsymbol{\xi} = (\xi_1, \dots, \xi_I)$ is a vector of such measures, put

$$\langle f, \boldsymbol{\xi} \rangle := (\langle f, \xi_1 \rangle, \dots, \langle f, \xi_I \rangle).$$

Let $\mathcal{M}$ be the space of finite non-negative Borel measures on $\mathbb{R}^2_+$ endowed with the weak topology: $\xi^k \overset{\text{w}}{\to} \xi$ in $\mathcal{M}$ as $k \to \infty$ if and only if $\langle f, \xi^k \rangle \to \langle f, \xi \rangle$ for all continuous bounded function $f\colon \mathbb{R}^2_+ \to \mathbb{R}$. Weak convergence of elements of $\mathcal{M}$ is equivalent to convergence in the Prokhorov metric given by: for $\xi, \varphi \in \mathcal{M}$,

$$d_{\mathcal{M}}(\xi, \varphi) := \inf\{\varepsilon\colon \xi(B) \le \varphi(B^\varepsilon) + \varepsilon,\ \varphi(B) \le \xi(B^\varepsilon) + \varepsilon \\ \text{for all non-empty closed } B \subseteq \mathbb{R}^2_+\},$$

where $B^\varepsilon := \{x \in \mathbb{R}^2_+ \colon \inf_{y \in B} \|x - y\| < \varepsilon\}$.

For $\boldsymbol{\xi}, \boldsymbol{\varphi} \in \mathcal{M}^I$, define

$$d_{\mathcal{M}^I}(\boldsymbol{\xi}, \boldsymbol{\varphi}) := \max_{1 \le i \le I} d_{\mathcal{M}}(\xi_i, \varphi_i).$$

Equipped with the metric $d_{\mathcal{M}^I}(\cdot, \cdot)$, the space $\mathcal{M}^I$ is separable and complete.

3.2 Stochastic model

This section contains a detailed description of the model under consideration. In particular, it specifies the structure of the network, the policy it operates under and the stochastic dynamical assumptions. Also, a stochastic process is introduced that keeps track of the state of the network, see the state descriptor paragraph.

Network structure Consider a network that consists of a finite number of links labeled by $j = 1, \dots, J$. Traffic offered to the network is represented by elastic flows coming from a finite number of classes labeled by $i = 1, \dots, I$. All class i flows are transferred through a certain subset of links, which is called *route* i. Transfer of a flow starts immediately upon its arrival and is continuous with all links on the route of the flow being traversed simultaneously. Let A be the $J \times I$ incidence matrix, where $A_{j,i} = 1$ if route i contains link j and $A_{j,i} = 0$ otherwise.

Suppose that at a particular time t the population of the network is $\mathbf{z} = (z_1, \dots, z_I) \in \mathbb{Z}_+^I$, where z_i stands for the number of flows on route i. All flows on route i are transferred at the same rate $\lambda_i(\mathbf{z})$ that is at most $m_i \in (0, \infty)$. If $z_i = 0$, put $\lambda_i(\mathbf{z}) := 0$. We refer to $\Lambda_i(\mathbf{z}) := \lambda_i(\mathbf{z}) z_i$ as the *bandwidth allocated to route* i. The sum of the bandwidths allocated to the routes that contain link j is the *bandwidth allocated through link* j and is at most $C_j \in (0, \infty)$. We call C_j the *capacity of link* j. Hence, the vectors $\boldsymbol{\lambda}(\mathbf{z}) = (\lambda_1(\mathbf{z}), \dots, \lambda_I(\mathbf{z}))$ and $\boldsymbol{\Lambda}(\mathbf{z}) = (\Lambda_1(\mathbf{z}), \dots, \Lambda_I(\mathbf{z}))$ must satisfy

$$A(\boldsymbol{\lambda}(\mathbf{z}) \times \mathbf{z}) = A\boldsymbol{\Lambda}(\mathbf{z}) \le \mathbf{C}, \quad \boldsymbol{\lambda}(\mathbf{z}) \le \mathbf{m},$$

where $\mathbf{C} = (C_1, \dots, C_J)$ and $\mathbf{m} = (m_1, \dots, m_I)$ are the vectors of link capacities and rate constraints.

Bandwidth-sharing policy At each point in time, the link capacities should be distributed among the routes in such a way that the network utility is maximized. Namely, to each flow on route i we assign a utility $\mathcal{U}_i(\cdot)$ that is a function of the rate allocated to that flow. Assume that the functions $\mathcal{U}_i(\cdot)$ are strictly increasing and concave in $\mathbb{R}_+$, and twice differentiable in $(0, \infty)$ with $\lim_{x \downarrow 0} \mathcal{U}_i'(x) = \infty$. We also allow $\lim_{x \downarrow 0} \mathcal{U}_i(x) = -\infty$ as, for example, in the case of a logarithmic function. Then, for $\mathbf{z} \in \mathbb{R}_+^I$, the vector $\boldsymbol{\lambda}(\mathbf{z})$ of rates is the unique optimal solution to

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^I z_i \mathcal{U}_i(\lambda_i) \\ \text{subject to} & A(\boldsymbol{\lambda} \times \mathbf{z}) \le \mathbf{C}, \boldsymbol{\lambda} \le \mathbf{m}, \end{array} \tag{3.1}$$

where, by convention, $0 \times (-\infty) := 0$. Although the population vector has integer-valued coordinates, we assume that $\boldsymbol{\lambda}(\mathbf{z})$ and $\boldsymbol{\Lambda}(\mathbf{z}) := \boldsymbol{\lambda}(\mathbf{z}) \times \mathbf{z}$ are defined via (3.1) in the entire orthant $\mathbb{R}_+^I$ to accommodate fluid analogues of the population process later.

The utility maximization procedure (3.1) implies that $\lambda_i(\mathbf{z}) = \Lambda_i(\mathbf{z}) = 0$ if $z_i = 0$. The assumption $\lim_{x \downarrow 0} \mathcal{U}_i'(x) = \infty$ guarantees non-idling, that is $\lambda_i(\mathbf{z}), \Lambda_i(\mathbf{z}) > 0$ if $z_i > 0$. Reed and Zwart [88] proved that the functions $\boldsymbol{\lambda}(\cdot)$ and $\boldsymbol{\Lambda}(\cdot)$ are differentiable in any direction and, in particular, locally Lipschitz continuous in the interior of $\mathbb{R}_+^I$. We also show continuity of $\boldsymbol{\Lambda}(\cdot)$ on the boundary of $\mathbb{R}_+^I$ (see Section 3.9).

Lemma 3.1. *The bandwidth allocation function* $\boldsymbol{\Lambda}(\cdot)$ *is continuous in* $\mathbb{R}_+^I$.

Stochastic assumptions All stochastic primitives introduced in this paragraph are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

Suppose at time zero there is an a.s. finite number of flows in the network, we call them *initial flows*. A random vector $\mathbf{Z}^0 = (Z_1^0, \dots, Z_I^0) \in \mathbb{R}_+^I$ represents the initial population, and Z_i^0 is the number of initial flows on route i. New flows arrive to the network according to a stochastic process $\mathbf{E}(\cdot) = (E_1, \dots, E_I)(\cdot)$ with sample paths in the Skorokhod space $\mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I)$. The coordinates of the arrival process are independent counting processes. Recall that a *counting process* is a non-decreasing non-negative integer-valued process starting from zero. For $t \in \mathbb{R}_+$, $E_i(t)$ represents the number of flows that have arrived to route i during the time interval $(0, t]$. The kth such arrival occurs at time $U_{i,k} = \inf\{t \in \mathbb{R}_+ : E_i(t) \geq k\}$, it is called *flow k on route i*, $k \in \mathbb{N}$. Simultaneous arrivals are allowed.

Flows leave the network due to transfer completions or because they run out of patience, depending on what happens earlier for each particular flow. Flow sizes and patience times are drawn from sequences $\{(B_{i,l}^0, D_{i,l}^0)\}_{l\in\mathbb{N}}$, $\{(B_{i,k}, D_{i,k})\}_{k\in\mathbb{N}}$, $i = 1, \dots, I$, of $(0,\infty)^2$-valued r.v.'s. For $l = 1, \dots, Z_i^0$, B_{il}^0 and D_{il}^0 represent the residual size and residual patience time at time zero of initial flow l on route i. For $k \in \mathbb{N}$, B_{ik} and D_{ik} represent the initial size and initial patience time of flow k on route i, where "initial" means as upon arrival at time $U_{i,k}$. Let $(B_{i,k}, D_{i,k})$, $k \in \mathbb{N}$, be i.i.d. copies of a r.v. (B_i, D_i) with distribution law θ_i; and let the mean values $\mathbb{E}B_i =: 1/\mu_i$ and $\mathbb{E}D_i = 1/\nu_i$ be finite. Assume that the sequences $\{(B_{i,k}, D_{i,k})\}_{k\in\mathbb{N}}$ are independent and do not depend on the arrival process $\mathbf{E}(\cdot)$. For the moment, we do not make any specific assumptions about the sequences $\{(B_{i,l}^0, D_{i,l}^0)\}_{l\in\mathbb{N}}$.

State descriptor We denote the *population process* by $\mathbf{Z}(\cdot) = (Z_1(\cdot), \dots, Z_I(\cdot))$, where $Z_i(t)$ is the number of flows on route i at time t. As can be seen from what follows, $\mathbf{Z}(\cdot)$ is a random element of the Skorokhod space $\mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I)$.

For $i = 1, \dots, I$, introduce operators $S_i \colon \mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I) \to \mathbf{C}(\mathbb{R}_+^2, \mathbb{R}_+)$ defined by

$$S_i(\mathbf{z}, s, t) := \int_s^t \lambda_i(\mathbf{z}(u))\, du,$$

For $t \geq s \geq 0$, $S_i(\mathbf{Z}, s, t)$ is the *cumulative bandwidth allocated per flow on route i during time interval* $[s, t]$. The *residual size* and *residual lead time at time t* of initial flow $l = 1, \dots, Z_i^0$ on route i are given by

$$B_{i,l}^0(t) := (B_{i,l}^0 - S_i(\mathbf{Z}, 0, t))^+,$$
$$D_{i,l}^0(t) := (D_{i,l}^0 - t)^+,$$

and those of flow $k = 1, \dots, E_i(t)$ on route i by

$$B_{i,k}(t) := (B_{i,k} - S_i(\mathbf{Z}, U_{i,k}, t))^+$$
$$D_{i,k}(t) := (D_{i,k} - (t - U_{i,k}))^+.$$

The state of the network at any time t is defined by the residual sizes and residual patience times of the flows present in the network. With each flow, we associate a dot

in $\mathbb{R}^2_+$, whose coordinates are the residual size and residual patience time of the flow (see Figure 3.1). As a flow is getting transferred, the corresponding dot moves toward the axis: to the left at the transfer rate (which is $\lambda_i(\mathbf{Z}(t))$ for a flow on route i) and downward at the constant rate of 1. As a dot hits the vertical axis, the corresponding flow leaves due to completion of its transfer. As a dot hits the horizontal axis, the corresponding flow leaves due to impatience. We combine these moving dots into the stochastic process $\boldsymbol{\mathcal{Z}}(\cdot) = (\mathcal{Z}_1, \dots, \mathcal{Z}_I)(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathcal{M}^I)$ with

$$\mathcal{Z}_i(t) := \sum_{l=1}^{Z_i^0} \delta^+_{(B^0_{i,l}(t), D^0_{i,l}(t))} + \sum_{k=1}^{E_i(t)} \delta^+_{(B_{i,k}(t), D_{i,k}(t))}, \tag{3.2}$$

where, for $(x_1, x_2) \in \mathbb{R}^2_+$, $\delta^+_{(x_1,x_2)} \in \mathcal{M}$ is the Dirac measure at (x_1, x_2) if $x_1 \wedge x_2 > 0$ and zero measure otherwise (i.e. assigns a zero mass to any Borel subset of $\mathbb{R}^2_+$). That is, $\mathcal{Z}_i(t)$ is a counting measure on $\mathbb{R}^2_+$ that assigns a unit mass to each of the dots representing class i flows except those on the axes. The process $\boldsymbol{\mathcal{Z}}(\cdot)$ given by (3.2) is called the *state descriptor*. Note that the total mass of the state descriptor coincides with the network population, $\langle 1, \boldsymbol{\mathcal{Z}}(\cdot)\rangle = \mathbf{Z}(\cdot)$.

When proving the results of the chapter, we decompose the state descriptors into two parts keeping track of initial and newly arriving flows, respectively. That is,

$$\boldsymbol{\mathcal{Z}}(\cdot) = \boldsymbol{\mathcal{Z}}^{\text{init}}(\cdot) + \boldsymbol{\mathcal{Z}}^{\text{new}}(\cdot),$$

where

$$\mathcal{Z}_i^{\text{init}}(t) := \sum_{l=1}^{Z_i(0)} \delta^+_{(B^0_{i,l}(t), D^0_{i,l}(t))},$$
$$\mathcal{Z}_i^{\text{new}}(t) := \sum_{k=1}^{E_i(t)} \delta^+_{(B_{i,k}(t), D_{i,k}(t))}.$$

We also define the corresponding total mass processes

$$\mathbf{Z}^{\text{init}}(\cdot) := \langle 1, \boldsymbol{\mathcal{Z}}^{\text{init}}(\cdot)\rangle, \quad \mathbf{Z}^{\text{new}}(\cdot) := \langle 1, \boldsymbol{\mathcal{Z}}^{\text{new}}(\cdot)\rangle.$$

3.3 Fluid model

In this section we define and investigate a fluid model that is a deterministic analogue of the stochastic model described in the previous section. Later on the fluid model will be shown to arise as the limit of the stochastic model under a proper fluid scaling. This convergence implies, in particular, existence of the fluid model.

To define the fluid model we need data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0) \in (0, \infty)^I \times \mathcal{M}^I \times \mathcal{M}^I$. The coordinates of $\boldsymbol{\eta}$ play the role of arrival rates. As in the previous section, θ_i (the ith coordinate of $\boldsymbol{\theta}$) is the joint distribution of the generic size B_i and patience time D_i of a newly arrived flow on route i with finite expectations $\mathbb{E}B_i = 1/\mu_i$ and $\mathbb{E}D_i = 1/\nu_i$. We also introduce

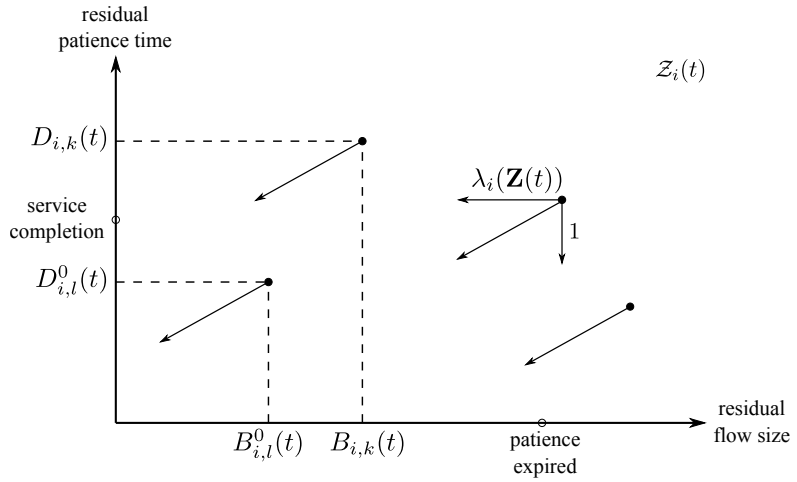

Figure 3.1: The i-th coordinate $\mathcal{Z}_i(\cdot)$ of the state descriptor puts a unit mass to the dots representing class i flows except those on the axes.

the constants

$$\rho_i := \eta_i/\mu_i, \quad \sigma_i := \eta_i/\nu_i,$$

and the vectors $\boldsymbol{\rho}, \boldsymbol{\sigma} \in (0,\infty)^I$,

$$\boldsymbol{\rho} := (\rho_1, \ldots, \rho_I), \quad \boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_I).$$

Finally, the measure-valued vector $\boldsymbol{\zeta}^0$ characterises the initial state of the network. Put $\mathbf{z}^0 := \langle 1, \boldsymbol{\zeta}^0 \rangle$ and, for all i, take a r.v. (B_i^0, D_i^0) that is degenerate at $(0,0)$ if $z_i^0 = 0$ and has distribution ζ_i^0/z_i^0 otherwise. Then $\mathbf{z}^0$ represents the initial population, and (B_i^0, D_i^0) the generic size and patience time of an initial flow on route i. We only consider initial conditions $\boldsymbol{\zeta}^0$ such that the (marginal) distributions of B_i^0 and D_i^0 have no atoms. This restriction is necessary because we require the fluid model to be continuous, see Definition 3.1 below.

Denote by $\mathcal{C}$ the collection of corner sets,

$$\mathcal{C} := \{[x,\infty) \times [y,\infty) \colon (x,y) \in \mathbb{R}_+^2\}.$$

Definition 3.1. A pair $(\boldsymbol{\zeta}, \mathbf{z}) \in \mathbf{C}(\mathbb{R}_+, \mathcal{M}^I) \times \mathbf{C}(\mathbb{R}_+, \mathbb{R}_+^I)$ is called a *fluid model solution (FMS) for the data* $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0)$ if $\mathbf{z}(\cdot) = \langle 1, \boldsymbol{\zeta}(\cdot) \rangle$ and, for all i, $t \in \mathbb{R}_+$ and $A \in \mathcal{C}$,

$$\begin{aligned}\zeta_i(t)(A) = {} & z_i^0\, \mathbb{P}\{(B_i^0, D_i^0) \in A + (S_i(\mathbf{z}, 0, t), t)\} \\ & + \eta_i \int_0^t \mathbb{P}\{(B_i, D_i) \in A + (S_i(\mathbf{z}, s, t), t-s)\}\, ds. \end{aligned} \tag{3.3}$$

In particular, for all i and $t \in \mathbb{R}_+$,

$$z_i(t) = \zeta_i(t)(\mathbb{R}_+^2) = z_i^0 \, \mathbb{P}\{B_i^0 \geq S_i(\mathbf{z}, 0, t), D_i^0 \geq t\} + \eta_i \int_0^t \mathbb{P}\{B_i \geq S_i(\mathbf{z}, s, t), D_i \geq t - s\} \, ds. \tag{3.4}$$

The function $\boldsymbol{\zeta}(\cdot)$ is called a *measure-valued fluid model solution (MVFMS)* and the function $\mathbf{z}(\cdot)$ a *numeric fluid model solution (NFMS)*

Equations (3.3) and (3.4) have appealing physical interpretations. For example, (3.4) simply means that a flow is still in the network at time t if its size and patience exceed, respectively, the amount of service it has received and the time that has passed since its arrival up to time t.

Remark 3.1. By Dynkin's π-λ theorem (see [47, Section 2.3]), FMS's satisfy (3.3) with any Borel set $A \subseteq \mathbb{R}_+^2$.

Remark 3.2. FMS's are invariant with respect to time shifts in the sense that, if $(\boldsymbol{\zeta}, \mathbf{z})(\cdot)$ is an FMS, then, for any $\delta > 0$, $(\boldsymbol{\zeta}^\delta, \mathbf{z}^\delta)(\cdot) := (\boldsymbol{\zeta}, \mathbf{z})(\cdot + \delta)$ is an FMS for the data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}(\delta))$. That is, for all i, $t \geq \delta$ and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) = \zeta_i(\delta)(A + (S_i(\mathbf{z}, \delta, t), t - \delta)) + \eta_i \int_\delta^t \mathbb{P}\{(B_i, D_i) \in A + (S_i(\mathbf{z}, s, t), t - s)\} \, ds, \tag{3.5a}$$

$$z_i(t) = \zeta_i(\delta)([S_i(\mathbf{z}, \delta, t), \infty) \times [t - \delta, \infty)) + \eta_i \int_\delta^t \mathbb{P}\{B_i \geq S_i(\mathbf{z}, s, t), D_i \geq t - s\} ds. \tag{3.5b}$$

Remark 3.3. The measure-valued and numeric components of an FMS uniquely define each other. In particular, uniqueness of an NFMS implies uniqueness of an MVFMS, and the other way around.

As was mentioned earlier, the existence of FMS's is guaranteed by Theorem 3.5 that follows in the next section. In the rest of this section, we discuss sufficient conditions for an FMS to be unique and for an invariant (i.e. constant) FMS to be unique and asymptotically stable. To prove the stability result, we derive relations for asymptotic bounds for FMS's, which seems to be a novel approach since we have not seen analogous results in the related literature. We also give an example of multiple invariant FMS's.

Uniqueness of an FMS The proof of the following theorem follows along the lines of the proofs of similar results [15, Proposition 4.2] and [47, Theorem 3.5], see Section 3.6.

Theorem 3.1. *Suppose that either* (i) $z_i^0 = 0$ *for all* i, *or* (ii) $\mathbf{z}^0 \in (0, \infty)^I$ *and the first projection of* $\boldsymbol{\zeta}^0$ *is Lipschitz continuous, i.e. there exists a constant* $L \in (0, \infty)$ *such that for all* i, $x < \widetilde{x}$ *and* y,

$$\zeta_i^0([x, \widetilde{x}] \times [y, \infty)) \leq L(\widetilde{x} - x).$$

Then an FMS for the data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0)$ *is unique.*

Uniqueness of an invariant FMS Let $(\boldsymbol{\zeta}^*, \mathbf{z}^*)$ be an invariant FMS (we also call it a *fixed point*). By Lemma 3.3 in Section 3.6, all of the coordinates of $\mathbf{z}^*$ are positive, and the fluid model equations (3.3) and (3.4) for $(\boldsymbol{\zeta}^*, \mathbf{z}^*)$ look as follows: for all i, Borel subsets $A \subseteq \mathbb{R}_+^2$ and $t \in \mathbb{R}_+$,

$$\begin{aligned} \zeta_i^*(A) &= \zeta_i^*(A + (\lambda_i(\mathbf{z}^*)t, t)) \\ &\quad + \eta_i \int_0^t \theta_i(A + (\lambda_i(\mathbf{z}^*)s, s))\, ds, \qquad (3.6) \\ z_i^* &= \zeta_i^*([\lambda_i(\mathbf{z}^*)t, \infty) \times [t, \infty)) \\ &\quad + \eta_i \int_0^t \mathbb{P}\{B_i \geq \lambda_i(\mathbf{z}^*)s, D_i \geq s\}\, ds. \qquad (3.7) \end{aligned}$$

Letting $t \to \infty$ in (3.6) and (3.7), we obtain the equations

$$\zeta_i^*(A) = \eta_i \int_0^\infty \theta_i(A + (\lambda_i(\mathbf{z}^*)s, s))ds, \qquad (3.8)$$

$$z_i^* = \eta_i \mathbb{E}(B_i/\lambda_i(\mathbf{z}^*) \wedge D_i), \qquad (3.9)$$

which are actually equivalent to (3.6) and (3.7).

Thus, we have the closed-form equation (3.9) for the numeric components of invariant FMS's, and the corresponding measure-valued components are defined by (3.8). In particular, uniqueness of an invariant FMS is equivalent to uniqueness of a solution to (3.9).

Multiplying the coordinates of (3.9) by the corresponding rates $\lambda_i(\mathbf{z}^*)$, we obtain the equivalent equation

$$\Lambda_i(\mathbf{z}^*) = g_i(\lambda_i(\mathbf{z}^*)) \quad \text{for all } i, \qquad (3.10)$$

where

$$g_i(x) := \eta_i \mathbb{E}(B_i \wedge xD_i), \quad x \in \mathbb{R}_+.$$

We suggest a sufficient condition for uniqueness of a solution to (3.10) (i.e. of an invariant NFMS) that involves the left most points of supports of certain distributions.

Definition 3.2. For an $\mathbb{R}$-valued r.v. X, denote by $\inf X$ the left most point of its support. Recall that the *support* of X is the minimal (in the sense of inclusion) closed interval S such that $\mathbb{P}\{X \in S\} = 1$.

As we show later (see Lemmas 3.6 and 3.7 in Section 3.6), if $m_i \leq 1/\inf(D_i/B_i)$, where $1/0 := \infty$ by definition, the function $g_i(\cdot)$ is continuous and strictly increasing in the interval $[0, m_i]$, implying that its inverse is well-defined in $[0, g_i(m_i)]$. Then we can prove the following result.

Theorem 3.2. *Let*

$$\inf(D_i/B_i) \leq 1/m_i \quad \textit{for all } i. \qquad (3.11)$$

Then there exists a unique invariant FMS $(\boldsymbol{\zeta}^*, \mathbf{z}^*)$*, and the bandwidth allocation vector* $\boldsymbol{\Lambda}(\mathbf{z}^*)$ *is the unique solution to the optimization problem*

$$\begin{array}{ll} \textit{maximize} & \sum_{i=1}^{I} G_i(\Lambda_i) \\ \textit{subject to} & A\boldsymbol{\Lambda} \le \mathbf{C},\ \Lambda_i \le g_i(m_i) \textit{ for all } i, \end{array} \tag{3.12}$$

with strictly concave functions $G_i(\cdot)$ *such that* $G_i'(\cdot) = \mathcal{U}_i'(g_i^{-1}(\cdot))$ *in* $[0, g_i(m_i)]$.

Remark 3.4. Note that it is realistic to assume that flows do not abandon if they are always served at the maximum rate, i.e. that $D_i \ge B_i/m_i$. For such routes, we have $g_i(m_i) = \rho_i$, and the sufficient uniqueness condition (3.11) reads as $\inf(D_i/B_i) = 1/m_i$.

The complete proof of Theorem 3.2 is postponed to Section 3.6, here we only discuss the main ideas. As we plug the fixed point equation (3.10) into the optimization problem (3.1) for the rate vector $\boldsymbol{\lambda}(\mathbf{z}^*)$, the problem (3.12) follows through two applications of the KKT conditions — necessary and sufficient conditions for optimality. The problem (3.12) is strictly concave and does not depend on $\mathbf{z}^*$. Hence, $\boldsymbol{\Lambda}(\mathbf{z}^*) =: \boldsymbol{\Lambda}^*$ is the same for all invariant points $\mathbf{z}^*$. This idea of transforming the optimization problem defining the rate vector combined with the fixed point equation into an independent problem for the bandwidth allocation vector we adopted from Borst et al. [15, Lemma 5.2]. Now, since the functions $g_i(\cdot)$ are invertible in the feasible rate intervals $[0, m_i]$, it follows from (3.10) that the fixed point is unique and given by

$$z_i^* = \Lambda_i^* / g_i^{-1}(\Lambda_i^*). \tag{3.13}$$

Note that this method not only proves uniqueness of an invariant FMS, but also suggests a two-step algorithm to compute it: first we need to solve the strictly concave optimization problem (3.12) for $\boldsymbol{\Lambda}^*$, which can be done with any desired accuracy in polynomial time, and then we can compute the fixed point $\mathbf{z}^*$ itself by formula (3.13).

Example 3.1 (Single link). The sufficient condition for uniqueness of an invariant FMS given by Theorem 3.2 is sometimes also necessary. Consider, for example, processor sharing in critical load, that is $J = I = 1$ and (omitting the link and class indices) $\rho = C$. In this case, the fixed point equation (3.9) looks like

$$z^* = \eta\, \mathbb{E}\left(\frac{B}{C/z^* \wedge m} \wedge D \right),$$

which, for z^* such that $C/z^* \le m$ and $Bz^*/C \le D$ a.s., reduces to

$$z^* = \eta \mathbb{E}(Bz^*/C \wedge D) = \eta \mathbb{E} Bz^*/C = \rho z^*/C = z^*.$$

That is, any $z^* \in [C/m, C \inf D/B]$ is an invariant NFMS. In particular, if $\inf D/B > 1/m$, which violates the assumption of Theorem 3.2, then there is a continuum of invariant FMS's.

For a single link critically loaded by multiple classes of flows, we have an analogous result, which is more complicated to derive and therefore the proof is postponed to

Section 3.6.

Theorem 3.3. *Assume that $J = 1$ (in what follows we omit the link index), and that the utility functions are $\mathcal{U}_i(x) = \kappa_i \log x$. If $\sum_{i=1}^{I} \eta_i \mathbb{E}(B_i/m_i \wedge D_i) \neq C$, then there is a unique invariant FMS. Otherwise there might be a continuum of invariant FMS's.*

Asymptotic bounds for FMS's Here we derive asymptotic bounds for NFMS's that, for a wide class of bandwidth-sharing networks, imply convergence to the invariant NFMS provided it is unique.

Theorem 3.4. *There exist constants $\mathbf{l} = (l_1, \ldots, l_I), \mathbf{u} = (u_1, \ldots, u_I) \in (0, \infty)^I$ such that, for any NFMS $\mathbf{z}(\cdot)$ and all i,*

$$0 < l_i \leq \underline{\lim}_{t\to\infty} z_i(t) \leq \overline{\lim}_{t\to\infty} z_i(t) \leq u_i.$$

These constants satisfy the following relations: for all i,

$$\begin{aligned} l_i &= \eta_i \mathbb{E}(B_i/R_i(\mathbf{l}, \mathbf{u}) \wedge D_i), \\ u_i &= \eta_i \mathbb{E}(B_i/r_i(\mathbf{l}, \mathbf{u}) \wedge D_i), \end{aligned} \tag{3.14}$$

where the functions $\mathbf{r}(\cdot,\cdot) = (r_1, \ldots, r_I)(\cdot,\cdot)$ and $\mathbf{R}(\cdot,\cdot) = (R_1, \ldots, R_I)(\cdot,\cdot)$ are defined by: for all i and $\mathbf{x} \leq \widetilde{\mathbf{x}}$,

$$r_i(\mathbf{x}, \widetilde{\mathbf{x}}) := \inf_{\mathbf{x}\leq\mathbf{z}\leq\widetilde{\mathbf{x}}} \lambda_i(\mathbf{z}), \quad R_i(\mathbf{x}, \widetilde{\mathbf{x}}) := \sup_{\mathbf{x}\leq\mathbf{z}\leq\widetilde{\mathbf{x}}} \lambda_i(\mathbf{z}).$$

Remark 3.5. There could be more than one pair of vectors $(\mathbf{l}, \mathbf{u})$ solving (3.14). The asymptotic bounds $\mathbf{l}$ and $\mathbf{u}$ for NFMS's given by Theorem 3.4 form one of such pairs.

We now proceed with the proof of Theorem 3.4.

Proof. Note that if, for all i,

$$0 < \widetilde{l}_i \leq \underline{\lim}_{t\to\infty} z_i(t) \leq \overline{\lim}_{t\to\infty} z_i(t) \leq \widetilde{u}_i, \tag{3.15}$$

then also

$$\begin{aligned} \underline{\lim}_{t\to\infty} z_i(t) &\geq \eta_i \mathbb{E}(B_i/R_i(\widetilde{\mathbf{l}}, \widetilde{\mathbf{u}}) \wedge D_i), \\ \overline{\lim}_{t\to\infty} z_i(t) &\leq \eta_i \mathbb{E}(B_i/r_i(\widetilde{\mathbf{l}}, \widetilde{\mathbf{u}}) \wedge D_i). \end{aligned} \tag{3.16}$$

Indeed, by (3.15), for any $\varepsilon \in (0, \min_{1\leq i\leq I} \widetilde{l}_i)$, there exists a t_ε such that, for all i and $t \geq t_\varepsilon$,

$$\widetilde{l}_i - \varepsilon \leq z_i(t) \leq \widetilde{u}_i + \varepsilon.$$

Put $\boldsymbol{\varepsilon} := (\varepsilon, \ldots, \varepsilon) \in \mathbb{R}_+^I$, then we have, for $t \geq s \geq t_\varepsilon$,

$$r_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon})(t-s) \leq S_i(\mathbf{z}, s, t) \leq R_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon})(t-s),$$

which, when plugged into the shifted fluid model equation (3.5b), implies that, for all $t \geq t_\varepsilon$,

$$\begin{aligned} z_i(t) &\geq \eta_i \int_{t_\varepsilon}^{t} \mathbb{P}\{B_i \geq R_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon})(t-s), D_i \geq (t-s)\}\, ds, \\ z_i(t) &\leq \zeta_i(t_\varepsilon)\big([S_i(\mathbf{z}, t_\varepsilon, t), \infty) \times [t - t_\varepsilon, \infty)\big) \\ &\quad + \eta_i \int_{t_\varepsilon}^{t} \mathbb{P}\{B_i \geq r_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon})(t-s), D_i \geq (t-s)\}\, ds, \end{aligned}$$

where $\boldsymbol{\zeta}(\cdot)$ is the corresponding MVFMS. Taking $t \to \infty$ in the last two inequalities, we obtain

$$\begin{aligned} \underline{\lim}_{t\to\infty} z_i(t) &\geq \eta_i \mathbb{E}(B_i / R_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon}) \wedge D_i), \\ \overline{\lim}_{t\to\infty} z_i(t) &\leq \eta_i \mathbb{E}(B_i / r_i(\widetilde{\mathbf{l}} - \boldsymbol{\varepsilon}, \widetilde{\mathbf{u}} + \boldsymbol{\varepsilon}) \wedge D_i), \end{aligned}$$

and then (3.16) follows as $\varepsilon \to 0$.

Now we will iterate (3.15)–(3.16). The rate constraints plugged into (3.4) imply the initial bounds: for all i,

$$\begin{aligned} \underline{\lim}_{t\to\infty} z_i(t) &\geq \eta_i \mathbb{E}(B_i / m_i \wedge D_i) =: l_i^0 > 0, \\ \overline{\lim}_{t\to\infty} z_i(t) &\leq \eta_i \mathbb{E} D_i =: u_i^0, \end{aligned}$$

and then (3.15)–(3.16) yield the recursive bounds: for all $k \in \mathbb{N}$ and i,

$$\begin{aligned} \underline{\lim}_{t\to\infty} z_i(t) &\geq \eta_i \mathbb{E}(B_i / R_i(\mathbf{l}^{k-1}, \mathbf{u}^{k-1}) \wedge D_i) =: l_i^k, \\ \overline{\lim}_{t\to\infty} z_i(t) &\leq \eta_i \mathbb{E}(B_i / r_i(\mathbf{l}^{k-1}, \mathbf{u}^{k-1}) \wedge D_i) =: u_i^k. \end{aligned} \tag{3.17}$$

The sequence $\{\mathbf{l}^k\}_{k\in\mathbb{N}}$ is non-decreasing and bounded from above by $\mathbf{u}^0$. The sequence $\{\mathbf{u}^k\}_{k\in\mathbb{N}}$ is non-increasing and bounded from below by $\mathbf{l}^0$. Hence, there exist the limits $\lim \mathbf{l}^k =: \mathbf{l}$ and $\lim \mathbf{u}^k =: \mathbf{u}$. In (3.17), let $k \to \infty$, then (3.14) follows.

Note finally that the recursive bounds $\{\mathbf{l}^k\}_{k\in\mathbb{N}}$ and $\{\mathbf{u}^k\}_{k\in\mathbb{N}}$ as well as their limits $\mathbf{l}$ and $\mathbf{u}$ do not depend on a particular NFMS. □

Asymptotic stability of an invariant FMS It is natural to assume that transfer rates in a bandwidth-sharing network decrease as its population grows. In particular, tree networks satisfy this property, see [15].

Definition 3.3. If $\widetilde{\mathbf{z}} \geq \mathbf{z} \in (0, \infty)^I$ implies $\boldsymbol{\lambda}(\widetilde{\mathbf{z}}) \leq \boldsymbol{\lambda}(\mathbf{z})$, the network is called *monotone*.

For monotone networks, the system of equations (3.14) decomposes into two independent systems of equations for the lower bound $\mathbf{l}$ and for the upper bound $\mathbf{u}$: for all i,

$$l_i = \eta_i \mathbb{E}(B_i/\lambda_i(\mathbf{l}) \wedge D_i), \tag{3.18a}$$
$$u_i = \eta_i \mathbb{E}(B_i/\lambda_i(\mathbf{u}) \wedge D_i), \tag{3.18b}$$

which implies the following result.

Corollary 3.1. *Suppose that the network is monotone and has a unique invariant FMS* $(\boldsymbol{\zeta}^*, \mathbf{z}^*)$. *Then any FMS* $(\boldsymbol{\zeta}, \mathbf{z})(t) \to (\boldsymbol{\zeta}^*, \mathbf{z}^*)$ *as* $t \to \infty$.

Indeed, both (3.18a) and (3.18b) coincide with the fixed point equation (3.9), and since Corollary 3.1 assumes that the latter equation has a unique solution $\mathbf{z}^*$, it immediately follows by Theorem 3.4 that, for any NFMS $\mathbf{z}(\cdot)$ and all i,

$$l_i = \underline{\lim}_{t\to\infty} z_i(t) = \overline{\lim}_{t\to\infty} z_i(t) = u_i = z_i^*.$$

In Section 3.9 we also show that $\mathbf{z}(t) \to \mathbf{z}^*$ implies $\boldsymbol{\zeta}(t) \to \boldsymbol{\zeta}^*$.

3.4 Fluid limit theorem

In this section we study the asymptotic behavior of the stochastic network described in Section 3.2 as its global parameters — capacities and arrival rates — grow large, while the characteristics of an individual flow remain of a fixed order. This is a fluid scaling, which we refer to as the large capacity regime.

Large capacity fluid scaling With each positive number r, we associate a stochastic model as defined in Section 3.2. We label all parameters associated with the r-th model with a superscript r. In particular, model r is defined on a probability space $(\Omega^r, \mathcal{F}^r, \mathbb{P}^r)$ with expectation operator $\mathbb{E}^r$. We assume the following.

Assumption 3.1. *Network structure, rate constraints and utility function are the same in all models:* $A^r = A$, $\mathbf{m}^r = \mathbf{m}$ *and* $\mathcal{U}_i^r(\cdot) = \mathcal{U}_i(\cdot)$ *for all* i.

Assumption 3.2. *Link capacities grow linearly in* r: $\mathbf{C}^r = r\mathbf{C}$.

Assumption 3.3. *Arrival rates grow linearly in* r: $\overline{\mathbf{E}}^r(\cdot) := \mathbf{E}^r(\cdot)/r \Rightarrow \boldsymbol{\eta}(\cdot)$ *as* $r \to \infty$, *where* $\boldsymbol{\eta}(t) := t\boldsymbol{\eta}$ *and* $\boldsymbol{\eta} \in (0,\infty)^I$.

Assumption 3.4. *Flow sizes and patience times remain of a fixed order: for all* i, $(B_i^r, D_i^r) \Rightarrow (B_i, D_i)$ *as* $r \to \infty$, *where* (B_i, D_i) *are* $(0,\infty)^I$*-valued r.v.'s with distributions* θ_i *and finite mean values* $(1/\mu_i, 1/\nu_i)$, *and also* $(1/\mu_i^r, 1/\nu_i^r) \to (1/\mu_i, 1/\nu_i)$.

Assumption 3.5. *The scaled initial configuration converges in distribution to a random vector of finite measures:* $\overline{\boldsymbol{\mathcal{Z}}}^r(0) := \boldsymbol{\mathcal{Z}}^r(0)/r \Rightarrow \boldsymbol{\zeta}^0$, *where, for all* i, *the projections* $\zeta_i^0(\cdot \times \mathbb{R}_+)$ *and* $\zeta_i^0(\mathbb{R}_+ \times \cdot)$ *are a.s. free of atoms.*

Fluid limit theorem In the large capacity regime, the stochastic model defined is Section 3.2 converges to the fluid model defined in Section 3.3. More precisely, introduce the scaled versions of the state descriptors and population processes:

$$\overline{\boldsymbol{\mathcal{Z}}}^r(\cdot) := \boldsymbol{\mathcal{Z}}^r(\cdot)/r, \quad \overline{\mathbf{Z}}^r(\cdot) := \langle 1, \overline{\boldsymbol{\mathcal{Z}}}^r(\cdot)\rangle = \mathbf{Z}^r(\cdot)/r,$$

and also the scaled versions of their two components:

$$\begin{aligned}
\overline{\boldsymbol{\mathcal{Z}}}^{r,\text{init}}(\cdot) &:= \boldsymbol{\mathcal{Z}}^{r,\text{init}}(\cdot)/r,\\
\overline{\mathbf{Z}}^{r,\text{init}}(\cdot) &:= \langle 1, \overline{\boldsymbol{\mathcal{Z}}}^{r,\text{init}}(\cdot)\rangle = \mathbf{Z}^{r,\text{init}}(\cdot)/r,\\
\overline{\boldsymbol{\mathcal{Z}}}^{r,\text{new}}(\cdot) &:= \boldsymbol{\mathcal{Z}}^{r,\text{new}}(\cdot)/r,\\
\overline{\mathbf{Z}}^{r,\text{new}}(\cdot) &:= \langle 1, \overline{\boldsymbol{\mathcal{Z}}}^{r,\text{new}}(\cdot)\rangle = \mathbf{Z}^{r,\text{new}}(\cdot)/r.
\end{aligned}$$

Remark 3.6. Let $\boldsymbol{\lambda}(\cdot)$ be the rate allocation function in the unscaled network, then

$$\begin{aligned}
\boldsymbol{\lambda}^r(\mathbf{z}) &=: \arg\max_{\substack{A(\boldsymbol{\lambda}\times\mathbf{z})\le r\mathbf{C}\\ \boldsymbol{\lambda}\le\mathbf{m}}} \textstyle\sum_{i=1}^I z_i\,\mathcal{U}_i(\lambda_i)\\
&= \arg\max_{\substack{A(\boldsymbol{\lambda}\times\mathbf{z}/r)\le\mathbf{C}\\ \boldsymbol{\lambda}\le\mathbf{m}}} \textstyle\sum_{i=1}^I (z_i/r)\,\mathcal{U}_i(\lambda_i) =: \boldsymbol{\lambda}(\mathbf{z}/r),
\end{aligned}$$

and

$$S_i^r(\mathbf{Z}^r, s, t) := \int_s^t \lambda_i^r(\mathbf{Z}^r(u))du = \int_s^t \lambda_i(\overline{\mathbf{Z}}^r(u))du =: S_i(\overline{\mathbf{Z}}^r, s, t).$$

We refer to weak limits along convergent subsequences of the processes $(\overline{\boldsymbol{\mathcal{Z}}}^r, \overline{\mathbf{Z}}^q)(\cdot)$, $r \to \infty$, as *fluid limits*. Now we present one of the main results of this chapter.

