HAL
open science

Mobility can drastically improve the heavy traffic performance from 1/(1-rho) to log(1/(1-rho))

Florian Simatos, Alain Simonian

To cite this version:

Florian Simatos, Alain Simonian. Mobility can drastically improve the heavy traffic performance from 1/(1-rho) to log(1/(1-rho)). Queueing Systems, 2020, 95, pp.1-28. 10.1007/s11134-020-09652-0 . hal-03102446

HAL Id: hal-03102446

https://hal.science/hal-03102446v1

Submitted on 7 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive Toulouse Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of some Toulouse researchers and makes it freely available over the web where possible.

This is an author's version published in: https://oatao.univ-toulouse.fr/27123

Official URL: http://doi.org/10.1007/s11134-020-09652-0

To cite this version :

Simatos, Florian and Simonian, Alain Mobility can drastically improve the heavy traffic performance from 1/(1-rho) to log(1/(1-rho)). (2020) Queueing Systems, 95. 1-28. ISSN 0257-0130

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr

Mobility can drastically improve the heavy traffic performance from $\frac{1}{1-\varrho}$ to $\log(1/(1-\varrho))$

Florian Simatos[1] · **Alain Simonian**[2]

Abstract
We study a model of wireless networks where users move at speed $\theta \geq 0$, which has the original feature of being defined through a fixed-point equation. Namely, we start from a two-class processor-sharing queue to model one representative cell of this network: class 1 users are patient (non-moving) and class 2 users are impatient (moving). This model has five parameters, and we study the case where one of these parameters is set as a function of the other four through a fixed-point equation. This fixed-point equation captures the fact that the considered cell is in balance with the rest of the network. This modeling approach allows us to alleviate some drawbacks of earlier models of mobile networks. Our main and surprising finding is that for this model, mobility drastically improves the heavy traffic behavior, going from the usual $\frac{1}{1-\varrho}$ scaling without mobility (i.e., when $\theta = 0$) to a logarithmic scaling $\log(1/(1-\varrho))$ as soon as $\theta > 0$. In the high load regime, this confirms that the performance of mobile systems benefits from the spatial mobility of users. Finally, other model extensions and complementary methodological approaches to this heavy traffic analysis are discussed.

Keywords Heavy traffic · Mobile network · Large deviation

Mathematics Subject Classification 60F05 · 60K25

✉ Florian Simatos
florian.simatos@isae.fr

Alain Simonian
alain.simonian@orange.com

1 ISAE SUPAERO and Université de Toulouse, Toulouse, France

2 ORANGE LABS, Châtillon, France

1 Introduction

1.1 Background and undesirable ergodicity assumption

Since the emergence of wireless networks and following their continual development, the impact of user mobility on network performance has attracted significant attention. In [10], the authors showed that mobility creates a multi-user diversity leading to a significant improvement in per-user throughput. Since this seminal work, the observation that mobility increases throughput has been confirmed in a wide variety of situations captured by various stochastic models; see [2,3,5–8,17,23]. These different works also show that various reasons can lead to an increase in performance, for instance opportunistic channel-aware scheduling or the mobility itself which acts as a distributed load balancing algorithm. Interestingly, to the best of our knowledge, the first paper to show that mobility could under certain circumstances actually degrade delay only appeared recently [1].

In all these models, user mobility is represented by an ergodic process on a finite region of the plane. For instance, in [10] users follow a stationary and ergodic trajectory on the unit disk; in [3,5–7], users follow an irreducible Markovian trajectory in a network consisting of a finite number of cells. In our view, one of the limitations of such a modeling assumption is the highly unrealistic behavior it displays under congestion. Indeed, in the congestion regime, users stay in the network for a long time, so that if their trajectory is ergodic, they necessarily visit the same place a large number of times, as if they were walking circularly.

1.2 High-level model description and motivation

In the present paper, we pursue the modeling approach started in [18,24]. The main idea to alleviate the aforementioned drawback resulting from the ergodic trajectory assumption is to focus on a single cell and abstract the rest of the network as a single state. By doing so, we only keep track of the precise location of users when they are located in the considered cell: when located elsewhere (either outside the network or in the rest of the network) we do not track them precisely. This simple model could be generalized by focusing on several cells rather than a single one (see the discussion in Sect. 5). Users can thus be in one of three "places", as pictured in Fig. 1:

(1) outside the network, meaning that they do not require service (the left fluffy shape);
(2) in the considered cell (the middle hexagon);
(3) in the network but not in the considered cell, i.e., in the rest of the network (the right fluffy shape).