Theorem 3.5. *Under Assumptions 3.1–3.5, the family of the scaled processes $(\overline{\boldsymbol{\mathcal{Z}}}^r, \overline{\mathbf{Z}}^r)(\cdot)$ is $\mathbf{C}$-tight in $\mathbf{D}(\mathbb{R}_+, \mathcal{M}^I) \times \mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I)$, and all fluid limits are a.s. FMS's for the data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0)$. In particular, if there is a unique FMS $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$ for the data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0)$, then $(\overline{\boldsymbol{\mathcal{Z}}}^r, \overline{\mathbf{Z}}^r)(\cdot) \Rightarrow (\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$ as $r \to \infty$.*

The proof follows in Section 3.7. To show tightness we extend the techniques of Gromoll and Williams [46] to the two-dimensional case, since in [46] flows are patient and state descriptors are vectors of measures on $\mathbb{R}_+$. The proof of convergence to FMS's follows the lines of that in Gromoll et al. [47]. It uses the boundedness of fluid limits away from zero, and the key difference is that in [47] this property is guaranteed by the overload regime, while in our model it holds in any load regime due to the rate constraints.

3.5 Fixed-point approximations for the stationary distribution

Assume that, in the stochastic model defined in Section 3.2, the arrival processes are Poisson of rates $\eta_1, \ldots, \eta_I$. Then there exists a unique stationary (and also limiting as

$t \to \infty$) distribution of the state descriptor $\boldsymbol{\mathcal{Z}}(\cdot)$. Indeed, without loss of generality, there are i.i.d. r.v.'s $\{\widetilde{D}_{i,k}\}_{k\in\mathbb{N},1\le i\le I}$ distributed as $\max_{1\le i\le I} D_i$ and such that a.s. $D_{i,k} \le \widetilde{D}_{i,k}$ for all k and i. Then the total population $\|\mathbf{Z}(\cdot)\|_1$ of the network is a.s. and within the whole time horizon $\mathbb{R}_+$ bounded from above by the length of the $M/G/\infty$ queue (see Remark 1.2) with the following parameters. At time $t = 0$, there are $\|\mathbf{Z}(0)\|_1$ customers in the queue whose service times are patience times of initial flows in the network. The input process for the queue is the superposition of those for the network, and hence is Poisson of rate $\|\boldsymbol{\eta}\|_1$. Service times of new customers in the queue are drawn from the sequence $\{\widetilde{D}_{i,k}\}_{k\in\mathbb{N},1\le i\le I}$ of upper bounds for patience times of new flows in the network. As any other $M/G/\infty$ queue, the queue under consideration is regenerative. The instants when a customer enters the empty queue form an embedded renewal process whose cycle length is non-lattice and has a finite mean value $\exp(\|\boldsymbol{\eta}\|_1 \mathbb{E}\widetilde{D}_{1,1})/\|\boldsymbol{\eta}\|_1$. With respect to this renewal process , the state descriptor $\boldsymbol{\mathcal{Z}}(\cdot)$ is also regenerative. Then, by Asmussen [5, Chapter V.I, Theorem 1.2], there exists a limiting distribution for $\boldsymbol{\mathcal{Z}}(t)$ as $t \to \infty$.

Now consider a sequence of stochastic models as defined in Section 3.2 that satisfies Assumptions 3.1, 3.2, 3.4 from Section 3.4 and Assumptions 3.6, 3.7 below.

Assumption 3.6. *The input processes $E_1^r(\cdot), \dots, E_I^r(\cdot)$ are mutually independent Poisson processes of rates $\eta_1^r, \dots, \eta_I^r$, and $\boldsymbol{\eta}^r/r \to \boldsymbol{\eta} \in (0,\infty)^I$ as $r \to \infty$.*

Assumption 3.7. *On all routes i, the size B_i^r of a flow and its patience time D_i^r are independent.*

Let $\boldsymbol{\mathcal{Y}}^r$ have the stationary distribution of $\boldsymbol{\mathcal{Z}}^r(\cdot)$ and put $\mathbf{Y}^r := \langle 1, \boldsymbol{\mathcal{Y}}^r\rangle$. Introduce also the scaled versions

$$\overline{\boldsymbol{\mathcal{Y}}}^r := \boldsymbol{\mathcal{Y}}^r/r, \quad \overline{\mathbf{Y}}^r := \langle 1, \overline{\boldsymbol{\mathcal{Y}}}^r\rangle = \mathbf{Y}^r/r.$$

We have proven the following result.

Theorem 3.6. *Under Assumptions 3.1, 3.2, 3.4, 3.6 and 3.7, the family of the fluid scaled stationary distributions $(\overline{\boldsymbol{\mathcal{Y}}}^r, \overline{\mathbf{Y}}^r)$ is tight, and any weak limit point $(\boldsymbol{\mathcal{Y}}, \mathbf{Y})$ is a weak invariant FMS, i.e. there exists a stationary FMS $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(t) \stackrel{d}{=} (\boldsymbol{\mathcal{Y}}, \mathbf{Y})$, $t \in \mathbb{R}_+$. In particular, by Corollary 3.1, if the network is monotone and has a unique invariant FMS $(\boldsymbol{\zeta}^*, \mathbf{z}^*)$, then $(\overline{\boldsymbol{\mathcal{Y}}}^r, \overline{\mathbf{Y}}^r) \Rightarrow (\boldsymbol{\zeta}^*, \mathbf{z}^*)$ as $r \to \infty$.*

The general strategy of the proof is adopted from Kang and Ramanan [53, Theorem 3.3]: we check that any convergent sequence of initial conditions $\overline{\boldsymbol{\mathcal{Z}}}^q(0) \stackrel{\mathrm{d}}{=} \overline{\boldsymbol{\mathcal{Y}}}^q \Rightarrow \boldsymbol{\mathcal{Y}}$, $q \to \infty$, satisfies the assumptions of the fluid limit theorem (we only need to check Assumption 3.5). Then the corresponding sequence $\{\overline{\boldsymbol{\mathcal{Z}}}^q(\cdot)\}_{q\to\infty}$ of the scaled state descriptors converges to an MVFMS that is stationary (i.e. $\boldsymbol{\mathcal{Y}}$ is a weak invariant MVFMS) since all $\overline{\boldsymbol{\mathcal{Z}}}^q(\cdot)$ are stationary.

The techniques we use to implement this strategy are different from the techniques of [53], though. Our key instruments for establishing tightness are $M/G/\infty$ bounds, see Section 3.8. Below we present an elegant proof of weak limit points of the family of the scaled stationary distributions $\overline{\boldsymbol{\mathcal{Y}}}^r$ satisfying Assumption 3.5.

Lemma 3.2. *The weak limit $\boldsymbol{\mathcal{Y}}$ along any convergent sequence $\{\overline{\boldsymbol{\mathcal{Y}}}^q\}_{q\to\infty}$ has both projections $\boldsymbol{\mathcal{Y}}(\cdot\times\mathbb{R}_+)$ and $\boldsymbol{\mathcal{Y}}(\mathbb{R}_+\times\cdot)$ a.s. free of atoms.*

Proof. The key idea is the following. Consider the network in its stationary regime. Then, on one hand, it always has the same distribution, and on the other hand, all initial flows are gone at some point, and newly arriving flows do not accumulate along horizontal and vertical lines.

Let $\boldsymbol{\mathcal{Y}}$ be the weak limit along a subsequence $\{\overline{\boldsymbol{\mathcal{Y}}}^q\}_{q\to\infty}$, and run the q-th network starting from $\overline{\boldsymbol{\mathcal{Z}}}^q(0)\stackrel{\mathrm{d}}{=}\overline{\mathcal{Y}}^q$. By [47, Lemma 6.2], it suffices to show that, for any $\delta>0$ and $\varepsilon>0$, there exists an $a>0$ such that

$$\underline{\lim}_{q\to\infty}\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\boldsymbol{\mathcal{Y}}}^q(H_x^{x+a})\|\vee\|\overline{\boldsymbol{\mathcal{Y}}}^q(V_x^{x+a})\|\le\delta\}\ge 1-\varepsilon, \tag{3.19}$$

where $H_a^b:=\mathbb{R}_+\times[a,b]$ and $V_a^b:=[a,b]\times\mathbb{R}_+$ for all $b\ge a\ge 0$.

First we estimate the time when there are only a few (when scaled) initial flows left. The initial flows whose initial patience times are less than t are already gone at time t. Then Lemma 3.16 (see Section 3.8) implies that (in what follows $\mathrm{Pois}(\alpha)$ stands for a r.v. with Poisson distribution with parameter α)

$$\begin{aligned}\overline{Z}_i^{q,\,\mathrm{init}}(t)&\le\overline{\mathcal{Z}}_i^q(0)(\mathbb{R}_+\times[t,\infty))\stackrel{\mathrm{d}}{=}\overline{\mathcal{Y}}_i^q(\mathbb{R}_+\times[t,\infty))\\&\le_{\mathrm{st}}\frac{1}{q}\mathrm{Pois}\left(\eta_i^q\int_t^\infty\mathbb{P}^q\{D_i^q>y\}dy\right)\\&\Rightarrow\eta_i\int_t^\infty\mathbb{P}\{D_i>y\}dy.\end{aligned}$$

Take T such that $\max_{1\le i\le I}\eta_i\int_t^\infty\mathbb{P}\{D_i>y\}dy<\delta/2$, then

$$\lim_{q\to\infty}\mathbb{P}^r\{\|\overline{\mathbf{Z}}^{q,\,\mathrm{init}}(T)\|\le\delta/2\}=1.$$

Now, in Lemma 3.11 (see Section 3.7), we prove that newly arriving customers do not accumulate in thin horizontal and vertical strips, i.e. there exists an $a>0$ such that

$$\begin{aligned}\underline{\lim}_{q\to\infty}\mathbb{P}^q\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\|\overline{\boldsymbol{\mathcal{Z}}}^{q,\,\mathrm{new}}(t)(H_x^{x+a})\|\\\vee\|\overline{\boldsymbol{\mathcal{Z}}}^{q,\,\mathrm{new}}(t)(V_x^{x+a})\|\le\delta/2\}\ge 1-\varepsilon.\end{aligned}$$

Finally, because of stationarity of $\boldsymbol{\mathcal{Y}}^q$,

$$\begin{aligned}&\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\boldsymbol{\mathcal{Y}}}^q(H_x^{x+a})\|\vee\|\overline{\boldsymbol{\mathcal{Y}}}^q(V_x^{x+a})\|\le\delta\}\\&=\mathbb{P}^q\{\sup_{x\in\mathbb{R}_+}\|\overline{\boldsymbol{\mathcal{Z}}}^q(T)(H_x^{x+a})\|\vee\|\overline{\boldsymbol{\mathcal{Z}}}^q(T)(V_x^{x+a})\|\le\delta\}\\&\ge\mathbb{P}^q\{\|\overline{\mathbf{Z}}^{q,\,\mathrm{init}}(T)\|\le\delta/2,\ \sup_{x\in\mathbb{R}_+}\|\overline{\boldsymbol{\mathcal{Z}}}^{q,\,\mathrm{new}}(T)(H_x^{x+a})\|\\&\qquad\vee\|\overline{\boldsymbol{\mathcal{Z}}}^{q,\,\mathrm{new}}(T)(V_x^{x+a})\|\le\delta/2\},\end{aligned}$$

which implies (3.19) by the choice of T and a. □

3.6 Proof of fluid model properties

Here we prove the results of Section 3.3.

3.6.1 Proof of Theorem 3.1

In the proof of Theorem 3.1, we exploit boundedness of NFMS's away from zero and in the norm (see Lemma 3.3), and Lipschitz continuity of MVFMS's in the first coordinate (see Lemma 3.5). We also use the auxiliary Lemma 3.4.

Recall the notations $\sigma_i = \eta_i \mathbb{E} D_i$ and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_I)$.

Lemma 3.3. *Let $\mathbf{z}(\cdot)$ be an NFMS. Then $\sup_{t\in\mathbb{R}_+} \|\mathbf{z}(t)\| \le \|\mathbf{z}(0)\| + \|\boldsymbol{\sigma}\| < \infty$ and, for any $\delta > 0$, $\inf_{t\ge\delta} \min_{1\le i\le I} z_i(t) > 0$. In particular, if $z_i(0) > 0$, then $\inf_{t\in\mathbb{R}_+} z_i(t) > 0$.*

Prrof. By the rate constraints, $S_i(\mathbf{z}, s, t) \le m_i(t-s)$ for all $s \le t$, which, when plugged into the fluid model equation (3.4), implies the following lower bound:

$$z_i(t) \ge \eta_i \int_0^t \mathbb{P}\{B_i/m_i \wedge D_i \ge s\} ds.$$

Since

$$f_i(s) := \mathbb{P}\{B_i/m_i \wedge D_i \ge s\} \uparrow \mathbb{P}\{B_i/m_i \wedge D_i > 0\} = 1 \quad \text{as } s \downarrow 0,$$

it follows that $f_i(\cdot) \ge 1/2$ in a small enough interval $[0, \varepsilon]$. Then, for $t \ge \delta$,

$$z_i(t) \ge \eta_i \int_0^{\delta\wedge\varepsilon} f_i(s) ds \ge \eta_i(\delta \wedge \varepsilon)/2.$$

The upper bound follows from (3.4) directly:

$$z_i(t) \le z_i(0) + \eta_i \int_0^t \mathbb{P}\{D_i \ge s\} ds \uparrow z_i(0) + \sigma_i \quad \text{as } t \uparrow \infty.$$ □

Lemma 3.4. *For an $\mathbb{R}$-valued r.v. ξ and $x \le \widetilde{x}$, $\int_{\mathbb{R}} \mathbb{P}\{u + x \le \xi \le u + \widetilde{x}\} du \le \widetilde{x} - x$.*

See Section 3.9 for the proof.

Lemma 3.5. *Under assumption* (ii) *of Theorem 3.1, any MVFMS $\boldsymbol{\zeta}(\cdot)$ at any time $t \in \mathbb{R}_+$ has a Lipschitz continuous first projection, i.e. there exists a constant $L(\boldsymbol{\zeta}, t) \in (0, \infty)$ such that for all i, $x < \widetilde{x}$ and y,*

$$\zeta_i(t)([x, \widetilde{x}] \times [y, \infty)) \le L(\boldsymbol{\zeta}, t)(\widetilde{x} - x).$$

Proof. For an FMS $(\boldsymbol{\zeta}, \mathbf{z})(\cdot)$, for all i, $t \in \mathbb{R}_+$, $x < \widetilde{x}$ and y,

$$\zeta_i(t)([x, \widetilde{x}] \times [y, \infty)) \leq f_i(x, \widetilde{x}, y) + \eta_i g_i(x, \widetilde{x}, y),$$

where

$$\begin{aligned} f_i(x, \widetilde{x}, y) &:= \zeta_i^0([x + S_i(\mathbf{z}, 0, t), \widetilde{x} + S_i(\mathbf{z}, 0, t)] \times [y + t, \infty)), \\ g_i(x, \widetilde{x}, y) &:= \int_0^t \mathbb{P}\{x + S_i(\mathbf{z}, s, t) \leq B_i \leq \widetilde{x} + S_i(\mathbf{z}, s, t)\} ds. \end{aligned} \tag{3.20}$$

By Lipschitz continuity of the initial condition, $f_i(x, \widetilde{x}, y) \leq L(\widetilde{x} - x)$. In (3.20), change the variable of integration to $v = V(s) := S_i(\mathbf{z}, s, t)$. Then

$$g_i(x, \widetilde{x}, y) = \int_0^{S_i(\mathbf{z}, 0, t)} \mathbb{P}\{x + v \leq B_i \leq \widetilde{x} + v\} / \lambda_i(\mathbf{z}(V^{-1}(v))) dv \leq M(\boldsymbol{\zeta}, t)(\widetilde{x} - x),$$

where $M(\boldsymbol{\zeta}, t) := \sup_{s \in [0,t]} \max_{1 \leq i \leq I} 1/\lambda_i(\mathbf{z}(s))$. By Lemma 3.3, the functions $1/\lambda_i(z(\cdot))$ are continuous in $[0, t]$. Hence $M(\boldsymbol{\zeta}, t)$ is finite and the first projection of $\boldsymbol{\zeta}(t)$ is Lipschitz continuous with the constant $L(\boldsymbol{\zeta}, t) := L + \|\boldsymbol{\eta}\| M(\boldsymbol{\zeta}, t)$. □

Now we are in a position to prove Theorem 3.1.

Proof of Theorem 3.1. Let $(\boldsymbol{\zeta}^1, \mathbf{z}^1)(\cdot)$ and $(\boldsymbol{\zeta}^2, \mathbf{z}^2)(\cdot)$ be two FMS's for the data $(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\zeta}^0)$.

(i) We show that the two FMS's coincide in an interval $[0, \delta]$. We check that $\mathbf{z}^1(\delta) = \mathbf{z}^2(\delta) \in (0, \infty)^I$ and that the first projection of $\boldsymbol{\zeta}^1(\delta) = \boldsymbol{\zeta}^2(\delta)$ is Lipschitz continuous. Then, by Remark 3.2 and the second part of the theorem, the two FMS's coincide everywhere.

Note that, for a vector $\mathbf{z} \in (0, \infty)^I$ of a small enough norm, $\lambda_i(\mathbf{z}) = m_i$ for all i. Lemma 3.3 and the fluid model equation (3.4) imply that $0 < z_i^1(t), z_i^2(t) \leq \eta_i t$ for all i and $t > 0$. Then, for all i and $s, t \in [0, \delta]$, where δ is small enough,

$$S_i(\mathbf{z}^1, s, t) = S_i(\mathbf{z}^2, s, t) = m_i(t - s). \tag{3.21}$$

Plugging (3.21) into (3.4), we obtain, for $t \in [0, \delta]$ and all i,

$$z_i^1(t) = z_i^2(t) = \eta_i \int_0^t \mathbb{P}\{B_i / m_i \wedge D_i \geq s\} ds.$$

By Remark 3.3, $\boldsymbol{\zeta}^1(\cdot)$ and $\boldsymbol{\zeta}^2(\cdot)$ coincide in $[0, \delta]$, too. Lipschitz continuity of the first projection of $\boldsymbol{\zeta}^1(\delta) = \boldsymbol{\zeta}^2(\delta)$ follows as we plug (3.21) into the fluid model equation (3.3)

(recall that it is valid for all Borel sets): for all i, $x < \widetilde{x}$ and y,

$$\begin{aligned}\zeta_i^j(\delta)([x,\widetilde{x}]\times[y,\infty)) &= \eta_i \int_0^\delta \mathbb{P}\{x + m_i s \le B_i \le \widetilde{x} + m_i s, D_i \ge y + s\} ds \\ &\le \eta_i \int_0^\delta \mathbb{P}\{x/m_i + s \le B_i/m_i \le \widetilde{x}/m_i + s\} ds \\ &\le \eta_i (\widetilde{x} - x)/m_i, \quad j = 1, 2,\end{aligned}$$

where the last inequality holds by Lemma 3.4.

(ii) Suppose that the two FMS's are different, that is $t_* := \inf\{t > 0 \colon \mathbf{z}^1(t) \ne \mathbf{z}^2(t)\} < \infty$.

Without loss of generality we may assume that $t_* = 0$. Indeed, otherwise we can consider the time-shifted FMS's $(\boldsymbol{\zeta}^j, \mathbf{z}^j)(t_* + \cdot)$, $j = 1, 2$. By Lemmas 3.3 and 3.5, they start from $\mathbf{z}^1(t_*) = \mathbf{z}^2(t_*) \in (0,\infty)^I$ and $\boldsymbol{\zeta}^1(t_*) = \boldsymbol{\zeta}^2(t_*)$ with a Lipschitz continuous first projection.

By Lemma 3.3, the two NFMS never leave a compact set $[\delta, \Delta]^I \subset (0,\infty)^I$. Since the rate functions $\lambda_i(\mathbf{z})$ are Lipschitz continuous in such sets, there exists a constant $K \in (0,\infty)$ such that, for all i and $s \le t$,

$$|S_i(\mathbf{z}^1, s, t) - S_i(\mathbf{z}^2, s, t)| \le K t \underbrace{\sup_{s\in[0,t]} \|\mathbf{z}^1(s) - \mathbf{z}^2(s)\|}_{=:\,\varepsilon(t)}.$$

Then, by Lipschitz continuity of the initial condition, we have, for all i and $t \in \mathbb{R}_+$,

$$\begin{aligned}&|z_i^1(t) - z_i^2(t)| \\ &\le\ LKt\varepsilon(t) + \eta_i \int_0^t \mathbb{P}\{S_i(\mathbf{z}^1, s, t) - Kt\varepsilon(t) \le B_i \le S_i(\mathbf{z}^1, s, t) + Kt\varepsilon(t)\} ds.\end{aligned}$$

In the last equation, change the variable of integration to $v = S_i(\mathbf{z}^1, s, t)$ (*cf.* the proof of Lemma 3.5) and put $M = \sup_{z\in[\delta,\Delta]^I} \max_{1\le i\le I} 1/\lambda_i(\mathbf{z})$. Then, for all i,

$$|z_i^1(t) - z_i^2(t)| \le LKt\varepsilon(t) + \eta_i M 2Kt\varepsilon(t),$$

and hence,

$$\varepsilon(t) \le (L + 2\|\boldsymbol{\eta}\|M)Kt\varepsilon(t).$$

The last inequality implies that $\varepsilon(t) = 0$ for small enough t, and we arrive at a contradiction with $t_* = 0$. □

3.6.2 Proof of Theorem 3.2

Before proceeding with the proof of the theorem, we state some properties of the functions $g_i(\cdot)$ in the auxiliary Lemmas 3.6 and 3.7. Recall that these functions are given

by

$$g_i(x) = \eta_i \mathbb{E}(B_i \wedge xD_i), \quad x \in \mathbb{R}_+.$$

Lemma 3.6. *The function $g_i(\cdot)$ is continuous. Also $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i]$ and constant in $[\alpha_i, \infty)$, where*

$$\alpha_i := \inf\{x \colon g_i(x) = \rho_i\} > 0,$$

and the infimum over the empty set is defined to be ∞.

Proof. Continuity of $g_i(\cdot)$ follows by the dominated convergence theorem.

The situation $\alpha_i = 0$ is not possible since in that case $g_i(x) = \rho_i$ for all $x > 0$ by the definition of α_i. But $g_i(\cdot)$ is continuous and $g_i(x) \to g_i(0) = 0$ as $x \to 0$.

If $\alpha_i < \infty$, then, again by the definition of α_i and continuity of $g_i(\cdot)$, we have $g_i(x) = \rho_i$ for all $x \geq \alpha_i$ and $g_i(x) < \rho_i = g_i(\alpha_i)$ for all $x < \alpha_i$.

It is left to check that $g_i(\cdot)$ is strictly increasing in $[0, \alpha_i)$. Assume that $0 \leq x < y < \alpha_i$, but $g_i(x) = g_i(y)$. Then

$$\begin{aligned}
0 &= g_i(y)/\eta_i - g_i(x)/\eta_i \\
&= \mathbb{E}B_i\mathbb{I}\{B_i \leq xD_i\} + \mathbb{E}B_i\mathbb{I}\{xD_i < B_i \leq yD_i\} \\
&\quad + \mathbb{E}yD_i\mathbb{I}\{B_i > yD_i\} - \mathbb{E}B_i\mathbb{I}\{B_i \leq xD_i\} \\
&\quad - \mathbb{E}xD_i\mathbb{I}\{xD_i < B_i \leq yD_i\} - \mathbb{E}xD_i\mathbb{I}\{B_i > yD_i\}) \\
&= \mathbb{E}\underbrace{(B_i - xD_i)\mathbb{I}\{xD_i < B_i \leq yD_i\}}_{=:\,X} + (y-x)\mathbb{E}\underbrace{D_i\mathbb{I}\{B_i > yD_i\}}_{=:\,Y},
\end{aligned}$$

where the r.v.'s X and Y are non-negative, so they must a.s. equal zero. In particular, we have $B_i \leq yD_i$ and $g_i(y) = \rho_i$, which contradicts the definition of α_i since $y < \alpha_i$. □

The stabilisation points α_i of the functions $g_i(\cdot)$ are related with the r.v.'s (B_i, D_i) in the following way.

Lemma 3.7. *If $\alpha_i < \infty$, then $\inf D_i/B_i = 1/\alpha_i$. If $\alpha_i = \infty$, then $\inf D_i/B_i = 0$.*

Proof. First assume $\alpha_i < \infty$. Rewrite the relation $g_i(x) = \rho_i$ as $\mathbb{E}B_i(1 - (1 \wedge xD_i/B_i)) = 0$, which, for $x > 0$, is equivalent to $D_i/B_i \geq 1/x$ a.s. Hence $\alpha_i = \inf\{x > 0 \colon D_i/B_i \geq 1/x \text{ a.s.}\}$ and $1/\alpha_i = \sup\{y > 0 \colon D_i/B_i \geq y \text{ a.s.}\}$. In the right-hand side of the latter equation we see the definition of $\inf D_i/B_i$.

Now consider the case $\alpha_i = \infty$. Assume that $\inf D_i/B_i = y > 0$, then $D_i/y \geq B_i$ a.s. and $g_i(1/y) = \rho_i$. On the other hand, since $\alpha_i = \infty$, there is no $x > 0$ such that $g_i(x) = \rho_i$. Hence $y = 0$. □

Having established the above properties of the $g_i(\cdot)$'s, we now can prove Theorem 3.2 by adapting a technique developed by Borst et al. [15, Lemma 5.2].

Proof of Theorem 3.2. We first show uniqueness. Let $\mathbf{z}^* \in (0,\infty)^I$ be an invariant NFMS, i.e. satisfy (3.10). Recall that $\boldsymbol{\lambda}(\mathbf{z}^*)$ is the unique optimal solution to the concave optimization problem (3.1). The necessary and sufficient conditions for that are given by the Karush-Kuhn-Tucker (KKT) theorem (see e.g. Balder [8, Theorem 3.1]): there exist $\mathbf{p} \in \mathbb{R}^J_+$ and $\widetilde{\mathbf{q}} \in \mathbb{R}^I_+$ such that

$$\begin{aligned} z_i^* \, \mathcal{U}_i'(\lambda_i(\mathbf{z}^*)) = z_i^* \textstyle\sum_{j=1}^J A_{j,i} p_j + \widetilde{q}_i \quad \text{for all } i, \\ p_j \left(\textstyle\sum_{i=1}^I A_{j,i} \lambda_i(\mathbf{z}^*) z_i^* - C_j \right) = 0 \quad \text{for all } j, \\ \widetilde{q}_i (\lambda_i(\mathbf{z}^*) - m_i) = 0 \quad \text{for all } i. \end{aligned} \tag{3.22}$$

or equivalently, there exist $\mathbf{p} \in \mathbb{R}^J_+$ and $\mathbf{q} \in \mathbb{R}^I_+$ ($q_i = \widetilde{q}_i / z_i^*$) such that

$$\mathcal{U}_i'(\lambda_i(\mathbf{z}^*)) = \textstyle\sum_{j=1}^J A_{j,i} p_j + q_i \quad \text{for all } i, \tag{3.23a}$$

$$p_j \left(\textstyle\sum_{i=1}^I A_{j,i} \Lambda_i(\mathbf{z}^*) - C_j \right) = 0 \quad \text{for all } j, \tag{3.23b}$$

$$q_i (\lambda_i(\mathbf{z}^*) - m_i) = 0 \quad \text{for all } i. \tag{3.23c}$$

The theorem assumes that $m_i \le \alpha_i$. So, by Lemmas 3.6 and 3.7, the functions $g_i(\cdot)$ are strictly increasing in the intervals $[0, m_i]$ (see also the left graph in Figure 3.2), which implies two things. First, the fixed point equation (3.10) can be rewritten as $\lambda_i(\mathbf{z}^*) = g_i^{-1}(\Lambda_i(\mathbf{z}^*))$ for all i, and we plug that into (3.23a). Second, the second multiplier $(\lambda_i(\mathbf{z}^*) - m_i)$ in (3.23c) is zero if and only if $g_i(\lambda_i(\mathbf{z}^*)) = g_i(m_i)$, and that, by (3.10), is equivalent to $\Lambda_i(\mathbf{z}^*) = g_i(m_i)$. Hence, $\boldsymbol{\Lambda}(\mathbf{z}^*)$ satisfies

$$\mathcal{U}_i'(g_i^{-1}(\Lambda_i(\mathbf{z}^*))) = \textstyle\sum_{j=1}^J A_{j,i} p_j + q_i \quad \text{for all } i, \tag{3.24a}$$

$$p_j \left(\textstyle\sum_{i=1}^I A_{j,i} \Lambda_i(\mathbf{z}^*) - C_j \right) = 0 \quad \text{for all } j, \tag{3.24b}$$

$$q_i (\Lambda_i(\mathbf{z}^*) - g_i(m_i)) = 0 \quad \text{for all } i. \tag{3.24c}$$

Now note that the last three equations form the KKT conditions for another optimization problem. Indeed, take functions $\widetilde{g}_i(\cdot)$ that are continuous and strictly increasing in $\mathbb{R}_+$ and coincide with $g_i(\cdot)$ in $[0, m_i]$ (and hence, the inverse functions $\widetilde{g}_i^{-1}(\cdot)$ and $g_i^{-1}(\cdot)$ coincide in $[0, g_i(m_i)]$). Also take functions $G_i(\cdot)$ such that $G_i'(\cdot) = \mathcal{U}_i'(\widetilde{g}_i^{-1}(\cdot))$ in $(0,\infty)$. Then (3.24) gives necessary and sufficient conditions for $\boldsymbol{\Lambda}(\mathbf{z}^*)$ to solve (3.12). Since the functions $\mathcal{U}_i(\cdot)$ are strictly concave, their derivatives $\mathcal{U}_i'(\cdot)$ are strictly decreasing. Then, since the $\widetilde{g}_i^{-1}(\cdot)$'s are strictly increasing, the $G_i'(\cdot)$'s are strictly decreasing and, equivalently, the $G_i(\cdot)$'s are strictly concave, which implies that $\boldsymbol{\Lambda}(\mathbf{z}^*) = \boldsymbol{\Lambda}^*$ is actually the unique solution to (3.12) and does not depend on $\mathbf{z}^*$. Then we invert the $g_i(\cdot)$'s in the fixed point equation (3.10), which implies that the fixed point $z_i^* = \Lambda_i^* / g_i^{-1}(\Lambda_i^*)$ is unique because $\boldsymbol{\Lambda}^*$ is unique.

The existence result follows similarly. There exists a unique optimal solution $\boldsymbol{\Lambda}^*$ to (3.12) and it satisfies the KKT conditions (3.24). Put $\lambda_i^* = g_i^{-1}(\Lambda_i^*)$ for all i. Then $\boldsymbol{\lambda}^*$ and $\boldsymbol{\Lambda}^*$ satisfy the KKT conditions (3.23) and (3.22) , i.e., for the vector $\mathbf{z}^*$ with $z_i^* := \Lambda_i^* / \lambda_i^*$, we

have $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}(\mathbf{z}^*)$ and $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}(\mathbf{z}^*)$. Plugging the last two relations into the definition of $\boldsymbol{\lambda}^*$, we get the fixed point equation. □

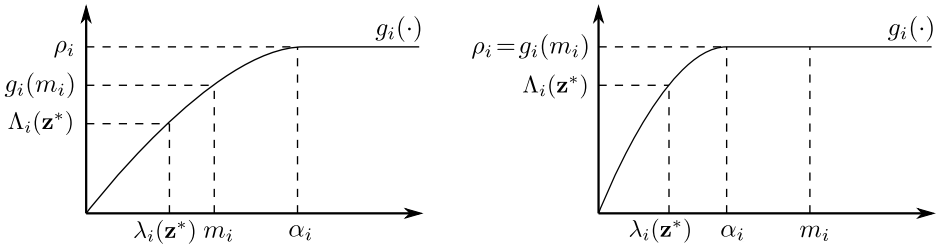

Figure 3.2: Graph of the function $g_i(\cdot)$ in the two possible cases: when $m_i \leq \alpha_i$ (left) and when $m_i > \alpha_i$ (right); $\mathbf{z}^*$ is a invariant NFMS.

3.6.3 Proof of Theorem 3.3

The fixed point equation (3.10) and the monotonicity of the functions $g_i(\cdot)$ imply that the bandwidth class i gets in an equilibrium is at most $g_i(m_i)$. Therefore, we refer to the scenarios $\sum_{i=1}^I g_i(m_i) < C$, $\sum_{i=1}^I g_i(m_i) = C$ and $\sum_{i=1}^I g_i(m_i) > C$ as underloaded, critically loaded and overloaded, respectively. Below we calculate the invariant NFMS's in the three cases.

Summing up (3.9), the KKT conditions (3.23) for (3.1) and the capacity and rate constraints, a $\mathbf{z}^* \in (0,\infty)^I$ is an invariant NFMS if and only if there exist $p \in \mathbb{R}_+$ and $\mathbf{q} \in \mathbb{R}_+^I$ such that (we omit the argument of the rates $\lambda_i(\mathbf{z}^*)$ and bandwidth allocations $\Lambda_i(\mathbf{z}^*)$)

$$\Lambda_i = g_i(\lambda_i) \quad \text{for all } i, \tag{3.25a}$$
$$\kappa_i/\lambda_i = p + q_i \quad \text{for all } i, \tag{3.25b}$$
$$p\left(\sum_{i=1}^I \Lambda_i - C\right) = 0 \tag{3.25c}$$
$$q_i(\lambda_i - m_i) = 0 \quad \text{for all } i, \tag{3.25d}$$
$$\sum_{i=1}^I \Lambda_i \leq C, \tag{3.25e}$$
$$\lambda_i \leq m_i \quad \text{for all } i. \tag{3.25f}$$

Underload In this case, there is no interaction between the classes, they do not compete but all get the maximum rate allowed. Indeed, (3.25c) and (3.25b) imply that $p = 0$ and all $q_i > 0$. Then, by (3.25d) and (3.25a), all $\lambda_i = m_i$ and all $\Lambda_i = g_i(m_i)$. Hence, there is a unique invariant NFMS given by

$$z_i^* = g_i(m_i)/m_i \quad \text{for all } i.$$

Critical load First note that

$$\Lambda_i = g_i(m_i) \quad \text{for all } i. \tag{3.26}$$

Indeed, there are two possibilities: either $p = 0 \overset{(3.25b)}{\Rightarrow}$ all $q_i > 0 \overset{(3.25d)}{\Rightarrow}$ all $\lambda_i = m_i \overset{(3.25a)}{\Rightarrow}$ (3.26), or $p > 0 \overset{(3.25c)}{\Rightarrow} \sum_{i=1}^I \Lambda_i = C \Rightarrow$ (3.26), where the last implication is due to $\Lambda_i \le g_i(m_i)$ and $\sum_{i=1}^I g_i(m_i) = C$.

Recall from Lemma 3.7 that

$$\alpha_i := \inf\{x\colon g_i(x) = \rho_i\} = 1/\inf(D_i/B_i).$$

By (3.26), the relations (3.25a) and (3.25f) are equivalent to $m_i \wedge \alpha_i \le \lambda_i \le m_i$ (see Figure 3.2). Hence, (3.25) reduces to

$$\kappa_i/\lambda_i = p + q_i, \tag{3.27a}$$
$$q_i(\lambda_i - m_i) = 0, \tag{3.27b}$$
$$m_i \wedge \alpha_i \le \lambda_i \le m_i. \tag{3.27c}$$

Let

$$\mathcal{I}_{\text{crit}} := \{i\colon m_i \le \alpha_i\}.$$

For $i \in \mathcal{I}_{\text{crit}}$, by (3.27c), we have $\lambda_i = m_i$ and $z_i = g_i(m_i)/m_i$. Then (3.27b) is satisfied, and (3.27a) implies that $p \le \kappa_i/m_i$.

Now divide $\{1,\dots,I\} \setminus \mathcal{I}_{\text{crit}}$ into two subsets $\mathcal{I}_1 \cap \mathcal{I}_2 = \varnothing$. For $i \in \mathcal{I}_1$, put $\lambda_i = m_i$, then (as for $i \in \mathcal{I}_{\text{crit}}$) $z_i^* = g_i(m_i)/m_i$, (3.27b) is satisfied, and (3.27a) implies that $p \le \kappa_i/m_i$. For $i \in \mathcal{I}_2$, assume $\alpha_i \le \lambda_i < m_i$. Then $q_i = 0$ by (3.27b), $\kappa_i/\lambda_i = p$ by (3.27a), and $\kappa_i/m_i < p \le \kappa_i/\alpha_i$. Also $z_i^* = g_i(m_i)p/\kappa_i$.

Summing up everything said above, the set of invariant NFMS's is given by

$$S_{\mathbf{z}^*} := \bigcup\nolimits_{\mathcal{I} \supseteq \mathcal{I}_{\text{crit}}} \left\{ \mathbf{z}^*\colon z_i^* = \begin{cases} g_i(m_i)/m_i, & i \in \mathcal{I}, \\ g_i(m_i)p/\kappa_i, & i \notin \mathcal{I}, p \in S_{p,\mathcal{I}} \end{cases} \right\},$$

where

$$S_{p,\mathcal{I}} = (\max_{i\notin\mathcal{I}} \kappa_i/m_i, \min_{i\in\mathcal{I}} \kappa_i/m_i \wedge \min_{i\notin\mathcal{I}} \kappa_i/\alpha_i]$$

Equivalent descriptions of $S_{\mathbf{z}^*}$ are:

$$S_{\mathbf{z}^*} = \left\{ \mathbf{z}^*\colon \text{for } p \in S_p, z_i^* = \begin{cases} g_i(m_i)/m_i, & p \le \kappa_i/m_i, \\ g_i(m_i)p/\kappa_i, & p > \kappa_i/m_i \end{cases} \right\},$$

where

$$S_p := (0, \min_{i\in\mathcal{I}_{\text{crit}}} \kappa_i/m_i \wedge \min_{i\notin\mathcal{I}_{\text{crit}}} \kappa_i/\alpha_i] = (0, \min_{1\le i\le I} \kappa_i/(m_i \wedge \alpha_i)],$$

and

$$S_{\mathbf{z}^*} = \{\mathbf{z}^*\colon z_i^* = g_i(m_i)/(m_i \wedge \kappa_i x),\ x \in S_x\},$$

where

$$S_x := [\max_{1\le i\le I}(m_i \wedge \alpha_i)/\kappa_i, \infty).$$

We now apply the last formula in a couple of simple examples.

Example 3.2. If $m_i \le \alpha_i$ for all i, then $S_x = [\max_{1\le i\le I} m_i/\kappa_i, \infty)$, and $\kappa_i x \ge m_i$ for all $x \in S_x$ and all i. Hence, there is a unique invariant NFMS given by $z_i^* = g_i(m_i)/m_i$ for all i, which agrees with Theorem 3.2.

Example 3.3. If $m_1 > \alpha_1$, $m_i \le \alpha_i$ for $i \ne 1$ and $\alpha_1/\kappa_1 \ge \max_{i\ne 1} m_i/\kappa_i$, then, for any $\lambda_1 \in [\alpha_1, m_1]$, $\mathbf{z}^* = (g_1(m_1)/\lambda_1, g_2(m_2)/m_2, \dots, g_I(m_I)/m_I)$ is an invariant NFMS.

Overload In this situation, by the capacity constraint (3.25e), at least one class of flows does not receive the maximum service rate, i.e. at least one $\Lambda_i < g_i(m_i)$. We first determine which classes get the maximum service rate and which do not, and then calculate the unique invariant NFMS.

Who gets the maximum service rate. Since at least one $\Lambda_i < g_i(m_i)$, at least one $\lambda_i < m_i \wedge \alpha_i$ (see Figure 3.2). Then (3.25d), (3.25b) and (3.25c) imply that at least one $q_i = 0$, $p > 0$ and $\sum_{i=1}^I \Lambda_i = C$. At this point, we can equivalently rewrite (3.25) as follows: there exist $x > 0$ and $\varepsilon \in \mathbb{R}_+^I$ such that (the functions $\widetilde{g}_i(\cdot)$ are introduced below)

$$\Lambda_i = g_i(\lambda_i) \Leftrightarrow \Lambda_i = \widetilde{g}_i(\lambda_i), \tag{3.28a}$$

$$\textstyle\sum_{i=1}^I g_i(\lambda_i) = C \Leftrightarrow \sum_{i=1}^I \widetilde{g}_i(\lambda_i) = C \tag{3.28b}$$

$$\lambda_i = \kappa_i(x - \varepsilon_i), \tag{3.28c}$$

$$\varepsilon_i(\lambda_i - m_i) = 0, \tag{3.28d}$$

$$\lambda_i \le m_i. \tag{3.28e}$$

For all i and $x \in \mathbb{R}_+$, put

$$\widetilde{g}_i(x) := g_i(m_i \wedge x).$$

By the rate constraints (3.28e), in (3.28), we can equivalently replace $g_i(\cdot)$ by $\widetilde{g}_i(\cdot)$.

If $\lambda_i < m_i$, then, by (3.28d) and (3.28c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$, and hence

$$\widetilde{g}_i(\lambda_i) = \widetilde{g}_i(\kappa_i x). \tag{3.29}$$

If $\lambda_i = m_i$, then, by (3.28c), $\kappa_i x \ge m_i$ and $\widetilde{g}_i(\kappa_i x) = g_i(m_i)$, and, again, (3.29) holds.

Plugging (3.29) into (3.28b), we get

$$\textstyle\sum_{i=1}^I \widetilde{g}_i(\kappa_i x) = C. \tag{3.30}$$

The function $\widetilde{g}(x) := \sum_{i=1}^I \widetilde{g}_i(\kappa_i x)$ is continuous everywhere, strictly increasing in the interval

$$0 \le x \le \max_{1\le i\le I}(m_i \wedge \alpha_i)/\kappa_i =: x_0$$

and constant for $x \geq x_0$, and also $\widetilde{g}(0) = 0$ and $\widetilde{g}(x_0) = \sum_{i=1}^{I} g_i(m_i) > C$. So there exists a unique x solving (3.30) and $x \in (0, x_0)$.