Moreover, our work is motivated by presently rolled-out LTE networks where cells can be small in range (pico, femto cells). In this context, users experience similar radio conditions and we will therefore assume below that they receive the same transmission capacity, independently of their location within the cell. While focusing on the spatial mobility aspect of users, the present study consequently ignores the possible spatial variations of transmission capacity inevitably presented by larger cells. In the following, this equal capacity is denoted by $1/\mu$.

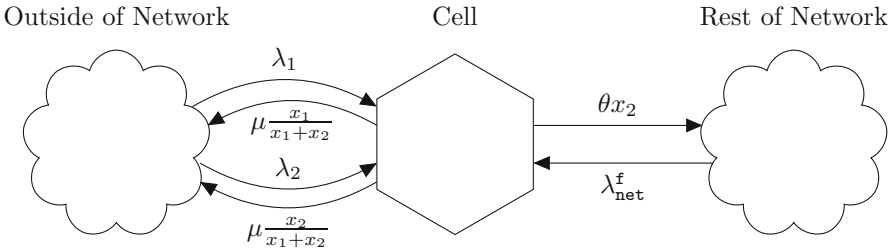

Fig. 1 Description of the model considered in the paper. Without imposing the balance condition corresponding to the fixed-point equation (FP), this is a two-class processor-sharing queue with one impatient class, namely the class-2 of mobile users, with arrival rate $\lambda_2 + \lambda^{\mathtt{f}}_{\mathtt{net}} = \lambda^{\mathtt{f}}_{\mathtt{tot}}$. The balance equation (FP) accounts for the fact that a typical cell at equilibrium is considered, with equal flows from and to the rest of the network

1.3 Mathematical model and results

Our mathematical model is introduced in two steps. At this stage, we only give a high-level description of our model in order to give the big picture: details are provided in Sect. 2.

We first introduce a "free" model $\mathbf{X}^{\mathtt{f}}$, which is simply a two-class processor-sharing queue with one impatient class: from the mobile network perspective, patient users correspond to static users who do not move, and impatient users to mobile users who move and thus potentially leave the cell for the rest of the network. The non-zero transition rates of the Markov process $\mathbf{X}^{\mathtt{f}}$ are given by

$$\mathbf{x} \in \mathbb{N}^2 \longrightarrow \begin{cases} \mathbf{x} + \mathbf{e}_1 & \text{at rate } \lambda_1, \\ \mathbf{x} + \mathbf{e}_2 & \text{at rate } \lambda_2 + \lambda^{\mathtt{f}}_{\mathtt{net}}, \\ \mathbf{x} - \mathbf{e}_1 & \text{at rate } \mu \dfrac{x_1}{x_1 + x_2}, \\ \mathbf{x} - \mathbf{e}_2 & \text{at rate } \mu \dfrac{x_2}{x_1 + x_2} + \theta x_2, \end{cases} \tag{1.1}$$

with $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ (see Sect. 2.1 for a detailed interpretation of these parameters); as specified below, θ represents the impatience/mobility rate.

In a second step, we introduce our full model which is obtained from the free model (1.1) by enforcing a balance condition in the form of the fixed-point equation (FP). This fixed-point equation means that the flows of mobile users to and from the rest of the network must balance out. This condition consequently means that the considered cell is "typical", in that the cell imposes a load on the rest of the network equal to the reciprocal load from the rest of the network to the considered cell.

If $\varrho_1 = \lambda_1/\mu$ denotes the load of static (i.e., patient) users and $\varrho_2 = \lambda_2/\mu$ the load of mobile (i.e., impatient) users, the stability condition without enforcing this balance equation is $\varrho_1 < 1$ since class-2 users are impatient and thus cannot accumulate (see

Lemma 2.1). From the mobile network perspective, the interpretation is that mobile users can always escape to the rest of the network where they are not tracked. The stability condition $\varrho_1 < 1$ is therefore clearly fictitious, because even if we do not keep track of the precise location of mobile users in the rest of the network, they still impose a load on the network which should be accounted for. When enforcing the balance equation (FP), the stability condition then becomes $\varrho_1 + \varrho_2 < 1$ which is the natural expected stability condition since, considering the cell as a representative cell of a larger network, $\varrho_1 + \varrho_2$ is the normalized load per cell (see Lemma 2.4).

The study of this model is driven by the desire to understand the impact of mobility on performance. We wish, in particular, to address questions such as: given the total load $\varrho = \varrho_1 + \varrho_2 < 1$, does the network perform better if the proportion ϱ_2/ϱ of mobile users increases? Answering such a question being generally difficult, we here resort to the approximation obtained in the heavy traffic regime where $\varrho \uparrow 1$. In addition to providing useful insight into the impact of mobility on performance, this model turns out to exhibit a highly original heavy traffic behavior, whereby the number of users in system scales like $\log(1/(1-\varrho))$ as $\varrho \uparrow 1$. If all users were static, we would have the usual $(1-\varrho)^{-1}$ scaling; our model therefore suggests that not only throughput but also delay is improved with mobility.

To the best of our knowledge, this unusual heavy traffic scaling only appeared earlier [15] in the case of the shortest-remaining-processing-time service discipline with heavy tailed service distribution. In this case, such an improvement is conceivable: indeed, since the service distribution is heavy tailed, very long jobs are not so rare. If the service discipline is FIFO, then these jobs impose a very large delay on the numerous smaller jobs that arrive after them. With SRPT, in contrast, only the large jobs spend a long time in the network, essentially due to their large service requirement. As regards the impact of mobility in wireless networks, it has been already observed [24], through an approximate analysis and extensive simulation, that the performance gain due to mobility can be related to an "opportunistic" displacement of mobile users within the network; in fact, any local increase in traffic in one given cell induces the displacement of the moving users to a neighboring cell in order to complete their transmission, hence alleviating the traffic for remaining (static or moving) users in the original cell. Our contribution in this paper is to theoretically justify this statistical behavior in the heavy traffic regime.

1.4 Organization of the paper

We start by introducing our model and Theorem 2.5, the main result of the paper, in Sect. 2. In this section, we will also present a conjecture refining our main result, which is discussed in Sect. 5. Sections 3 and 4 are devoted to the proof of Theorem 2.5.

2 Model description and main result

We now introduce our model in details: as above, we first address a "free" model simply represented by a two-class processor-sharing queue with one impatient class;

further, we introduce the full model which derives from the free model by enforcing a balance condition in the form of a fixed-point equation (FP). We then state our main result and explain the main steps of the proof.

2.1 Free model

In the free model represented by the Markov process $\mathbf{X}^{\mathtt{f}}$, with non-zero transition rates (1.1), we consider two classes of users:

(1) class-1 users are static: they arrive to the cell from the outside at rate λ_1, require a service which is exponentially distributed with parameter μ and are served according to the processor-sharing service discipline. They consequently leave the network (to the outside) at an aggregate rate $\mu x_1/(x_1+x_2)$, with x_i the number of class-i users;

(2) class-2 users are mobile: they arrive to the cell from the outside at rate λ_2, require a service which is exponentially distributed with parameter μ and are served according to the processor-sharing service discipline. As for class-1 users, they leave the network to the outside upon completing service at an aggregate rate $\mu x_2/(x_1+x_2)$; the difference with class-1 users is that they are mobile and can thus leave the cell (now, to the rest of the network and not the outside) before completing service. We assume that each mobile user leaves the cell at rate θ, and so class-2 users leave the cell for the rest of the network at an aggregate rate θx_2. Finally, mobility can also make users enter the cell from outside the network and we assume that this happens at rate $\lambda^{\mathtt{f}}_{\mathtt{net}}$.

As to the processor-sharing discipline considered for both user classes, we recall that it accounts at flow level for the fair sharing of the total capacity of the base station [4].

At this stage, it is apparent from rates (1.1) that differentiating the outside and the rest of the network is artificial and bears no consequence on the distribution of this Markov process. All that matters is the total arrival rate $\lambda^{\mathtt{f}}_{\mathtt{tot}} := \lambda_2 + \lambda^{\mathtt{f}}_{\mathtt{net}}$ and the total service rate $\mu x_2/(x_1+x_2)+\theta x_2$ of class-2 users. This distinction, however, will become crucial later.