By (3.28a) and (3.29), $\Lambda_i = \widetilde{g}_i(\kappa_i x)$. Then (see Figure 3.3) $\Lambda_i = g_i(m_i)$ if $(m_i \wedge \alpha_i)/\kappa_i \leq x$ and $\Lambda_i < g_i(m_i)$ if $(m_i \wedge \alpha_i)/\kappa_i > x$. Hence, the set of classes that get the maximum service rate is

$$\mathcal{I}_{\text{over}} := \{i\colon (m_i \wedge \alpha_i)/\kappa_i \leq x\}. \tag{3.31}$$

Invariant NFMS. For $i \notin \mathcal{I}_{\text{over}}$, $\Lambda_i = g_i(\lambda_i) < g_i(m_i)$, which implies that (see Figure 3.2) $\lambda_i < m_i \wedge \alpha_i$. Then, by (3.28d) and (3.28c), $\varepsilon_i = 0$ and $\lambda_i = \kappa_i x$ (meeting the rate constraint (3.28e)), and $z_i = \Lambda_i/\lambda_i = g_i(\kappa_i x)/(\kappa_i x)$.

For $i \in \mathcal{I}_{\text{over}}$, consider the two possible cases: $\kappa_i x < m_i$ and $\kappa_i x \geq m_i$. If $\kappa_i x < m_i$, then, by (3.28c) and (3.28d), $\lambda_i \leq \kappa_i x < m_i$ and $\varepsilon_i = 0$, and, again by (3.28c), $\lambda_i = \kappa_i x$. If $\kappa_i x \geq m_i$, then $\lambda_i = m_i$ because otherwise we would arrive at a contradiction: $\lambda_i < m_i \overset{(3.28d)}{\Rightarrow} \varepsilon_i = 0 \overset{(3.28c)}{\Rightarrow} \lambda_i = \kappa_i x \geq m_i$. Hence, for $i \in \mathcal{I}_{\text{over}}$, $\lambda_i = m_i \wedge \kappa_i x$ and $z_i = \Lambda_i/\lambda_i = g_i(m_i)/(m_i \wedge \kappa_i x)$.

Summing up, the unique invariant NFMS is given by

$$z_i^* = \begin{cases} g_i(m_i)/(m_i \wedge \kappa_i x), & i \in \mathcal{I}_{\text{over}} \\ g_i(\kappa_i x)/(\kappa_i x), & i \notin \mathcal{I}_{\text{over}}, \end{cases}$$

where x is the unique solution to (3.30) and $\mathcal{I}_{\text{over}}$ is defined by (3.31).

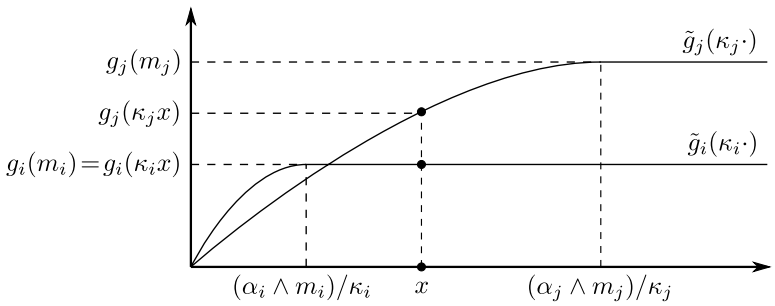

Figure 3.3: Graphs of the functions $\widetilde{g}_i(\cdot)$; x is the unique solution to (3.30).

3.7 Proof of Theorem 3.5

We prove the **C**-tightness of the family of the scaled processes $\overline{\boldsymbol{Z}}^r(\cdot)$ by Proposition 1.3, that is we check the conditions of compact containment (Section 3.7.2) and oscillation control (Section 3.7.4). In Section 3.7.6, we show that fluid limits satisfy the fluid model equation (3.3).

To establish these main steps of the proof, we develop a number of auxiliary results. Section 3.7.1 contains a law of large numbers result for the load process. Section 3.7.3

proves that, for large r, $\overline{\mathcal{Z}}^r(\cdot)$ puts arbitrarily small mass to thin horizontal and vertical strips, which in particular implies that fluid limits have both projections free of atoms. In Section 3.7.5, fluid limits are shown to be coordinate-wise bounded away from zero outside $t = 0$.

Recall that model r is defined on a probability space $(\Omega^r, \mathcal{F}^r, \mathbb{P}^r)$ with expectation operator $\mathbb{E}^r$.

3.7.1 Load process

Introduce the measure-valued load processes: for $t \geq s \geq 0$,

$$\boldsymbol{\mathcal{L}}^r(t) = (\mathcal{L}^r_1, \ldots, \mathcal{L}^r_I)(t), \quad \boldsymbol{\mathcal{L}}^r(s,t) = \boldsymbol{\mathcal{L}}^r(t) - \boldsymbol{\mathcal{L}}^r(s),$$

where, for all i,

$$\mathcal{L}^r_i(t) := \textstyle\sum_{k=1}^{E^r_i(t)} \delta_{(B^r_{i,k}, D^r_{i,k})}.$$

The corresponding fluid scaled versions are: for $t \geq s \geq 0$,

$$\overline{\boldsymbol{\mathcal{L}}}^r(t) := \boldsymbol{\mathcal{L}}^r(rt)/r, \quad \overline{\boldsymbol{\mathcal{L}}}^r(s,t) := \overline{\boldsymbol{\mathcal{L}}}^r(t) - \overline{\boldsymbol{\mathcal{L}}}^r(s).$$

The following property is useful when proving other results of the section. Only minor adjustments in the proof of Gromoll and Williams [46, Theorem 5.1] are needed to establish it.

Lemma 3.8. *By Assumptions 3.3 and 3.4, as $r \to \infty$,*

$$(\overline{\boldsymbol{\mathcal{L}}}^r(\cdot), \langle \chi_1, \overline{\boldsymbol{\mathcal{L}}}^r(\cdot)\rangle, \langle \chi_2, \overline{\boldsymbol{\mathcal{L}}}^r(\cdot)\rangle) \Rightarrow (\boldsymbol{\eta}(\cdot) \times \boldsymbol{\theta}, \boldsymbol{\rho}(\cdot), \boldsymbol{\sigma}(\cdot)),$$

where $\chi_1(x_1, x_2) := x_1$, $\chi_2(x_1, x_2) := x_2$, and $\boldsymbol{\eta}(t) := t\boldsymbol{\eta}$, $\boldsymbol{\rho}(t) := t\boldsymbol{\rho}$, $\boldsymbol{\sigma}(t) := t\boldsymbol{\sigma}$.

3.7.2 Compact containment

The property we prove here, together with the oscillation control result that follows in Section 3.7.4, implies **C**-tightness of the scaled state descriptors.

Lemma 3.9. *By Assumptions 3.3–3.5, for any $T > 0$ and $\varepsilon > 0$, there exists a compact set $K \subset \mathcal{M}^I$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\boldsymbol{\mathcal{Z}}}^r(t) \in K \text{ for all } t \in [0,T]\} \geq 1 - \varepsilon.$$

Proof. Fix T and ε. It suffices to show that, for each i, there exist a compact set $K_i \subset \mathcal{M}$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}^r_i(t) \in K_i \text{ for all } t \in [0,T]\} \geq 1 - \varepsilon/I. \tag{3.32}$$

We use the following criterion (see e.g. Kallenberg [52, Theorem 15.7.5]).

Proposition 3.1. *A set $K \subset \mathcal{M}$ is relatively compact if and only if* $\sup_{\xi\in K} \xi(\mathbb{R}^2_+) < \infty$ *and* $\sup_{\xi\in K} \xi(\mathbb{R}^2_+ \setminus [0,n]^2) \to 0$ *as* $n \to \infty$.

Note that

$$\overline{\mathcal{Z}}^r_i(t)(\mathbb{R}^2_+) = \overline{Z}^r_i(t) \le \overline{Z}^r_i(0) + \overline{E}^r_i(T) = \overline{\mathcal{Z}}^r_i(0)(\mathbb{R}^2_+) + \overline{\mathcal{L}}^r_i(T)(\mathbb{R}^2_+). \tag{3.33}$$

Also note that, if the residual size (patience time) of a flow at time t exceeds n, then its initial size (patience time), must have exceeded n, too, which implies the following bound:

$$\overline{\mathcal{Z}}^r_i(t)(\mathbb{R}^2_+ \setminus [0,n]^2) \le \overline{\mathcal{Z}}^r_i(0)(\mathbb{R}^2_+ \setminus [0,n]^2) + \overline{\mathcal{L}}^r_i(T)(\mathbb{R}^2_+ \setminus [0,n]^2). \tag{3.34}$$

The family of $\overline{\mathcal{Z}}^r_i(0) + \overline{\mathcal{L}}^r_i(T)$ converges as $r \to \infty$ and hence in tight, i.e. there exists a compact set $K'_i \subset \mathcal{M}$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\mathcal{Z}}^r_i(0) + \overline{\mathcal{L}}^r_i(T) \in K'_i\} \ge 1 - \varepsilon/I. \tag{3.35}$$

Put

$$\begin{aligned} K''_i := \{\xi \in \mathcal{M} &: \text{for some } \xi' \in K'_i,\ \xi(\mathbb{R}^2_+) \le \xi'(\mathbb{R}^2_+) \\ &\text{and } \xi(\mathbb{R}^2_+ \setminus [0,n]^2) \le \xi'(\mathbb{R}^2_+ \setminus [0,n]^2),\ n \in \mathbb{N}\}. \end{aligned}$$

Then the criterion of relative compactness for K''_i follows from that for K'_i, and (3.33)–(3.35) imply (3.32) with K_i taken as the closure of K''_i. □

3.7.3 Asymptotic regularity

This section contains three Lemmas. Lemmas 3.10 and 3.11 prove that neither initial nor newly arriving flows concentrate along horizontal and vertical lines. These two results are combined in Lemma 3.12 which implies the oscillation control result of the next section, and also is useful when deriving the limit equations for the state descriptors in Section 3.7.6.

Recall from Section 3.5 that, for $b \ge a \ge 0$,

$$H^b_a = \mathbb{R}_+ \times [a,b], \quad V^b_a = [a,b] \times \mathbb{R}_+,$$

and introduce similar notations

$$H^\infty_a := \mathbb{R}_+ \times [a,\infty), \quad V^\infty_a := [a,\infty) \times \mathbb{R}_+.$$

Lemma 3.10. *By Assumption 3.5, for any $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\sup_{x\in\mathbb{R}_+} \|\overline{\boldsymbol{\mathcal{Z}}}^r(0)(H^{x+a}_x)\| \vee \|\overline{\boldsymbol{\mathcal{Z}}}^r(0)(V^{x+a}_x)\| \le \delta\} \ge 1 - \varepsilon.$$

Proof. Fix δ and ε. Since, for any $\boldsymbol{\xi} \in \mathcal{M}^I$ and $a > 0$,

$$\sup_{x\in\mathbb{R}_+} \|\boldsymbol{\xi}(H_x^{x+a})\| \vee \|\boldsymbol{\xi}(V_x^{x+a})\| \le 2 \sup_{n\in\mathbb{N}} \|\boldsymbol{\xi}(H_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}(V_{(n-1)a}^{na})\|,$$

it suffices to find an a such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\boldsymbol{\mathcal{Z}}}^r(0) \in \mathcal{M}_a^I\} \ge 1 - \varepsilon,$$

where

$$\mathcal{M}_a^I := \{\boldsymbol{\xi} \in \mathcal{M}^I \colon \sup_{n\in\mathbb{N}} \|\boldsymbol{\xi}(H_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}(V_{(n-1)a}^{na})\| < \delta/2\}.$$

The set $\mathcal{M}_a^I$ is open because $\boldsymbol{\xi}^k \xrightarrow{\mathrm{w}} \boldsymbol{\xi} \in \mathcal{M}_a^I$ implies that $\boldsymbol{\xi}^k \in \mathcal{M}_a^I$ for k large enough. Indeed, pick an $N \in \mathbb{N}$ such that $\|\boldsymbol{\xi}(H_{Na}^{\infty})\| \vee \|\boldsymbol{\xi}(V_{Na}^{\infty})\| < \delta/2$. Then, by the Portmanteau theorem,

$$\begin{aligned}
&\overline{\lim}_{k\to\infty} \sup_{n\in\mathbb{N}} \|\boldsymbol{\xi}^k(H_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}^k(V_{(n-1)a}^{na})\| \\
&\le \overline{\lim}_{k\to\infty} \max_{1\le n\le N} \|\boldsymbol{\xi}^k(H_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}^k(V_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}^k(H_{Na}^{\infty})\| \vee \|\boldsymbol{\xi}^k(V_{Na}^{\infty})\| \\
&\le \max_{1\le n\le N} \|\boldsymbol{\xi}(H_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}(V_{(n-1)a}^{na})\| \vee \|\boldsymbol{\xi}(H_{Na}^{\infty})\| \vee \|\boldsymbol{\xi}(V_{Na}^{\infty})\| < \delta/2.
\end{aligned}$$

By Assumption 3.5 and [46, Lemma A.1], there exists an a such that $\mathbb{P}\{\boldsymbol{\zeta}^0 \in \mathcal{M}_a^I\} \ge 1 - \varepsilon$. Then, again by the Portmanteau theorem,

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\overline{\boldsymbol{\mathcal{Z}}}^r(0) \in \mathcal{M}_a^I\} \ge \mathbb{P}\{\boldsymbol{\zeta}^0 \in \mathcal{M}_a^I\} \ge 1 - \varepsilon. \qquad \square$$

Besides being used in the proof of the fluid limit theorem, the following result is also used when establishing convergence of the stationary distributions of the scaled state descriptors, see Section 3.5.

Lemma 3.11. *By Assumptions 3.3 and 3.4, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\begin{aligned}
\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} &\|\overline{\boldsymbol{\mathcal{Z}}}^{r,new}(t)(H_x^{x+a})\| \\
&\vee \|\overline{\boldsymbol{\mathcal{Z}}}^{r,new}(t)(V_x^{x+a})\| \le \delta\} \ge 1 - \varepsilon.
\end{aligned}$$

Proof. Fix T, δ and ε. Denote the event in the last equation by Ω_*^r. We first construct auxiliary events Ω_0^r such that $\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\Omega_0^r\} \ge 1 - \varepsilon$, and then show that $\Omega_*^r \supseteq \Omega_0^r$ for all r, which implies the theorem.

Definition of Ω_0^r. By Lemma 3.9, there exists a compact set $K \subset \mathcal{M}^I$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\underbrace{\overline{\boldsymbol{\mathcal{Z}}}^r(t) \in K \text{ for all } t \in [0,T]}_{=:\,\Omega_1^r}\} \ge 1 - \varepsilon,$$

and by Proposition 3.1, $M := \sup_{\boldsymbol{\xi}\in K} \|\boldsymbol{\xi}(\mathbb{R}_+^2)\| < \infty$ and $\sup_{\boldsymbol{\xi}\in K} \|\boldsymbol{\xi}(\mathbb{R}_+^2) \setminus [0,L]^2\| \le \delta/4$ for a large enough L.

For each i, the rate function $\lambda_i(\cdot)$ is positive on $\{\mathbf{z} \in \mathbb{R}_+^I : z_i > 0\}$ and, by Lemma 3.1, it is continuous there. Hence,

$$\lambda_* := \min_{1\le i\le I} \inf\{\lambda_i(\mathbf{z}) : z_i \ge \delta/4,\ \|\mathbf{z}\| \le M\} > 0. \tag{3.36}$$

Put

$$\gamma := \frac{\delta}{72\|\boldsymbol{\eta}\|} \wedge T \quad \text{and} \quad a := \frac{\gamma(\lambda_* \wedge 1)}{3}.$$

Also pick an N large enough so that

$$Na > L + (\|\mathbf{m}\| \vee 1)T.$$

For $n_1, n_2 \in \mathbb{N}$, define the sets

$$I_{n_1,n_2} := [(n_1-1)a, n_1a) \times [(n_2-1)a, n_2a),$$
$$I^{n_1,n_2} := [(n_1-2)^+a, (n_1+1)a) \times [(n_2-2)^+a, (n_2+1)a),$$

and pick functions $g_{n_1,n_2} \in \mathbf{C}(\mathbb{R}_+^2, [0,1])$ such that

$$\mathbb{I}_{I_{n_1,n_2}}(\cdot) \le g_{n_1,n_2}(\cdot) \le \mathbb{I}_{I^{n_1,n_2}}(\cdot).$$

Since $\boldsymbol{\theta}$ is a vector of probability measures,

$$\sum_{n_1,n_2\in\mathbb{N}} \|\langle g_{n_1,n_2}, \boldsymbol{\theta}\rangle\| \le \|\sum_{n_1,n_2\in\mathbb{N}} \boldsymbol{\theta}(I^{n_1,n_2})\| \le 9. \tag{3.37}$$

By Lemma 3.8 and the continuous mapping theorem, for all $n_1, n_2 \in \mathbb{N}$, $\langle g_{n_1,n_2}, \overline{\boldsymbol{\mathcal{L}}}^r(\cdot)\rangle \Rightarrow \boldsymbol{\eta}(\cdot)\langle g_{n_1,n_2}, \boldsymbol{\theta}\rangle$ as $r \to \infty$. Since the limits are deterministic, we have convergence in probability. Since the limits are continuous, we have uniform convergence on compact sets. Hence,

$$\lim_{r\to\infty} \mathbb{P}^r\{\max_{1\le n_1,n_2\le N} \sup_{t\in[0,T]} \|\langle g_{n_1,n_2}, \overline{\boldsymbol{\mathcal{L}}}^r(t)\rangle - t\boldsymbol{\eta} \times \langle g_{n_1,n_2}, \boldsymbol{\theta}\rangle\| \le \delta/(16N^2)\} = 1.$$

We denote the event in the last equation by Ω_2^r.

Similarly, by Assumption 3.3,

$$\lim_{r\to\infty} \mathbb{P}^r \underbrace{\{\sup_{t\in[0,T]} \|\overline{\mathbf{E}}^r(t) - t\boldsymbol{\eta}\| \le \delta/16\}}_{=:\,\Omega_3^r} = 1.$$

For all r, put

$$\Omega_0^r := \Omega_1^r \cap \Omega_2^r \cap \Omega_3^r,$$

then $\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\Omega_0^r\} \ge 1 - \varepsilon$, and it is left to show that $\Omega_0^r \subseteq \Omega_*^r$.

Proof of $\Omega_0^r \subseteq \Omega_*^r$. Fix an r, $t \in [0,T]$, $x \in \mathbb{R}_+$ and i. Also fix an outcome $\omega \in \Omega_0^r$. All random objects in the rest of the proof will be evaluated at this ω. We have to check that

$$\overline{\mathcal{Z}}_i^{r,\mathrm{new}}(t)(H_x^{x+a}) \le \delta, \tag{3.38a}$$
$$\overline{\mathcal{Z}}_i^{r,\mathrm{new}}(t)(V_x^{x+a}) \le \delta. \tag{3.38b}$$

We will show (3.38a), (3.38b) follows similarly.

Define the random time $\tau := \sup\{s \le t\colon \overline{Z}_i^{r,\mathrm{new}}(s) < \delta/4\}$ (supremum over the empty set equals 0 by convention). Although in general τ is not a continuity point for $\overline{Z}_i^{r,\mathrm{new}}(\cdot)$, we still can estimate $\overline{Z}_i^{r,\mathrm{new}}(\tau)$:

$$\overline{Z}_i^{r,\mathrm{new}}(\tau) \le \delta/2. \tag{3.39}$$

Indeed, if $\tau = 0$, then $\overline{Z}_i^{r,\mathrm{new}}(\tau) = 0$, and (3.39) holds. If $\tau > 0$, pick a $\widetilde{\tau} \in [(\tau-\gamma)^+, \tau]$ such that $\overline{Z}_i^{r,\mathrm{new}}(\widetilde{\tau}) < \delta/4$. Then, by the definition of Ω_3^r,

$$\begin{aligned}\overline{Z}_i^{r,\mathrm{new}}(\tau) &\le \overline{Z}_i^{r,\mathrm{new}}(\widetilde{\tau}) + (\overline{E}_i^r(\tau) - \overline{E}_i^r(\widetilde{\tau}))\\ &\le \delta/4 + \eta_i(\tau-\widetilde{\tau}) + \delta/8 \le \|\boldsymbol{\eta}\|\gamma + 3\delta/8,\end{aligned}$$

and (3.39) holds by the choice of γ.

Now, if $\tau = t$, then (3.39) implies (3.38a), and the proof is finished. Assume that $\tau < t$. Then, by the choice of L and (3.39),

$$\begin{aligned}\overline{\mathcal{Z}}_i^{r,\mathrm{new}}(t)(H_x^{x+a}) &\le \overline{\mathcal{Z}}_i^{r,\mathrm{new}}(t)(H_x^{x+a}\cap[0,L]^2) + \delta/4\\ &\le \underbrace{\overline{Z}_i^{r,\mathrm{new}}(\tau)}_{\le \delta/2} + \frac{1}{r}\sum\nolimits_{E_i^r(\tau)+1}^{E_i^r(t)} s_k + \delta/4,\end{aligned}$$

where

$$s_k := \mathbb{I}_{H_x^{x+a}\cap[0,L]^2}(B_{i,k}^r - S_i(\overline{\mathbf{Z}}^r, U_{i,k}^r, t), D_{i,k}^r - (t - U_{i,k}^r)).$$

At this point, in order to have (3.38a), it suffices to show that

$$\begin{aligned}\Sigma &:= \frac{1}{r}\sum\nolimits_{E_i^r(\tau)+1}^{E_i^r(t)} s_k\\ &= \sum\nolimits_{n_1,n_2\in\mathbb{N}} \underbrace{\frac{1}{r}\sum\nolimits_{E_i^r(\tau)+1}^{E_i^r(t)} s_k \mathbb{I}_{I_{n_1,n_2}}(B_{i,k}^r, D_{i,k}^r)}_{=:\,\Sigma_{n_1,n_2}} \le \delta/4.\end{aligned} \tag{3.40}$$

First note that

$$\Sigma_{n_1,n_2} = 0 \quad \text{if } n_1 > N \text{ or } n_2 > N. \tag{3.41}$$

Indeed, consider a flow on route i that arrived at $U_{i,k}^r \in (\tau, t]$ with $(B_{i,k}^r, D_{i,k}^r) \in I_{n_1,n_2}$. If $n_1 > N$, then $B_{i,k}^r > L + \|\mathbf{m}\|T$ by the choice of N, $B_{i,k}^r - S_i(\overline{\mathbf{Z}}^r, U_{i,k}^r, t) > L$ by the rate constraints, and $s_k = 0$. If $n_2 > N$, then $D_{i,k}^r > L + T$ by the choice of N, $D_{i,k}^r - (t - U_{i,k}^r) > L$ and again $s_k = 0$.

Now we estimate Σ_{n_1,n_2} for $1 \le n_1, n_2 \le N$. Fix n_1, n_2. Consider two flows $k < l$ such that $U^r_{i,k}, U^r_{i,l} \in (\tau, t]$ and $(B^r_{i,k}, D^r_{i,k}), (B^r_{i,l}, D^r_{i,l}) \in I_{n_1,n_2}$. In $(\tau, t]$, $\overline{Z}^r_i(\cdot) \ge \overline{Z}^{r,\mathrm{new}}_i(\cdot) \ge \delta/4$ and $\|\overline{\mathbf{Z}}^r(\cdot)\| \le M$, and then (3.36) implies that

$$\inf_{s\in(\tau,t]} \lambda_i(\overline{\mathbf{Z}}^r(s)) \ge \lambda_*.$$

If $U^r_{i,l} - U^r_{i,k} \ge \gamma$, then

$$(B^r_{i,l} - S_i(\overline{\mathbf{Z}}^r, U^r_{i,l}, t)) - (B^r_{i,k} - S_i(\overline{\mathbf{Z}}^r, U^r_{i,k}, t)) \ge \underbrace{\gamma\lambda_*}_{\ge 3a} - \underbrace{(B^r_{i,k} - B^r_{i,l})}_{\le a} \ge 2a,$$

and

$$(D^r_{i,l} - (t - U^r_{i,l}) - (D^r_{i,k} - (t - U^r_{i,k})) \ge \underbrace{\gamma}_{\ge 3a} - \underbrace{(D^r_{i,k} - D^r_{i,l})}_{\le a} \ge 2a,$$

Hence, at most one of s_k and s_l is non-zero. This implies that all arrivals to route i during $(\tau, t]$ that correspond to non-zero summands in Σ_{n_1,n_2} occur actually during a smaller interval $(t_{n_1,n_2}, t_{n_1,n_2} + \gamma] \subseteq (\tau, t]$. Then, by the definition of Ω^r_2,

$$\begin{aligned}
\Sigma_{n_1,n_2} &\le \frac{1}{r} \sum_{k=E^r_i(t_{n_1,n_2})+1}^{E^r_i(t_{n_1,n_2}+\gamma)} \mathbb{I}_{I_{n_1,n_2}}(B^r_{i,k}, D^r_{i,k}) \\
&\le \sup_{s\in[0,T-\gamma]} \frac{1}{r} \sum_{k=E^r_i(s)+1}^{E^r_i(s+\gamma)} g_{n_1,n_2}(B^r_{i,k}, D^r_{i,k}) \\
&= \sup_{s\in[0,T-\gamma]} (\langle g_{n_1,n_2}, \overline{\mathcal{L}}^r_i(s+\gamma)\rangle - \langle g_{n_1,n_2}, \overline{\mathcal{L}}^r_i(s)\rangle) \\
&\le \gamma\eta_i \langle g_{n_1,n_2}, \theta_i\rangle + \delta/(8N^2).
\end{aligned}$$

We plug the last inequality and (3.41) into $\Sigma = \sum_{n_1,n_2\in\mathbb{N}} \Sigma_{n_1,n_2}$, then (3.40) follows by (3.37) and the choice of γ. □

The previous two lemmas are summed up into the following result.

Lemma 3.12. *By Assumptions 3.3–3.5, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $a > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\{\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \|\overline{\boldsymbol{\mathcal{Z}}}^r(t)(H^{x+a}_x)\| \vee \|\overline{\boldsymbol{\mathcal{Z}}}^r(t)(V^{x+a}_x)\| \le \delta\} \ge 1 - \varepsilon.$$

Proof. Note that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}_+} \|\overline{\boldsymbol{\mathcal{Z}}}^{r,\mathrm{init}}(t)(H^{x+a}_x)\| \vee \|\overline{\boldsymbol{\mathcal{Z}}}^{r,\mathrm{init}}(t)(V^{x+a}_x)\| \\
&\le \sup_{x\in\mathbb{R}_+} \|\overline{\boldsymbol{\mathcal{Z}}}^r(0)(H^{x+a}_x)\| \vee \|\overline{\boldsymbol{\mathcal{Z}}}^r(0)(V^{x+a}_x)\|.
\end{aligned}$$

Indeed,

$$\overline{\mathcal{Z}}_i^{r,\,\mathrm{init}}(t)(H_x^{x+a}) \le \overline{\mathcal{Z}}_i^r(0)(H_{x+t}^{x+a+t})$$

and

$$\overline{\mathcal{Z}}_i^{r,\,\mathrm{init}}(t)(V_x^{x+a}) \le \overline{\mathcal{Z}}_i^r(0)\left(V_{x+S_i(\overline{\mathbf{Z}}^r,0,t)}^{x+a+S_i(\overline{\mathbf{Z}}^r,0,t)}\right).$$

Then the lemma follows by the representation $\overline{\boldsymbol{\mathcal{Z}}}^r(\cdot) = (\overline{\boldsymbol{\mathcal{Z}}}^{r,\,\mathrm{init}} + \overline{\boldsymbol{\mathcal{Z}}}^{r,\,\mathrm{new}})(\cdot)$ and Lemmas 3.10 and 3.11. □

3.7.4 Oscillation control

Here we establish the second key ingredient of tightness of the scaled state descriptors, the first one is proven in Section 3.7.2.

Lemma 3.13. *By Assumptions 3.3–3.5, for any $T > 0$, $\delta > 0$ and $\varepsilon > 0$, there exists an $h > 0$ such that*

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\underbrace{\{\omega(\overline{\boldsymbol{\mathcal{Z}}}^r, h, T) \le \delta\}}_{=:\,\Omega_*^r} \ge 1-\varepsilon,$$

where $\omega(\overline{\boldsymbol{\mathcal{Z}}}^r, h, T) := \sup\{d_{\mathcal{M}^I}(\overline{\boldsymbol{\mathcal{Z}}}^r(s), \overline{\boldsymbol{\mathcal{Z}}}^r(t)) \colon s,t \in [0,T],\ |s-t| < h\}$.

Proof. Fix T, δ and ε. By Assumption 3.3,

$$\lim_{r\to\infty} \mathbb{P}^r\underbrace{\{\sup_{t\in[0,T]} \|\overline{\mathbf{E}}^r(t) - t\boldsymbol{\eta}\| \le \delta/4\}}_{=:\,\Omega_1^r} = 1.$$

By Lemma 3.12, there exists an $a > 0$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}^r\underbrace{\{\sup_{t\in[0,T]} \|\overline{\boldsymbol{\mathcal{Z}}}^r(t)(H_0^a \cup V_0^a)\| \le \delta\}}_{=:\,\Omega_2^r} \ge 1-\varepsilon.$$

Pick an h such that $h(\|\mathbf{m}\| \vee 1) \le \delta \wedge a$ and $h\|\boldsymbol{\eta}\| \le \delta/2$. We now show that, for all r, $\Omega_*^r \supseteq \Omega_1^r \cap \Omega_2^r$, and then the lemma follows.

Fix an r, i and $s,t \in [0,T]$ such that $s < t$, $t - s < h$. Also fix an outcome $\omega \in \Omega_1^r \cap \Omega_2^r$. All random objects in the rest of the proof will be evaluated at this ω. We have to check that, for any non-empty closed Borel subset $B \subseteq \mathbb{R}_+^2$,

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(t)(B^\delta) + \delta, \tag{3.42a}$$

$$\overline{\mathcal{Z}}_i^r(t)(B) \le \overline{\mathcal{Z}}_i^r(s)(B^\delta) + \delta. \tag{3.42b}$$

First we check (3.42a). Note that it suffices to show

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(\tau)(B^\delta) + \delta, \tag{3.43}$$

where $\tau := \inf\{u \in [s,t] \colon \overline{Z}_i^r(u) = 0\}$ and infimum over the empty set equals t by definition. Indeed, if $\tau = t$, then (3.43) implies (3.42a). If $\tau < t$, then by the right-continuity of $\overline{Z}_i^r(\cdot)$, $\overline{\mathcal{Z}}_i^r(\tau)(B^\delta) = \overline{Z}_i^r(\tau) = 0$, and again (3.43) implies (3.42a).

Now we prove (3.43). If $\tau = s$, then (3.43) holds. Assume that $\tau > s$. By the definition of Ω_2^r,

$$\overline{\mathcal{Z}}_i^r(s)(B) \le \overline{\mathcal{Z}}_i^r(s)(B \cap [a,\infty)^2) + \delta. \tag{3.44}$$

Since $S_i(\overline{\mathbf{Z}}^r, s, \tau) < \|\mathbf{m}\|h \le \delta \wedge a$ and $\tau - s < h \le \delta \wedge a$,

$$\overline{\mathcal{Z}}_i^r(s)(B \cap [a,\infty)^2) \le \overline{\mathcal{Z}}_i^r(\tau)(B^\delta),$$

which together with (3.44) implies (3.43).

It is left to check (3.42b). Since $S_i(\overline{\mathbf{Z}}^r, s, \tau) < \|\mathbf{m}\|h \le \delta$ and $\tau - s < h \le \delta$,

$$\overline{\mathcal{Z}}_i^r(t)(B) \le \overline{\mathcal{Z}}_i^r(s)(B^\delta) + (\overline{E}_i^r(t) - \overline{E}_i^r(s)),$$

and (3.42b) follows by the definition of Ω_1^r. □

3.7.5 Fluid limits are bounded away from zero

Rate constraints provide infinite-server-queue lower bounds for bandwidth-sharing networks. First we show that properly scaled infinite-server queues are bounded away from zero, and then the same follows for bandwidth-sharing networks with rate constraints.

Consider a sequence of $G/G/\infty$ queues (see Remark 1.2) labeled by positive numbers r. At $t = 0$, the queues are empty. To the r-th queue, customers arrive according to a counting process $A^r(\cdot)$ and have i.i.d. service times $\{B_k^r\}_{k\in\mathbb{N}}$ distributed as B^r. Let $\overline{A}^r(\cdot) := A^r(\cdot)/r \Rightarrow \alpha(\cdot)$, where $\alpha(t) := t\alpha$ and $\alpha > 0$. Also let $B^r \Rightarrow B$, where $\mathbb{P}\{B > 0\} > 0$. Denote by $Q^r(\cdot)$ the population process of the r-th queue and put $\overline{Q}^r(\cdot) := Q^r(\cdot)/r$.

Lemma 3.14. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,*

$$\mathbb{P}^r\{\inf_{\delta \le t \le \Delta} \overline{Q}^r(t) \ge C(\delta)\} \to 1 \quad \text{as } r \to \infty.$$

Proof. Let us first explain the result heuristically. Consider the arrivals with long service times, i.e. exceeding a $b > 0$. During $(0, b/2]$, there are $r\alpha\mathbb{P}\{B > b\}b/2$ such arrivals to the r-th queue. They will leave the queue after $t = b$, and hence, in $(b/2, b]$, the scaled queue length $\overline{Q}^r(\cdot)$ is bounded from below by $\alpha\mathbb{P}\{B > b\}b/2$. Similarly, $\overline{Q}^r(\cdot) \ge \alpha\mathbb{P}\{B > b\}b/2$ in any interval $((n-1)b/2, nb/2]$, $n \in \mathbb{N}$.

We now proceed more formally. Pick a $b \in (0, \delta)$ such that b is a continuity point for the distribution of B, and

$$p := \mathbb{P}\{B \geq b\} > 0.$$

Then, as $r \to \infty$,

$$p_r := \mathbb{P}^r\{B^r \geq b\} \to p.$$

Partition $(0, \Delta]$ into subintervals of length $b/2$,

$$(0, \Delta] \subseteq \bigcup\nolimits_{1 \leq n \leq N(\Delta)} ((n-1)b/2, nb/2].$$

Denote by $\overline{A}^r_n$ the scaled number of arrivals during $((n-1)b/2, nb/2]$, and by $\overline{A}^r_n(b)$ the scaled number of arrivals during $((n-1)b/2, nb/2]$ with service times at least b,

$$\overline{A}^r_n := \overline{A}^r(nb/2) - \overline{A}^r((n-1)b/2),$$
$$\overline{A}^r_n(b) := \frac{1}{r} \sum\nolimits_{k=A^r((n-1)b/2)+1}^{A^r(nb/2)} \mathbb{I}\{B^r_k \geq b\}.$$

By $\overline{A}^r \Rightarrow \alpha(\cdot)$ and $p_r \to p$ as $r \to \infty$,

$$(\overline{A}^r_1, \ldots, \overline{A}^r_{N(\Delta)}) \Rightarrow (\alpha b/2, \ldots, \alpha b/2),$$
$$(\overline{A}^r_1(b), \ldots, \overline{A}^r_{N(\Delta)}(b)) \Rightarrow (\alpha p b/2, \ldots, \alpha p b/2).$$

Pick a $C(\delta) < \alpha p b/2$, then

$$\begin{aligned}
&\mathbb{P}^r\{\inf\nolimits_{\delta \leq t \leq \Delta} \overline{Q}^r(t) \geq C(\delta)\} \\
&\geq \mathbb{P}^r\{\inf\nolimits_{t \in ((n-1)b/2, nb/2]} \overline{Q}^r(t) \geq C(\delta),\ n = 2, \ldots, N(\Delta)\} \\
&\geq \mathbb{P}^r\{\overline{A}^r_n(b) \geq C(\delta),\ n = 1, \ldots, N(\Delta) - 1\} \to 1 \quad \text{as } r \to \infty.
\end{aligned}$$

□

We can now prove easily that all fluid limits are bounded away from zero outside $t = 0$.

Lemma 3.15. *For any $\delta > 0$, there exists a $C(\delta) > 0$ such that, for any fluid limit $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$,*

$$\text{a.s.} \quad \inf\nolimits_{t \geq \delta} \min\nolimits_{1 \leq i \leq I} Z_i(t) \geq C(\delta).$$

Proof. Consider a flow k on route i in the r-th network. By the rate constraints, this flow will stay in the network at least for $B^r_{i,k}/m_i \wedge D^r_{i,k}$ after its arrival. Hence, the route i population process $Z^r_i(\cdot)$ is bounded from below by the length $Q^r_i(\cdot)$ of the infinite server queue with arrivals $E^r_i(\cdot)$ and i.i.d. service times $\{B^r_{i,k}/m_i \wedge D^r_{i,k}\}_{k \in \mathbb{N}}$. Assume that $Q^r_i(0) = 0$ and, as before, $\overline{Q}^r_i(\cdot) = Q^r_i(\cdot)/r$. Then, by Lemma 3.14, for any $\delta > 0$ there exists a $C(\delta) > 0$ such that, for any $\Delta > \delta$,

$$\begin{aligned}
&\mathbb{P}^r\{\inf\nolimits_{t \in [\delta, \Delta]} \min\nolimits_{1 \leq i \leq I} \overline{Z}^r_i(t) \geq C(\delta)\} \\
&\geq \mathbb{P}^r\{\inf\nolimits_{t \in [\delta, \Delta]} \min\nolimits_{1 \leq i \leq I} \overline{Q}^r_i(t) \geq C(\delta)\} \to 1 \quad \text{as } r \to \infty.
\end{aligned}$$

Now consider a fluid limit $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$ along a sequence $\{(\overline{\boldsymbol{\mathcal{Z}}}^q, \overline{\mathbf{Z}}^q)(\cdot)\}_{q\to\infty}$. For any compact set $K \subset \mathbb{R}_+$, the mapping $\varphi_K : \mathbf{D}(\mathbb{R}_+, \mathbb{R}^I) \to \mathbb{R}$, $\varphi_K(\mathbf{x}) := \inf_{t\in K} \min_{1\le i\le I} x_i(t)$ is continuous at continuous $\mathbf{x}(\cdot)$. Hence, $\varphi_{[\delta,\Delta]}(\overline{\mathbf{Z}}^q) \Rightarrow \varphi_{[\delta,\Delta]}(\mathbf{Z})$ as $q \to \infty$ and, by the Portmanteau theorem,

$$\mathbb{P}\{\varphi_{[\delta,\Delta]}(\mathbf{Z}) \ge C(\delta)\} \ge \overline{\lim}_{q\to\infty} \mathbb{P}^q\{\varphi_{[\delta,\Delta]}(\overline{\mathbf{Z}}^q) \ge C(\delta)\} = 1,$$

where $\Delta > \delta$ is arbitrary. Then the lemma follows.

Note also that $C(\delta)$ does not depend on a particular fluid limit $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$. □

3.7.6 Fluid limits as FMS's

Here we show that fluid limits a.s. satisfy the fluid model equation (3.3).

Let $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$ be a fluid limit along a sequence $\{(\overline{\boldsymbol{\mathcal{Z}}}^q, \overline{\mathbf{Z}}^q)(\cdot)\}_{q\to\infty}$. Lemma 3.12 implies that (*cf.* the proof of Gromoll et al. [47, Lemma 6.2])

$$\text{a.s., for all } t \in \mathbb{R}_+, \text{ all } i \text{ and } A \in \mathcal{C}, \quad \mathcal{Z}_i(t)(\partial_A) = 0, \tag{3.45}$$

where ∂_A denotes the boundary of A. Then, when proving (3.3) for $(\boldsymbol{\mathcal{Z}}, \mathbf{Z})(\cdot)$, it suffices to consider sets A from

$$\mathcal{C}^+ := \{[x,\infty) \times [y,\infty) \colon x \wedge y > 0\}.$$

It also suffices to consider t from a finite interval $[0, T]$.

The rest of the proof splits into two parts. First we derive dynamic equations for the prelimit processes $(\overline{\boldsymbol{\mathcal{Z}}}^q, \overline{\mathbf{Z}}^q)(\cdot)$, and then show that these equations converge to (3.3).

Prelimit equations Fix a q and $t \in [0, T]$. Also fix a coordinate i and a set $A \in \mathcal{C}^+$. What follows up to and including equation (3.48), holds for all possible outcomes $w \in \Omega^q$ of the probability space $(\Omega^q, \mathcal{F}^q, \mathbb{P}^q)$ on which system q is defined.