The distribution of the Markov process $\mathbf{X}^{\mathtt{f}}$ with non-zero transition rates (1.1) thus depends on the five parameters λ_1, λ_2, $\lambda^{\mathtt{f}}_{\mathtt{net}}$, θ and μ (and more precisely, on λ_2 and $\lambda^{\mathtt{f}}_{\mathtt{net}}$ only through their sum $\lambda^{\mathtt{f}}_{\mathtt{tot}} = \lambda_2 + \lambda^{\mathtt{f}}_{\mathtt{net}}$). The superscript $\mathtt{f}$ refers to "free", as the "full" process in that we will be mainly interested in belongs to this class, but with $\lambda^{\mathtt{f}}_{\mathtt{net}}$ chosen as a function of the other four parameters λ_1, λ_2, θ and μ.

In the rest of the paper, we write $\varrho_i = \lambda_i/\mu$ and $\varrho = \varrho_1 + \varrho_2$. The following result describes the stability region of $\mathbf{X}^{\mathtt{f}}$, which depends on whether $\theta = 0$ or $\theta > 0$. Whenever $\mathbf{X}^{\mathtt{f}}$ is positive recurrent, we denote by $\mathbf{X}^{\mathtt{f}}(\infty)$ its stationary distribution. Here and throughout the paper, vector inequalities are understood component-wise, so for instance $\mathbb{E}(\mathbf{X}^{\mathtt{f}}(\infty)) < \infty$ means that $\mathbb{E}(X^{\mathtt{f}}_i(\infty)) < \infty$ for $i \in \{1,2\}$. Note finally that $X^{\mathtt{f}}$ is not reversible.

Lemma 2.1 *Stability of $\mathbf{X}^{\mathtt{f}}$ depends on whether $\theta = 0$ or $\theta > 0$ in the following way:*

- *if $\theta = 0$, then $\mathbf{X}^{\mathtt{f}}$ is positive recurrent if $\varrho + \lambda^{\mathtt{f}}_{net}/\mu < 1$, null recurrent if $\varrho + \lambda^{\mathtt{f}}_{net}/\mu = 1$ and transient if $\varrho + \lambda^{\mathtt{f}}_{net}/\mu > 1$;*

- *if $\theta > 0$, then $\mathbf{X}^{\mathtt{f}}$ is positive recurrent if $\varrho_1 < 1$, null recurrent if $\varrho_1 = 1$ and transient if $\varrho_1 > 1$.*

In either case, when the process is positive recurrent, then we have $\mathbb{E}(\mathbf{X}^{\mathtt{f}}(\infty)) < \infty$.

These results can be proved with Lyapunov-type arguments and the comparison with suitable $M/M/1$ queues. Such arguments are standard and the proof is therefore omitted.

2.2 Constrained model

The previous result formalizes the behavior pointed out in the introduction, namely that in the presence of mobile users (i.e., when $\theta > 0$), mobile users do not matter as regards to stability. In fact, if they accumulate, they can then escape to the rest of the network where they are not tracked. However, this is only an artifact of our modeling approach since mobile users that escape to the rest of the network should somehow be accounted for. The constrained model $\mathbf{X}$ that we now introduce aims at doing this; it is obtained by taking $\lambda^{\mathtt{f}}_{\mathtt{net}}$ as a function of the other four parameters through a fixed-point equation.

2.2.1 The fixed-point equation

In the free model, the three parameters λ_1, λ_2 and μ govern the transitions involving the outside, while the two parameters θ and $\lambda^{\mathtt{f}}_{\mathtt{net}}$ govern transitions within the network. Out of these five parameters, all but $\lambda^{\mathtt{f}}_{\mathtt{net}}$ can be considered as exogenous and dictated by the users' behavior: how often they arrive, how fast they move, etc. In contrast, $\lambda^{\mathtt{f}}_{\mathtt{net}}$ is hard to directly tie down with users' behavior and is more an artifact of our modeling approach.

In order to fix the value of $\lambda^{\mathtt{f}}_{\mathtt{net}}$ in an exogenous way, the idea is to impose a balance condition. Roughly speaking, we assume that the cell is in equilibrium (see Sect. 5 for a discussion on this assumption) and that the flows of mobile users to and from the rest of the network balance each other. Provided that $\mathbf{X}^{\mathtt{f}}$ is positive recurrent, we thus want to impose the balance equation

$$\lambda^{\mathtt{f}}_{\mathtt{net}} = \theta \cdot \mathbb{E}\left(X^{\mathtt{f}}_2(\infty)\right). \tag{FP}$$

We note that (FP) is a fixed-point equation, as $\mathbb{E}(X^{\mathtt{f}}_2(\infty))$ is a function of $\lambda^{\mathtt{f}}_{\mathtt{net}}$, the other four parameters being kept fixed. Provided that there exists a unique solution to (FP) with the four parameters λ_1, λ_2, μ and θ given (necessary and sufficient conditions for this will be stated below), this unique solution is denoted by $\Lambda_{\mathtt{net}}$. We then consider the process $\mathbf{X}$ with the same transition rates (1.1) as the free process, but where the value of the parameter $\lambda^{\mathtt{f}}_{\mathtt{net}}$ has been set to $\Lambda_{\mathtt{net}}$, chosen as a function of $\lambda_1, \lambda_2, \mu$ and θ via (FP). The process $\mathbf{X}$ will be the main object of investigation in this paper.

Definition 1 Provided that there exists a unique solution to (FP), denoted by $\Lambda_{\mathtt{net}} = \Lambda_{\mathtt{net}}(\lambda_1, \lambda_2, \mu, \theta)$, the constrained model $\mathbf{X}$ is the $\mathbb{N}^2$-valued Markov process with non-zero transition rates given by (1.1) with $\lambda^{\mathtt{f}}_{\mathtt{net}} = \Lambda_{\mathtt{net}}$.

Remark 2.2 The balance equation (FP) can also be interpreted in the stand-alone context of the free process. In the free process, $\theta\mathbb{E}(X_2^{\mathtt{f}}(\infty))$ is the rate at which impatient users leave the system because of impatience. If we allow these customers to retry, we can then interpret $\lambda_{\mathtt{net}}^{\mathtt{f}}$ as the retrial rate, and (FP) then has the natural meaning that these customers will eventually re-enter the queue.

Our main result is that even a slight amount of mobility (i.e., $\theta > 0$ even very small, instead of $\theta = 0$) dramatically increases the performance of the network and leads to an unusual $\log(1/(1-\varrho))$ heavy traffic scaling. To explain this we first discuss the case $\theta = 0$ with no mobility.

2.2.2 Heavy traffic regime

When we say $\varrho \uparrow 1$, we mean that we consider a sequence of systems indexed by n, where the parameters λ_1^n, λ_2^n, μ^n and, θ^n in the n-th system satisfy $\varrho^n < 1$, (where $\varrho_i^n = \lambda_i^n/\mu^n$, $\varrho^n = \varrho_1^n + \varrho_2^n$) and, as $n \to \infty$, we have $\lambda_i^n \to \lambda_i$, $\mu^n \to \mu$, $\theta^n \to \theta$ with $\lambda_1, \lambda_2, \mu, \theta \in (0, \infty)$, $\varrho = 1$, where $\varrho = \varrho_1 + \varrho_2$ and $\varrho_i = \lambda_i/\mu$. We then use the notation $\Rightarrow_\varrho$ to mean weak convergence as $\varrho \uparrow 1$.

We will also consider convergence when other parameters vary. We use, in particular, the notation $\Rightarrow_{\lambda_{\mathtt{tot}}^{\mathtt{f}}}$ to mean weak convergence as $\lambda_{\mathtt{tot}}^{\mathtt{f}} \to \infty$, and also introduce another parameter $\varepsilon > 0$ and use the notation $\Rightarrow_{\lambda_{\mathtt{tot}}^{\mathtt{f}},\varepsilon}$ to mean weak convergence first as $\lambda_{\mathtt{tot}}^{\mathtt{f}} \to \infty$ and then as $\varepsilon \downarrow 0$. To be more precise, $Z \Rightarrow_{\lambda_{\mathtt{tot}}^{\mathtt{f}},\varepsilon} Z'$ means that for any continuous and bounded function f we have

$$\limsup_{\lambda_{\mathtt{tot}}^{\mathtt{f}} \to \infty} \left|\mathbb{E}\left(f(Z)\right) - \mathbb{E}\left(f(Z')\right)\right| \xrightarrow[\varepsilon \to 0]{} 0.$$

2.2.3 The case $\theta = 0$

Consider now the case $\theta = 0$. We distinguish two cases:

- if $\varrho \geq 1$, then the free process is transient or null recurrent, and so (FP) is not defined;
- if $\varrho < 1$, 0 is the only solution to (FP) because $\mathbb{E}(X_2^{\mathtt{f}}(\infty)) < \infty$ by Lemma 2.1.