We have

$$\begin{aligned}\overline{\mathcal{Z}}_i^q(t)(A) = \overline{\mathcal{Z}}_i^q(0)(A + (S_i(\overline{\mathbf{Z}}^q, 0, t), t)) \\ + \underbrace{\frac{1}{q}\sum_{k=1}^{E_i^q(t)} \overbrace{\mathbb{I}_A(B_{i,k}^q - S_i(\overline{\mathbf{Z}}^q, U_{i,k}^q, t), D_{i,k}^q - (t - U_{i,k}^q))}^{=:\, s_k}}_{=:\, \Sigma}.\end{aligned} \tag{3.46}$$

Fix a partition $0 < t_0 < t_1 < ... < t_N = t$, then

$$\Sigma = \frac{1}{q}\sum_{k=1}^{E_i^q(t_0)} s_k + \frac{1}{q}\sum_{j=0}^{N-1} \sum_{k=E_i^q(t_j)+1}^{E_i^q(t_{j+1})} s_k.$$

Suppose that a function $y(\cdot)$ is non-increasing in $[t_0, t]$ and that, for some δ,

$$\sup_{s\in[t_0,t]} |S_i(\overline{\mathbf{Z}}^q, s, t) - y(s)| \le \delta.$$

Now we can estimate Σ. If $U^q_{ik} \in (t_j, t_{j+1}]$, then

$$B^q_{i,k} - (y(t_j) + \delta) \le B^q_{i,k} - S(\overline{\mathbf{Z}}^q, U^q_{i,k}, t) \le B^q_{i,k} - (y(t_{j+1}) - \delta),$$
$$D^q_{i,k} - (t - t_j) \le D^q_{i,k} - (t - U^q_{i,k}) \le D^q_{i,k} - (t - t_{j+1}),$$

and

$$\Sigma \ge \sum_{j=0}^{N-1} \frac{1}{q} \sum_{k=E^q_i(t_j)+1}^{E^q_i(t_{j+1})} \mathbb{I}_A(B^q_{i,k} - (y(t_j) + \delta), D^q_{i,k} - (t - t_j)),$$
$$\Sigma \le \overline{E}^q_i(t_0) + \sum_{j=0}^{N-1} \frac{1}{q} \times \sum_{k=E^q_i(t_j)+1}^{E^q_i(t_{j+1})} \mathbb{I}_A(B^q_{i,k} - (y(t_{j+1}) - \delta), D^q_{i,k} - (t - t_{j+1})),$$

which can be rewritten as

$$\begin{aligned} \Sigma &\ge \sum_{j=0}^{N-1} \overline{\mathcal{L}}^q_i(t_j, t_{j+1})(A + (y(t_j) + \delta, t - t_j)), \\ \Sigma &\le \overline{E}^q_i(t_0) + \sum_{j=0}^{N-1} \overline{\mathcal{L}}^q_i(t_j, t_{j+1})(A + (y(t_{j+1}) - \delta, t - t_{j+1})). \end{aligned} \tag{3.47}$$

Put

$$X^q := \sup_{A\in\mathcal{C}} \sup_{0\le s\le t\le T} \|(\overline{\boldsymbol{\mathcal{L}}}^q(s,t)(A) - (t-s)\boldsymbol{\eta} \times \boldsymbol{\theta}^q(A)\|,$$

then, by (3.47) and (3.46),

$$\begin{aligned} &\sum_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta^q_i(A + (y(t_j) + \delta, t - t_j)) - X^q\Big) \\ &\le \overline{\mathcal{Z}}^q_i(t)(A) - \overline{\mathcal{Z}}^q_i(0)(A + (S_i(\overline{\mathbf{Z}}^q, 0, t), t)) \\ &\le \eta_i t_0 + X^q + \sum_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta^q_i(A + (y(t_{j+1}) - \delta, t - t_{j+1}) + X^q\Big). \end{aligned} \tag{3.48}$$

To summarize, we have shown that, for all q and all possible outcomes $\omega \in \Omega^q$,

$$(\overline{\boldsymbol{\mathcal{Z}}}^q(\cdot), X^q) \in \mathcal{A}^q, \tag{3.49}$$

where $\mathcal{A}^q \subset \mathbf{D}(\mathbb{R}_+, \mathcal{M}^I) \times \mathbb{R}_+$ is the set of pairs $(\boldsymbol{\zeta}(\cdot), x)$ such that, for any set $A \in \mathcal{C}^+$, any partition $0 < t_0 < t_1 < \ldots < t_N = t \le T$ and any function $y(\cdot)$ that is non-increasing in $[t_0, t]$ and that satisfies $\sup_{s\in[t_0,t]} |S_i(\langle 1, \boldsymbol{\zeta}\rangle, s, t) - y(s)| \le \delta$ for some i and δ,

$$\begin{aligned} &\sum_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta^q_i(A + (y(t_j) + \delta, t - t_j)) - x\Big) \\ &\le \zeta_i(t)(A) - \zeta_i(0)(A + (S_i(\langle 1, \boldsymbol{\zeta}\rangle, 0, t), t)) \\ &\le \eta_i t_0 + x + \sum_{j=0}^{N-1} \Big(\eta_i(t_{j+1} - t_j)\theta^q_i(A + (y(t_{j+1}) - \delta, t - t_{j+1}) + x\Big). \end{aligned}$$

Limit equations By Assumptions 3.3 and 3.4 (*cf.* the proof of Gromoll et al. [47, Lemma 5.1]),

$$X^q \Rightarrow 0 \quad \text{as } q \to \infty.$$

Since the limit of X^q is deterministic, the joint convergence $(\overline{\boldsymbol{\mathcal{Z}}}^q(\cdot), X^q) \Rightarrow (\boldsymbol{\mathcal{Z}}(\cdot), 0)$ holds. By the Skorokhod representation theorem, there exist random elements

$$\{\widetilde{\boldsymbol{\mathcal{Z}}}^q(\cdot)\}_{q\to\infty}, \quad \widetilde{\boldsymbol{\mathcal{Z}}}(\cdot), \quad \{\widetilde{X}^q\}_{q\to\infty}$$

defined on a common probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ such that

$$(\widetilde{\boldsymbol{\mathcal{Z}}}^q(\cdot), \widetilde{X}^q) \overset{\mathrm{d}}{=} (\overline{\boldsymbol{\mathcal{Z}}}^q(\cdot), X^q) \quad \text{for all } q, \quad \widetilde{\boldsymbol{\mathcal{Z}}}(\cdot) \overset{\mathrm{d}}{=} \boldsymbol{\mathcal{Z}}(\cdot),$$

and

$$\text{a.s., as } q \to \infty, \quad (\widetilde{\boldsymbol{\mathcal{Z}}}^q(\cdot), \widetilde{X}^q) \to (\widetilde{\boldsymbol{\mathcal{Z}}}(\cdot), 0). \tag{3.50}$$

Introduce also the total mass processes $\widetilde{\mathbf{Z}}^q(\cdot) := \langle 1, \widetilde{\boldsymbol{\mathcal{Z}}}^q(\cdot)\rangle$ for all q, and $\widetilde{\mathbf{Z}}(\cdot) := \langle 1, \widetilde{\boldsymbol{\mathcal{Z}}}(\cdot)\rangle$. By Lemma 3.15, (3.45) and (3.49),

$$\text{a.s., for all } t > 0 \text{ and all } i, \quad \widetilde{Z}_i(t) > 0, \tag{3.51a}$$

$$\text{a.s., for all } t \in \mathbb{R}_+, \text{ all } i \text{ and } A \in \mathcal{C}, \quad \widetilde{\mathcal{Z}}_i(t)(\partial_A) = 0, \tag{3.51b}$$

$$\text{a.s., for all } q, \quad (\widetilde{\boldsymbol{\mathcal{Z}}}^q(\cdot), \widetilde{X}^q) \in \mathcal{A}^q. \tag{3.51c}$$

Denote by $\widetilde{\Omega}_*$ the set of outcomes $w \in \widetilde{\Omega}$ for which (3.50) and (3.51) hold. We will show that, for all $\omega \in \widetilde{\Omega}_*$, all i, $t \in [0, T]$ and $A \in \mathcal{C}^+$,

$$\widetilde{\mathcal{Z}}_i(t)(A) = \widetilde{\mathcal{Z}}_i(0)(A + (S_i(\widetilde{\mathbf{Z}}, 0, t), t)) + \eta_i \int_0^t \theta_i(A + (S_i(\widetilde{\mathbf{Z}}, s, t), t - s))ds, \tag{3.52}$$

and that will complete the proof of Theorem 3.5.

Fix $t \in [0, T]$, i and $A \in \mathcal{C}^+$. Also fix an outcome $\omega \in \widetilde{\Omega}_*$. All random elements in the rest of the proof are evaluated at this ω.

By (3.50) and (3.51b),

$$\widetilde{\mathcal{Z}}_i^q(t)(A) \to \widetilde{\mathcal{Z}}_i(t)(A) \quad \text{as } q \to \infty. \tag{3.53}$$

By (3.51a), the rate constraints and the dominated convergence theorem,

$$S_i(\widetilde{\mathbf{Z}}^q, s, t) \to S_i(\widetilde{\mathbf{Z}}, s, t) \quad \text{for all } s \in [0, t] \text{ as } q \to \infty, \tag{3.54}$$

which in particular implies that, as $q \to \infty$,

$$\widetilde{\mathcal{Z}}_i^q(0)(A + (S_i(\widetilde{\mathbf{Z}}^q, 0, t), t)) \to \widetilde{\mathcal{Z}}_i(0)(A + (S_i(\widetilde{\mathbf{Z}}, 0, t), t)). \tag{3.55}$$

Fix $t_0 \in (0, t)$ and $\delta > 0$. By (3.51a), the function $S_i(\widetilde{\mathbf{Z}}, \cdot, t)$ is continuous in $[t_0, t]$, and

the functions $S_i(\widetilde{\mathbf{Z}}^q, \cdot, t)$ are monotone. Then the point-wise convergence (3.54) implies uniform convergence in $[t_0, t]$, and for q large enough,

$$\sup_{s\in[t_0,t]} |S_i(\widetilde{\mathbf{Z}}^q, s, t) - S_i(\widetilde{\mathbf{Z}}, s, t)| \le \delta. \tag{3.56}$$

Now fix a partition $t_0 < t_1 < \ldots < t_N = t$. The bound (3.56) and (3.51c) imply that (in the definition of $\mathcal{A}^q$ we take $y(\cdot) = S_i(\widetilde{\mathbf{Z}}, \cdot, t)$)

$$\begin{aligned}
&\textstyle\sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1}-t_j)\theta_i^q(A + (S_i(\widetilde{\mathbf{Z}}, t_j, t) + \delta, t - t_j)) - \widetilde{X}^q\Big)\\
&\le \widetilde{\mathcal{Z}}_i^q(t)(A) - \widetilde{\mathcal{Z}}_i^q(A + (S_i(\widetilde{\mathbf{Z}}^q, 0, t), t))\\
&\le \eta_i t_0 + \widetilde{X}^q + \textstyle\sum_{j=0}^{N-1}\Big(\eta_i(t_{j+1}-t_j)\theta_i^q(A + (S_i(\widetilde{\mathbf{Z}}, t_{j+1}, t) - \delta, t - t_{j+1}) + \widetilde{X}^q\Big).
\end{aligned} \tag{3.57}$$

Since $\theta_i(\cdot \times \mathbb{R}_+)$ and $\theta_i(\mathbb{R}_+ \times \cdot)$ are probability measures, the set of $B \in \mathcal{C}$ for which $\theta_i(\partial_B) > 0$ is at most countable. By (3.51), $S_i(\widetilde{\mathbf{Z}}, \cdot, t)$ is strictly monotone in $[t_0, t]$. Hence, the set D of $s \in [t_0, t]$ for which $\theta_i\left(\partial_{A+(S_i(\widetilde{\mathbf{Z}},s,t)+\delta,t-s)}\right) > 0$ or $\theta_i\left(\partial_{A+(S_i(\widetilde{\mathbf{Z}},s,t)-\delta,t-s)}\right) > 0$ is at most countable, too. In (3.57), let $q \to \infty$ assuming that the partition contains no points from D. Then, by (3.50), (3.53) and (3.55),

$$\begin{aligned}
&\textstyle\sum_{j=0}^{N-1}\eta_i(t_{j+1}-t_j)\theta_i(A + (S_i(\widetilde{\mathbf{Z}}, t_j, t) + \delta, t - t_j))\\
&\le \widetilde{\mathcal{Z}}_i(t)(A) - \widetilde{\mathcal{Z}}_i(0)(A + (S_i(\widetilde{\mathbf{Z}}, 0, t), t))\\
&\le \eta_i t_0 + \textstyle\sum_{j=0}^{N-1}\eta_i(t_{j+1}-t_j)\theta_i(A + (S_i(\widetilde{\mathbf{Z}}, t_{j+1}, t) - \delta, t - t_{j+1}).
\end{aligned} \tag{3.58}$$

Now, in (3.58), let the diameter of the partition go to 0 keeping t_0 fixed. Then

$$\begin{aligned}
&\eta_i \int_{t_0}^{t} \theta_i(A + (S_i(\widetilde{\mathbf{Z}}, s, t) + \delta, t - s))ds\\
&\le \widetilde{\mathcal{Z}}_i(t)(A) - \widetilde{\mathcal{Z}}_i(0)(A + (S_i(\widetilde{\mathbf{Z}}, 0, t), t))\\
&\le \eta_i t_0 + \eta_i \int_{t_0}^{t} \theta_i(A + (S_i(\widetilde{\mathbf{Z}}, s, t) - \delta, t - s)ds.
\end{aligned}$$

Finally, in the last inequality, let $\delta \to 0$ (recall (3.51b)) and $t_0 \to 0$, then (3.52) follows.

3.8 Proof of Theorem 3.6

By the discussion following Theorem 3.6 and Lemma 3.2, it is left to show tightness of the scaled stationary distributions. It suffices to show coordinate-wise tightness, so fix i. By Jakubowski [51, Theorem 2.1] and Kallenberg [52, Theorem 15.7.5], a sequence

$\{\overline{\mathcal{Y}}_i^q, \overline{Y}_i^q\}_{q\to\infty}$ is tight if

$$\sup_q \mathbb{E}^q \overline{Y}_i^q < \infty, \tag{3.59a}$$

$$\lim_{n\to\infty} \sup_q \mathbb{E}^q \overline{\mathcal{Y}}_i^q(V_n^\infty) = 0, \tag{3.59b}$$

$$\lim_{n\to\infty} \sup_q \mathbb{E}^q \overline{\mathcal{Y}}_i^q(H_n^\infty) = 0, \tag{3.59c}$$

where $V_n^\infty = [n,\infty)\times\mathbb{R}_+$ and $H_n^\infty = \mathbb{R}_+\times[n,\infty)$.

First check (3.59a). For each q, the route i population process $Z_i^q(\cdot)$ is bounded from above by the length $Q_i^q(\cdot)$ of the $M/G/\infty$ queue with the following parameters:

(Q.1) at $t=0$, there are $Z_i^q(0)$ customers whose service times are patience times of the initial flows on route i of the q-th network;

(Q.2) the input process is the route i input process of the q-th network;

(Q.3) service times of newly arriving customers are patience times of newly arriving flows on route i of the q-th network.

Throughout the proof, $\mathrm{Pois}(\alpha)$ stands for a Poisson r.v. with parameter α.

For all q and t, $Z_i^q(t) \le Q_i^q(t)$. As $t\to\infty$, $Z_i^q(t) \Rightarrow Y_i^q$ and $Q_i^q(t) \Rightarrow \mathrm{Pois}(\eta_i^q \mathbb{E}^q D_i^q)$. Hence, $Y_i^q \le_{\mathrm{st}} \mathrm{Pois}(\eta_i^q \mathbb{E}^q D_i^q)$ for all q, and $\mathbb{E}^q \overline{Y}_i^q \le \eta_i^q \mathbb{E}^q D_i^q/q \to \eta_i \mathbb{E} D_i$ as $q\to\infty$, which implies (3.59a).

Now we check (3.59b). Note that, if at some point the residual flow size is at least n, then the initial flow size was at least n, too. Hence, $\mathcal{Z}_i^q(\cdot)(V_n^\infty)$ is bounded from above by the length $Q_i^{q,n}(\cdot)$ of the $M/G/\infty$ queue whose initial state is as in (Q.1), newly arriving customers are newly arriving flows on route i of the q-th network with initial sizes at least n, and service times of newly arriving customers are patience times of the corresponding flows. In particular, the input process for this queue is Poisson with intensity $\eta_i^q \mathbb{P}^q\{B_i^q \ge n\}$ and, by Assumption 3.7, it does not depend on service times.

Let $f_n(\cdot)$ be a continuous function on $\mathbb{R}_+^2$ such that

$$\mathbb{I}_{V_{n+1}^\infty}(\cdot) \le f_n(\cdot) \le \mathbb{I}_{V_n^\infty}(\cdot).$$

Then, for all q and t,

$$\langle f_n, \mathcal{Z}_i^q(t)\rangle \le \mathcal{Z}_i^q(t)(V_n^\infty) \le Q_i^{q,n}(t)$$

Letting $t\to\infty$, we obtain

$$\mathcal{Y}_i^q(V_{n+1}^\infty) \le \langle f_n, \mathcal{Y}_i^q\rangle \le_{\mathrm{st}} \mathrm{Pois}(\eta_i^q\, \mathbb{P}^q\{B_i^q \ge n\}\mathbb{E}^q D_i^q),$$
$$\mathbb{E}^q \overline{\mathcal{Y}}_i^q(V_{n+1}^\infty) \le \eta_i^q\, \mathbb{P}^q\{B_i^q \ge n\}\mathbb{E}^q D_i^q/q,$$

and then (3.59b) follows.

Finally, (3.59c) is valid due to the following lemma.

Lemma 3.16. *For any q, i and Borel set $S \subseteq \mathbb{R}_+$,*

$$\mathcal{Y}_i^q(\mathbb{R}_+ \times S) \leq_{\text{st}} \text{Pois}(\eta_i^q \, \mathbb{E}^q D_i^q \, \mathbb{P}^q\{\widetilde{D}_i^q \in S\}),$$

where $\widetilde{D}_i^q$ has density $\mathbb{P}^q\{D_i^q > x\}/\mathbb{E}^q D_i^q$, $x \geq 0$.

Proof. Fix q, i and a Borel set $S \subseteq \mathbb{R}_+$. It suffices to show that, for any $\delta > 0$,

$$\mathcal{Y}_i^q(\mathbb{R}_+ \times S) \leq_{\text{st}} \text{Pois}(\eta_i^q \mathbb{E}^q D_i^q \, \mathbb{P}^q\{\widetilde{D}_i^q \in S^\delta\}),$$

so fix $\delta > 0$.

Consider the upper bound queue $Q_i^q(\cdot)$ with parameters (Q.1)–(Q.3). We denote by $Q_i^q(t)(S^\delta)$ the number of customers in this queue whose residual service times at time t are in S^δ. Then

$$\mathcal{Z}_i^q(\cdot)(\mathbb{R}_+ \times S^\delta) \leq Q_i^q(\cdot)(S^\delta).$$

Given at time t there are k customers in the queue, denote by $D_1(t), \ldots, D_k(t)$ their residual service times. By Takács [102, Chapter 3.2, Theorem 2],

$$\begin{aligned}\lim_{t\to\infty} \mathbb{P}^q\{D_1(t) \leq x_1, \ldots, D_k(t) \leq x_k | Q_i^q(t) = k\} \\ = \mathbb{P}^q\{\widetilde{D}_i^q \leq x_1\} \ldots \mathbb{P}^q\{\widetilde{D}_i^q \leq x_k\},\end{aligned}$$

which together with $Q_i^q(t) \Rightarrow \text{Pois}(\eta_i^q \mathbb{E}^q D_i^q)$ as $t \to \infty$ implies that

$$Q_i^q(t)(S^\delta) \Rightarrow \text{Pois}(\eta_i^q \, \mathbb{E}^q D_i^q \, \mathbb{P}^q\{\widetilde{D}_i^q \in S^\delta\}) \quad \text{as } t \to \infty.$$

Let g_δ be a continuous function on $\mathbb{R}_+^2$ such that

$$\mathbb{I}_{\mathbb{R}_+ \times S}(\cdot) \leq g_\delta(\cdot) \leq \mathbb{I}_{\mathbb{R}_+ \times S^\delta}(\cdot).$$

Then, for any t,

$$\langle g^\delta, \mathcal{Z}_i^q(t)\rangle \leq \mathcal{Z}_i^q(t)(\mathbb{R}_+ \times S^\delta) \leq Q_i^q(t)(S^\delta),$$

and as $t \to \infty$,

$$\mathcal{Y}_i^q(\mathbb{R}_+ \times S) \leq \langle g^\delta, \mathcal{Y}_i^q\rangle \leq_{\text{st}} \text{Pois}(\eta_i^q \, \mathbb{E}^q D_i^q \, \mathbb{P}^q\{\widetilde{D}_i^q \in S^\delta\}). \qquad \square$$

3.9 Proofs of auxiliary results

Proof of Lemma 3.1. It suffices to show that, for a vector $\mathbf{z} \in \mathbb{R}_+^I$ with the first $\widetilde{I} < I$ coordinates positive and the rest of them zero, and a sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}} \subset (0,\infty)^I$ such that $\mathbf{z}^k \to \mathbf{z}$, we have $\mathbf{\Lambda}(\mathbf{z}^k) \to \mathbf{\Lambda}(\mathbf{z})$.

Suppose that $\mathbf{z}^k \to \mathbf{z}$ but $\mathbf{\Lambda}(\mathbf{z}^k) \not\to \mathbf{\Lambda}(\mathbf{z})$. Since $\{\mathbf{\Lambda}(\mathbf{z}^k)\}_{k\in\mathbb{N}}$ is a subset of the compact set $\{\mathbf{\Lambda} \in \mathbb{R}_+^I \colon \|\mathbf{\Lambda}\| \leq \|\mathbf{C}\|\}$, without loss of generality we may assume that $\mathbf{\Lambda}(\mathbf{z}^k) \to$

$\widetilde{\mathbf{\Lambda}} \neq \mathbf{\Lambda}(\mathbf{z})$.

Recall that $\mathbf{\Lambda}(\mathbf{z})$ is the unique optimal solution to

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{I} z_i \mathcal{U}_i(\Lambda_i/z_i) \\ \text{subject to} & A\mathbf{\Lambda} \leq \mathbf{C}, \mathbf{\Lambda} \leq \mathbf{m} \times \mathbf{z}, \end{array} \tag{3.60}$$

where, by convention, $\Lambda_i/0 := 0$ and $0 \times (-\infty) := 0$.

For all k, $A\mathbf{\Lambda}(\mathbf{z}^k) \leq \mathbf{C}$ and $\mathbf{\Lambda}(\mathbf{z}^k) \leq \mathbf{m} \times \mathbf{z}^k$. Hence, $\widetilde{\mathbf{\Lambda}}$ is feasible for (3.60) and $\widetilde{\Lambda}_i = 0 = \Lambda_i(\mathbf{z})$ for $i > \widetilde{I}$. Since $\widetilde{\mathbf{\Lambda}} \neq \mathbf{\Lambda}(\mathbf{z})$ is not optimal for (3.60),

$$l := \sum_{i=1}^{\widetilde{I}} z_i \mathcal{U}_i(\Lambda_i(\mathbf{z})/z_i) > \sum_{i=1}^{\widetilde{I}} z_i \mathcal{U}_i(\widetilde{\Lambda}_i/z_i) =: r. \tag{3.61}$$

Now we construct a sequence $\mathbf{\Lambda}^k \to \mathbf{\Lambda}(\mathbf{z})$ such that $\mathbf{\Lambda}^k$ is feasible for the optimization problem (3.60) with $\mathbf{z}^k$ in place of $\mathbf{z}$. Introduce vectors $\mathbf{C}^k \in \mathbb{R}_+^J$ with $C_j^k = \sum_{i=\widetilde{I}+1}^{I} A_{ji}\Lambda_i(\mathbf{z}^k)$. Put the first $\widetilde{I}$ coordinates of $\mathbf{\Lambda}^k$ to be $\Lambda_i^k = (\Lambda_i(\mathbf{z}) - \|\mathbf{C}^k\|)^+ \wedge m_i z_i^k$, and the rest of them $\Lambda_i^k = \Lambda_i(\mathbf{z}^k)$. That is, in the bandwidth allocation $\mathbf{\Lambda}(\mathbf{z})$, the bandwidth $\mathbf{C}^k$, which is required for the last $I - \widetilde{I}$ routes, is taken away from the first $\widetilde{I}$ routes.

Since $\mathbf{z}^k \to \mathbf{z}$, $\mathbf{\Lambda}^k \to \mathbf{\Lambda}(\mathbf{z})$ and $\mathbf{\Lambda}(\mathbf{z}^k) \to \widetilde{\mathbf{\Lambda}}$, we have

$$\sum_{i=1}^{\widetilde{I}} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) \to l, \quad \sum_{i=1}^{\widetilde{I}} z_i^k \mathcal{U}_i(\Lambda_i(\mathbf{z}^k)/z_i^k) \to r.$$

Also, for all k,

$$\sum_{i=\widetilde{I}+1}^{I} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) = \sum_{i=\widetilde{I}+1}^{I} z_i^k \mathcal{U}_i(\Lambda_i(\mathbf{z}^k)/z_i^k).$$

Then, by (3.61), for k big enough,

$$\sum_{i=1}^{I} z_i^k \mathcal{U}_i(\Lambda_i^k/z_i^k) > \sum_{i=1}^{I} z_i^k \mathcal{U}_i(\Lambda_i(\mathbf{z}^k)/z_i^k),$$

which contradicts to $\mathbf{\Lambda}(\mathbf{z}^k)$ being optimal for (3.60) with $\mathbf{z}^k$ in place of $\mathbf{z}$. □

Proof of Corollary 3.1. Fix an FMS $(\boldsymbol{\zeta}, \mathbf{z})(\cdot)$. In Section 3.3, we discussed how Theorem 3.4 implies that $\mathbf{z}(t) \to \mathbf{z}^*$ as $t \to \infty$. Here we prove that $\mathbf{z}(t) \to \mathbf{z}^*$ implies $\boldsymbol{\zeta}(t) \to \boldsymbol{\zeta}^*$. It suffices to show that, for any $\varepsilon > 0$, there exists a t_ε such that, for all $t \geq t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) \leq \zeta_i^*(A^\varepsilon) + \varepsilon, \tag{3.62a}$$

$$\zeta_i^*(A) \leq \zeta_i(t)(A^\varepsilon) + \varepsilon, \tag{3.62b}$$

so fix $\varepsilon > 0$.

For any $\delta \in (0, \min_{1\le i\le I} z_i^*)$, there exists a τ_δ such that, for all $t \ge \tau_\delta$,

$$\mathbf{z}^* - \boldsymbol{\delta} \le z(t) \le \mathbf{z}^* + \boldsymbol{\delta},$$

where $\boldsymbol{\delta} := (\delta, \dots, \delta) \in \mathbb{R}^I$. Then, for all $t \ge s \ge \tau_\delta$ and i, we have

$$r_i^\delta(t-s) \le \lambda_i(\mathbf{z}^*)(t-s),\ S_i(\mathbf{z}, s, t) \le R_i^\delta(t-s), \tag{3.63}$$

where $r_i^\delta := \inf_{\mathbf{z}^*-\delta\le\mathbf{z}\le\mathbf{z}^*+\delta} \lambda_i(\mathbf{z})$ and $R_i^\delta := \sup_{\mathbf{z}^*-\delta\le\mathbf{z}\le\mathbf{z}^*+\delta} \lambda_i(\mathbf{z})$.

Recall from Section 3.3 that, for all i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i^*(A) = \eta_i \int_0^\infty \theta_i(A + (\lambda_i(\mathbf{z}^*)s, s))ds. \tag{3.64}$$

From the shifted fluid model equation (3.5a) it follows that, for all $t \ge \tau_\delta$, i and Borel sets $A \subseteq \mathbb{R}_+^2$,

$$\zeta_i(t)(A) \le \zeta_i(\tau_\delta)(\mathbb{R}_+ \times [t-\tau_\delta, \infty)) + \eta_i \int_{\tau_\delta}^t \theta_i(A + (S_i(\mathbf{z}, s, t), t-s))ds, \tag{3.65}$$

where, by (3.63), the second summand admits the estimate

$$\begin{aligned}
&\eta_i \int_{\tau_\delta}^t \theta_i(A + (S_i(\mathbf{z}, s, t), t-s))ds \\
&\le \eta_i \int_{\tau_\delta}^t \theta_i(A^{(R_i^\delta - r_i^\delta)(t-s)} + (\lambda_i(\mathbf{z}^*)(t-s), (t-s)))ds \\
&= \eta_i \int_0^{t-\tau_\delta} \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(\mathbf{z}^*)s, s))ds.
\end{aligned} \tag{3.66}$$

Take $\widetilde{\tau}_\varepsilon$ and $\delta \in (0, \min_{1\le i\le I} z_i^*)$ such that

$$\eta_i \int_{\widetilde{\tau}_\varepsilon}^\infty \mathbb{P}\{D_i \ge s\}ds \le \varepsilon/2, \tag{3.67}$$

$$\|\mathbf{R}^\delta - \mathbf{r}^\delta\|\widetilde{\tau}_\varepsilon \le \varepsilon/2, \tag{3.68}$$

and take $t_\varepsilon \ge \tau_\delta + \widetilde{\tau}_\varepsilon$ such that, for all i,

$$\zeta_i(\tau_\delta)(\mathbb{R}_+ \times [t_\varepsilon - \tau_\delta, \infty)) \le \varepsilon/2. \tag{3.69}$$

Now, it follows from (3.65)–(3.66) by the choice of δ, τ_δ, $\widetilde{\tau}_\varepsilon$ and t_ε that, for all $t \ge t_\varepsilon$, i and

Borel sets $A \subseteq \mathbb{R}^2_+$,

$$\begin{aligned}
\zeta_i(t)(A) &\overset{(3.69)}{\leq} \varepsilon/2 + \eta_i \int_0^{\widetilde{\tau}_\varepsilon} \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(\mathbf{z}^*)s, s))ds \\
&\quad + \eta_i \int_{\widetilde{\tau}_\varepsilon}^{\infty} \theta_i(A^{(R_i^\delta - r_i^\delta)s} + (\lambda_i(\mathbf{z}^*)s, s))ds \\
&\overset{(3.68)}{\leq} \varepsilon/2 + \eta_i \int_0^{\widetilde{\tau}_\varepsilon} \theta_i(A^\varepsilon + (\lambda_i(\mathbf{z}^*)s, s))ds \\
&\quad + \eta_i \int_{\widetilde{\tau}_\varepsilon}^{\infty} \mathbb{P}\{D_i \geq s\}ds \\
&\overset{(3.64),(3.67)}{\leq} \zeta_i^*(A^\varepsilon) + \varepsilon,
\end{aligned}$$

i.e. we have (3.62a).

Similarly, we show (3.62b): for all $t \geq t_\varepsilon$, i and Borel sets $A \subseteq \mathbb{R}^2_+$,

$$\begin{aligned}
\zeta_i^*(t)(A) &\overset{(3.64)}{\leq} \eta_i \int_0^{\widetilde{\tau}_\varepsilon} \theta_i(A + (\lambda_i(\mathbf{z}^*)s, s))ds + \eta_i \int_{\widetilde{\tau}_\varepsilon}^{\infty} \mathbb{P}\{D_i \geq s\}ds \\
&\overset{(3.67)}{\leq} \eta_i \int_{t-\widetilde{\tau}_\varepsilon}^{t} \theta_i(A + (\lambda_i(\mathbf{z}^*)(t-s), (t-s)))ds + \varepsilon/2 \\
&\overset{(3.63)}{\leq} \eta_i \int_{t-\widetilde{\tau}_\varepsilon}^{t} \theta_i(A^{(R_i^\delta - r_i^\delta)(t-s)} + (S_i(\mathbf{z}, s, t), (t-s)))ds + \varepsilon/2 \\
&\overset{(3.68)}{\leq} \eta_i \int_{t-\widetilde{\tau}_\varepsilon}^{t} \theta_i(A^\varepsilon + (S_i(\mathbf{z}, s, t), (t-s)))ds + \varepsilon/2 \\
&\overset{(3.5a)}{\leq} \zeta_i(t)(A^\varepsilon) + \varepsilon/2.
\end{aligned}$$

□

Proof of Lemma 3.4. For all $s \leq t$ and $\varepsilon > 0$,

$$\begin{aligned}
&\int_s^t \mathbb{P}\{u + x \leq \xi < u + \widetilde{x} + \varepsilon\}du \\
&= \int_{s+x}^{t+x} \mathbb{P}\{\xi \geq u\}du - \int_{s+\widetilde{x}+\varepsilon}^{t+\widetilde{x}+\varepsilon} \mathbb{P}\{\xi \geq u\}du \\
&\leq \int_{s+x}^{s+\widetilde{x}+\varepsilon} \mathbb{P}\{\xi \geq u\}du \leq \widetilde{x} - x + \varepsilon.
\end{aligned}$$

The lemma follows as we first let $\varepsilon \to 0$ (applying the dominated convergence theorem) and then $s \to -\infty$, $t \to \infty$. □

Chapter 4

Random Fluid Limit of an Overloaded Polling Model

4.1 Introduction

This chapter is dedicated to stochastic networks called polling models. Broadly speaking, a polling model can be defined as multiple queues served one at a time by a single server. As for further details — service disciplines at the queues, routing of the server, and its walking times from one queue to another — there exist numerous variations motivated by the wide range of applications. The earliest polling study to appear in the literature seems to be by Mack et al. [68] (1957), who investigated a problem in the British cotton industry involving a single repairman cyclically patrolling multiple machines, inspecting them for malfunctioning and repairing them. Over the past few decades, polling techniques have been of extensive use in the areas of computer and communication networks as well as manufacturing and maintenance. Along with that, a vast body of related literature has grown. For overviews of the available results on polling models and their analysis methodologies, we refer the reader to Takagi [103, 104, 105], Boxma [18], Yechiali [120] and Borst [16].

Across the great variety of polling models, there exists the "classical" one, which was first used in the analysis of time-sharing computer systems in the early 70's. This model is *cyclic*, i.e. if there are I queues in total, they are visited by the server in the cyclic order $1, 2, \ldots, I, 1, 2, \ldots$. All of the queues are supposed to be infinite-buffer queues, and to each of them there is a Poisson stream of arriving customers with i.i.d. service times. After all visits to a queue, i.i.d. walking, or switchover, times are incurred. All interarrival times, service times and switchover times are mutually independent, and their distributions may vary from queue to queue as well as the service disciplines. Examples of the most common service disciplines are *exhaustive* (the queue is served until it becomes empty), *gated* (in the course of a visit, only those customers get served who are present in the queue when the server arrives to, or *polls*, the queue), and *k-limited* (at most k customers get served per visit). This chapter is also centered around the classical polling model. We assume zero switchover times and allow a wide class

of service disciplines that includes both exhaustive and gated policies and is discussed later in more detail.

Amongst desirable properties of any service system, the first one is stability. So, naturally, the major part of the polling related literature is focused on the performance of stable models. Foss and Kovalevskii [42] obtained an interesting result of null recurrence over a thick region of the parameter space for a two-server polling-like system. MacPhee et al. [69, 70] have recently observed the same phenomenon for a hybrid polling/Jackson network, where the service rate and customer rerouting probabilities are randomly updated each time the server switches from one queue to another.

The study of critically loaded polling models was initiated about two decades ago by Coffman et al. [28, 29], who proved a so called averaging principle: in the diffusion heavy traffic limit, certain functionals of the joint workload process can be expressed via the limit total workload, which was shown to be a reflected Brownian motion and a Bessel process in the case of zero and non-zero switchover times, respectively. In subsequent years, the work has been carried on by Kroese [63], Vatutin and Dyakonova [110], Altman and Kushner [3], Van der Mei [106] and others. In particular, heavy-traffic approximations of the steady state and waiting time distributions have been derived.

Although overloaded service systems are an existing reality and it is of importance to control or predict how fast they blow up over time, to the best of our knowledge, for polling models this problem has not been addressed in the literature so far. This chapter aims to fill in the gap. Moreover, this appears to be a really exciting problem because it reveals the following unusual phenomenon. Our interest is in fluid approximations of the system, namely we look for the a.s. limit of the scaled joint queue length process

$$(Q_1, \dots, Q_I)(x_n \cdot)/x_n$$

along a deterministic sequence $x_n \to \infty$. Remarkably, in contrast to the many basic queueing systems with deterministic fluid limits, overloaded polling models preserve some randomness under passage to the fluid dynamics. Other examples of simplistic designs combined with random fluid limits are two-queue two-server models of Foss and Kovalevskii [42] and Kovalevski et al. [62]. We refer to the latter work [62] for an insightful discussion of the nature of randomness in fluid limits in general and for an overview of the publications on the topic.

To illustrate the key idea that has led us to the result, consider the simple, symmetric model of $I = 2$ queues with exhaustive service, zero switchover times and empty initial condition (without the last assumption, the analysis becomes much simpler). In isolation, the queues are stable, and the whole system is overloaded, i.e. $1/2 < \lambda/\mu < 1$, where λ and $1/\mu$ are the arrival rate and the mean service time, respectively (in both queues). Denote the supposedly existing limit queue length process by $(\overline{Q}_1, \overline{Q}_2)(\cdot)$. Note that, given the limit size of the queue in service at any non-zero time instant, the entire limit trajectories of both queues can be restored by the SLLN. Indeed, the limit total population $(\overline{Q}_1 + \overline{Q}_2)(\cdot)$ grows at rate $2\lambda - \mu$. Because of the symmetry, at any fixed time $T > 0$, the limit queues are in service with equal probabilities. Let queue 1 be in service at time T. Then in Figure 4.1 the limit queues 1 and 2 follow the solid and dashed trajectories, respectively. Starting from time T, the limit queue 1 gets cleared up

at rate $\lambda - \mu$ until it becomes empty, say, at time $\bar{t}_1$. Since $\bar{t}_1$, when the limit total population $(2\lambda - \mu)\bar{t}_1$ comes from queue 2 alone, queue 2 gets cleared up at rate $\lambda - \mu$ until it becomes empty at time $\bar{t}_2$, while queue 1 grows at the arrival rate λ. Moving forward and backward in this way, one can continue the two trajectories onto $[T, \infty)$ and $(0, T]$, respectively, and see that they oscillate at an infinite rate when approaching zero. Now, the same principle applies if, instead of the queue sizes at time T, we know $\bar{t}_1$, which is the first switching instant after T. It turns out that $\bar{t}_1$ is random, and that makes the whole fluid limit random. The following crucial observation makes it possible to find the distribution of $\bar{t}_1$. Let customer 2 to be a descendant of customer 1 if customer 2 arrives to the system while customer 1 is receiving service, or customer 2 is a descendant of a descendant of customer 1. Then the size of the non-empty queue at switching instants form a branching process.

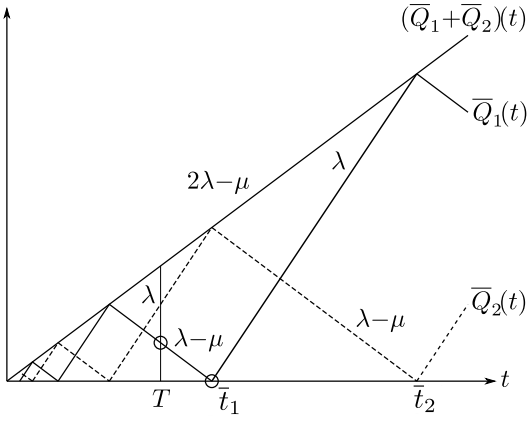

Figure 4.1: Fluid limit of a symmetric two-queue model with exhaustive service

The idea of representing arriving customers as descendants of the customer in service, has appeared in Foss [41] in the studies of an extension of Klimov's μc-rule, and in Fuhrmann and Cooper [44] who proved a decomposition result for the stationary distribution of the $M/G/1$ queue with server vacations. Then Resing [94] introduced a wide class of service disciplines that, for the classical polling model (more general periodic server routing is also allowed), guarantee the joint queue length at the successive polling instants of a fixed queue to form a multitype branching process (MTBP). This embedded MTBP is the cornerstone of the analysis that we carry out in this chapter.

We now describe the class of service disciplines that we allow. It is a subclass of the MTBP-policies, and we call them *multigated* meaning that each visit of each queue consists of a number of consecutive gated service phases. The upper bound on the number of phases, called the *gating index*, comes from the input data (together with the interarrival and service times). Gating indices for different visits of the same queue are i.i.d. random variables whose distribution may vary from queue to queue, and gating indices for different queues are mutually independent. Gating indices equal to 1 and ∞ correspond to exhaustive and conventional gated service, respectively. Multigated policies with deterministic gating indices were studied (and, in fact, introduced) recently by Van Wijk et al. [109] with the purpose of balancing fairness and efficiency of polling models.

Van der Mei and co-authors [107, 108] consider multi-stage gated policies, but those are different than in [109] and here.

Throughout the chapter, we consider the case of zero switchover times. The case of non-zero switchover times can be treated with similar methods.

As for the proofs, multiple asymmetric queues with non-exhaustive service create more work compared to the simple two-queue example discussed above. Knowing the limit total population is of little use now since it only reduces the dimension of the problem by one. We show that, in the general situation, the fluid limit queue length trajectory $(\overline{Q}_1, \dots, \overline{Q}_I)(\cdot)$ is determined by $2I$ random parameters: the earliest polling instants $\bar{t}_1, \dots, \bar{t}_I$ that, in the limit, follow a fixed time instant, and the limit sizes $\overline{Q}_1(\bar{t}_1), \dots, \overline{Q}_I(\bar{t}_I)$ of the corresponding polled queues. The overload assumption and multigated policies provide the framework of supercritical MTBP's, and we can apply the Kesten-Stigum theorem [60, 66] (the classical result on asymptotics of supercritical MTBP's) to find the distribution of, for example, $(\overline{Q}_1, \dots, \overline{Q}_I)(\bar{t}_1)$. Then suitable SLLN's imply that the other parameters $\bar{t}_1, \dots, \bar{t}_I$, $\overline{Q}_2(\bar{t}_2), \dots, \overline{Q}_I(\bar{t}_I)$ can be expressed either via the Kesten-Stigum limit $(\overline{Q}_1, \dots, \overline{Q}_I)(\bar{t}_1)$ or via each other. Note also that the Kesten-Stigum theorem requires certain moments of the offspring distribution to be finite. The visit at a queue is the longest when service is exhaustive, implying more customers in the other queues in the end of the cycle. So attempts to satisfy the moment conditions of the Kesten-Stigum theorem boil down to proving finiteness of the corresponding moment for the busy period of an $M/G/1$ queue (see Remark 1.1), which is an interesting and novel result by itself. Besides, we obtain an estimate for this moment, and our approach is valid for a wide class of regularly varying convex functions, in particular power and logarithmic functions.

The chapter is organised as follows. Section 4.2 describes the cyclic polling model and the class of service disciplines. Section 4.3 explains the connection between the model and MTBP's, gives some preliminaries from the theory of MTBP's and derives characteristics of the embedded MTBP. In Section 4.4, we state our main result — the fluid limit theorem — and discuss the optimal representation of the fluid limit from the computational point of view (Remark 4.5). Section 4.5 proves the results of Section 4.3, see the proof of Lemma 4.3 for estimates on the moments of the busy period of an $M/G/1$ queue. Section 4.6 proves the fluid limit theorem. Proofs of some auxiliary statements are given in Section 4.7. In the remainder of the section, we list the specific notations of this chapter.