Thus, the constrained model is only defined for $\varrho < 1$; in this case, it corresponds to the free process with $\lambda_{\mathtt{net}}^{\mathtt{f}} = 0$ and is in particular positive recurrent. The following result, taken from [20], states that its heavy traffic behavior obeys the usual $(1-\varrho)^{-1}$ scaling.

Lemma 2.3 *If $\theta = 0$, then $(1-\varrho)\mathbf{X}(\infty) \Rightarrow_\varrho (E, E)$ with E an exponential random variable with parameter* 2.

2.2.4 The case $\theta > 0$

We now show that, whatever the value of $\theta > 0$, the behavior changes dramatically and leads to an unusual $\log(1/(1-\varrho))$ scaling. We first investigate the existence and uniqueness to the fixed-point equation (FP). The proof relies on monotonicity and continuity arguments detailed in [18] and it is thus only briefly recalled here.

Lemma 2.4 *Assume that $\theta > 0$. If $\varrho < 1$, then there exists a unique solution to* (FP). *If $\varrho_1 < 1$ but $\varrho \geq 1$, then there is no solution to* (FP).

This result is comforting: indeed, $\varrho < 1$ is the "natural" stability condition. Comparing Lemmas 2.1 and 2.4, we see that imposing (FP) changes the stability condition from $\varrho_1 < 1$ (mobile users do not matter) to $\varrho < 1$ (mobile users matter). Moreover, we observe the peculiar feature that, whenever the stability condition is violated, the Markov process is not defined at all, and not simply transient as is usually the case. This is due to the fact that we seek to impose a long-term balance equation through (FP), which cannot be sustained for a system out of equilibrium.

For completeness and since the key equation (2.2) will be useful later, we provide a short sketch of the proof of Lemma 2.4. So consider $\theta > 0$ and assume $\varrho_1 < 1$, since otherwise $\mathbf{X}^{\mathtt{f}}(\infty)$ is not defined. Let

$$Q(\lambda^{\mathtt{f}}_{\mathtt{net}}) = \mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}),$$

the other four parameters being fixed. The balance of flow for the free system entails $\lambda_1 + \lambda_2 + \lambda^{\mathtt{f}}_{\mathtt{net}} = \mu\, \mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) \neq \mathbf{0}) + \theta\, \mathbb{E}(X^{\mathtt{f}}_2(\infty))$, or equivalently,

$$Q(\lambda^{\mathtt{f}}_{\mathtt{net}}) = 1 - \varrho - \frac{\lambda^{\mathtt{f}}_{\mathtt{net}}}{\mu} + \frac{\theta}{\mu}\, \mathbb{E}(X^{\mathtt{f}}_2(\infty)). \tag{2.1}$$

In particular, (FP) is equivalent to

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}) = 1 - \varrho. \tag{2.2}$$

Since $\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}) > 0$, this relation shows that no solution can exist for $\varrho \geq 1$. Assume now that $\varrho < 1$. It is intuitively clear that Q is continuous and strictly decreasing to 0: as class-2 users arrive at a higher rate, the probability of the system being empty decreases strictly and continuously to 0. As $Q(0) > 1 - \varrho$ after (2.1), this entails the existence and uniqueness of solutions to (FP). We recall that this unique solution is written $\Lambda_{\mathtt{net}}$ and define

$$\Lambda_{\mathtt{tot}} = \lambda_2 + \Lambda_{\mathtt{net}}$$

as the total arrival rate of class-2 users in the constrained model.

According to Lemma 2.4, the heavy traffic behavior consists of letting $\varrho \uparrow 1$ when $\theta > 0$. The following result is the main result of the paper. Extensions of this result are discussed in Sect. 5.