Notation For a real number x, along with its maximum integer lower bound $\lfloor x \rfloor$, we also operate with its fractional part and minimum integer upper bound given by

$$\{x\} := x - \lfloor x \rfloor, \quad \lceil x \rceil := \min\{n \in \mathbb{Z} \colon n \geq x\}.$$

All vectors in this chapter are I-dimensional. For vectors $\mathbf{x} \in (0, \infty)^I, \mathbf{y} \in \mathbb{R}^I$, we define the power operation

$$\mathbf{x}^{\mathbf{y}} = \prod_{i=1^I} x_i^{y_i},$$

and for vectors $\mathbf{l}, \mathbf{k} \in \mathbb{Z}_+^I$, $\mathbf{l} \leq \mathbf{k}$, the binomial coefficient

$$\binom{\mathbf{x}}{\mathbf{y}} = \prod_{i=1}^{I} \binom{x_i}{y_i} = \prod_{i=1}^{I} \frac{x_i!}{y_i!(x_i - y_i)!}.$$

Finally, for $i = 1, \ldots, I$, we introduce the vectors $\mathbf{e}_i$ with coordinate i equal to 1 and the other coordinates equal to 0. Recall that $\mathbf{0}$ denotes the zero vector, and $\mathbf{1}$ denotes the vector of ones.

4.2 Stochastic model

This section contains a detailed description of the cyclic polling model and the class of service disciplines that we allow for this model. It also specifies the stochastic assumptions. All stochastic primitives introduced throughout the chapter are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$.

Cyclic polling Consider a system that consists of multiple infinite-buffer queues labeled by $i = 1, \ldots, I$, where I is finite, and a single server. There are external arrivals of customers to the queues that line up in the corresponding buffers in the order of arrival. The server idles if and only if the entire system is empty. While the system is non-empty, the server works at unit rate serving one queue at a time and switching from one queue to another in the cyclic order: after a period of serving queue i, called a *visit to queue* i, a visit to queue $i \bmod I + 1$ follows. Note that, while the system is non-empty, empty queues get visited as well in the sense that, once the server arrives to (or, *polls*) an empty queue, say at time t, it has to leave immediately, and the visit in this case is defined to be the empty interval $[t, t)$. Now suppose that, at a particular time instant, the system empties upon completion of a non-empty visit to queue i. For mathematical convenience, we assume that such an instant is followed by a single (empty) visit to each of the empty queues $i + 1, \ldots, I$. Then the server idles until the first arrival into the empty system. If that arrival is to queue i, a single (empty) visit to each of the empty queues $1, \ldots, i - 1$ precedes the visit to queue i. In the course of a visit, a number of customers at the head of the queue get served in the order of arrival and depart. The service disciplines at the queues specify how many customers should get served per visit, we now proceed with their description.

Multigated service With multigated service in a queue we mean that each visit of that queue consisits of a number of consecutive gated service phases. More formally, we say that *the server gates a queue* at a particular time instant meaning that the queue is in service at the moment, and all the customers found in the queue at the moment are guaranteed to receive service during the current visit. Customers gated together are served in the order of arrival. For each visit, its *gating index* is defined: it is the upper bound on the number of times the server is supposed to gate the queue in the course of the visit. The gating indices for different queues and for different visits of the same

queue might be different. The first time during a visit when the server gates the queue is upon polling the queue. The other gating instants are defined by induction: as soon as the customers found in the queue the last time it was gated have been served, the queue is gated again provided that the total number of gating procedures is not going to exceed the gating index. If the queue is empty upon gating, the server switches to the next queue, and thus the actual number of gating procedures performed during a visit might differ from the gating index for that visit. Now we define a generic multigated service discipline.

Definition 4.1. Let a random variable X take values in $\mathbb{Z}_+ \cup \{\infty\}$. The service discipline at a particular queue is called *X-gated* if the gating indices for different visits of this queue are i.i.d. copies of X. If a gating index equals 0, the server should leave immediately after polling the queue. The values 1 and ∞ of a gating index correspond to conventional gated and exhaustive service, respectively.

Remark 4.1. Multigated service disciplines guarantee the population of the polling system at polling instants of a fixed queue to form an MTBP, laying the foundation for the analysis that we carry out in this chapter. We discuss this connection with MTBP's in detail in the next section.

Stochastic assumptions We assume the cyclic polling system described above to evolve in the continuous time horizon $t \in \mathbb{R}_+$. At $t = 0$, the system is empty. Arrivals of customers to queue i form a Poisson process $A_i(\cdot)$ of rate λ_i. Introduce also the vector of arrival rates

$$\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_I).$$

Service times of queue i customers are drawn from a sequence $\{B_{i,n}\}_{n\in\mathbb{N}}$ of i.i.d. copies of a positive random variable B_i with a finite mean value $1/\mu_i$. Gating indices for queue i are drawn from a sequence $\{X_{i,n}\}_{n\in\mathbb{N}}$ of i.i.d. copies of a random variable X_i taking values in $\mathbb{Z}_+ \cup \{\infty\}$. The random elements $A_i(\cdot)$, $\{B_{i,n}\}_{n\in\mathbb{N}}$ and $\{X_{i,n}\}_{n\in\mathbb{N}}$, $i = 1, \ldots, I$, are mutually independent. Additionally, we impose the following conditions on the load intensities and service times.

Assumption 4.1. *For all i, $\lambda_i/\mu_i < 1$, and $\sum_{i=1}^{I} \lambda_i/\mu_i > 1$.*

Assumption 4.2. *For all i, $\mathbb{E}B_i \log B_i < \infty$.*

We study the system behavior in terms of its queue length process

$$\mathbf{Q}(\cdot) = (Q_1, \ldots, Q_I)(\cdot),$$

where $Q_i(t)$ is the number of customers in queue i at time $t \in \mathbb{R}_+$.

4.3 Connection with MTBP's

This section is devoted to a multitype branching process (MTBP) embedded in the queue length process $\mathbf{Q}(\cdot)$ and enabling its further analysis.

To start with, we divide the time horizon into pairwise-disjoint finite intervals in such a way that each interval includes a single (possibly empty) visit of the server to each of the queues starting from the first one. Let

$$\mathbb{R}_+ = \bigcup_{n\in\mathbb{Z}_+} [t_n, t_{n+1}),$$
$$[t_n, t_{n+1}) = [t_n, t_{1,n}) \bigcup_{i=1}^{I} [t_{i,n}, t_{i+1,n}),$$

where

- $t_0 = 0$ and $t_n \le t_{1,n} \le \ldots \le t_{I+1,n} = t_{n+1}$;
- if the system is empty at t_n, then the interval $[t_n, t_{1,n})$ is the period of waiting until the first arrival, otherwise $t_n = t_{1,n}$;
- the interval $[ti, n, t_{i+1,n})$ is the visit to queue i following t_n, with $t_{i,n} = t_{i+1,n}$ if the visit is empty.

The interval $[t_n, t_{n+1})$ is called *session n*. The interval $[t_{i,n}, t_{i+1,n})$ is called *visit n to queue i*, and the gating index for this visit is $X_{i,n}$.

For multigated service disciplines that we consider in this chapter, the following holds.

Property 4.1. *For all $i = 1, \ldots, I$, the customers found in queue i at a polling instant get replaced during the course of the visit by i.i.d. copies of a random vector $\mathbf{L}_i^{\mathrm{v}} = (L_{i,1}^{\mathrm{v}}, \ldots, L_{i,I}^{\mathrm{v}})$ that has the distribution of $\mathbf{Q}(t_{i+1,n})$ given that $\mathbf{Q}(t_{i,n}) = \mathbf{e}_i$ (this distribution does not depend on n).*

By Resing [94], Property 4.1 implies that the sequence

$$\{\mathbf{Q}(t_n)\}_{n\in\mathbb{Z}_+}$$

forms an MTBP with immigration in state $\mathbf{0}$. In the rest of the section, we introduce a number of objects associated with this MTBP and discuss some of its properties.

The random vector $\mathbf{L}_i^{\mathrm{v}}$ mentioned in Property 4.1 is called the *visit offspring of a queue i customer*. Define also the *visit duration at queue i* to be a random variable V_i equal in distribution to $(t_{i+1,n} - t_{i,n})$ given that $Q_i(t_{i,n}) = 1$, and the *session offspring of a queue i customer* to be a random vector $\mathbf{L}_i = (L_{i,1}, \ldots, L_{i,I})$ that has the distribution of $\mathbf{Q}(t_{n+1})$ given that $\mathbf{Q}(t_n) = \mathbf{e}_i$. Then the immigration distribution is given by

$$G(\mathbf{k}) := \mathbb{P}\{\mathbf{Q}(t^{(n+1)}) = \mathbf{k} | \mathbf{Q}(t_n) = \mathbf{0}\} = \textstyle\sum_{i=1}^{I} \lambda_i \mathbb{P}\{\mathbf{L}_i = \mathbf{k}\} / \sum_{i=1}^{I} \lambda_i, \quad \mathbf{k} \in \mathbb{Z}_+^I.$$

The following lemma computes the mean values

$$\begin{aligned}\gamma_i &:= \mathbb{E}V_i,\\ \mathbf{m}_i^{\mathrm{v}} = (m_{i,1}^{\mathrm{v}},\dots,m_{i,I}^{\mathrm{v}}) &:= \mathbb{E}\mathbf{L}_i^{\mathrm{v}},\\ \mathbf{m}_i = (m_{i,1},\dots,m_{i,I}) &:= \mathbb{E}\mathbf{L}_i.\end{aligned}$$

Lemma 4.1. *For all i,*

$$m_{i,i}^{\mathrm{v}} = \mathbb{E}(\lambda_i/\mu_i)^{X_i}, \quad \gamma_i = \frac{1-m_{i,i}^{\mathrm{v}}}{\mu_i-\lambda_i},$$

and, for $i \neq j$,

$$m_{i,j}^{\mathrm{v}} = \lambda_j\gamma_i.$$

For the $m_{i,j}$'s, there is a recursive formula:

$$m_{I,j} = m_{I,j}^{\mathrm{v}} \quad \textit{for all } j,$$

and, for $i \leq I-1$, $\mathbf{m}_i$ is computed via $\mathbf{m}_{i+1}$,

$$m_{i,j} = m_{i,j}^{\mathrm{v}}\mathbb{I}\{i \geq j\} + \textstyle\sum_{k=i+1}^{I} m_{i,k}^{\mathrm{v}}m_{k,j} \quad \textit{for all } j.$$

The proof follows in Section 4.5.1.

By the Perron-Frobenius theorem (see e.g. Harris [49, Theorem 5.1]), the *mean session offspring matrix* $M := \{m_{i,j}\}_{i,j=1}^{I}$ has a positive eigenvalue ρ that is greater in absolute value than any other eigenvalue of M. The eigenspace associated with ρ is one-dimensional and parallel to a vector with all coordinates positive. Then there exist (row) vectors $\mathbf{u} = (u_1,\dots,u_I)$ and $\mathbf{v} = (v_1,\dots,v_I)$ with all coordinates positive such that

$$M\mathbf{u}^T = \rho\mathbf{u}^T, \quad \mathbf{v}M = \rho\mathbf{v} \quad \text{and} \quad \mathbf{v}\mathbf{u}^T = 1,$$

where $\mathbf{u}^T$ is the transpose of $\mathbf{u}$.

Now introduce an auxiliary MTBP $\{\mathbf{Z}(n)\}_{n\in\mathbb{Z}_+}$ with no immigration and such that, given $\mathbf{Z}(n) = \mathbf{e}_i$, the next generation $\mathbf{Z}(n+1)$ is equal in distribution to $\mathbf{L}_i$. Denote by q_i the *extinction probability* for the process $\{\mathbf{Z}(n)\}_{n\in\mathbb{Z}_+}$ given that $\mathbf{Z}(0) = \mathbf{e}_i$, and introduce the vector of extinction probabilities

$$\mathbf{q} := (q_1,\dots,q_I).$$

Then the probability for the process $\{\mathbf{Q}(t_n)\}_{n\in\mathbb{Z}_+}$ to return to $\mathbf{0}$ is given by

$$q_G := \textstyle\sum_{\mathbf{k}\in\mathbb{Z}_+^I} G(\mathbf{k})\mathbf{q}^{\mathbf{k}}.$$

Remark 4.2. Since all time instants t such that $\mathbf{Q}(t) = \mathbf{0}$ are contained among the t_n's, the probability for the process $\mathbf{Q}(\cdot)$ to return to $\mathbf{0}$ equals q_G, too.

By Assumption 4.1, the MTBP's $\{\mathbf{Q}(t_n)\}_{n\in\mathbb{Z}_+}$ and $\{\mathbf{Z}(n)\}_{n\in\mathbb{Z}_+}$ are supercritical (the proof is postponed to Section 4.7).

Lemma 4.2. *For the Perron-Frobenius eigenvalue ρ and the extinction probabilities q_i, we have $\rho > 1$ and $q_i < 1$ for all i. By the latter, $q_G < 1$, too.*

Assumption 4.2 guarantees finiteness of the corresponding moments for the offspring distribution of the MTBP's $\{\mathbf{Q}(t_n)\}_{n\in\mathbb{Z}_+}$ and $\{\mathbf{Z}(n)\}_{n\in\mathbb{Z}_+}$ (see Section 4.5.2 for the proof).

Lemma 4.3. *For all i and j, $\mathbb{E}L_{i,j}\log L_{i,j} < \infty$, where $0\log 0 := 0$ by convention.*

Finally, we quote the Kesten-Stigum theorem for supercritical MTBP's (see e.g. [60, 66]), which is our starting point when proving the convergence results of the next section. By that theorem and Lemmas 4.2 and 4.3, the auxiliary process $\{\mathbf{Z}(n)\}_{n\in\mathbb{Z}_+}$ has the following asymptotics.

Proposition 4.1. *Given $\mathbf{Z}(0) = \mathbf{e}_i$,*

$$\mathbf{Z}(n)/\rho^n \to \zeta_i \mathbf{v} \quad \text{a.s. as } n \to \infty,$$

where the distribution of the random variable ζ_i has a jump of magnitude $q_i < 1$ at 0 and a continuous density function on $(0,\infty)$, and $\mathbb{E}\zeta_i = u_i$.

4.4 Fluid limit theorem

In this section, we present our main result which concerns the behavior of the system under study on a large time scale.

For each $n \in \mathbb{Z}_+$, introduce the scaled queue length process

$$\overline{\mathbf{Q}}^n(t) := \mathbf{Q}(\rho^n t)/\rho^n, \quad t \in \mathbb{R}_+. \tag{4.1}$$

We are interested in the a.s. limit of the processes (4.1) as $n \to \infty$, which we call the *fluid limit* of the model. It appears that, in order to precisely describe the fluid limit, the information provided by the following theorem is sufficient.

For $n \in \mathbb{Z}$, let

$$\eta_n := \begin{cases} \min\{k\colon t_k \geq \rho^n\}, & n \geq 0, \\ 0, & n < 0. \end{cases}$$

Theorem 4.1. *There exist constants $\overline{b}_i \in (0,\infty)$ and $\overline{\mathbf{a}}_i = (\overline{a}_{i,1},\dots,\overline{a}_{i,I}) \in \mathbb{R}_+^I$, $i = 1,\dots,I+1$, and a random variable ξ with values in $[1,\rho)$ such that, for all $k \in \mathbb{Z}_+$ and i,*

$$\text{a.s. as } n \to \infty, \quad t_{i,\eta_n+k}/\rho^n \to \rho^k \overline{b}_i \xi, \quad \mathbf{Q}(t_{i,\eta_n+k})/\rho^n \to \xi \rho^k \overline{\mathbf{a}}_i. \tag{4.2}$$

The $\overline{b}_i$'s and $\overline{\mathbf{a}}_i$'s are given by

$$\begin{aligned} \overline{b}_1 &= 1, \\ \overline{b}_{i+1} &= \overline{t}_i + (v_i/\alpha + \lambda_i(\overline{b}_i - \overline{b}_1))\gamma_i, \quad i \leq I, \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} \overline{\mathbf{a}}_1 &= \mathbf{v}/\alpha, \\ \overline{\mathbf{a}}_{i+1} &= \overline{\mathbf{a}}_i + (\overline{b}_{i+1} - \overline{b}_i)\boldsymbol{\lambda} - (\overline{b}_{i+1} - \overline{b}_i)\mu_i \mathbf{e}_i, \quad i \le I, \end{aligned} \tag{4.4}$$

where

$$\alpha = \frac{\sum_{i=1}^I v_i/\mu_i}{\sum_{i=1}^I \lambda_i/\mu_i - 1}.$$

The distribution of ξ is given by: for $x \in [1, \rho)$,

$$\begin{aligned} \mathbb{P}\{\xi \ge x\} = &\frac{1}{1-q_G} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^I, \\ ||\mathbf{k}||_1 \ge 1}} G(\mathbf{k}) \sum_{\substack{\mathbf{l}\le\mathbf{k}, \\ ||\mathbf{l}||_1 \ge 1}} \binom{\mathbf{k}}{\mathbf{l}} (\mathbf{1}-\mathbf{q})^{\mathbf{l}} \mathbf{q}^{\mathbf{k}-\mathbf{l}} \\ &\times \mathbb{P}\{\{\log_\rho(\alpha \sum_{i=1}^I \sum_{n=1}^{l_i} \xi_{i,n})\} \ge \log_\rho x\}, \end{aligned}$$

where $\xi_{i,n}$, $n \in \mathbb{N}$, are i.i.d. r.v.'s with the distribution of ζ_i given that $\zeta_i > 0$, and the sequences $\{\xi_{i,n}\}_{n\in\mathbb{N}}$, $i = 1, \dots, I$, are mutually independent.

The proof of Theorem 4.1 combines the Kesten-Stigum theorem with various dynamic equations and laws of large numbers, see Section 4.6.

Remark 4.3. Since the system is overloaded (Assumption 4.1), it never empties after a finite period of time. Hence, for all n big enough, we have $t_{I+1,n} = t_{1,n+1}$, and then Theorem 4.1 implies that

$$\overline{b}_{I+1} = \rho \overline{b}_1, \quad \overline{\mathbf{a}}_{I+1} = \rho \overline{\mathbf{a}}_1.$$

Remark 4.4. There is an alternative way to compute the $\overline{a}_i$'s:

$$\begin{aligned} \overline{\mathbf{a}}_1 &= \mathbf{v}/\alpha, \\ \overline{\mathbf{a}}_{i+1} &= \overline{\mathbf{a}}_i - \overline{a}_{i,i}\mathbf{e}_i + \overline{a}_{i,i}\mathbf{m}_i^{\mathrm{v}}, \quad i \le I, \end{aligned}$$

which implies that $\overline{a}_{i,j} > 0$ if $|i-j| \ne 1$ and $\overline{a}_{i,i+1} = 0$ if and only if the service discipline at queue i is exhaustive. See Lemma 4.7 and Remark 4.7 in Section 4.6.2.

Based on the results of Theorem 4.1, Theorem 4.2 below derives the fluid limit equations from the suitable dynamic equations, see Section 4.6 for the proof.

Theorem 4.2. *There exists a deterministic function $\overline{\mathbf{Q}}(\cdot) = (\overline{Q}_1, \dots, \overline{Q}_I)(\cdot) \colon \mathbb{R}_+ \to \mathbb{R}_+^I$ such that,*

$$\text{a.s. as } n \to \infty, \quad \overline{\mathbf{Q}}^n(\cdot) \to \xi \overline{\mathbf{Q}}(\cdot/\xi) \quad \text{u.o.c.},$$

where the r.v. ξ is defined in Theorem 4.1 *and the abbreviation "u.o.c." stands for uniform convergence on compact sets.*

The function $\overline{\mathbf{Q}}(\cdot)$ *is continuous and piecewise linear and given by*

$$\overline{\mathbf{Q}}(t) = \begin{cases} \mathbf{0}, & t = 0, \\ \rho^k \overline{\mathbf{a}}_i + (t - \rho^k \overline{b}_i)\boldsymbol{\lambda} - (t - \rho^k \overline{b}_i)\mu_i \mathbf{e}_i, & t \in [\rho^k \overline{b}_i, \rho^k \overline{b}_{I+1}), \\ & i = 1, \ldots, I, \\ & k \in \mathbb{Z}, \end{cases} \tag{4.5}$$

or, equivalently, by

$$\text{for all } i, \quad \overline{Q}_i(t) = \begin{cases} 0, & t = 0, \\ \rho^k \overline{a}_{i,i} + (\lambda_i - \mu_i)(t - \rho^k \overline{b}_i), & t \in [\rho^k \overline{b}_i, \rho^k \overline{b}_{i+1}), \\ & k \in \mathbb{Z}, \\ \rho^{k+1} \overline{a}_{i,i} - \lambda_i(\rho^{k+1} \overline{b}_i - t), & t \in [\rho^k \overline{b}_{i+1}, \rho^{k+1} \overline{b}_i), \\ & k \in \mathbb{Z}. \end{cases} \tag{4.6}$$

Remark 4.5. By (4.6), the whole process $\overline{\mathbf{Q}}(\cdot)$ is defined by the constants $\overline{b}_i$ and $\overline{a}_{i,i}$. The fastest way to compute the $\overline{b}_i$'s and $\overline{a}_{i,i}$'s is using the simultaneous recursion

$$\overline{b}_1 = 1, \quad \text{for } i \leq I, \quad \overline{a}_{i,i} = v_i/\alpha + \lambda_i(\overline{b}_i - \overline{b}_1), \quad \overline{b}_{i+1} = \overline{b}_i + \overline{a}_{i,i}\gamma_i.$$

See the last part of the proof of Lemma 4.7 (namely, (4.34) and (4.35)) and Remark 4.7 in Section 4.6.2.

Remark 4.6. Since the fluid limit is continuous, it approximates the fluid-scaled queue length process both in the metric of uniform convergence on compact sets and in the J_1-metric of the Skorokhod space $\mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I)$, see Proposition 1.2.

Finally, Figure 4.2 depicts a trajectory of the limiting process $\xi\overline{Q}(\cdot/\xi)$.

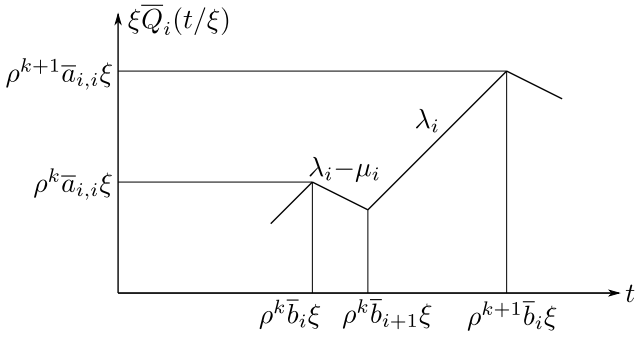

Figure 4.2: Fluid limit of queue i

4.5 Proofs for Section 4.3

In this section, we prove the properties of the offspring distribution of the embedded MTBP $\{\mathbf{Q}(t_n)\}_{n\in\mathbb{Z}_+}$.

4.5.1 Proof of Lemma 4.1

First we compute the γ_i's. For $k \in \mathbb{Z}_+ \cup \{\infty\}$, let a random variable $V_i(k)$ be the visit duration at queue i given that the service discipline at queue i is k-gated. Recall that the gating index equal ∞ corresponds to exhaustive service, and hence

$$\mathbb{E}V_i(\infty) = 1/(\mu_i - \lambda_i).$$

Now note that

$$\begin{aligned} V_i(0) &= 0, \\ V_i(k+1) &\stackrel{\mathrm{d}}{=} B_i + \textstyle\sum_{n=1}^{A_i(B_i)} V_{i,n}(k), \quad k \in \mathbb{Z}_+. \end{aligned} \tag{4.7}$$

where the random elements B_i, $A_i(\cdot)$ and $\{V_{i,n}(k)\}_{n\in\mathbb{N}}$ are mutually independent, and $V_{i,n}(k)$, $n \in \mathbb{N}$, are i.i.d. copies of $V_i(k)$. Then, for $k \in \mathbb{Z}_+$,

$$\begin{aligned} \mathbb{E}V_i(k+1) &= \frac{1}{\mu_i} + \frac{\lambda_i}{\mu_i}\mathbb{E}V_i(k) = \frac{1}{\mu_i}\left(1 + \frac{\lambda_i}{\mu_i}\right) + \left(\frac{\lambda_i}{\mu_i}\right)^2 \mathbb{E}V_i(k-1) = \dots \\ &= \frac{1}{\mu_i}\left(1 + \frac{\lambda_i}{\mu_i} + \dots + \left(\frac{\lambda_i}{\mu_i}\right)^k\right) + \left(\frac{\lambda_i}{\mu_i}\right)^{k+1} \mathbb{E}V_i(0) \\ &= \frac{1}{\mu_i}\frac{1 - (\lambda_i/\mu_i)^{k+1}}{1 - \lambda_i/\mu_i}, \end{aligned} \tag{4.8}$$

and

$$\gamma_i = \textstyle\sum_{k\in\mathbb{Z}_+\cup\{\infty\}} \mathbb{P}\{X_i = k\}\mathbb{E}V_i(k) = \dfrac{1 - \mathbb{E}(\lambda_i/\mu_i)^{X_i}}{\mu_i - \lambda_i}.$$

In a similar way, we compute the $m^{\mathrm{v}}_{i,i}$'s. For $k \in \mathbb{Z}_+ \cup \{\infty\}$, let a random variable $L^{\mathrm{v}}_{i,i}(k)$ be the queue i visit offspring of a queue i customer given that the service discipline at queue i is k-gated. We have

$$\begin{aligned} L^{\mathrm{v}}_{i,i}(\infty) &= 0, \\ L^{\mathrm{v}}_{i,i}(0) &= 1, \\ L^{\mathrm{v}}_{i,i}(k+1) &\stackrel{\mathrm{d}}{=} \textstyle\sum_{n=1}^{A_i(B_i)} L^{\mathrm{v}}_{i,i,n}(k), \quad k \in \mathbb{Z}_+, \end{aligned}$$

where the random elements B_i, $A_i(\cdot)$ and $\{L^{\mathrm{v}}_{i,i,n}(k)\}_{n\in\mathbb{N}}$ are mutually independent, and

$L^{\mathrm{v}}_{i,i,n}(k)$, $n \in \mathbb{N}$, are i.i.d. copies of $L^{\mathrm{v}}_{i,i}(k)$. Then, for $k \in \mathbb{Z}_+$,

$$\mathbb{E}L^{\mathrm{v}}_{i,i}(k+1) = (\lambda_i/\mu_i)\mathbb{E}L^{\mathrm{v}}_{i,i}(k) = \ldots = (\lambda_i/\mu_i)^{k+1},$$

and hence,

$$m^{\mathrm{v}}_{i,i} = \mathbb{E}(\lambda_i/\mu_i)^{X_i}.$$

The formulas for the $m^{\mathrm{v}}_{i,j}$'s, $i \neq j$, and the $m_{i,j}$'s follow, respectively, by the representations

$$L^{\mathrm{v}}_{i,j} \stackrel{\mathrm{d}}{=} A_j(V_i), \quad i \neq j, \tag{4.9}$$

where V_i and $A_j(\cdot)$ are independent, and

$$L_{i,j} \stackrel{\mathrm{d}}{=} L^{\mathrm{v}}_{i,j}\mathbb{I}\{i \geq j\} + \textstyle\sum_{n=1}^{L^{\mathrm{v}}_{i,i+1}} L_{i+1,j,n} + \ldots + \sum_{n=1}^{L^{\mathrm{v}}_{i,I}} L_{I,j,n}, \tag{4.10}$$

where $L_{i,j,n}$, $n \in \mathbb{N}$, are i.i.d. copies of $L_{i,j}$, and the sequences $\{L_{i,j,n}\}_{n\in\mathbb{N}}$, $i,j = 1,\ldots,I$, are mutually independent and do not depend on the vectors $\mathbf{L}^{\mathrm{v}}_i$, $i = 1,\ldots,I$.

4.5.2 Proof of Lemma 4.3

The cornerstone of this proof is finiteness of the corresponding moments for the busy periods of the queues in isolation, which we check with the help of the auxiliary Lemmas 4.4 and 4.5 that follow below together with their proofs.

Lemma 4.4. *Suppose that a function $f(\cdot)\colon \mathbb{R}_+ \to \mathbb{R}_+$ is bounded in a finite interval $[0,T]$ and non-decreasing in $[T,\infty)$, and that $f(x) \to \infty$ as $x \to \infty$. Suppose also that, for some (and hence for all) $c > 1$,*

$$\overline{\lim}_{x\to\infty} f(cx)/f(x) < \infty. \tag{4.11}$$

Consider an i.i.d. sequence $\{Y_n\}_{n\in\mathbb{N}}$ of non-negative, non-degenerate at zero random variables, and the renewal process

$$Y(t) = \max\{n \in \mathbb{Z}_+\colon \textstyle\sum_{k=1}^{n} Y_k \leq t\}, \quad t \in \mathbb{R}_+.$$

Let τ be a non-negative random variable which may depend on the sequence $\{Y_n\}_{n\in\mathbb{N}}$. Assume that $\mathbb{E}f(\tau) < \infty$. Then $\mathbb{E}f(Y(\tau))$ is finite too.

Proof. Without loss of generality, we can assume that the function $f(\cdot)$ is non-decreasing in the entire domain $\mathbb{R}_+$ and right-continuous (otherwise, instead of $f(\cdot)$, one can consider $\widetilde{f}(x) = \sup_{0\leq y\leq x} f(y)$, $x \in \mathbb{R}_+$).

First we show that, if (4.11) holds for some $c > 1$, then it holds for any $\widetilde{c} > 1$. For $\widetilde{c} = c^k$, $k \in \mathbb{N}$, we have

$$\overline{\lim}_{x\to\infty}\frac{f(c^k x)}{f(x)} \leq \overline{\lim}_{x\to\infty}\frac{f(c^k x)}{f(c^{k-1}x)}\overline{\lim}_{x\to\infty}\frac{f(c^{k-1}x)}{f(c^{k-2}x)}\ldots\overline{\lim}_{x\to\infty}\frac{f(cx)}{f(x)} < \infty.$$

Then, for $\widetilde{c} > 1$ other than powers of c, (4.11) follows by the monotonicity of $f(\cdot)$.

Condition (4.11) also implies that

$$\lim_{x\to\infty} \log(f(x))/x = 0. \tag{4.12}$$

Indeed, in (4.11) take $c = e$, the exponent. Since $M := \overline{\lim}_{x\to\infty} f(ex)/f(x) < \infty$, there exists a large enough $\widetilde{T} > 0$ such that $\sup_{x\in[\widetilde{T},\infty)} f(ex)/f(x) \le 2M$. Note that any $x \in [e\widetilde{T},\infty)$ admits a unique representation $x = e^{k(x)}y(x)$, where $y(x) \in [\widetilde{T}, e\widetilde{T})$ and $k(x) \in \mathbb{N}$. Hence, for any $x \in [e\widetilde{T},\infty)$,

$$f(x) = \frac{f(e^{k(x)}y(x))}{f(e^{k(x)-1}y(x))}\frac{f(e^{k(x)-1}y(x))}{f(e^{k(x)-2}y(x))}\cdots\frac{f(ey(x))}{f(y(x))}f(y(x)) \le (2M)^{k(x)}f(e\widetilde{T})$$

and

$$\frac{\log(f(x))}{x} \le \frac{k(x)\log(2M) + \log(f(e\widetilde{T}))}{\widetilde{T}e^{k(x)}},$$

implying (4.12).

Now define the pseudo-inverse function

$$f^{-1}(y) := \inf\{x \in \mathbb{R}_+ : f(x) \ge y\}, \quad y \in [0,\infty).$$

For any $c > 0$, we have

$$\begin{aligned}\mathbb{E}f(Y(\tau)) &\le \textstyle\sum_{n\in\mathbb{Z}_+} \mathbb{P}\{f(Y(\tau)) \ge n\} \le \sum_{n\in\mathbb{Z}_+} \mathbb{P}\{Y(\tau) \ge f^{-1}(n)\}\\ &\le \textstyle\sum_{n\in\mathbb{Z}_+} \mathbb{P}\{\sum_{k=1}^{\lceil f^{-1}(n)\rceil} Y_k \le \tau\} \le \Sigma_1(c) + \Sigma_2(c),\end{aligned}$$

where

$$\begin{aligned}\Sigma_1(c) &:= \textstyle\sum_{n\in\mathbb{Z}_+} \mathbb{P}\{\sum_{k=1}^{\lceil f^{-1}(n)\rceil} Y_k \le c\lceil f^{-1}(n)\rceil\},\\ \Sigma_1(c) &:= \textstyle\sum_{n\in\mathbb{Z}_+} \mathbb{P}\{cf^{-1}(n) < \tau\}.\end{aligned}$$

By condition (4.11), $\mathbb{E}f(\tau/c) < \infty$ for any $c > 0$, and hence

$$\Sigma_2(c) \le \textstyle\sum_{n\in\mathbb{Z}_+} \mathbb{P}\{f(\tau/c) \ge n\} \le 1 + \mathbb{E}f(\tau/c) < \infty.$$

We now pick a c such that $\Sigma_1(c) < \infty$, and this will finish the proof. By Markov's inequality, $\mathbb{P}\{\sum_{k=1}^n Y_k \le cn\} = \mathbb{P}\{\exp(-\sum_{k=1}^n Y_k) \ge e^{-cn}\} \le (e^c\,\mathbb{E}e^{-Y_1})^n$. Let c be small enough so that $\widetilde{c} := e^c\,\mathbb{E}e^{-Y_1} < 1$. Since $\lceil f^{-1}(n)\rceil = m$ implies $n \le f(m+1)$, we have

$$\Sigma_1(c) \le \textstyle\sum_{m\in\mathbb{Z}_+} \mathbb{P}\{\sum_{k=1}^m Y_k \le c'm\}f(m+1) \le \frac{1}{\widetilde{c}}\sum_{m\in\mathbb{N}} \widetilde{c}^m f(m).$$

Take an $\varepsilon \in (0, |\log(\widetilde{c})|)$. By (4.12), there exists a large enough $N \in \mathbb{N}$ such that $f(m) \le e^{m\varepsilon}$ for $m > N$. Then

$$\Sigma_1(c) \le \frac{1}{\widetilde{c}} \sum_{m=1}^{N} \widetilde{c}^m f(m) + \frac{1}{\widetilde{c}} \sum_{m=N+1}^{\infty} (\widetilde{c}e^{\varepsilon})^m,$$

where $\widetilde{c}e^{\varepsilon} = e^{\varepsilon - |\log(\widetilde{c})|} < 1$ by the choice of ε, and hence $\Sigma_1(c) < \infty$. □

Lemma 4.5. *Consider a sequence $\{Y_n\}_{n\in\mathbb{N}}$ of non-negative random variables that are identically distributed (but not necessarily independent), and also a $\mathbb{Z}_+$-valued random variable η that does not depend on $\{Y_n\}_{n\in\mathbb{N}}$. If $f(\cdot)\colon \mathbb{R}_+ \to \mathbb{R}$ is a convex function, then*

$$\mathbb{E}f(\textstyle\sum_{k=1}^{\eta} Y_k) \le \mathbb{E}f(\eta Y_1).$$

Proof. By the convexity of $f(\cdot)$, for any $n \in \mathbb{Z}_+$,

$$\mathbb{E}f(\textstyle\sum_{k=1}^{n} Y_k) = \mathbb{E}f(\sum_{k=1}^{n} \frac{1}{n}(nY_k)) \le \sum_{k=1}^{n} \frac{1}{n}\mathbb{E}f(nY_k) = \mathbb{E}f(nY_1).$$

Then Lemma 4.5 follows by the independence between $\{Y_n\}_{n\in\mathbb{N}}$ and η:

$$\begin{aligned}\mathbb{E}f(\textstyle\sum_{k=1}^{\eta} Y_k) &= \sum_{n\in\mathbb{Z}_+} \mathbb{P}\{\eta = n\} f(\sum_{k=1}^{n} Y_k) \\ &\le \sum_{n\in\mathbb{Z}_+} \mathbb{P}\{\eta = n\}\, \mathbb{E}f(nY_1) = \mathbb{E}f(\eta Y_1).\end{aligned}$$ □

Now we proceed with the proof of Lemma 4.3. It suffices to show that

$$\mathbb{E}f(L_{i,j}) < \infty, \quad \text{for all } i \text{ and } j,$$

where

$$f(x) = \begin{cases} 0, & x \in [0,1], \\ x\log x, & x \in [1,\infty). \end{cases}$$

Note that the function $f(\cdot)$ is convex: in $(1,\infty)$, its derivative $\log(\cdot)+1$ is non-decreasing, and in the other points, it is easy to check the definition of convexity. Also note that

$$f(xy) \le xf(y) + yf(x), \quad x, y \in \mathbb{R}_+. \tag{4.13}$$

The rest of the proof is divided into three parts. The two key steps are to show that the f-moments of the visit duration V_i and the same type visit offspring $L^{\mathrm{v}}_{i,i}$ are finite. Then the finiteness of the f-moments of the session offspring $L_{i,j}$ follows easily.

Finiteness of $\mathbb{E}f(V_i)$ It suffices to show that, in the $M/G/1$ queue (see Remark 1.1) with the arrival process $A_i(\cdot)$ and service times $B_{i,n}$, $n \in \mathbb{N}$, the f-moment of the busy period is finite. Suppose that at time $t = 0$, there is one customer in the queue, and his service time B^0_i is equal in distribution to B_i and independent from $A_i(\cdot)$ and $\{B_{i,n}\}_{n\in\mathbb{N}}$.

Let

$$\begin{aligned}\tau_i &= \min\{t > 0\colon \text{the queue is empty at } t\},\\ \tau_i(0) &= 0,\\ \tau_i(1) &= B_i^0,\\ \tau_i(k+2) &= \tau_i(k+1) + \textstyle\sum_{n=E_i(\tau_i(k))+1}^{E_i(\tau_i(k+1))} B_{i,n}, \quad k \in \mathbb{Z}_+.\end{aligned}$$

Whilst τ_i is a busy period, $\tau_i(k)$ is equal in distribution to the visit duration in queue i of the polling system given that the service discipline in that queue is k-gated, and

$$\tau_i(k) \uparrow \tau_i \quad \text{a.s. as } k \to \infty.$$

Now we show that the moments $\mathbb{E}f(\tau_i(k))$, $k \in \mathbb{Z}_+$, are bounded. Then the finiteness of $\mathbb{E}f(\tau_i)$ follows by the continuity of $f(\cdot)$ and the dominated convergence theorem.

Mimicking (4.7) , we have

$$\tau_i(k+1) \stackrel{\mathrm{d}}{=} B_i^0 + \textstyle\sum_{n=1}^{A_i(B_i^0)} \tau_{i,n}(k), \quad k \in \mathbb{N},$$

where $\tau_{i,n}(k)$, $n \in \mathbb{N}$, are i.i.d. copies of $\tau_i(k)$ that are independent from B_i^0 and $A_i(\cdot)$. Then, by the monotonicity and convexity of $f(\cdot)$, and the auxiliary Lemma 4.5 combined with (4.13),

$$\begin{aligned}\mathbb{E}f(\tau_i(k)) \le \mathbb{E}f(\tau_i(k+1)) &\le \frac{1}{2}\mathbb{E}f(2B_i^0) + \frac{1}{2}\mathbb{E}f(2\textstyle\sum_{n=1}^{A_i(B_i^0)} \tau_{i,n}(k))\\ &\le \frac{1}{2}\mathbb{E}f(2B_i^0) + \frac{1}{2}\mathbb{E}f(2A_i(B_i^0)\tau_{i,1}(k))\\ &\le \frac{1}{2}\mathbb{E}f(2B_i^0) + \frac{\lambda_i}{\mu_i}\mathbb{E}f(\tau_i(k)) + \frac{1}{2}\mathbb{E}\tau_i(k)\mathbb{E}f(2A_i(B_i^0)),\end{aligned}$$

where $\mathbb{E}f(2A_i(B_i^0)) < \infty$ by the auxiliary Lemma 4.4, and $\mathbb{E}\tau_i(k) \le 1/(\mu_i - \lambda_i)$ by (4.8). Thus, we have

$$\mathbb{E}f(\tau_i(k)) \le \frac{C_i}{1 - \lambda_i/\mu_i} \quad \text{for all } k \ge 2,$$

where

$$C_i = \frac{\mathbb{E}f(2B_i^0)}{2} + \frac{\mathbb{E}f(2A_i(B_i^0))}{2(\mu_i - \lambda_i)} < \infty.$$

Finiteness of $\mathbb{E}f(L_{i,i}^{\mathrm{v}})$ Note that $L_{i,i}$ is bounded stochastically from above by the number of service completions during the busy period of the $M/G/1$ queue introduced when proving the finiteness of $\mathbb{E}f(V_i)$. The number of service completions during the first busy period τ_i is given by $1 + A_i(\tau_i)$, and the finiteness of $\mathbb{E}f(1 + A_i(\tau_i))$ follows by the auxiliary Lemma 4.4.

Finiteness of $\mathbb{E}f(L_{i,j})$ This part of the proof uses mathematical induction. Now that we have shown the finiteness of the moments $\mathbb{E}f(L^{\mathrm{v}}_{i,i})$, (4.9) and Lemma 4.4 imply that

$$\mathbb{E}f(L^{\mathrm{v}}_{i,j}) < \infty \quad \text{for all } i \text{ and } j. \tag{4.14}$$

Then we have the basis of induction: $\mathbb{E}f(L_{I,j}) = \mathbb{E}f(L^{\mathrm{v}}_{I,j}) < \infty$ for all j. Suppose that $\mathbb{E}f(L_{k,j}) < \infty$ for $k = i+1, \dots, I$ and all j. Then the induction step (from $i+1$ to i) follows by (4.10), the convexity of $f(\cdot)$, Lemma (4.5) combined with (4.13), and (4.14).

4.6 Proofs for Section 4.4

First we make preparations in Sections 4.6.1 and 4.6.2, and then proceed with the proofs of Theorems 4.1 and 4.2 in Sections 4.6.3 and 4.6.4, respectively.

4.6.1 Additional notation

In this section we introduce a number of auxiliary random objects that we operate with when proving the a.s. convergence results of the chapter.

Queue length dynamics Define the renewal processes

$$B_i(t) := \max\{n \in \mathbb{Z}_+ \colon \textstyle\sum_{k=1}^{n} B_{i,k} \le t\}, \quad t \in \mathbb{R}_+,$$

and the processes

$$I_i(t) := \int_0^t \mathbb{I}\{\text{queue } i \text{ is in service at time } s\}\, ds, \quad t \in \mathbb{R}_+,$$

which keep track of how much time the server has spent in each of the queues. Then the number of queue i customers that have departed up to time t is given by

$$D_i(t) := B_i(I_i(t)).$$

Most of the a.s. convergence results of this chapter we derive from the basic equations

$$Q_i(\cdot) = A_i(\cdot) - D_i(\cdot).$$

The preliminary results of Section 4.6.2 depend on when the system empties for the last time. The number of indices n such that $\mathbf{Q}(t_n) = 0$ has a geometric distribution with parameter $q_G < 1$ (see Lemma 4.2). Denote by ν the last such index, i.e.

$$\mathbf{Q}(t_\nu) = \mathbf{0}, \quad \mathbf{Q}(t_n) \neq \mathbf{0} \quad \text{for all } n > \nu.$$

Ancestor-descendant relationships between customers By the following three rules, we define the binary relation "*is a descendant of*" on the set of customers:

- each customer is a descendant of himself;
- if customer 2 arrives while customer 1 is receiving service (the two customers are allowed to come from different queues), then customer 2 is a descendant of customer 1;
- if customer 2 is a descendant of customer 1, and customer 3 a descendant of customer 2, then customer 3 is a descendant of customer 1.

Now suppose that a customer is in position k in queue i at the beginning of visit n to queue i. Denote by $V_{i,n,k}$ the amount of time during the visit that his descendants are in service, and by $L^{\mathrm{v}}_{i,j,n,k}$ the number of his descendants in queue j at the end of the visit. If a customer is in position k in queue i at the beginning of session n, denote by $L_{i,j,n,k}$ the number of his descendants in queue j at the end of the session. Introduce also the random vectors

$$\begin{aligned}\mathbf{L}^{\mathrm{v}}_{i,n,k} &:= (L^{\mathrm{v}}_{i,1,n,k}, \ldots, L^{\mathrm{v}}_{i,I,n,k}),\\ \mathbf{L}_{i,n,k} &:= (L_{i,1,n,k}, \ldots, L_{i,I,n,k}).\end{aligned}$$

4.6.2 Preliminary results

In this section, we characterize the asymptotic behavior of the system at the switching instants $t_{i,n}$, laying the basis for Theorem 4.1 that concerns the bigger scale times t_{i,η_n}.

From the Kesten-Stigum theorem, we derive the following result for the $t_{1,n}$'s.

Lemma 4.6. *There exists a positive random variable ζ such that*

$$\text{a.s. as } n \to \infty, \quad \mathbf{Q}(t_n)/\rho^n \to \zeta \mathbf{v}, \quad \mathbf{Q}(t_{1,n})/\rho^n \to \zeta \mathbf{v}.$$

The distribution of ζ is given by: for $x \in (0, \infty)$,

$$\begin{aligned}\mathbb{P}\{\zeta \geq x\} =& \frac{1}{1-q_G} \sum\nolimits_{n \in \mathbb{Z}_+} \mathbb{P}\{\nu = n\} \sum\nolimits_{\substack{\mathbf{k} \in \mathbb{Z}^I_+,\\ ||\mathbf{k}||_1 \geq 1}} G(\mathbf{k})\\ &\times \sum\nolimits_{\substack{\mathbf{l} \leq \mathbf{k},\\ ||\mathbf{l}||_1 \geq 1}} \binom{\mathbf{k}}{\mathbf{l}} (\mathbf{1}-\mathbf{q})^l \mathbf{q}^{\mathbf{k}-\mathbf{l}} \mathbb{P}\{\textstyle\sum_{i=1}^I \sum_{m=1}^{l_i} \xi_{i,m} \geq \rho^{n+1} x\},\end{aligned} \tag{4.15}$$

where the random variables $\xi_{i,m}$ are the same as in Theorem 4.1.

Proof. Since $t_{1,n} = t_n$ for $n > \nu$, it suffices to find the a.s. limit of $\mathbf{Q}(t_n)/\rho^n$.

First we find the asymptotics of the auxiliary MTBP $\{\mathbf{Z}(n)\}_{n \in \mathbb{N}}$ (without immigration) under the assumption that $\mathbf{Z}(0)$ is distributed according to $\{G(\mathbf{k})\}_{\mathbf{k} \in \mathbb{Z}^I_+}$ (the immigration distribution for the MTBP $\{\mathbf{Q}(t_n)\}_{n \in \mathbb{N}}$).

By Proposition 4.1, if the distribution of $\mathbf{Z}(0)$ is $\{G(\mathbf{k})\}_{\mathbf{k}\in\mathbb{Z}_+^I}$, then, as $n\to\infty$,

$$\mathbf{Z}(n)/\rho^n \overset{\text{a.s.}}{\to} \Big(\underbrace{\textstyle\sum_{\mathbf{k}\in\mathbb{Z}_+^I}\mathbb{I}\{\mathbf{Z}(0)=\mathbf{k}\}\sum_{i=1}^I\sum_{m=1}^{k_i}\zeta_{i,m}}_{=:\,\zeta_G}\Big)\,\mathbf{v},$$

where $\zeta_{i,m}$, $m\in\mathbb{N}$, are i.i.d. copies of ζ_i, and the sequences $\{\zeta_{i,m}\}_{m\in\mathbb{N}}$, $i=1,\dots,I$, are mutually independent and also independent from $\mathbf{Z}(0)$.

The distribution of ζ_G is given by

$$\begin{aligned}
&\mathbb{P}\{\zeta_G=0\}=q_G,\\
&\mathbb{P}\{\zeta_G\ge x\}=\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^I,\\ ||\mathbf{k}||_1\ge 1}} G(\mathbf{k})\sum_{\substack{\mathbf{l}\le\mathbf{k},\\ ||\mathbf{l}||_1\ge 1}}\binom{\mathbf{k}}{\mathbf{l}}p(\mathbf{l})\\
&\qquad\qquad\times\underbrace{\textstyle\prod_{i=1}^I\left(\mathbb{P}\{\zeta_i>0\}\right)^{l_i}\left(\mathbb{P}\{\zeta_i=0\}\right)^{k_i-l_i}}_{=(\mathbf{1}-\mathbf{q})^{\mathbf{k}-\mathbf{l}}\mathbf{q}^{\mathbf{l}}},\quad x\in(0,\infty),
\end{aligned}\tag{4.16}$$

where

$$\begin{aligned}
p(\mathbf{l})&=\mathbb{P}\{\textstyle\sum_{i=1}^I\sum_{m=1}^{l_i}\zeta_{i,m}\ge x\;\big|\;\zeta_{i,m}>0\text{ for all }i\text{ and }m=1,\dots,l_i\}\\
&=\mathbb{P}\{\textstyle\sum_{i=1}^I\sum_{m=1}^{l_i}\xi_{i,m}\ge x\}
\end{aligned}$$

with the random variables $\xi_{i,m}$ defined in Theorem 4.1.

Now, on the event $\{\nu=N\}$, as $n\to\infty$,

$$\mathbf{Q}(t_{N+1+n})/\rho^n\overset{\text{a.s.}}{\to}\zeta_N\mathbf{v},$$

where

$$\mathbb{P}\{\zeta_N\in\cdot\}=\mathbb{P}\{\zeta_G\in\cdot\mid\mathbf{Z}(n)\ne\mathbf{0}\text{ for all }n\in\mathbb{Z}_+\}=\mathbb{P}\{\zeta_G\in\cdot\mid\zeta_G>0\}.$$

Then, as $n\to\infty$,

$$\mathbf{Q}(t_n)/\rho^n\overset{\text{a.s.}}{\to}\Big(\underbrace{\textstyle\sum_{N\in\mathbb{Z}_+}\mathbb{I}\{\nu=N\}\zeta_N/\rho^{N+1}}_{=:\,\zeta}\Big)\,\mathbf{v},$$

and it is left to check that the distribution of ζ is given by (4.15).

For $x\in(0,\infty)$, we have

$$\mathbb{P}\{\zeta\ge x\}=\textstyle\sum_{N\in\mathbb{Z}_+}\mathbb{P}\{\nu=N\}\mathbb{P}\{\zeta_G\ge\rho^{N+1}x\}/\mathbb{P}\{\zeta_G>0\},$$

and then (4.15) follows by (4.16). □

To deal with the other $t_{i,n}$'s, we combine the previous lemma with LLN's.

Lemma 4.7. *For $i = 1, \dots, I+1$, there exist constants $b_i \in (0, \infty)$ and $\mathbf{a}_i = (a_{i,1}, \dots, a_{i,I}) \in \mathbb{R}_+^I$ such that*

$$\text{a.s. as } n \to \infty, \quad t_{i,n}/\rho^n \to b_i\zeta, \quad \mathbf{Q}(t_{i,n}/\rho^n) \to \zeta \mathbf{a}_i.$$

The b_i's and $\mathbf{a}_i$'s are given by

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^I v_i/\mu_i}{\sum_{i=1}^I \lambda_i/\mu_i - 1}, \\ b_{i+1} &= b_i + (v_i + \lambda_i(b_i - b_1))\gamma_i, \quad i \le I, \end{aligned} \tag{4.17}$$

and

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{v}, \\ \mathbf{a}_{i+1} &= \mathbf{a}_i + (b_{i+1} - b_i)\boldsymbol{\lambda} - (b_{i+1} - b_i)\mu_i \mathbf{e}_i, \quad i \le I. \end{aligned} \tag{4.18}$$

The $\mathbf{a}_i$'s also satisfy

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{v}, \\ \mathbf{a}_{i+1} &= \mathbf{a}_i - a_{i,i}\mathbf{e}_i + a_{i,i}\mathbf{m}_i^{\mathrm{v}}, \quad i \le I. \end{aligned} \tag{4.19}$$

Remark 4.7. As we compare (4.17)–(4.18) with (4.3)–(4.4), it immediately follows that

$$b_i = \alpha \bar{b}_i \quad \mathbf{a}_i = \alpha \bar{\mathbf{a}}_i$$

Proof of Lemma 4.7. First we show that the sequences of $t_{i,n}/\rho^n$ and $\mathbf{Q}(t_{i,n})/\rho^n$ converge a.s., and that their limits satisfy the relations (4.18). Then we derive equations (4.17) and (4.19) relying on an LLN that, generally speaking, guarantees the b_i's to be in-probability-limits only.

Asymptotics of $t_{1,n}$ By the definition of ν, which is a.s. finite, we have, for $n > \nu$,

$$t_{1,n} = t_n = \Sigma + \sum_{i=1}^I \sum_{k=1}^{D_i(t_n)} B_{i,k}, \tag{4.20}$$

where

$$\Sigma := \sum_{m=0}^{\nu} (t_{1,m} - t_m) \overset{\text{a.s.}}{<} \infty.$$

Equation (4.20) with $D_i(t_n) = A_i(t_n) - Q_i(t_n)$ plugged in can be transformed into

$$t_{1,n} = t_n = \Sigma + t_n \Sigma_1(n) - \rho^n \Sigma_2(n),$$

where

$$\Sigma_1(n) := \sum_{i=1}^{I} \frac{\sum_{k=1}^{D_i(t_n)} B_{i,k}}{D_i(t_n)} \frac{A_i(t_n)}{t_n},$$
$$\Sigma_2(n) := \sum_{i=1}^{I} \frac{\sum_{k=1}^{D_i(t_n)} B_{i,k}}{D_i(t_n)} \frac{Q_i(t_n)}{\rho^n}.$$

Then we have

$$t_{1,n}/\rho^n = t_n/\rho^n = \frac{\Sigma_2(n) - \Sigma/\rho^n}{\Sigma_1(n) - 1}. \tag{4.21}$$

By the SLLN and Lemma 4.6, as $n \to \infty$,

$$\Sigma_1(n) \overset{\text{a.s.}}{\to} \sum_{i=1}^{I} \lambda_i/\mu_i, \quad \Sigma_2(n) \overset{\text{a.s.}}{\to} (\sum_{i=1}^{I} v_i/\mu_i)\zeta,$$

which, together with (4.21), implies that

$$t_n/\rho^n \overset{\text{a.s.}}{\to} b_1\zeta, \quad t_{1,n}/\rho^n \overset{\text{a.s.}}{\to} b_1\zeta, \tag{4.22}$$

where the value of b_1 is the one claimed in the lemma.

Convergence of $t_{i,n}/\rho^n$ Note that

$$t_{i+1,n} - t_{i,n} = I_i(t_{n+1}) - I_i(t_n),$$

and hence,

$$\frac{t_{i+1,n}}{\rho^n} = \frac{t_{i,n}}{\rho^n} + \frac{I_i(t_{n+1})}{B_i(I_i(t_{n+1}))} \frac{D_i(t_{n+1})}{\rho^{n+1}}\rho - \frac{I_i(t_n)}{B_i(I_i(t_n))} \frac{D_i(t_n)}{\rho^n}. \tag{4.23}$$

By the SLLN, as $n \to \infty$,

$$\frac{B_i(I_i(t_n))}{I_i(t_n)} \overset{\text{a.s.}}{\to} \mu_i. \tag{4.24}$$

By the SLLN, (4.21) and Lemma 4.6,

$$\frac{D_i(t_n)}{\rho^n} = \frac{A_i(t_n)}{t_n} \frac{t_n}{\rho^n} - \frac{Q_i(t_n)}{\rho^n} \overset{\text{a.s.}}{\to} (\lambda_i b_1 - v_i)\zeta. \tag{4.25}$$

As we put (4.22)–(4.25) together, it follows that there exist positive numbers b_i such that, as $n \to \infty$,

$$t_{i,n}/\rho^n \overset{\text{a.s.}}{\to} b_i\zeta, \quad i = 1, \ldots, I+1. \tag{4.26}$$

(The value of b_1 is the one claimed in the Lemma, and the equations for the other b_i's that follow from (4.22)–(4.25) are not given here since they will not be used anywhere in the proofs.)

Convergence of $\mathbf{Q}(t_{i,n})/\rho^n$ and (4.18) Since, during the time interval $[t_{i,n}, t_{i+1,n})$, there are no departures from queues other than i, we have

$$\begin{aligned} Q_j(t_{i+1,n}) - Q_j(t_{i,n}) = A_j(t_{i+1,n}) - A_j(t_{i,n}) \\ - \mathbb{I}\{j=i\}(B_i(I_i(t_{n+1})) - B_i(I_i(t_n))). \end{aligned} \tag{4.27}$$

By the SLLN and (4.26), as $n \to \infty$,

$$\frac{A_j(t_{i+1,n}) - A_j(t_{i,n})}{\rho^n} \overset{\text{a.s.}}{\to} \lambda_j(b_{i+1} - b_i)\zeta. \tag{4.28}$$

By (4.26),

$$\frac{I_i(t_n)}{\rho^n} = \frac{\sum_{k=1}^{n-1}(t_{i+1,k} - t_{i,k})}{\rho^n} \overset{\text{a.s.}}{\to} \frac{b_{i+1} - b_i}{\rho - 1}\zeta, \tag{4.29}$$

which, together with the SLLN, implies that

$$\frac{B_i(I_i(t_{n+1})) - B_i(I_i(t_n))}{\rho^n} \overset{\text{a.s.}}{\to} \mu_i(b_{i+1} - b_i)\zeta. \tag{4.30}$$

As we put Lemma 4.6 and (4.27)–(4.30) together, it follows that

$$\mathbf{Q}(t_{i,n})/\rho^n \overset{\text{a.s.}}{\to} \zeta \mathbf{a}_i, \quad i = 1, \dots, I+1, \tag{4.31}$$

where the vectors $\mathbf{a}_i = (a_{i,1}, \dots, a_{i,I})$ are given by (4.18).

Proof of (4.17) and (4.19) We derive (4.17) from the equations

$$t_{i+1,n} = t_{i,n} + \sum_{k=1}^{Q_i(t_{i,n})} V_{i,n,k}, \tag{4.32}$$

$$Q_i(t_{i,n}) = Q_i(t_{1,n}) + A_i(t_{i,n}) - A_i(t_{1,n}). \tag{4.33}$$

To (4.32), we apply the following form of the LLN (the proof is postponed to Section 4.7).

Statement 4.1. *Let a random variable Y have a finite mean value and, for each $n \in \mathbb{N}$, let $Y_{n,k}$, $k \in \mathbb{N}$, be i.i.d. copies of Y. Let τ_n, $n \in \mathbb{N}$, be $\mathbb{Z}_+$-valued random variables such that τ_n is independent of the sequence $\{Y_{n,k}\}_{k\in\mathbb{N}}$ for each n and $\tau_n \to \infty$ in probability as $n \to \infty$. Finally, let a sequence $\{T_n\}_{n\in\mathbb{N}}$ of positive numbers increase to ∞. If there exists an a.s. finite random variable τ such that $\tau_n/T_n \to \tau$ in probability as $n \to \infty$, then*

$$\sum_{k=1}^{\tau_n} Y_{n,k}/T_n \to \tau \mathbb{E}Y \quad \textit{in probability as } n \to \infty.$$

By (4.32) and Statement 4.1,

$$b_{i+1} - b_i = a_{i,i}\gamma_i. \tag{4.34}$$

By (4.33) and the SLLN,

$$a_{i,i} = v_i + \lambda_i(b_i - b_1). \tag{4.35}$$

Then (4.17) follows as we plug (4.35) into (4.34).

Finally, (4.19) follows as we apply Statement 4.1 to the equation

$$\mathbf{Q}(t_{i+1,n}) = \mathbf{Q}(t_{i,n}) - Q_i(t_{i,n})\mathbf{e}_i + \sum_{k=1}^{Q_i(t_{i,n})} \mathbf{L}^{\mathrm{v}}_{i,n,k}. \qquad \square$$

4.6.3 Proof of Thereom 4.1

This proof converts the results of Lemma 4.7 using the following tool.

Lemma 4.8. *Suppose that random variables Y_n, $n \in \mathbb{Z}$, and Y are such that*

$$Y_n/\rho^n \to Y\zeta \quad \text{a.s. as } n \to \infty.$$

Then, for all $k \in \mathbb{Z}$,

$$Y_{\eta_n+k}/\rho^n \to Y\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}\rho^k/\alpha \quad \text{a.s. as } n \to \infty.$$

Proof. First we show that,

$$\text{a.s., for all } n \text{ big enough}, \quad n - \eta_n = \lfloor \log_\rho(\alpha\zeta)\rfloor. \tag{4.36}$$

Indeed, we have $\log_\rho(t_n) - n = \log_\rho(\alpha\zeta) + \delta_n$, where $\delta_n \to 0$ a.s. as $n \to \infty$. Then $\eta_n = \min\{k\colon \log_\rho(t_k) \ge n\} = \min\{k\colon k + \delta_k \ge n - \log_\rho(\alpha\zeta)\}$. Introduce the event $\Omega_* := \{\delta_n \to 0,\ \log_\rho(\alpha\zeta) \notin \mathbb{Z}\}$. When estimated at any $\omega \in \Omega_*$, $\eta_n = \lceil n - \log_\rho(\alpha\zeta)\rceil = n - \lfloor \log_\rho(\alpha\zeta)\rfloor$ for all n big enough. Also we have $\mathbb{P}\{\Omega_*\} = 1$ because the distribution function of ζ is continuous in $(0,\infty)$ (see (4.15), where the r.v.'s $\xi_{i,m}$ have continuous densities on $(0,\infty)$ by Proposition 4.1).

Now fix a $k \in \mathbb{Z}$. By (4.36),

$$\frac{Y_{\eta_n+k}}{\rho^n} = \frac{Y_{\eta_n+k}}{\rho^{\eta_n+k}}\frac{\rho^k}{\rho^{n-\eta_n}} \overset{\text{a.s.}}{\to} Y\zeta\frac{\rho^k}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}},$$

where

$$\frac{\zeta}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}} = \frac{\rho^{\log_\rho(\alpha\zeta)}}{\rho^{\lfloor \log_\rho(\alpha\zeta)\rfloor}}\frac{1}{\alpha} = \rho^{\{\log_\rho(\alpha\zeta)\}}/\alpha,$$

and hence Lemma 4.8 is proven. $\square$

Now we proceed with the proof of Theorem 4.1.

Lemmas 4.7 and 4.8 imply that the convergence (4.2) holds with

$$\xi := \rho^{\{\log_\rho(\alpha\zeta)\}}.$$

By definition, ξ takes values in $[1,\rho)$, and it is left to calculate its distribution.

Fix an $x \in [1, \rho)$. Since

$$\begin{aligned}\mathbb{P}\{\xi \geq x\} &= \mathbb{P}\{\{\log_\rho(\alpha\zeta)\} \geq \log_\rho x\} \\ &= \textstyle\sum_{j\in\mathbb{Z}} \mathbb{P}\{j + \log_\rho x \leq \log_\rho(\alpha\zeta) < j+1\} \\ &= \textstyle\sum_{j\in\mathbb{Z}} \mathbb{P}\{\rho^j x/\alpha \leq \zeta < \rho^{j+1}/\alpha\},\end{aligned}$$

we have, by Lemma 4.6,

$$\begin{aligned}\mathbb{P}\{\xi \geq x\} =& \frac{1}{1-q_G} \sum_{n\in\mathbb{Z}_+} \mathbb{P}\{\nu = n\} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^I, \\ ||\mathbf{k}||_1 \geq 1}} G(\mathbf{k}) \\ &\times \sum_{\substack{\mathbf{l}\leq\mathbf{k}, \\ ||\mathbf{l}||_1\geq 1}} \binom{\mathbf{k}}{\mathbf{l}} (\mathbf{1}-\mathbf{q})^{\mathbf{l}} \mathbf{q}^{\mathbf{k}-\mathbf{l}} \Sigma(n, \mathbf{l}),\end{aligned}$$

where

$$\Sigma(n,\mathbf{l}) := \textstyle\sum_{j\in\mathbb{Z}} \mathbb{P}\{\rho^{j+n+1} x/\alpha \leq \sum_{i=1}^I \sum_{m=1}^{l_i} \xi_{i,m} < \rho^{j+n+2}/\alpha\}.$$

Note that $\Sigma_{n,\mathbf{l}}$ does not depend on n:

$$\begin{aligned}\Sigma_{n,\mathbf{l}} &= \textstyle\sum_{j\in\mathbb{Z}} \mathbb{P}\{\rho^j x/\alpha \leq \sum_{i=1}^I \sum_{m=1}^{l_i} \xi_{i,m} < \rho^{j+1}/\alpha\} \\ &= \textstyle\sum_{j\in\mathbb{Z}} \mathbb{P}\{j + \log_\rho x \leq \log_\rho(\alpha \sum_{i=1}^I \sum_{j=1}^{l_i} \xi_{i,m}) < j+1\} \\ &= \mathbb{P}\{\{\log_\rho(\alpha \textstyle\sum_{i=1}^I \sum_{j=1}^{l_i} \xi_{i,m})\} \geq \log_\rho x\},\end{aligned}$$

and this finishes the proof of Theorem 4.1.

4.6.4 Proof of Theorem 4.2

The proof consists of several steps. Throughout the proof, we assume that the function $\overline{\mathbf{Q}}(\cdot)$ is defined by (4.6). First we show that the process $\xi\overline{\mathbf{Q}}(\cdot/\xi)$ coincides a.s. with the pointwise limit of the scaled processes $\overline{\mathbf{Q}}^n(\cdot)$. Then we check that $\overline{\mathbf{Q}}(\cdot)$ satisfies (4.5) and is continuous. Finally, we prove that the pointwise convergence of the processes $\overline{\mathbf{Q}}^n(\cdot)$ implies their uniform convergence on compact sets.

Pointwise convergence To start with, we define the auxiliary event Ω_* on which, as $n \to \infty$,

$$\frac{t_{i,\eta_n+k}}{\rho^n} \to \rho^k \overline{b}_i \xi, \quad i = 1, \ldots, I+1, \quad k \in \mathbb{Z},$$
$$\frac{\mathbf{Q}(t_{i,\eta_n+k})}{\rho^n} \to \xi \rho^k \overline{a}_i, \quad i = 1, \ldots, I+1, \quad k \in \mathbb{Z},$$
$$\frac{I(t_{i,\eta_n+k})}{\rho^n} \to \rho^k \frac{\overline{b}_{i+1} - \overline{b}_i}{\rho(\rho-1)} \xi, \quad i = 1, \ldots, I, \quad k \in \mathbb{Z},$$

and, as $t \to \infty$,

$$A_i(t)/t \to \lambda_i, \quad B_i(t)/t \to \mu_i, \quad i = 1, \ldots, I.$$

By theorem 4.1, (4.29) and the SLLN, $\mathbb{P}\{\Omega_*\} = 1$.

We will now show that on Ω_*, as $n \to \infty$,

$$\overline{\mathbf{Q}}^n(t) \to \xi \overline{\mathbf{Q}}(t/\xi) \quad \text{for all } t \in \mathbb{R}_+, \tag{4.37}$$

where $\overline{\mathbf{Q}}(\cdot)$ is given by (4.6).

Fix a queue number i and an outcome $\omega \in \Omega_*$. All random objects in the rest of this part of the proof will be evaluated at this ω.

For $t = 0$, the convergence (4.37) holds since the system starts empty. For $t > 0$, we consider the three possible cases.

Case 1: $t \in [\rho^k \overline{b}_i \xi, \rho^k \overline{b}_{i+1} \xi)$ *for a* $k \in \mathbb{Z}$. By the definition of Ω_*, for all n big enough,

$$t_{i,\eta_n+k}/\rho^n < t < t_{i+1,\eta_n+k}/\rho^n,$$

implying that queue i is in service during $[t_{i,\eta_n+k}, \rho^n t)$, and hence

$$Q_i(\rho^n t) = Q_i(t_{i,\eta_n+k}) + (A_i(\rho^n t) - A_i(t_{i,\eta_n+k}) - (D_i(\rho^n t) - D_i(t_{i,\eta_n+k})),$$

where

$$D_i(\rho^n t) - D_i(t_{i,\eta_n+k}) = B_i\Big(I_i(t_{i,\eta_n+k}) + (\rho^n t - t_{i,\eta_n+k})\Big) - B_i(I_i(t_{i,\eta_n+k})).$$

Again by the definition of Ω_*, the last two equations imply that, as $n \to \infty$,

$$\overline{Q}_i^n(t) \to \rho^k \overline{a}_{i,i} \xi + \lambda_i(t - \rho^k \overline{b}_i \xi) - \mu_i(t - \rho^k \overline{b}_i \xi) = \xi \overline{Q}_i(t/\xi).$$

Case 2: $t \in [\rho^k \overline{b}_{i+1} \xi, \rho^{k+1} \overline{b}_i \xi)$ *for a* $k \in \mathbb{Z}$. In this case, for all n big enough,

$$t_{i+1,\eta_n+k}/\rho^n < t < t_{i,\eta_n+k+1}/\rho^n,$$

and hence, queue i is not in service during $[\rho^n t, t_{i,\eta_n+k+1})$, i.e.

$$Q_i(t_{i,\eta_n+k+1}) = Q_i(\rho^n t) + A_i(t_{i,\eta_n+k+1}) - A_i(\rho^n),$$

implying that

$$\overline{Q}_i^n(t) \to \rho^{k+1} a_{i,i}\xi - \lambda_i(\rho^{k+1}\overline{b}_i\xi - t) = \xi\overline{Q}_i(t/\xi).$$

Case 3: $t = \rho^k \overline{b}_i \xi$ *for a* $k \in \mathbb{Z}$. We have, as $n \to \infty$,

$$\begin{aligned} t_{i+1,\eta_n+k-1}/\rho^n &\to \rho^{k-1}\overline{b}_{i+1}\xi, \\ t_{i,\eta_n+k}/\rho^n &\to t, \\ t_{i+1,\eta_n+k}/\rho^n &\to \rho^k \overline{b}_{i+1}\xi, \end{aligned}$$

where the limits satisfy the inequality

$$\rho^{k-1}\overline{b}_{i+1}\xi < t < \rho^k \overline{b}_{i+1}\xi.$$

Hence, all n big enough fall into the two sets

$$\begin{aligned} \mathcal{N}_1 &:= \{n\colon t_{i,\eta_n+k} \le \rho^n t < t_{i+1,\eta_n+k}\} \\ \mathcal{N}_2 &:= \{n\colon t_{i+1,\eta_n+k-1} < \rho^n t < t_{i,\eta_n+k}\}. \end{aligned}$$

For $l = 1, 2$, we have to check that, if the set $\mathcal{N}_l$ is infinite, then

$$\overline{Q}_i^n(t) \to \rho^k \overline{a}_{i,i}\xi \quad \text{as } n \to \infty,\ n \in \mathcal{N}_l. \tag{4.38}$$

For $l = 1$, (4.38) follows along the lines of Case 1. For $l = 2$, we can prove (4.38) following the lines of Case 2 and replacing $k+1$ with k.

Equivalence of (4.5) **and** (4.6) Let $\widetilde{\mathbf{Q}}(\cdot) = (\widetilde{Q}_1, \ldots, \widetilde{Q}_I)(\cdot)$ be the unique solution to (4.5), whereas $\overline{\mathbf{Q}}(\cdot)$, as before, is given by (4.5). Fix a queue number i. The slopes of $\overline{Q}_i(\cdot)$ and $\widetilde{Q}_i(\cdot)$ coincide everywhere. Also $\overline{Q}_i(0) = 0 = \widetilde{Q}_i(0)$, and $\overline{Q}_i(\rho^k\overline{b}_i) = \rho^k\overline{a}_{i,i} = \widetilde{Q}_i(\rho^k\overline{b}_i)$, $k \in \mathbb{Z}$. Then it is left to check that

$$\overline{Q}_i(\rho^k\overline{b}_j) = \rho^k\overline{a}_{j,i} = \widetilde{Q}_i(\rho^k\overline{b}_j), \quad j \ne i, \quad k \in \mathbb{Z}. \tag{4.39}$$

If $j < i$, we have,

$$\begin{aligned} \rho^k\overline{b}_j &\in [\rho^{k-1}\overline{b}_{i+1}, \rho^k\overline{b}_i), \\ \overline{Q}_i(\rho^k\overline{b}_j) &= \rho^k(\overline{b}_i - \lambda_i(\overline{b}_i - \overline{b}_j)), \end{aligned}$$

and, if $j > i$,

$$\rho^k \overline{b}_j \in [\rho^k \overline{b}_{i+1}, \rho^{k+1} \overline{b}_i)$$
$$\overline{Q}_i(\rho^k \overline{b}_j) = \rho^k(\rho \overline{b}_i - \lambda_i(\rho \overline{b}_i - \overline{b}_j)).$$

Then (4.39) follows from the equations

$$\begin{aligned} Q_i(t_{i,n}) &= Q_i(t_{j,n}) + A_i(t_{i,n}) - A_i(t_{j,n}), \qquad j < i, \\ Q_i(t_{i,n+1}) &= Q_i(t_{j,n}) + A_i(t_{i,n+1}) - A_i(t_{j,n}), \quad j > i, \end{aligned}$$

by Lemma 4.7 and Remark 4.7.

Continuity of $\overline{Q}(\cdot)$ Fix a queue number i. As defined by (4.6), the function $\overline{Q}_i(\cdot)$ might have discontinuities only at $t = 0$ and $t = \rho^k \overline{b}_{i+1}$, $k \in \mathbb{Z}$.

Note that $\sup_{t \in [\rho^{k-1} \overline{b}_i, \rho^k \overline{b}_i)} \overline{Q}_i(t) = \rho^k \overline{a}_{i,i}$, $k \in \mathbb{Z}$. Then

$$\sup_{t \in (0, \rho^k \overline{b}_i)} \overline{Q}_i(t) = \sup_{l \in \mathbb{Z}, l \le k} \rho^k \overline{a}_{i,i} \to 0 \quad \text{as } k \to -\infty,$$

and hence, $\overline{Q}_i(t) \to 0 = \overline{Q}_i(0)$ as $t \to 0$.

At $t = \rho^k \overline{b}_{i+1}$, $k \in \mathbb{Z}$, the function $\overline{Q}_i(\cdot)$ is right-continuous with the left limit given by

$$\lim_{t \uparrow \rho^k \overline{b}_{i+1}} \overline{Q}_i(t) = \rho^k(\overline{a}_{i,i} + (\lambda_i - \mu_i)(\overline{b}_{i+1} - \overline{b}_i)).$$

By (4.4) and (4.39), we have

$$\lim_{t \uparrow \rho^k \overline{b}_{i+1}} \overline{Q}_i(t) = Q_i(\rho^k \overline{b}_{i+1}) = \rho^k a_{i+1,i}.$$

Uniform convergence on compact sets Define the auxiliary event $\widetilde{\Omega}$ on which, as $n \to \infty$, $\overline{\mathbf{Q}}^n(\cdot) \to \xi \overline{\mathbf{Q}}(\cdot/\xi)$ pointwise, and $A_i(\rho^n \cdot)/\rho^n \to \lambda_i(\cdot)$ u.o.c., $i = 1, \ldots, I$, where $\lambda_i(x) = \lambda_i x$ for all x. As follows from the first part of the proof and the functional SLLN, $\mathbb{P}\{\widetilde{\Omega}\} = 1$. For the rest of the proof, we estimate random objects at an outcome $\omega \in \widetilde{\Omega}$. Consider the scaled departure processes $D_i(\rho^n \cdot)/\rho^n = A_i(\rho^n \cdot)/\rho^n - \overline{Q}_i^n(\cdot)$. These processes are monotone and, by the definition of $\widetilde{\Omega}$, converge pointwise to the continuous functions $\lambda_i(\cdot) - \xi \overline{Q}_i(\cdot/\xi)$. Then they converge u.o.c., and the same is true for the processes $\overline{Q}_i^n(\cdot)$.

4.7 Proofs of auxiliary results

Proof of Lemma 4.2. Suppose $\rho \le 1$. Then, by Harris [49, Theorem 7.1], we have $q_i = 1$ for all i and $q_G = 1$. The latter implies that the queue length process $\mathbf{Q}(\cdot)$ hits $\mathbf{0}$ infinitely many times, and the same holds the workload process. Let $\{t_{n_k}\}_{k \in \mathbb{Z}_+}$ be the

sequence of consecutive time instants such that $\mathbf{Q}(t_{n_k}) = \mathbf{0}$. For different k, the differences $(t_{n_{k+1}} - t_{n_k})$ are bounded from below by the waiting times until the first arrival into the empty system, which are i.i.d. r.v.'s distributed exponentially with parameter $\sum_{i=1}^{I} \lambda_i$. Therefore, $t_{n_k} \to \infty$ a.s. as $k \to \infty$. This leads to a contradiction with the fact that the system is overloaded and its total workload grows infinitely large with time (by the SLLN, $(\sum_{i=1}^{I} \sum_{k=1}^{A_i(t)} B_{i,k} - t)/t \to \sum_{i=1}^{I} \lambda_i/\mu_i - 1 > 0$ a.s. as $t \to \infty$). Hence, $\rho > 1$, and then [49, Theorem 7.1] implies that $q_i < 1$ for all i and $q_G < 1$. □

Proof of Statement 4.1. First we show that

$$\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n \to \mathbb{E}Y \quad \text{in probability as } n \to \infty. \tag{4.40}$$

By the independence between τ_n and $\{Y_{n,k}\}_{k\in\mathbb{N}}$, for all $N \in \mathbb{Z}_+$,

$$\begin{aligned}&\mathbb{P}\{|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n - \mathbb{E}Y| \geq \varepsilon, \tau_n = N\}\\ &= \mathbb{P}\{|\textstyle\sum_{k=1}^{N} Y_{1,k}/N - \mathbb{E}Y| \geq \varepsilon\}\mathbb{P}\{\tau_n = N\}.\end{aligned}$$

Then, for any $M \in \mathbb{Z}_+$,

$$\begin{aligned}&\mathbb{P}\{|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n - \mathbb{E}Y| \geq \varepsilon\}\\ &\leq \mathbb{P}\{\tau_n \leq M\} + \sup_{N>M} \mathbb{P}\{|\textstyle\sum_{k=1}^{N} Y_{1,k}/N - \mathbb{E}Y| \geq \varepsilon\},\end{aligned}$$

and (4.40) follows as we first let $n \to \infty$, and then let $M \to \infty$.

Now that we have shown (4.40), the Statement follows by

$$\begin{aligned}&\mathbb{P}\{|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/T_n - \tau\mathbb{E}Y| \geq \varepsilon\}\\ &\leq \mathbb{P}\{|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n||\tau_n/T_n - \tau| \geq \varepsilon/2\} + \mathbb{P}\{\tau|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n - \mathbb{E}Y| \geq \varepsilon/2\}\\ &\leq \mathbb{P}\{x_1|\tau_n/T_n - \tau| \geq \varepsilon/2\} + \mathbb{P}\{|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n| > x_1\}\\ &\quad + \mathbb{P}\{x_2|\textstyle\sum_{k=1}^{\tau_n} Y_{n,k}/\tau_n - \mathbb{E}Y|\} + \mathbb{P}\{\tau > x_2\}\end{aligned}$$

as we first let $n \to \infty$, and then let $x_1 \to \infty$, $x_2 \to \infty$. □

Chapter 5

PS-queue with Multistage Service

5.1 Introduction

The original motivation for this chapter lies in a model different than the one claimed in the title. Namely, we are interested in freelance job websites, which have two kinds of visitors: customers offering jobs and freelancers, or servers, looking for jobs. The key feature of such websites is that multiple servers compete for a single job there. The most common situation is competition at the stage of application, i.e. to get the job. Along with that, the applicants might have to do the job, and then the one who has done it best gets paid — this is, for example, the way websites for finding the cheapest flight connections, such as `flightfox.com`, work.

To start with, we designed a basic model of a freelance job website, where there is a Poisson stream of customers and a Poisson stream of freelancers of rates λ and μ, respectively. Each customer upon arrival posts a job on the website main page and sets a patience clock that is distributed exponentially with parameter ν. Each freelancer upon arrival picks a job from the main page at random and applies for it, in the form of leaving a comment. If there are no jobs, the freelancer leaves. At most I applications are allowed per job. Once a customer receives the I-th application, or his patience expires, he should remove the job from the main page and continue communication with the applicants via private messaging. The state of this system is represented by the vector composed of the numbers of jobs on the main page with $i = 0, \dots, I-1$ applications. Mathematically, the constraint on the number of applications per job is a plus since it makes the problem finite-dimensional. In practice, such a threshold is always present implicitly: jobs with too many applications are not attractive anymore since the chance to get them is small. We are not aware of websites with an explicit threshold, but we suggest it as a guarantee of a chance to get a job, which is necessary in case of the "do-it-best-get-paid" policy or in case the website aims to expand and thus encourage unexperienced freelancers to join.

Our next observation (inspired by Borst et al. [17]) was that the basic freelance model is actually equivalent to a PS-queue where

- arrivals are Poisson of rate λ,
- customers re-enter the queue for I times with independent service requirements distributed exponentially with parameter μ,
- patience times of customers are exponentially distributed with parameter ν.

The state of the PS-queue is, respectively, the vector composed of the numbers of customers who have entered the queue for $i = 1, \ldots, I$ times, or we also say "*customers at stage i of service*". Finally, we generalised the model by allowing service requirements at different stages of service have different parameters μ_i (the distribution is still exponential).

As in the previous chapters, we develop approximations for the model under study using the fluid limit approach. We show that trajectories of the per-stage population process, when scaled properly, converge to solutions of a system of differential equations, which in turn stabilise to the unique invariant solution over time. Convergence of the scaled trajectories follows because asymptotically they live on a compact set and their oscillations are small. To prove convergence of fluid limits to the invariant point, we use an equivalent description of fluid limits, which is a generalisation of the approximating equation suggested by Gromoll et al. [47] for a single-stage-service PS queue.

Driven by mathematical curiosity, we also tried another method to establish the asymptotic stability of the invariant point — the method of Lyapunov functions. Note that the multistage service can be interpreted as tandem routing, that is from class i to class $i+1$. Along with that, the model we consider assumes customers to be impatient. By Bramson [23], PS with Markovian routing and patient customers admits an entropy-like Lyapunov function. For PS with impatience and no routing, there exists a quadratic Lyapunov function, see Remark 5.2 in Section 5.8. To the best of our knowledge, no Lyapunov function is known for PS with both routing and impatience. We have two partial solutions to this problem. We have proved that the entropy Lyapunov function of Bramson [23] still works outside a compact neighbourhood of the invariant point if impatience is allowed with the same parameters for all classes. It guarantees that fluid limits get attracted into that neighbourhood of the invariant point over time. If there are only two classes, the entropy function works everywhere. We also suggest a quadratic Lyapunov function for PS with two classes with different impatience parameters and a Markovian routing.

In the future, we aim to build on the motivation behind this chapter. A next logical step would be to incorporate the service stage in addition to the application stage. We are mostly interested in the scenario when the same job is done multiple times, which mathematically is a special kind of dependence of job sizes. There are also optimization questions that arise in practice. For example, if freelancers are ranked in a way, what strategies should they follow to build and maintain a strong reputation? The majority of freelance websites exist at the cost of transaction fees, then what are the ways to increase website profits while keeping transaction fees affordable to visitors? With so many possible directions, we believe this area of research is promising.

The remainder of the chapter is organised as follows. In Section 5.2, we discuss in detail how the PS queue with multistage service arises from the basic freelance model. In

Section 5.3, we introduce two equivalent deterministic systems of equations that are analogues of the stochastic model, and check that, for any initial state, the solution to these systems is unique and stabilises to the unique invariant point in a long time run. Section 5.4 specifies the fluid scaling under which the stochastic model converges to its deterministic analogues. In Sections 5.5 and 5.7, the proofs for the results of Sections 5.3 and 5.4 are given. In Section 5.8, we discuss Lyapunov functions for PS with routing and impatience. Section 5.6 shows how the convergence to the invariant point in the single-stage-service case implies that for the multistage-service case.

All notations of this chapter are listed in the introduction to the thesis.

5.2 Stochastic model

As was mentioned in the introduction, our original interest is in the dynamics of freelance job websites. So we start out with a basic model of such a website, and then transform it into a more general model of PS with multistage service. This section describes both the original and more general models, and explains in what sense one is a particular case of the other.

Basic model of a freelance job website There are two types of visitors on a freelance job website: customers, who publish job descriptions, and freelancers, who apply for those jobs. We assume that new jobs appear on the website main page according to a Poisson process of rate λ, and that freelancers intending to find a job visit the website according to a Poisson process of rate μ. As a freelancer looking for a job visits the website, he picks a job from the main page at random and applies for it, say leaves a comment. Each job is allowed to collect at most I applications while its patience time lasts, measured from the moment the job was published, and exponentially distributed with parameter ν. Patience times for different jobs are mutually independent and also do not depend on the arrival processes of jobs and freelancers. As soon as a job either gets I applications, or its patience time expires, the customer removes the job description from the main page and continues communication with the applicants elsewhere. In this model, our focus is going to be on the process

$$\mathbf{Q}^{\mathrm{FL}}(\cdot) = (Q_0^{\mathrm{FL}}, \dots, Q_{I-1}^{\mathrm{FL}})(\cdot),$$

where $Q_i^{\mathrm{FL}}(t)$ is the number of jobs on the main page that have collected i applications up to time instant t.

PS-queue with multistage service Now consider a PS queue with Poisson arrivals of rate λ. We assume that each customer of this queue should undergo I stages of service, stage $i+1$ starting immediately upon completion of stage i and the service requirement at stage i being exponentially distributed with parameter μ_i. A customer is supposed to leave the queue upon service completion, but if his patience time expires earlier, he abandons then. As in the previous model, patience times are distributed exponentially

with parameter ν. Also the arrival process, service requirements of all customers at all stages and patience times of all customers are mutually independent. We are going to keep track of how populated different stages of service are, i.e. to analyse the process

$$\mathbf{Q}(\cdot) = (Q_1, \dots, Q_I)(\cdot),$$

where $Q_i(t)$ stands for the number of customers in stage i of service at time instant t.

Equivalence of the two models in case all μ_i's are the same Suppose that, in the second model, all service stages have the same distribution parameter μ. In this case, the processes $\mathbf{Q}^{\mathrm{FL}}(\cdot)$ and $\mathbf{Q}(\cdot)$ are distributed identically. The idea is that the jobs waiting for the i-th application (i.e. those with $i-1$ applications) can be viewed as customers of the PS-queue who are undergoing stage i of service, and the moments jobs receive applications — as completions of stages of service in the PS-queue. When a freelancer applies for a job, he picks one at random. Correspondingly, if there is a service stage completion in the PS-queue, all of the service stages that have been ongoing are equally likely to be the one that has finished. That is due to the memoryless property of the exponential distribution and because all the μ_i's are the same.

The above insight originally belongs to Borst et al. [17], who discussed the equivalence of PS and random order of service in the context of a $GI/M/1$ queue (see Remark 1.1). To formalise the idea they constructed a probabilistic coupling, which can be generalised in a straightforward way to the two models we consider here.

Dynamic equations From now on, we will be working with the more general model of PS with multistage service. We assume it is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$. Denote the arrival process of customers, which is Poisson of rate λ, by $A(\cdot)$. These are arrivals to stage 1 of service. Let $D_i^{\mathrm{s}}(\cdot)$ stand for the process of service completions at stage i. Note that, for $i \leq I-1$, $D_i^{\mathrm{s}}(\cdot)$ is the arrival process to stage $i+1$, and $D_I^{\mathrm{s}}(\cdot)$ is the process of departures due to total service completions. Finally, denote by $D_i^{\mathrm{a}}(\cdot)$ the process of abandonments due to impatience at stage i. Since service requirements at all stages and patience times of all customers are exponentially distributed, and since the exponential distribution is memoryless, the processes $D_i^{\mathrm{s}}(\cdot)$ and $D_i^{\mathrm{a}}(\cdot)$ are doubly stochastic Poisson with instantaneous rates $\mu_i Q_i(\cdot)/\|Q(\cdot)\|$ (zero by convention when the system is empty) and $\nu Q_i(\cdot)$, respectively. That is, the population process $\mathbf{Q}(\cdot) = (Q_1, \dots, Q_I)(\cdot)$ can be represented as the unique (see e.g. [72]) solution to the following system of equations: for $t \in \mathbb{R}_+$,

$$\begin{aligned} Q_1(t) &= Q_1(0) + A(t) - D_1^{\mathrm{s}}(t) - D_1^{\mathrm{a}}(t), \\ Q_i(t) &= Q_i(0) + D_{i-1}^{\mathrm{s}}(t) - D_i^{\mathrm{s}}(t) - D_i^{\mathrm{a}}(t), \quad i \geq 2, \end{aligned} \tag{5.1}$$

with

$$\begin{aligned} D_i^{\mathrm{s}}(t) &= \Pi_i^{\mathrm{s}}\Big(\mu_i \int_0^t \frac{Q_i(u)}{\|\mathbf{Q}(u)\|_1}\, du\Big), \\ D_i^{\mathrm{a}}(t) &= \Pi_i^{\mathrm{a}}\Big(\nu \int_0^t Q_i(u) du\Big), \end{aligned} \tag{5.2}$$

where $\Pi_i^{\mathrm{s}}(\cdot), \Pi_i^{\mathrm{a}}(\cdot)$ are Poisson processes of unit rate for all i, and also the initial state $\mathbf{Q}(0) = (Q_1, \ldots, Q_I)(0)$, the arrival process $A(\cdot)$ and the processes $\Pi_i^{\mathrm{s}}(\cdot)$, $\Pi_i^{\mathrm{a}}(\cdot)$ are mutually independent.

Finally, we assume the following throughout the rest of the chapter.

Assumption 5.1. *The system is overloaded, i.e.* $\lambda \sum_{i=1}^{I} 1/\mu_i > 1$.

The subsequent sections will characterise fluid limits of the process $\mathbf{Q}(\cdot)$. The overload regime guarantees that they are non-trivial, and hence make sensible approximations.

5.3 Fluid model

In this section, we define and analyse a fluid model which is a deterministic analogue of the PS-queue with multistage service introduced above. In the next section, this stochastic queue will be shown to converge to the fluid model under a proper scaling.

Definition 5.1. A function $\mathbf{z}(\cdot) = (z_1, \ldots, z_I)(\cdot) \colon \mathbb{R}_+ \to \mathbb{R}_+^I$ that is continuous and such that $\inf_{t \geq \delta} \|\mathbf{z}(t)\| > 0$ for any $\delta > 0$ is called a *fluid model solution (FMS)* if it solves the following system of differential equations: for $t > 0$,

$$\begin{aligned} z_1'(t) &= \lambda - \mu_1 \frac{z_1(t)}{\|\mathbf{z}(t)\|_1} - \nu z_1(t), \\ z_i'(t) &= \mu_{i-1} \frac{z_{i-1}(t)}{\|\mathbf{z}(t)\|_1} - \mu_i \frac{z_i(t)}{\|\mathbf{z}(t)\|_1} - \nu z_i(t), \quad i \geq 2. \end{aligned} \tag{5.3}$$

When investigating properties of FMS's, we will also use an alternative description of them. Let r.v.'s B_i, $i = 1, \ldots, I$, and D, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, be mutually independent and exponentially distributed, B_i with parameter μ_i for all i and D with parameter ν. Introduce also

$$B_j^i := \begin{cases} \sum_{l=j}^{i} B_l, & j \leq i, \\ 0, & j > i. \end{cases}$$

It turns out that (5.3) is equivalent to the following system of integral equations: for $i = 1, \ldots, I$ and $t \in \mathbb{R}_+$,

$$\begin{aligned} z_i(t) = {} & \textstyle\sum_{j=1}^{i} z_j(0)\, \mathbb{P}\Big\{ B_j^{i-1} \leq \int_0^t \frac{du}{\|\mathbf{z}(u)\|_1} < B_j^i,\, D > t \Big\} \\ & + \lambda \int_0^t \mathbb{P}\Big\{ B_1^{i-1} \leq \int_s^t \frac{du}{\|\mathbf{z}(u)\|_1} < B_1^i,\, D > t - s \Big\}\, ds. \end{aligned} \tag{5.4}$$

The two systems are equivalent in the sense that they have the same set of continuous, non-negative, non-zero outside $t = 0$ solutions.

While the differential equations (5.3) are direct analogues of the stochastic equations (5.1)–(5.2), it only takes a little thought to see that the integral equations (5.4) mimic the evolution of the stochastic system as well. Indeed, given a customer arrives at time instant s, he will be undergoing stage i of service at time instant $t \geq s$ if his patience time allows it, and if the amount of service he will have gotten up to t will cover the service requirements of the first $i-1$ stages completely and the service requirement of stage i only partially. That explains the second term in the RHS of (5.4). The first term has the same interpretation but in the context of customers who were present in the system at $t = 0$. Due to the memoryless property of the exponential distribution, the residual service requirements of the service stages that are ongoing at $t = 0$ are still exponentially distributed with the corresponding parameters.

A rigorous proof of the equivalence of the two descriptions of FMS's follows in Section 5.5. It exploits certain properties of the exponential and phase-type distributions.

We will now proceed with the analysis of FMS's.

Theorem 5.1. *For any initial state $\mathbf{z}(0)$, a FMS exists and is unique.*

Proof. Existence of FMS's is established in Sections 5.4 and 5.7: fluid limits of the population process $\mathbf{Q}(\cdot)$ are FMS's. When proving uniqueness, we distinguish between two cases. If the initial state is non-zero, the uniqueness follows from the description (5.3) by the Gronwall inequality — just as for the ALOHA-model of Chapter 2, see Theorem 2.1. In case the initial state is zero, we use the description (5.4). The summation of the equations of (5.4) where $\mathbf{z}(0) = \mathbf{0}$ implies that the norm $\|\mathbf{z}(\cdot)\|_1$ solves the following equation: for $t \in \mathbb{R}_+$,

$$x(t) = \lambda \int_0^t \mathbb{P}\Big\{ B_1^I > \int_s^t \frac{du}{x(u)}, D > t - s \Big\}\, ds. \tag{5.5}$$

The last equation is, in fact, the fluid model of a PS-queue with single-stage service; it is studied in Gromoll et al. [47] and shown to have a unique solution that is bounded away from zero outside $t = 0$, see Corollary 3.8. So the norm $\|\mathbf{z}(\cdot)\|_1$ is unique. Then a solution to (5.4) must be unique as well, since the individual coordinates $z_i(\cdot)$ are uniquely defined by the norm $\|\mathbf{z}(\cdot)\|_1$ in (5.4). □

The next question we are interested in is whether there are invariant, or constant, FMS's as they are candidates for the long-time limits of the rest of FMS's.

Theorem 5.2. *There exists a unique invariant FMS, which is given by*

$$\begin{aligned} z_1^* &= \frac{\lambda}{\mu_1 + \nu \|\mathbf{z}^*\|_1} \|\mathbf{z}^*\|_1, \\ z_i^* &= \frac{\mu_{i-1}}{\mu_i + \nu \|\mathbf{z}^*\|_1} z_{i-1}^*, \quad i \geq 2, \end{aligned} \tag{5.6}$$

where $\|\mathbf{z}^*\|_1$ *solves*

$$f(\|\mathbf{z}^*\|_1) := \lambda\left(\frac{1}{\mu_1+\nu\|\mathbf{z}^*\|_1}+\frac{\mu_1}{(\mu_1+\nu\|\mathbf{z}^*\|_1)(\mu_2+\nu\|\mathbf{z}^*\|_1)} + \cdots + \frac{\mu_1\dots\mu_{I-1}}{(\mu_1+\nu\|\mathbf{z}^*\|_1)\dots(\mu_I+\nu\|\mathbf{z}^*\|_1)}\right) = 1. \tag{5.7}$$

Proof. By definition, an invariant FMS must be non-zero. It follows from the description (5.3) that an invariant FMS $\mathbf{z}^* = (z_1^*,\dots,z_I^*)$ is defined by the following system of equations:

$$\begin{aligned} &\lambda - \mu_1\frac{z_1^*}{\|\mathbf{z}^*\|_1} - \nu z_1^* = 0,\\ &\mu_{i-1}\frac{z_{i-1}^*}{\|\mathbf{z}^*\|_1} - \mu_i\frac{z_i^*}{\|\mathbf{z}^*\|_1} - \nu z_i^* = 0, \quad i\geq 2. \end{aligned} \tag{5.8}$$

As we solve the i-th equation in (5.8) with respect to z_i^*, we obtain (5.6).

Now, (5.6) is equivalent to

$$\begin{aligned} z_1^* &= \frac{\lambda}{\mu_1+\nu\|\mathbf{z}^*\|_1}\,\|\mathbf{z}^*\|_1,\\ z_i^* &= \frac{\mu_{i-1}\dots\mu_1}{(\mu_i+\nu\|\mathbf{z}^*\|_1)\dots(\mu_2+\nu\|\mathbf{z}^*\|_1)}\,z_1^*, \quad i\geq 2. \end{aligned}$$

As we sum up over the last set of equations and divide by $\|\mathbf{z}^*\|_1$ on both sides, (5.7) follows.

Note that equations (5.6)–(5.7) have a unique solution. Indeed, the function $f(\cdot)$ is strictly decreasing in $(0,\infty)$ and takes all values between $\lambda\sum_{i=1}^I 1/\mu_i$ (which is bigger than 1 by Assumption 5.1) and 0 as it arguments runs from 0 to ∞. Hence (5.7) uniquely defines the norm $\|\mathbf{z}^*\|_1$, and then (5.7) uniquely defines the individual coordinates z_i via $\|\mathbf{z}^*\|_1$. □

Finally, we show the invariant FMS is asymptotically stable.

Theorem 5.3. *Any FMS* $\mathbf{z}(t)$ *converges to the unique invariant FMS* $\mathbf{z}^*$ *as* $t\to\infty$.

When proving the last theorem, we again refer to the paper [47] on PS with single-stage service. The equations of (5.4) summed up give: for $t\in\mathbb{R}_+$,

$$\begin{aligned} \|\mathbf{z}(t)\|_1 = \sum_{j=1}^I z_j(0)\,\mathbb{P}\Big\{B_j^I > \int_0^t \frac{du}{\|\mathbf{z}(u)\|_1}, D>t\Big\}\\ + \lambda\int_0^t \mathbb{P}\Big\{B_1^I > \int_s^t \frac{du}{\|\mathbf{z}(u)\|_1},\, D>t-s\Big\}\,ds. \end{aligned} \tag{5.9}$$

In the last equation, we put $\mathbf{z}(\cdot) \equiv \mathbf{z}^*$ and take $t \to \infty$, which implies that the norm $\|\mathbf{z}^*\|$ of the invariant FMS should solve the equation

$$x = \lambda \mathbb{E} \min\{x(B_1 + \ldots + B_I), D\}. \tag{5.10}$$

Gromoll et al. [47] show that (5.10) has a unique solution, so it must be $\|\mathbf{z}^*\|$; see Theorem 2.4 in [47]. It also follows from Theorem 2.4 that all solutions $\|\mathbf{z}(t)\|$ to (5.9) converge to the unique solution $\|\mathbf{z}^*\|$ of (5.10) as $t \to \infty$. To be precise, the theorem works with a slightly different equation than (5.9), but the difference is in the terms that represent the initial customers and vanish as $t \to \infty$. Now that we have the convergence of the norm $\|\mathbf{z}(t)\| \to \|\mathbf{z}^*\|$ for any FMS $\mathbf{z}(\cdot)$, the coordinate-wise convergence can be shown with the use of the same ideas as in Theorem 2.4 of Gromoll et al. [47]. We present the proof in Section 5.6 for completeness.

We conclude the section with a brief discussion of another approach to establishing stability of invariant solutions — the method of Lyapunov functions. It is known that a PS-queue with multiple classes of customers and Markovian routing admits an entropy Lyapunov function (see Bramson [20]), and a PS-queue with impatience and no routing — a quadratic Lyapunov function (see Remark 5.2 in Section 5.8). Whether there is a Lyapunov function for a PS-queue with both routing and impatience is an open problem. We have tried to solve it in the context of our model, where the routing is tandem: from class i to class $i+1$, and where all classes have the same impatience parameter. But these particular specifications do not seem to make the problem easier. However, we have come up with partial solutions: an entropy Lyapunov function that works everywhere except a compact set, and a quadratic function that works for two classes. We present these results in Section 5.8.

5.4 Fluid limit theorem

We have mentioned earlier in the chapter that the PS-queue with multistage service converges to the fluid model introduced in Section 5.3 when properly scaled. In the present section, we make those claims rigorous.

We use the same scaling as for the ALOHA-model of Chapter 2. That is, consider a family of models upper-indexed by positive numbers r, all of them defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let the arrival rate λ and the parameters μ_i of service stages be the same in all models (and satisfy Assumption 5.1), and the impatience parameter of model r be ν/r. Define the scaled population processes

$$\overline{\mathbf{Q}}^r(t) := \mathbf{Q}^r(rt)/r, \quad t \in \mathbb{R}_+. \tag{5.11}$$

We call weak limits along subsequences of the processes (5.11) *fluid limits*. The following result provides their characterisation.

Theorem 5.4. *Suppose that* $\overline{\mathbf{Q}}^r(0) \Rightarrow \mathbf{z}(0)$ *as* $r \to \infty$, *where* $\mathbf{z}(0)$ *is a random vector. Then the processes* $\overline{\mathbf{Q}}^r(\cdot)$ *converge weakly in the Skorokhod space* $\mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I)$ *to the unique FMS*

with initial state $\mathbf{z}(0)$.

The proof is given in Section 5.7. It follows the same, traditional strategy as in Chapters 2 and 3. First we show that fluid limits exist relying on the arguments of compact containment and oscillation control. Then we check that they are FMS's by deriving the fluid model equations (5.3) from the scaled stochastic dynamics (5.1)–(5.2).

5.5 Equivalence of the two fluid model descriptions

This proof partly relies on the ideas of the proof of a similar result in Chapter 2, see Lemma 2.1, but it is more involved. In particular, it uses and proves the special property (5.14) of the phase-type distribution.

Let a function $\mathbf{z}\colon \mathbb{R}_+ \to \mathbb{R}_+^I$ be continuous and non-zero outside $t = 0$.

Proof of (5.3) $\Rightarrow$ (5.4) Suppose that $\mathbf{z}(\cdot)$ is a solution to (5.3). Consider the following Cauchy problem with respect to $\mathbf{u}(\cdot)$: for $t > 0$,

$$\begin{aligned} u_1'(t) &= \lambda - \mu_1 \frac{u_1(t)}{\|\mathbf{z}(t)\|_1} - \nu u_1(t), \\ u_i'(t) &= \mu_{i-1} \frac{u_{i-1}(t)}{\|\mathbf{z}(t)\|_1} - \mu_i \frac{u_i(t)}{\|\mathbf{z}(t)\|_1} - \nu u_i(t), \quad i > 2, \\ \mathbf{u}(0) &= \mathbf{z}(0). \end{aligned} \tag{5.12}$$

This problem has at most one continuous solution. Indeed, let $\mathbf{u}(\cdot)$ and $\widetilde{\mathbf{u}}(\cdot)$ be two continuous solutions to (5.12). Then the difference $\mathbf{w}(\cdot) := (\mathbf{u} - \widetilde{\mathbf{u}})(\cdot)$ satisfies: for $t > 0$,

$$\begin{aligned} w_1'(t) &= -w_1(t) \left(\frac{\mu_1}{\|\mathbf{z}(t)\|_1} + \nu \right), \\ w_i'(t) &= w_{i-1}(t) \frac{\mu_{i-1}}{\|\mathbf{z}(t)\|_1} - w_i(t) \left(\frac{\mu_i}{\|\mathbf{z}(t)\|_1} + \nu \right), \quad i > 2, \\ \mathbf{w}(0) &= \mathbf{0}. \end{aligned}$$

Note that if $w_1(t) > 0$, then $w_1'(t) < 0$, and the other way around. Then $w_1(\cdot) \equiv 0$ by Lemma 2.1 and $w_2'(t) = -w_2(t)(\mu_i/\|\mathbf{z}(t)\|_1 + \nu)$, $t > 0$. To each pair $i, i+1$ of coordinates, we apply the same reasoning as we did to the first two, and thus obtain $\mathbf{w}(\cdot) \equiv \mathbf{0}$.

It is straightforward to check that the LHS and RHS of (5.4) both satisfy (5.12). Since a solution to (5.12) must be unique, the LHS and RHS of (5.4) must coincide.

Proof of (5.4) $\Rightarrow$ (5.3) Suppose now that $\mathbf{z}(\cdot)$ solves (5.4). As we differentiate the RHS of (5.4), it follows that, for $t > 0$,

$$\begin{aligned}
z_1'(t) &= \lambda - \mu_1 \frac{z_1(t)}{\|\mathbf{z}(t)\|_1} - \nu z_1(t), \\
z_i'(t) &= \sum_{j=1}^{i} \left(f_{B_j^{i-1}} - f_{B_j^i} \right) \left(\int_0^t \frac{du}{\|\mathbf{z}(u)\|_1} \right) \frac{\mathbb{P}\{D > t\}}{\|\mathbf{z}(t)\|_1} \\
&\quad + \lambda \int_0^t \left(f_{B_1^{i-1}} - f_{B_1^i} \right) \left(\int_s^t \frac{du}{\|\mathbf{z}(u)\|_1} \right) \frac{\mathbb{P}\{D > t-s\}}{\|\mathbf{z}(t)\|_1} \, ds \\
&\quad - \nu z_i(t), \quad i \geq 2,
\end{aligned}$$

where $f_{B_j^i}(\cdot)$ denotes the probability density function of the phase-type r.v.

$$B_j^i := \begin{cases} \sum_{l=j}^{i} B_l, & j \leq i, \\ 0, & j > i, \end{cases}$$

where B_l is exponentially distributed with parameter μ_l.

At this stage, in order to have (5.3), it suffices to show that, for $t > 0$,

$$\begin{aligned}
\mu_i z_i(t) = &\sum_{j=1}^{i} f_{B_j^i} \left(\int_0^t \frac{du}{\|\mathbf{z}(u)\|_1} \right) \mathbb{P}\{D > t\} \\
&+ \lambda \int_0^t f_{B_1^i} \left(\int_s^t \frac{du}{\|\mathbf{z}(u)\|_1} \right) \frac{\mathbb{P}\{D > t-s\}}{\|\mathbf{z}(t)\|_1} \, ds.
\end{aligned} \tag{5.13}$$

As we compare (5.13) to (5.4) (which we assume to hold), we conclude that, in order to have (5.13), it suffices to show that, for all $i \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$\frac{1}{\mu_i} f_{B_1^i}(x) = \mathbb{P}\{B_1^{i-1} \leq x < B_1^i\},$$

or equivalently,

$$\mathbb{P}\{B_1^i > x\} = \sum_{j=1}^{i} \frac{1}{\mu_j} f_{B_1^j}(x). \tag{5.14}$$

(To be precise, for $i > I$, we need to introduce r.v.'s B_i that are exponentially distributed with parameters μ_i, mutually independent with each other and with B_j, $j \leq I$.)

We prove (5.14) by induction: it holds for $i = 1$, assume that it holds for an $i \geq 1$, we have to check that it holds for $i + 1$ as well. By the convolution formula,

$$\begin{aligned}
&\mathbb{P}\{B_1^{i+1} > x\} \\
&= \int_0^\infty \mathbb{P}\{(y + B_2^{i+1} > x\} f_{B_1}(y) dy \\
&= \int_x^\infty f_{B_1}(y) dy + \int_0^x \mathbb{P}\{B_2^{i+1} > x - y\} f_{B_1}(y) dy.
\end{aligned}$$

Now we incorporate the induction hypothesis and obtain

$$\begin{aligned}\mathbb{P}\{B_1^{i+1} > x\} &= \mathbb{P}\{B_1 > x\} + \textstyle\sum_{j=2}^{i+1} \frac{1}{\mu_j} \int_0^x f_{B_2^j}(x-y) f_{B_1}(y) dy \\ &= \frac{1}{\mu_1} f_{B_1}(x) + \textstyle\sum_{j=2}^{i+1} \frac{1}{\mu_j} \int_{-\infty}^{\infty} f_{B_2^j}(x-y) f_{B_1}(y) dy \\ &= \frac{1}{\mu_1} f_{B_1}(x) + \textstyle\sum_{j=2}^{i+1} \frac{1}{\mu_j} f_{B_1^j}(x).\end{aligned}$$

So (5.14) indeed holds implying (5.13); and (5.13), in turn, implies (5.3).

5.6 Proof of Theorem 5.3

It follows from the fluid model description (5.4), that the coordinates of the invariant FMS $\mathbf{z}^*$ are uniquely defined by its norm via

$$z_i^* = \lambda \mathbb{E} \min\{\|\mathbf{z}^*\|_1 B_1^i, D\} - \lambda \mathbb{E} \min\{\|\mathbf{z}^*\|_1 B_1^{i-1}, D\} \quad \text{for all } i. \tag{5.15}$$

It is shown in Gromoll et al. [47, Theorem 2.4] that, for any FMS $\mathbf{z}(\cdot)$, we have $\|\mathbf{z}(t)\|_1 \to \|\mathbf{z}^*\|_1$ as $t \to \infty$. Here we will derive the coordinate-wise convergence from the convergence of norms.

As we compare (5.4) to (5.15), it follows that, in order to have $z_i(t) \to z_i^*$ as $t \to \infty$, it suffices to show that, for all i,

$$\int_0^t f_i(s,t)\, ds \to \mathbb{E} \min\{\|\mathbf{z}^*\|_1 B_1^i, D\}, \tag{5.16}$$

where

$$f_i(s,t) = \mathbb{P}\Big\{ B_1^i > \int_s^t \frac{du}{\|\mathbf{z}(u)\|_1}, D > t - s \Big\}.$$

Fix an $\varepsilon \in (0, \|\mathbf{z}^*\|_1)$ and let t_ε be such that

$$\|\mathbf{z}^*\|_1 - \varepsilon \le \|\mathbf{z}(t)\|_1 \le \|\mathbf{z}^*\|_1 + \varepsilon \quad \text{for all } t \ge t_\varepsilon.$$

For any fixed s, $f_i(s,t) \to 0$ as $t \to \infty$, and then, by the dominated convergence theorem,

$$\int_0^{t_\varepsilon} f_i(s,t)\, ds \to 0 \quad \text{as } t \to \infty. \tag{5.17}$$

For all $t \geq t_\varepsilon$, we have

$$\begin{aligned}\int_{t_\varepsilon}^t f_i(s,t)\,ds &\leq \int_{t_\varepsilon}^t \mathbb{P}\Big\{B_1^i > \int_s^t \frac{du}{\|\mathbf{z}^*\|_1+\varepsilon},\, D > t-s\Big\}\,ds \\ &\leq \int_0^{t-t_\varepsilon} \mathbb{P}\left\{\min\{(\|\mathbf{z}^*\|_1+\varepsilon)B_1^i,\, D\} \geq s\right\}\,ds,\end{aligned}$$

which, in combination with (5.17), implies that

$$\overline{\lim}_{t\to\infty} \int_0^t f_i(s,t)\,ds \leq \mathbb{E}\min\{(\|\mathbf{z}^*\|_1+\varepsilon)B_1^i,\, D\}.$$

Similarly, we obtain

$$\underline{\lim}_{t\to\infty} \int_0^t f_i(s,t)\,ds \geq \mathbb{E}\min\{(\|\mathbf{z}^*\|_1-\varepsilon)B_1^i,\, D\}.$$

As we take $\varepsilon \to 0$ in the last two equations, (5.16) follows.

5.7 Proof of Theorem 5.4

The proof consists of two parts. First we show that the family of the fluid scaled processes $\overline{\mathbf{Q}}^r(\cdot)$ is **C**-tight, i.e. that fluid limits exist and are continuous. Then we check that fluid limits are FMS's, i.e. that they are bounded away from zero outside $t = 0$ and solve the fluid model equations (5.3).

Throughout the proof, we use the following representation of the processes $\mathbf{Q}^r(\cdot)$ (they all are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$): for $t \in \mathbb{R}_+$,

$$\begin{aligned}Q_1^r(t) &= Q_1^r(0) + A(t) - D_1^{r,\mathrm{s}}(t) - D_1^{r,\mathrm{a}}(t), \\ Q_i^r(t) &= Q_i^r(0) + D_{i-1}^{r,\mathrm{s}}(t) - D_i^{r,\mathrm{s}}(t) - D_i^{r,\mathrm{a}}(t), \quad i \geq 2,\end{aligned} \tag{5.18}$$

with

$$\begin{aligned}D_i^{r,\mathrm{s}}(t) &= \Pi_i^{\mathrm{s}}\left(\mu_i \int_0^t \frac{Q_i^r(u)}{\|\mathbf{Q}^r(u)\|_1}\,du\right), \\ D_i^{r,\mathrm{a}}(t) &= \Pi_i^{\mathrm{a}}\left(\frac{\nu}{r}\int_0^t Q_i^r(u)du\right),\end{aligned} \tag{5.19}$$

where the processes $A(\cdot)$ and $\Pi_i^{\mathrm{s}}(\cdot), \Pi_i^{\mathrm{a}}(\cdot)$ are the same as in (5.1)–(5.2), except that this time we assume them to be independent from the family of the initial states $\mathbf{Q}^r(0)$.

C-tightness In order to prove that the family of the processes $\overline{\mathbf{Q}}^r(\cdot)$ is **C**-tight, it suffices to show that the following two properties hold (see Proposition 1.3): for any $T > 0$

and $\varepsilon > 0$, there exist an $M < \infty$ and a $\delta > 0$ such that

$$\underline{\lim}_{r\to\infty} \mathbb{P}\{\|\overline{\mathbf{Q}}^r(T)\| \le M\} \ge 1 - \varepsilon, \tag{5.20}$$

and

$$\underline{\lim}_{r\to\infty} \mathbb{P}\{\sup_{\substack{s,t\in[0,T],\\ |s-t|<\delta}} \|\overline{\mathbf{Q}}^r(s) - \overline{\mathbf{Q}}^r(t)\| \le \varepsilon\} \ge 1 - \varepsilon. \tag{5.21}$$

The compact containment condition (5.20) follows easily by the upper bound

$$\|\overline{\mathbf{Q}}^r(T)\| \le \|\overline{\mathbf{Q}}^r(0)\| + A(rT)/r \Rightarrow \|\mathbf{z}(0)\| + \lambda T \quad \text{as } r \to \infty.$$

Take an $\widetilde{M} < \infty$ that is a continuity point for the distribution of $\|\mathbf{z}(0)\|_1$ such that $\mathbb{P}\{\|\mathbf{z}(0)\| \le \widetilde{M}\} \ge 1 - \varepsilon$ and put $M = \widetilde{M} + \lambda T + 1$.

To establish the oscillation control condition (5.21), it is enough to have oscillations of the scaled departure processes $D_i^{r,\mathrm{s}}(r\cdot)/r$ and $D_i^{r,\mathrm{a}}(r\cdot)/r$ bounded.

Define the modulus of continuity for functions $x\colon \mathbb{R}_+ \to \mathbb{R}$,

$$\omega(x, T, \delta) := \sup\{|x(s) - x(t)|\colon s, t \in [0, T], |s - t| < \delta\}.$$

First we estimate oscillations of $D_i^{r,\mathrm{s}}(r\cdot)/r$. We have, for all $t \ge s \ge 0$,

$$\left|\frac{D_i^{r,\mathrm{s}}(rs)}{r} - \frac{D_i^{r,\mathrm{s}}(rt)}{r}\right| \le |G_i^{r,\mathrm{s}}(s)| + |G_i^{r,\mathrm{s}}(t)| + \mu_i \int_s^t \frac{\overline{Q}_i^r(u)}{\|\overline{\mathbf{Q}}^r(u)\|_1}\, du,$$

where, for all $t \in \mathbb{R}_+$,

$$G_i^{r,\mathrm{s}}(t) := \frac{1}{r}\Pi_i^{\mathrm{s}}\left(r\mu_i \int_0^t \frac{\overline{Q}_i^r(u)}{\|\overline{\mathbf{Q}}^r(u)\|_1}\, du\right) - \mu_i \int_0^t \frac{\overline{Q}_i^r(u)}{\|\overline{\mathbf{Q}}^r(u)\|_1}\, du. \tag{5.22}$$

Then

$$\omega\left(\frac{D_i^{r,\mathrm{s}}(r\cdot)}{r}, T, \delta\right) \le 2\sup_{t\in[0,\mu_i T]}\left|\frac{\Pi_i^{\mathrm{s}}(rt)}{r} - t\right| + \delta. \tag{5.23}$$

Now we switch to $D_i^{r,\mathrm{a}}(r\cdot)/r$. Consider a family of $M/M/\infty$ queues (as defined in Example 1.3) with a common arrival process $A(\cdot)$, queue r starting with $\|\mathbf{Q}^r(0)\|_1$ customers, and service times in queue r being patience times of the corresponding customers in the r-th PS-queue with multistage service. Denote the departure process of the r-th $M/M/\infty$-queue by $\widetilde{D}^r(\cdot)$. We have, for all i and $s, t \in \mathbb{R}_+$,

$$\left|\frac{D_i^{r,\mathrm{a}}(rs)}{r} - \frac{D_i^{r,\mathrm{a}}(rt)}{r}\right| \le \left|\frac{\widetilde{D}^r(rs)}{r} - \frac{\widetilde{D}^r(rt)}{r}\right|,$$

and hence,

$$\omega(D^{r,\mathrm{a}}(r\cdot)/r, T, \delta) \le \omega(\widetilde{D}^r(r\cdot)/r, T, \delta). \tag{5.24}$$

By e.g. Robert [95], the scaled processes $\widetilde{D}^r(r\cdot)/r$ converge weakly in the Skorokhod space $\mathbf{D}(\mathbb{R}_+, \mathbb{R}_+)$ to a continuous limit, which we denote by $\widetilde{D}(\cdot)$. (Although technically the fluid scalings considered in Robert [95] and here are different: arrival rates and space versus time and space, they result in the same distributions of the scaled processes, and hence the same limit.)

Since the modulus of continuity $\omega(\cdot, T, \delta)$ as a function on $\mathbf{D}(\mathbb{R}_+, \mathbb{R})$ is continuous at any continuous $x(\cdot)$, we have, by the continuous mapping theorem,

$$\omega(\widetilde{D}^r(r\cdot)/r, T, \delta) \Rightarrow \omega(\widetilde{D}(\cdot), T, \delta) \quad \text{as } r \to \infty. \tag{5.25}$$

Since continuity implies uniform continuity on compact sets, we also conclude that

$$\omega(\widetilde{D}(\cdot), T, \delta) \Rightarrow 0 \quad \text{as } \delta \to \infty \tag{5.26}$$

Finally, as we put together the FLLN for $A(\cdot)$, (5.23) and the FLLN for $\Pi_i^s(\cdot)$, and also (5.24)–(5.26), it follows that one can pick a δ such that (5.21) holds.

Fluid limits as FMS's Now that we know that fluid limits exist, it is left to check that they are FMS's. Consider a fluid limit $\widetilde{\mathbf{Q}}(\cdot)$ along a subsequence $\{\overline{\mathbf{Q}}^q(\cdot)\}_{q\to\infty}$. The **C**-tightness part of the proof implies that $\widetilde{\mathbf{Q}}(\cdot)$ is a.s. continuous. As we discussed before, the total population process of a PS-queue with multistage service behaves as an ordinary, single-stage-service PS-queue, whose fluid limits are studied by Gromoll et al. [47]. In particular, it follows from Assumption 5.1 and [47, Lemma 6.1] that a.s., for all $\delta > 0$, $\inf_{t\geq\delta} \|\widetilde{\mathbf{Q}}(t)\| > 0$. We will now show that $\widetilde{\mathbf{Q}}(\cdot)$ a.s. satisfies the fluid model equations (5.3), and this will finish the proof.

Consider the mappings $\varphi_i \colon \mathbf{D}(\mathbb{R}_+, \mathbb{R}_+^I) \to \mathbf{D}(\mathbb{R}_+, \mathbb{R})$, $i = 1, \ldots, I$, given by

$$\begin{aligned}
\varphi_1(\mathbf{x})(t) =: {} & x_1(t) - x_1(0) - \lambda t \\
& + \mu_1 \int_0^t \frac{x_1(u)}{\|\mathbf{x}(u)\|_1}\, du + \nu \int_0^t x_1(u)\, du, \\
\varphi_i(\mathbf{x})(t) =: {} & x_i(t) - x_i(0) - \mu_{i-1} \int_0^t \frac{x_{i-1}(u)}{\|\mathbf{x}(u)\|_1}\, du \\
& + \mu_i \int_0^t \frac{x_i(u)}{\|\mathbf{x}(u)\|_1}\, du + \nu \int_0^t x_i(u)\, du, \quad i \geq 2.
\end{aligned}$$

These mappings are continuous at any $\mathbf{x}(\cdot)$ that is continuous and non-zero outside $t = 0$. Then, by the continuous mapping theorem, for all i,

$$\varphi_i(\overline{\mathbf{Q}}^q) \Rightarrow \varphi_i(\widetilde{\mathbf{Q}}) \quad \text{as } q \to \infty. \tag{5.27}$$

On the other hand, it follows from the stochastic dynamics (5.18)–(5.19) that, for all q

and $t \in \mathbb{R}_+$,

$$\begin{aligned} \varphi_1(\overline{\mathbf{Q}}^q)(t) &= (A(qt)/q - \lambda t) - G_1^{q,\mathrm{s}}(t) - G_1^{q,\mathrm{a}}(t), \\ \varphi_i(\overline{\mathbf{Q}}^q)(t) &= G_{i-1}^{q,\mathrm{s}}(t) - G_i^{q,\mathrm{s}}(t) - G_i^{q,\mathrm{a}}(t), \quad i \geq 2, \end{aligned} \tag{5.28}$$

where, for all i and $t \in \mathbb{R}_+$,

$$G_i^{q,\mathrm{a}}(t) := \frac{1}{q}\Pi_i^{\mathrm{a}}\left(q\nu \int_0^t \overline{Q}_i^q(u)\,du\right) - \nu \int_0^t \overline{Q}_i^q(u)\,du,$$

and the processes $G_i^{q,\mathrm{s}}(\cdot)$ were defined earlier by (5.22).

We are going to use the following result (see e.g. Billingsley [9]).

Proposition 5.1 (Random time change theorem). *Consider stochastic processes $X^q(\cdot) \in \mathbf{D}(\mathbb{R}_+, S)$, where S is a complete and separable metric space, and non-decreasing stochastic processes $\Phi^q(\cdot) \in \mathbf{D}(\mathbb{R}_+, \mathbb{R}_+)$. Assume that the joint convergence $(X^q, \Phi^q)(\cdot) \Rightarrow (X, \Phi)(\cdot)$ holds as $q \to \infty$, and that the limits $X(\cdot)$ and $\Phi(\cdot)$ are a.s. continuous. Then $X^q(\Phi^q(\cdot)) \Rightarrow X(\Phi(\cdot))$ in $\mathbf{D}(\mathbb{R}_+, S)$ as $q \to \infty$.*

Now put $X^q(t) = \Pi_i^{\mathrm{s}}(qt)/q - t$ and $\Phi^q(t) = \nu \int_0^t \overline{Q}_i^q(u)\,du$ for all $t \in \mathbb{R}_+$. The marginal weak limits of these processes are $X(\cdot) \equiv 0$ and $\Phi(\cdot) = \nu \int_0^{\cdot} \widetilde{Q}_i(u)\,du$, respectively. Since one of the marginal limits is deterministic, we actually have the joint weak convergence, and then Proposition 5.1 implies that, as $q \to \infty$,

$$G_i^{q,\mathrm{a}}(\cdot) \Rightarrow 0 \quad \text{in } \mathbf{D}(\mathbb{R}_+, \mathbb{R}). \tag{5.29}$$

Similarly,

$$G_i^{q,\mathrm{s}}(\cdot) \Rightarrow 0 \quad \text{in } \mathbf{D}(\mathbb{R}_+, \mathbb{R}). \tag{5.30}$$

As we put (5.28)–(5.30) together with (5.27), it follows that

$$\text{a.s., for all } i, \quad \varphi_i(\widetilde{\mathbf{Q}}) \equiv 0,$$

which, after differentiation, gives (5.3).

5.8 Candidate Lyapunov functions for PS

An alternative way to establish the asymptotic stability of the invariant solution to the fluid model (5.3) would be to suggest a Lyapunov function, i.e. (see Proposition 1.1) a function $L(\cdot)$ defined on $(0, \infty)^I$ that is non-negative, such that $L(\mathbf{z}) \to \infty$ as $\|\mathbf{z}\| \to \infty$ and whose derivative with respect to (5.3) is non-positive. In this section, our attempts to find such a function are discussed.

We will consider a more general fluid model than (5.3), namely

$$z_i'(t) = \lambda_i + \sum_{j=1}^{I} P_{j,i}\mu_j \frac{z_j(t)}{\|\mathbf{z}(t)\|_1} - \mu_i \frac{z_i(t)}{\|\mathbf{z}(t)\|_1} - \nu_i z_i(t), \qquad i = 1, \ldots, I, \quad t > 0. \tag{5.31}$$

The system (5.31) is a deterministic analogue of a PS-queue with I classes of customers, where λ_i is the arrival rate to class i, $1/\mu_i$ and ν_i are the mean service time and abandonment rate of a class i customer, respectively, and $P_{i,j}$ is the probability that a class i customer, upon finishing service, is rerouted to class j. Naturally, we assume that the $P_{i,j}$'s form a sub-stochastic matrix:

$$\text{for any } i, \quad \sum_{j=1}^{I} P_{i,j} \leq 1.$$

Additionally, we assume that the system (5.31) is overloaded and has a unique invariant solution $\mathbf{z}^*$ (as in the particular case (5.3)), i.e. there exists a unique solution to

$$\lambda_i + \sum_{j=1}^{I} P_{j,i}\mu_j \frac{z_j^*}{\|\mathbf{z}^*\|_1} - \mu_i \frac{z_i^*}{\|\mathbf{z}^*\|_1} - \nu_i z_i^* = 0, \qquad i = 1, \ldots, I. \tag{5.32}$$

We have tested two kinds of candidates for a Lyapunov function for (5.31). The first one is an entropy function. It turned out to work for two classes with the same abandonment rates; and if there are more than two classes, it works everywhere except a compact set. The second candidate is a quadratic function. It works for two classes with different abandonment rates. The details follow below.

Entropy Lyapunov function Assume that the abandonment rates are the same in all classes: $\nu_1 = \ldots = \nu_I = \nu$. Consider the function $L_{\mathrm{lg}}(\cdot)$ defined on $(0, \infty)^I$ by

$$L_{\mathrm{lg}}(\mathbf{z}) := \sum_{i=1}^{I} z_i \log\left(\frac{z_i/\|\mathbf{z}\|_1}{z_i^*/\|\mathbf{z}^*\|_1}\right).$$

Since $L_{\mathrm{lg}}(\mathbf{z})/\|\mathbf{z}\|_1$ is the Kullback-Leibler distance between the distributions $\{z_i/\|\mathbf{z}\|_1\}_{i=1}^I$ and $\{z_i^*/\|\mathbf{z}^*\|_1\}_{i=1}^I$, the function $L_{\mathrm{lg}}(\cdot)$ is non-negative on $(0, \infty)^I$.

The following two lemmas check the sign of the derivative of $L_{\mathrm{lg}}(\cdot)$ with respect to (5.31), which is given by

$$L_{\mathrm{lg}}'(\mathbf{z}) := \sum_{i=1}^{I} R_i(\mathbf{z}) \log\left(\frac{z_i/\|\mathbf{z}\|_1}{z_i^*/\|\mathbf{z}^*\|_1}\right), \tag{5.33}$$

where

$$\begin{aligned} R_i(\mathbf{z}) &:= \lambda_i + \sum_{j=1}^I P_{j,i}\mu_j \frac{z_j}{\|\mathbf{z}\|_1} - \mu_i \frac{z_i}{\|\mathbf{z}\|_1} - \nu z_i \\ &= \sum_{j=1}^I P_{j,i}\mu_j \left(\frac{z_j(t)}{\|\mathbf{z}(t)\|_1} - \frac{z_j^*}{\|\mathbf{z}^*\|_1} \right) \\ &\quad - \mu_i \left(\frac{z_i(t)}{\|\mathbf{z}(t)\|_1} - \frac{z_i^*}{\|\mathbf{z}^*\|_1} \right) - \nu(z_i(t) - z_i^*). \end{aligned}$$

First we consider the case of two classes.

Lemma 5.1. *If $I = 2$, then $L'_{\mathrm{lg}}(\cdot) \le 0$ on $(0,\infty)^2$.*

Proof. Fix a $\mathbf{z} \in (0,\infty)$ and, to shorten notation, put

$$p_i := z_i/\|\mathbf{z}\|_1, \quad q_i := z_i^*/\|\mathbf{z}^*\|_1, \quad i,j = 1,2.$$

We make the following rearrangements in (5.33):

$$\begin{aligned} L'_{\mathrm{lg}}(z) &= \sum_{i=1}^2 \left(\sum_{j=1}^2 P_{j,i}\mu_j(p_j - q_j) - \mu_i(p_i - q_i) - \nu(z_i - z_i^*) \right) \log(p_i/q_i) \\ &= \Sigma_1 + \Sigma_2 - \nu\Sigma_3, \end{aligned}$$

where

$$\begin{aligned} \Sigma_1 &= \sum_{i=1}^2 (P_{i,i} - 1)\mu_i(p_i - q_i)\log(p_i/q_i), \\ \Sigma_2 &= P_{2,1}\mu_2(p_2 - q_2)\log(p_1/q_1) + P_{1,2}\mu_1(p_1 - q_1)\log(p_2/q_2) \\ \Sigma_3 &= \|\mathbf{z}\|_1 \sum_{i=1}^2 p_i \log(p_i/q_i) + \|\mathbf{z}^*\|_1 \sum_{i=1}^2 q_i \log(q_i/p_i). \end{aligned}$$

Since $(p_i - q_i)$ and $\log(p_i/q_i) = \log(p_i) - \log(q_i)$ are of the same sign, and $P_{i,i} \le 1$, we have $\Sigma_1 \le 0$. Now note that $p_i \le q_i$ implies that $p_{3-i} \ge q_{3-i}$. Then $(p_i - q_i)$ and $\log(p_{3-i}/q_{3-i})$ are of different signs, and hence $\Sigma_2 \le 0$. Finally, $\Sigma_3 \ge 0$ because it is the sum of two Kullback-Leibler distances with non-negative weights. □

In case there are more than two classes, we managed to check the sign of $L'_{\mathrm{lg}}(\cdot)$ everywhere except a compact set, and the proof becomes much trickier. We used the ideas of Bramson [20] here, who proved $L_{\mathrm{lg}}(\cdot)$ to be a Lyapunov function for (5.31) without impatience.

Lemma 5.2. *For $\mathbf{z} \in (0,\infty)^I$, $\|\mathbf{z}\|_1 \ge \|\mathbf{z}^*\|_1$, we have $L'_{\mathrm{lg}}(\mathbf{z}) \le 0$.*

Remark 5.1. We have run numerical tests on the fluid model of a freelance website, that is (5.3) with all μ_i's being the same. According to those tests, $L'_{\mathrm{lg}}(\mathbf{z})$ should be non-positive for $\|\mathbf{z}\|_1 \le \|\mathbf{z}^*\|_1$ as well. Moreover, as we omit the non-positive impatience

term

$$-\nu \sum_{i=1}^{I} z_i \log\left(\frac{z_i/\|\mathbf{z}\|_1}{z_i^*/\|\mathbf{z}^*\|_1}\right),$$

what is left should still be non-positive, i.e. for $\|\mathbf{z}\|_1 \le \|\mathbf{z}^*\|_1$,

$$\sum_{i=1}^{I}\left(\lambda_i + \sum_{j=1}^{I} P_{j,i}\mu_j \frac{z_j}{\|\mathbf{z}\|_1} - \mu_i \frac{z_i}{\|\mathbf{z}\|_1}\right)\log\left(\frac{z_i/\|\mathbf{z}\|_1}{z_i^*/\|\mathbf{z}^*\|_1}\right) \le 0.$$

On the other hand, the proof of Lemma 5.2 relies on the impatience term. So the two sets $\|\mathbf{z}\|_1 \le \|\mathbf{z}^*\|_1$ and $\|\mathbf{z}\|_1 \ge \|\mathbf{z}^*\|_1$ seem to need different approaches.

Proof of Lemma 5.2. Fix a $\mathbf{z} \in (0,\infty)^I$ such that $\|\mathbf{z}\|_1 \ge \|\mathbf{z}^*\|_1$ and, to shorten notation, put

$$a_i := \frac{z_i/\|\mathbf{z}\|_1}{z_i^*/\|\mathbf{z}^*\|_1}, \quad q_i := z_i^*/\|\mathbf{z}^*\|_1, \quad i = 1,\dots,I.$$

In the new notation,

$$L'_{\mathrm{lg}}(\mathbf{z}) = \Sigma - \nu \sum_{i=1}^{I} z_i \log(a_i), \tag{5.34}$$

where

$$\Sigma := \sum_{i=1}^{I}\left(\lambda_i + \sum_{j=1}^{I} P_{j,i}\mu_j q_j a_j - \mu_i q_i a_i\right)\log(a_i).$$

Also introduce

$$a_0 := 1, \quad q_0 := 1, \quad \mu_0 := \sum_{i=1}^{I}\lambda_i, \quad P_{0,0} := 0,$$
$$P_{0,i} := \lambda_i / \sum_{j=1}^{I}\lambda_j, \quad P_{i,0} := 1 - \sum_{j=1}^{I} P_{i,j}, \quad i = 1,\dots,I.$$

Note that, by the fixed point equation (5.32),

$$\sum_{j=0}^{I} P_{j,i}\mu_j q_j - \mu_i q_i = \gamma_i, \quad i = 0,\dots,I, \tag{5.35}$$

where

$$\gamma_0 := -\nu\|z^{\mathrm{e}}\|_1, \quad \gamma_i := \nu z_i^{\mathrm{e}}, \quad i = 1,\dots,I.$$

Now Σ can be rewritten as

$$\Sigma = \sum_{i=0}^{I}\left(\sum_{j=0}^{I} P_{j,i}\mu_j q_j a_j - \mu_i q_i a_i\right)\log(a_i). \tag{5.36}$$

Let $\sigma\colon \{0,\dots,I\} \to \{0,\dots,I\}$ be a permutation such that $a_{\sigma(i)}$ is non-decreasing in i. After reordering the classes according to σ, we apply to (5.36) the Abel partial summation rule, which reads as

$$\sum_{n=0}^{N}\alpha_n\beta_n = \alpha_N \sum_{m=0}^{N}\beta_m - \sum_{n=0}^{N-1}(\alpha_{n+1}-\alpha_n)\sum_{m=0}^{n}\beta_m.$$

Then we obtain

$$\Sigma = \log(a_{\sigma(I)})(B_I - C_I) - \sum_{i=0}^{I-1}(\log(a_{\sigma(i+1)}) - \log(a_{\sigma(i)}))(B_i - C_i). \tag{5.37}$$

where

$$B_i := \sum_{l=0}^{i} b_l, \quad C_i := \sum_{l=0}^{i} c_l,$$

and

$$\begin{aligned} b_l &:= \sum_{j=0}^{I} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)}, \\ c_l &:= \mu_{\sigma(l)} q_{\sigma(l)} a_{\sigma(l)}. \end{aligned}$$

Note that

$$\begin{aligned} B_I - C_I &= \sum_{i,j=0}^{I} P_{j,i} \mu_j q_j a_j - \sum_{i=1}^{I} \mu_i q_i a_i \\ &= \sum_{j=0}^{I} \mu_j q_j a_j \left(\sum_{i=1}^{I} P_{j,i} - 1 \right) = 0. \end{aligned} \tag{5.38}$$

We will prove that

$$\begin{aligned} &(\log(a_{\sigma(i+1)}) - \log(a_{\sigma(i)}))(B_i - C_i) \\ &\geq \Gamma_i (a_{\sigma(i+1)} \log(a_{\sigma(i+1)}) - a_{\sigma(i)} \log(a_{\sigma(i)})), \quad i = 0, \ldots, I-1, \end{aligned} \tag{5.39}$$

where

$$\Gamma_i := \sum_{l=0}^{i} \gamma_{\sigma(l)},$$

but first we demonstrate how this implies the lemma.

Combining (5.37)–(5.39) and the Abel partial summation rule, we get

$$\begin{aligned} \Sigma &\leq -\sum_{i=1}^{I-1} (a_{\sigma(i+1)} \log(a_{\sigma(i+1)}) - a_{\sigma(i)} \log(a_{\sigma(i)})) \Gamma_i \\ &= \sum_{i=0}^{I} a_{\sigma(i)} \log(a_{\sigma(i)}) \gamma_{\sigma(i)} - a_{\sigma(I)} \log(a_{\sigma(I)}) \underbrace{\Gamma_I}_{=0} \\ &= \sum_{i=0}^{I} \gamma_i a_i \log(a_i) = \sum_{i=1}^{I} \gamma_i a_i \log(a_i) \\ &= \nu \|\mathbf{z}^*\|_1 \sum_{i=1}^{I} \frac{z_i}{\|\mathbf{z}\|_1} \log(a_i). \end{aligned}$$

Now we plug the last bound for Σ into (5.34) and get:

$$\begin{aligned} L'(\mathbf{z}) &\leq \nu \|\mathbf{z}^*\|_1 \sum_{i=1}^{I} \frac{z_i}{\|\mathbf{z}\|_1} \log(a_i) - \nu \sum_{i=1}^{I} z_i \log(a_i) \\ &= \nu (\|\mathbf{z}^*\| - \|\mathbf{z}\|_1) \sum_{i=1}^{I} \frac{z_i}{\|\mathbf{z}\|_1} \log\left(\frac{z_i / \|\mathbf{z}\|_1}{z_i^* / \|\mathbf{z}^*\|_1} \right) \leq 0, \end{aligned}$$

where, in the second line, $\|\mathbf{z}^*\|_1 - \|\mathbf{z}\|_1 \leq 0$ by the lemma's assumption, and the summation term is non-negative as a Kullback-Leibler distance.

So it is left to show (5.39) in order to finish the proof. For $i = 0, \dots, I-1$, the following holds with $f_i = a_{\sigma(i)}$ and $f_i = a_{\sigma(i+1)}$:

$$\begin{aligned} B_i &= \sum_{l=0}^{i} \sum_{j=0}^{I} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)} \\ &= \sum_{l=0}^{i} \sum_{j=0}^{i} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)} + \sum_{l=0}^{i} \sum_{j=i+1}^{I} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)} \\ &\geq \sum_{l=0}^{i} \sum_{j=0}^{i} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)} + \underbrace{f_i \sum_{l=0}^{i} \sum_{j=i+1}^{I} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)}}_{=: \widetilde{B}_i}, \end{aligned} \tag{5.40}$$

where, by (5.35)

$$\begin{aligned} \widetilde{B}_i &= f_i \sum_{l=0}^{i} \left(\sum_{j=0}^{I} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} - \sum_{j=0}^{i} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} \right) \\ &= f_i \sum_{l=0}^{i} \left(\gamma_{\sigma(l)} + \mu_{\sigma(l)} q_{\sigma(l)} - \sum_{j=0}^{i} P_{\sigma(j),\sigma(l)} \mu_{\sigma(j)} q_{\sigma(j)} \right) \\ &= f_i \Gamma_i + f_i \sum_{j=0}^{i} \mu_{\sigma(j)} q_{\sigma(j)} - f_i \sum_{j=0}^{i} \mu_{\sigma(j)} q_{\sigma(j)} \sum_{l=0}^{i} P_{\sigma(j),\sigma(l)} \\ &= f_i \Gamma_i + f_i \sum_{j=0}^{i} \mu_{\sigma(j)} q_{\sigma(j)} \sum_{l=i+1}^{I} P_{\sigma(j),\sigma(l)} \\ &\geq f_i \Gamma_i + \sum_{j=0}^{i} \mu_{\sigma(j)} q_{\sigma(j)} a_{\sigma(j)} \sum_{l=i+1}^{I} P_{\sigma(j),\sigma(l)}, \end{aligned}$$

As we plug the last inequality back into (5.40), it follows that

$$B_i \geq C_i + f_i \Gamma_i,$$

and since $(\log(a_{\sigma(i+1)}) - \log(a_{\sigma(i)})) \geq 0$, we have

$$(\log(a_{\sigma(i+1)}) - \log(a_{\sigma(i)}))(B_i - C_i) \geq (\log(a_{\sigma(i+1)}) - \log(a_{\sigma(i)})) f_i \Gamma_i. \tag{5.41}$$

Let i_0 be the index for which $\sigma(i_0) = 0$. For $i < i_0$, we have $\log(a_{\sigma(i)}) \leq 0$ and $\Gamma_i \geq 0$. Then,

$$a_{\sigma(i+1)} \log(a_{\sigma(i)}) \leq a_{\sigma(i+1)} \log(a_{\sigma(i)})$$

and

$$\begin{aligned} &\Gamma_i (a_{\sigma(i+1)} \log(a_{\sigma(i+1)}) - a_{\sigma(i+1)} \log(a_{\sigma(i)})) \\ &\geq \Gamma_i (a_{\sigma(i+1)} \log(a_{\sigma(i+1)}) - a_{\sigma(i)} \log(a_{\sigma(i)})). \end{aligned}$$

The last equation, when compared to (5.41) with $f_i = a_{\sigma(i+1)}$, implies (5.39).

Similarly, we prove (5.39) for $i \geq i_0$. For such an i, $\log(a_{\sigma(i+1)}) \geq 0$ and $\Gamma_i \leq 0$. Hence

$$a_{\sigma(i)} \log(a_{\sigma(i+1)}) \leq a_{\sigma(i+1)} \log(a_{\sigma(i+1)})$$

and

$$\begin{aligned}&\Gamma_i(a_{\sigma(i)}\log(a_{\sigma(i+1)}) - a_{\sigma(i)}\log(a_{\sigma(i)}))\\ &\geq \Gamma_i(a_{\sigma(i+1)}\log(a_{\sigma(i+1)}) - a_{\sigma(i)}\log(a_{\sigma(i)})),\end{aligned}$$

As we compare the last inequality to (5.41) with $f_i = a_{\sigma(i)}$, (5.39) follows. □

Quadratic Lyapunov function In this paragraph, we assume that $I = 2$, and we will pick coefficients α_1, α_2 in such a way that the quadratic form

$$L_{\mathrm{qd}}(\mathbf{z}) = \frac{1}{2}[\alpha_1(z_1 - z_1^*)^2 + \alpha_2(z_2 - z_2^*)^2]$$

becomes a Lyapunov function for (5.31).

Lemma 5.3. *There exist $\alpha_1, \alpha_2 > 0$ such that the derivative of $L_{\mathrm{qd}}(\cdot)$ with respect to* (5.31) *is non-positive on $(0,\infty)^2$.*

Proof. Fix a $\mathbf{z} \in (0,\infty)^2$. Introduce the notations

$$\begin{aligned}\mathbf{y} = (y_1, y_2) &:= \mathbf{z} - \mathbf{z}^e,\\ p_i := z_i/\|\mathbf{z}\|_1, \quad q_i &:= z_i^*/\|\mathbf{z}^*\|_1, \quad i = 1,2,\end{aligned}$$

and note that

$$p_i - q_i = \frac{1}{\|\mathbf{z}\|_1}\left(y_i - q_i\sum_{j=1}^2 y_j\right). \tag{5.42}$$

Then

$$\begin{aligned}L'_{\mathrm{qd}}(\mathbf{z}) &= \sum_{i=1}^2 \alpha_i y_i\left(\lambda_i + \sum_{j=1}^2 P_{j,i}\mu_j p_j - \mu_i p_i - \nu_i z_i\right)\\ &= \sum_{i=1}^2 \alpha_i y_i\left(\sum_{j=1}^2 P_{j,i}\mu_j(p_j - q_j) - \mu_i(p_i - q_i) - \nu_i y_i\right)\\ &\overset{(5.42)}{=} -\alpha_1\nu_1 y_1^2 - \alpha_2\nu_2 y_2^2 - \frac{1}{\|\mathbf{z}\|_1}\Sigma(\mathbf{y}),\end{aligned}$$

where

$$\begin{aligned}\Sigma(\mathbf{y}) = \alpha_1 y_1&\left[(1 - P_{1,1})\mu_1\left(y_1 - q_1\sum_{j=1}^2 y_j\right) - P_{2,1}\mu_2\left(y_2 - q_2\sum_{j=1}^2 y_j\right)\right]\\ &+ \alpha_2 y_2\left[(1 - P_{2,2})\mu_2\left(y_2 - q_2\sum_{j=1}^2 y_j\right) - P_{1,2}\mu_1\left(y_1 - q_1\sum_{j=1}^2 y_j\right)\right].\end{aligned}$$

At this stage, it suffices to pick such α_i's that $\Sigma(\cdot) \geq 0$ on $\mathbb{R}^2$. The function $\Sigma(\cdot)$ is a quadratic form. Denote its coefficients in front of y_1^2, y_2^2 and $y_1 y_2$ by $a_{1,1}$, $a_{2,2}$ and $2a_{1,2}$,

respectively. Then

$$a_{1,1} = \alpha_1[(1-P_{1,1})\mu_1(1-q_1) + P_{2,1}\mu_2 q_2] = \alpha_1[(1-P_{1,1})\mu_1 + P_{2,1}\mu_2]q_2,$$
$$a_{2,2} = \alpha_2[(1-P_{2,2})\mu_2(1-q_2) + P_{1,2}\mu_1 q_1] = \alpha_2[(1-P_{2,2})\mu_2 + P_{1,2}\mu_1]q_1,$$

and

$$\begin{aligned} 2a_{1,2} &= -\alpha_1[(1-P_{1,1})\mu_1 q_1 + P_{2,1}\mu_2(1-q_2)] - \alpha_2[(1-P_{2,2})\mu_2 q_2 + P_{1,2}\mu_1(1-q_1)] \\ &= -\alpha_1[(1-P_{1,1})\mu_1 + P_{2,1}\mu_2]q_1 - \alpha_2[(1-P_{2,2})\mu_2 + P_{1,2}\mu_1]q_2. \end{aligned}$$

Now take

$$\alpha_1 = \frac{1}{[(1-P_{1,1})\mu_1 + P_{2,1}\mu_2]q_1}, \quad \alpha_2 = \frac{1}{[(1-P_{2,2})\mu_2 + P_{1,2}\mu_1]q_2}.$$

With this choice of the α_i's, we have

$$a_{1,1} = \frac{q_2}{q_1}, \quad a_{2,2} = \frac{q_1}{q_2}, \quad 2a_{1,2} = -2,$$

and

$$\Sigma(y) = \left(\sqrt{\frac{q_2}{q_1}}y_1 - \sqrt{\frac{q_1}{q_2}}y_2\right)^2 \geq 0,$$

which completes the proof. □

Remark 5.2. Following the lines of Theorem 2.2, one can check that the quadratic function

$$\widetilde{L}_{\mathrm{qd}}(\mathbf{z}) = \sum_{i=1}^{I} \frac{(z_i - z_i^*)^2}{\mu_i z_i^* / \|\mathbf{z}^*\|_1}$$

is a Lyapunov function for the system (5.31) in case there is no routing (the number I of classes can be arbitrary). We have done numerical tests which indicate that this function should also work as a Lyapunov function for the freelance fluid model (that is (5.3) with all μ_i's the same). Unlike for the entropy function, the impatience term

$$-\sum_{i=1}^{I} \frac{\nu_i}{\mu_i z_i^* / \|\mathbf{z}^*\|_1}(z_i - z_i^*)^2$$

is crucial in this case: without it the derivative can take positive values.

Bibliography

[1] N. Abramson. The ALOHA system — another alternative for computer communications. In *AFIPS Conference Proceedings*, volume 36, pages 295–298, 1970.

[2] I.J.B.F. Adan and J.A.C. Resing. Queueing theory, 2002. Lecture notes, `http://www.win.tue.nl/~iadan/queueing.pdf`.

[3] E. Altman and H. Kushner. Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM Journal on Control and Optimization*, 41:217–252, 2002.

[4] D. Andrews. Laws of Large Numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4(3):458–467, 1988.

[5] S. Asmussen. *Applied Probability and Queues*. Springer, 2nd edition, 2003.

[6] U. Ayesta and M. Mandjes. Bandwidth-sharing networks under a diffusion scaling. *Annals of Operations Research*, 170:41–58, 2009.

[7] F. Baccelli and P. Brémaud. *Elements of Queueing Theory*. Springer-Verlag, 2nd edition, 2003.

[8] E.J. Balder. On subdifferential calculus, 2010. Lecture notes, `http://www.staff.science.uu.nl/~balde101/cao10/cursus10_1.pdf`.

[9] P. Billingsley. *Convergence of Probability Measures*. Series in Probability and Statistics. Wiley, 2nd edition, 1999.

[10] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. In *Proceedings of ACM Sigmetrics & Performance Conference*, pages 82–91, Boston MA, USA, 2001.

[11] 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.

[12] Ch. Bordenave, S. Foss, and S. Shneer. A random multiple access protocol with spatial interactions. *Journal of Applied Probability*, 46(3):844–865, 2009.

[13] Ch. Bordenave, D. McDonald, and A. Proutière. Asymptotic stability region of slotted-Aloha. *IEEE Transactions on Information Theory*, 58(9):5841–5855, 2012.

[14] A.A. Borovkov. *Stochastic Processes in Queueing Theory*. Springer, 2nd edition, 1976.

[15] S. Borst, R. Egorova, and B. Zwart. Fluid limits for bandwidth-sharing networks in overload. To appear in *Mathematics of Operations Research*.

[16] S.C. Borst. *Polling Systems*. CWI, Amsterdam, 1996.

[17] S.C. Borst, O.J. Boxma, J.A. Morrison, and R. Nunez Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31(4):254–262, 2003.

[18] O.J. Boxma. Analysis and optimization of polling systems. *Queueing, Performance and Control in ATM*, pages 173–183, 1991. Eds. J.W. Cohen and C.D. Pack, North-Holland, Amsterdam.

[19] M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4(2):414–431, 1994.

[20] M. Bramson. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Systems*, 23(1–4):1–26, 1996.

[21] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998.

[22] M. Bramson. Stability of networks for max-min fair routing, 2005. Presentation at the *13th INFORMS Applied Probability Conference*, Ottawa ON, Canada.

[23] M. Bramson. *Stability of Queueing Networks*. Springer, 2008.

[24] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[25] H. Chen and A. Mandelbaum. Discrete flow networks: Bottleneck analysis and fluid approximations. *Mathematics of Operations Research*, 16(2):408–446, 1991.

[26] H. Chen and D.D. Yao. *Fundamentals of Queuing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag, 2001.

[27] M. Chiang, D. Shah, and A. Tang. Stochastic stability of network utility maximization: General file size distribution. In *Proceedings of Annual Allerton Conference*, pages 82–91, Monticello IL, USA, 2006.

[28] E.G. Coffman, A.A. Puhalskii, and M.I. Reiman. Polling systems with zero switchover times: A heavy-traffic principle. *Annals of Applied Probability*, 5:681–719, 1995.

[29] E.G. Coffman, A.A. Puhalskii, and M.I. Reiman. Polling systems in heavy-traffic: A Bessel process limit. *Mathematics of Operations Research*, 23:257–304, 1998.

[30] J.W. Cohen. *The Single Server Queue*. North Holland. North-Holland Series in Applied Mathematics and Mechanics, 2nd edition, 1992.

[31] J.G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid models. *Annals of Applied Probability*, 5(1):49–77, 1995.

[32] J.G. Dai. A fluid limit model criterion for instability of multiclass queueing networks. *Annals of Applied Probability*, 6:751–757, 1996.

[33] R.W.R. Darling and J.R. Norris. Differential equation approximations for Markov chains. *Probability Surveys*, 5:37–79, 2008.

[34] G. de Veciana, T.-L. Lee, and T. Konstantopoulos. Stability and performance analysis of network supporting services with rate control — could the Internet be unstable? In *Proceedings of IEEE Infocom*, pages 802–810, New York NY, USA, 1999.

[35] 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.

[36] S. Ding, M. Frolkova, R.D. van der Mei, and B. Zwart. Fluid approximation of a call center model with redials and reconnects. In progress.

[37] V. Dumas, F. Guillemin, and Ph. Robert. A Markovian analysis of additive-increase multiplicative-decrease algorithms. *Advances in Applied Probability*, 34 (1):85–111, 2002.

[38] R. Egorova, S.C. Borst, and B. Zwart. Bandwidth-sharing networks in overload. *Performance Evaluation*, 64:978–993, 2007.

[39] A. Ephremides and B. Hajek. Information theory and communication networks: An unconsummated union. *Transactions on Information Theory*, 44:2416–2434, 1998.

[40] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

[41] S. Foss. Queues with customers of several types. *Limit Theorems and Related Problems*, pages 348–377, 1984. Ed. A.A. Borovkov, *Optimization Software*.

[42] S. Foss and A. Kovalevskii. A stability criterion via fluid limits and its application to a polling model. *Queueing Systems*, 32:131–168, 1999.

[43] M. Frolkova, S. Foss, and B. Zwart. Fluid limits for an ALOHA-type model with impatient customers. *Queueing Systems*, 72:69–101, 2012.

[44] S.W. Fuhrmann and R.B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[45] H.C. Gromoll and R.J. Williams. Fluid model for a data network with alpha-fair bandwidth sharing and general document size distributions: Two examples of stability. *IMS Collections — Markov Processes and Related Topics*, 4:253–265, 2008.

[46] H.C. Gromoll and R.J. Williams. Fluid limits for networks with bandwidth sharing and general document size distributions. *Annals of Applied Probability*, 19: 243–280, 2009.

[47] H.C. Gromoll, Ph. Robert, and B. Zwart. Fluid limits for processor sharing queues with impatience. *Mathematics of Operations Research*, 33:375–402, 2008.

[48] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14:502–525, 1982.

[49] Th.E. Harris. *The Theory of Branching Processes*. Springer-Verlag, 1963.

[50] P. Hartman. *Ordinary Differential Equations*. Birkhäuser, 1982.

[51] A. Jakubowski. Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probability and Mathematical Statistics*, 9.1:95–114, 1988.

[52] O. Kallenberg. *Random Measures*. Akademie-Verlag, Berlin, 1983.

[53] W.N. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Annals of Applied Probability*, 22 (2):477–521, 2012.

[54] 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*, 19:1719–1780, 2009.

[55] F.P. Kelly. Networks of queues. *Advances in Applied Probability*, 8(2):416–432, 1976.

[56] F.P. Kelly. Loss networks. *Annals of Applied Probability*, 1(3):319–378, 1991.

[57] 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.

[58] F.P. Kelly and R.J. Williams. Heavy traffic on a controlled motorway. *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, pages 416–445, 2010.

[59] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, 1953.

[60] H. Kesten and B.P. Stigum. A limit theorem for multidimensional Galton-Watson processes. *Annals of Mathematical Statistics*, 37:1211–1223, 1966.

[61] A.Y. Khinchin. *Mathematical Methods in the Theory of Queueing*. Griffin, 1960.

[62] A.P. Kovalevski, V.A. Topchii, and S.G. Foss. On the stability of a queueing system with uncountable branching fluid limits. *Problems of Information Transmission*, 41: 254–279, 2005.

[63] D.P. Kroese. Heavy-traffic analysis for continuous polling models. *Journal of Applied Probability*, 34:720–732, 1997.

[64] P.R. Kumar and S.P. Meyn. Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 41:4–17, 1996.

[65] P.R. Kumar and T.I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, 35(3):289–298, 1990.

[66] Th.G. Kurtz, R. Lyons, R. Pemantle, and Yu. Peres. A conceptual proof of the Kesten-Stigum theorem for multi-type branching processes. *Classical and Modern branching processes*, 84:181–185, 1997.

[67] Y. Liu and W. Whitt. Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, 2013. Published online.

[68] C. Mack, T. Murphy, and N.L. Webb. The efficiency of n machines unidirectionally patrolled by one operative when walking times and repair times are constants. *Journal of the Royal Statistical Society*, 19:166–172, 1957.

[69] I. MacPhee, M. Menshikov, D. Petritis, and S. Popov. A Markov chain model of a polling system with parameter regeneration. *Annals of Applied Probability*, 17: 1447–1473, 2007.

[70] I. MacPhee, M. Menshikov, D. Petritis, and S. Popov. Polling systems with parameter regeneration, the general case. *Annals of Applied Probability*, 18:2131–2155, 2008.

[71] V.A. Malyshev. Networks and dynamical systems. *Advances in Applied Probability*, 25(1):140–175, 1993.

[72] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[73] B. Massey. Open networks of queues. *Advances in Applied Probability*, 16(1):176–201, 1984.

[74] L. Massoulié. Structural properties of proportional fairness: Stability and insensitivity. *Annals of Applied Probability*, 17:809–839, 2007.

[75] L. Massoulié and J.W. Roberts. Bandwidth sharing: Objectives & algorithms. In *Proceedings of IEEE Infocom*, Books in Statistics, pages 1395–1403, New York NY, USA, 1999.

[76] K. Maulik and B. Zwart. An extension of the square root law of TCP. *Annals of Operations Research*, 170(1):217–232, 2009.

[77] S.P. Meyn. Transience of multiclass queueing networks and their fluid models. *Annals of Applied Probability*, 5:946–957, 1995.

[78] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[79] V. Mikhailov. Geometric analysis of stability of Markov chains in $\mathbb{R}^n$ and its applications to throughput evaluation of the adaptive random multiple access algorithm. *Problems of Information Transmission*, 24:47–56, 1988.

[80] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8:556–567, 2000.

[81] J.R. Norris. *Markov Chains*. Cambridge University Press, 20th edition, 2008.

[82] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*, volume 83 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2004.

[83] G. Pang and W. Whitt. Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems*, 61:167–202, 2009.

[84] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*, 73(2):119–146, 2013.

[85] S.S. Panwar, D. Towsley, and J.K. Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *Journal of the Association for Computing Machinery*, 35:832–844, 1988.

[86] A.A. Puhalskii and A.N. Rybko. Nonergodicity of a queueing network under nonstability of its fluid model. *Problems of Information Transmission*, 36(1):23–41, 2000.

[87] J. Reed and A.R. Ward. Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, 2008.

[88] J. Reed and B. Zwart. Limit theorems for bandwidth sharing networks with rate constraints. Revised, preprint available at `http://people.stern.nyu.edu/jreed/Papers/BARevised.pdf`, 2010.

[89] J.E. Reed and R. Talreja. Distribution-valued heavy-traffic limits for the $G/GI/\infty$ queue. Submitted for publication, preprint available at `http://people.stern.nyu.edu/jreed/Papers/SubmittedVersionDistribution.pdf`, 2009.

[90] J.E. Reed and A.R. Ward. A diffusion approximation for a generalized Jackson network with reneging. In *Proceedings of the 42nd Annual Conference on Communication, Control, and Computing*, 2004.

[91] M. Remerova and B. Zwart. Fluid limits of a PS-queue with multistage service. In preparation, 2013.

[92] M. Remerova, J. Reed, and B. Zwart. Fluid limits for bandwidth-sharing networks with rate constraints. Accepted for publication in *Mathematics of Operations Research*, 2013.

[93] M. Remerova, S. Foss, and B. Zwart. Random fluid limit of an overloaded polling model. *Advances in Applied Probability*, 46:76–101, 2014.

[94] J.A.C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13:409–426, 1993.

[95] Ph. Robert. *Stochastic Networks and Queues*. Springer-Verlag, 2003.

[96] J.W. Roberts. A survey on statistical bandwidth sharing. *Computer Networks*, 45: 319–332, 2004.

[97] J.W. Roberts and L. Massoulié. Bandwidth sharing and admission control for elastic traffic. In *Proceedings of the ITC Specialist Seminar*, Yokohama, Japan, 1998.

[98] A.N. Rybko. Stationary distributions of time-homogeneous Markov processes describing the operation of networks transmitting messages. *Problems of Information Transmission*, 17(1):71–89, 1981. In Russian.

[99] A.N. Rybko and A.L. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problems of Information Transmission*, 28(3): 199–220, 1992.

[100] A.L. Stolyar. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes and Related Fields*, 1(4):491–512, 1995.

[101] X. Su and S.A. Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing and Service Operations Management*, 6(4):280–301, 2004.

[102] L. Takács. *Introduction to the Theory of Queues*. Oxford University Press, 1962.

[103] H. Takagi. *Analysis of Polling Systems*. MIT Press, Cambridge, MA, 1986.

[104] H. Takagi. Queueing analysis of polling systems: An update. *Stochastic Analysis of Computer and Communication Systems*, pages 267–318, 1990. Ed. H. Takagi, North-Holland, Amsterdam.

[105] H. Takagi. Queueing analysis of polling models: Progress in 1990–1994. *Frontiers in Queueing: Models and Applications in Science and Engineering*, pages 119–146, 1997. Ed. J.H. Dshalalow, CRC Press, Boca Raton, FL.

[106] R.D. van der Mei. Towards a unifying theory on branching-type polling models in heavy traffic. *Queueing Systems: Theory and Applicationss*, 57:29–46, 2007.

[107] R.D. van der Mei and J.A.C. Resing. Polling models with two-stage gated service: Fairness versus efficiency. In *Proceedings of the 20th International Teletraffic Congress*, Ottawa, Canada, 2007.

[108] R.D. van der Mei and A. Roubos. Polling systems with multi-phase gated service. *Annals of Operations Research*, 198:25–56, 2011.

[109] A.C.C. van Wijk, I.J.B.F. Adan, O.J. Boxma, and A. Wierman. Fairness and efficiency for polling models with the κ-gated service discipline. *Performance Evaluation*, 69:274–288, 2012.

[110] V.A. Vatutin and E.E. Dyakonova. Multitype branching processes and some queueing systems. *Journal of Mathematical Sciences*, 111:3901–3909, 2002.

[111] A.R. Ward. Asymptotic analysis of queueing systems with reneging: a survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science*, 16(1):1–14, 2011.

[112] A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

[113] A.R. Ward and P.W. Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking of reneging. *Queueing Systems*, 50(4):371–400, 2005.

[114] W. Whitt. An overview of Brownian and non-Brownian FCLTs for single-server queues. *Queueing Systems*, 36:39–70, 2000.

[115] W. Whitt. *Stochastic-Process Limits*. Springer, 2002.

[116] W. Whitt. A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research*, 52(6):922–941, 2004.

[117] R. Williams. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems*, 30:27–88, 1998.

[118] H. Ye and D.D. Yao. Heavy-traffic optimality of a stochastic network under utility-maximizing resource control. *Operations Research*, 56:453–470, 2008.

[119] H. Ye and D.D. Yao. Utility maximizing resource control: Diffusion limit and asymptotic optimality for a two-bottleneck model. *Operations Research*, 58:613–623, 2010.

[120] U. Yechiali. Analysis and control of polling systems. *Performance Evaluation of Computer and Communication Systems*, pages 630–650, 1993. Eds. L. Donatiello and R. Nelson, Springer, Berlin.

[121] J. Zhang, J.G. Dai, and B. Zwart. Law of large number limits of limited processor sharing queues. *Mathematics of Operations Research*, 34:937–970, 2009.

Nederlandse samenvatting

Vloeistoflimieten van stochastische netwerken

In wachtrijtheorie, is een exacte analyse alleen mogelijk voor eenvoudige modellen of onder beperkende aannames. In zulke gevallen, om ingewikkelder modellen te kunnen analyseren, moet men benaderingen gebruiken. Een optie is vloeistof-, of wet-van-de-grote- aantallen limieten die een macroscopische beschrijving geven van stochastische processen. Meestal (maar niet altijd) zijn vloeistoflimieten deterministisch en lossen ze differentiaal-/integralvergelijkingen op die de oorspronkelijke stochastische dynamiek nabootsen. Vloeistoflimieten zijn een krachtige en universele methode die eerst populariteit had gekregen als hulpmiddel in het onderzoeken van de stabiliteit van stochastische netwerken. In dit proefschrift, worden vloeistoflimieten gebruikt om overbelaste of op een andere manier dichtbevolkte systemen te benaderen.

In Hoofdstuk 2, bestuderen we een ALOHA-model waarin meerdere gebruikers hun datapakketten naar de hub doorzenden op dezelfde frequentie. Als twee of meer verzendingen botsen, moet elk van de gebruikers een periode wachten om het pakket opnieuw proberen te verzenden. Zulke protocolen komen bijvoorbeeld vaak in satellietcommunicatie voor. We breiden het conventionele (time-slotted en gecentraliseerde) ALOHA-model uit door meerdere soorten gebruikers te beschouwen die bovendien ongeduldig zijn. We ontwikkelen vloeistofbenaderingen die een systeem van deterministische differentiaalvergelijkingen oplossen. We laten ook zien dat de differentiaalvergelijkingen voor de vloeistoflimieten een uniek vast punt hebben en dat alle vloeistoflimieten dit vaste punt benaderen na verloop van tijd.

Hoofdstuk 3 behandelt bandwidth-sharing netwerken die de dynamische interactie modelleren van verkeer op bijvoorbeeld het Internet. Capaciteiten van de links worden tussen de verzendingen gedeeld op basis van een optimalisatie procedure. Dit hoofdstuk richt zich op een uitbreiding van het kleine aantal resultaten die met rate beperkingen voor individuele verzendingen rekening houden. Ook beschouwt dit hoofdstuk een onconventionele vloeistof scaling — de grote-capaciteit scaling. Zonder beperkende aannames op de verdeling van filegroottes karakteriseren we de vloeistoflimiet. Bovendien laten we zien dat het vaste punt van de limiet een oplossing is van een strict concaaf optimaliseringsprobleem, en dus in polynomiale tijd berekend kan worden. Verder bewijzen we onder een extra aanname dat het vaste punt van de vloeistof limiet dichtbij de stationaire verdeling van het netwerk zit. Dat is een nieuwe soort resultaat voor bandwidth-sharing netwerken.

In Hoofdstuk 4 komen we een verrassend fenomeen tegen — een stochastische vloeistoflimiet. Het model van dit hoofdstuk is een klassiek cyclisch polling systeem. De veronderstellingen die tot de stochastische vloeistoflimiet leiden zijn overbelasting en een lege begintoestand. We beschouwen een grote klasse van multigated bedieningsdisciplines die een connectie mogelijk maken met vertakkingsprocessen. Het onderliggende vertakkingsproces is superkritisch vanwege de overbelasting, en daardoor werkt de onzekerheid door in de vloeistoflimiet. Daarnaast hebben de paden van de vloeistoflimiet een interessante structuur: ze oscilleren oneindig vaak in de buurt van nul, en ze kunnen allemaal getransformeerd worden tot dezelfde functie door een lineare tijdruimte scaling. Een extra bijdrage van dit hoofdstuk zijn bovengrenzen op de f-momenten van de bezet periode in een $M/G/1$ wachtrij voor een grote klasse functies f, o.a. macht- en logarithmische functies.

Tenslotte, in Hoofdstuk 5, beginnen we met een studie van freelance websites. We stellen een basismodel van zo'n website voor en ontwikkelen vloeistoflimieten onder overbelasting. We laten ook zien dat alle vloeistoflimieten het unieke vaste punt benaderen over een lang tijdsinterval. De analyse in dit hoofdstuk is gebaseerd op een verband tussen het freelance model en een multistage processor sharing systeem.