Theorem 2.5 *Assume that $\theta > 0$. As $\varrho \uparrow 1$, the sequence*

$$\frac{\mathbf{X}(\infty)}{\log(1/(1-\varrho))}$$

is tight and any of its accumulation points is almost surely smaller than the point $\boldsymbol{\xi}^*$ *given by*

$$\boldsymbol{\xi}^* = (\boldsymbol{\xi}_1^*, \boldsymbol{\xi}_2^*) = \left(\frac{\varrho_1}{1-\varrho_1}, 1\right).$$

This result shows that adding even a slight amount of mobility, i.e., going from $\theta = 0$ to $\theta > 0$, dramatically changes the heavy traffic behavior, making $\mathbf{X}(\infty)$ scale like $\log(1/(1-\varrho))$ instead of $1/(1-\varrho)$. We could actually show that $\log(1/(1-\varrho))$ is indeed the right order, i.e., accumulation points are > 0 (see Sect. 5).

Remark 2.6 It is surprising that this upper bound does not depend on θ. Indeed, when $\theta = 0$, Lemma 2.3 implies that $\mathbf{X}(\infty)/\log(1/(1-\varrho)) \Rightarrow_\varrho \infty$ and so interchanging limits suggests that $\mathbf{X}(\infty)/\log(1/(1-\varrho))$ should converge to a limit $\boldsymbol{\xi}(\theta)$ that should blow up as $\theta \downarrow 0$. This is not the case, however, and we actually conjecture that $\mathbf{X}(\infty)/\log(1/(1-\varrho))$ converges to a limit independent of θ (see Sect. 5). That limits cannot be interchanged testifies to the subtlety of the result, which, we believe, is due to the fact that we need an unusual large deviation result for a system with two time-scales; see Sect. 5.2.

Let us now explain where this unusual $\log(1/(1-\varrho))$ scaling comes from: the idea is to reduce the problem to questions on the free process $\mathbf{X}^{\mathtt{f}}$ by writing

$$\frac{\mathbf{X}(\infty)}{\log(1/(1-\varrho))} = \frac{\Lambda_{\mathtt{tot}}}{\log(1/(1-\varrho))} \times \frac{\mathbf{X}(\infty)}{\Lambda_{\mathtt{tot}}}. \tag{2.3}$$

It is easy to see that $\Lambda_{\mathtt{tot}} \to \infty$ as $\varrho \uparrow 1$. Thus, as $\mathbf{X}$ is a particular case of $\mathbf{X}^{\mathtt{f}}$, understanding the asymptotic behavior of $\mathbf{X}(\infty)/\Lambda_{\mathtt{tot}}$ as $\varrho \uparrow 1$ amounts to understanding the asymptotic behavior of $\mathbf{X}^{\mathtt{f}}(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}}$ as $\lambda^{\mathtt{f}}_{\mathtt{tot}} \to \infty$. The following result specifies this behavior.

Lemma 2.7 *Assume that* $\theta > 0$ *and* $\varrho_1 < 1$. *Then as* $\lambda^{\mathtt{f}}_{tot} \to \infty$, *the sequence* $\mathbf{X}^{\mathtt{f}}(\infty)/\lambda^{\mathtt{f}}_{tot}$ *is tight and any accumulation point is almost surely smaller than the constant* $\theta^{-1}\boldsymbol{\xi}^*$ *with* $\boldsymbol{\xi}^*$ *given as in Theorem 2.5.*

As $\varrho \uparrow 1$, *in particular, the sequence* $\mathbf{X}(\infty)/\Lambda_{tot}$ *is tight and any accumulation point is almost surely smaller than the constant* $\theta^{-1}\boldsymbol{\xi}^*$.

Next, (2.2) shows that

$$\frac{\Lambda_{\mathtt{tot}}}{\log(1/(1-\varrho))} = \frac{\Lambda_{\mathtt{tot}}}{-\log \mathbb{P}(\mathbf{X}(\infty) = \mathbf{0})}$$

and so, for the same reason as above, understanding the asymptotic behavior of $\Lambda_{\mathtt{tot}}/\log(1/(1-\varrho))$ as $\varrho \uparrow 1$ amounts to understanding the asymptotic behavior of $-\log \mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0})/\lambda^{\mathtt{f}}_{\mathtt{tot}}$ as $\lambda^{\mathtt{f}}_{\mathtt{tot}} \to \infty$.

Lemma 2.8 *Assume that $\theta > 0$. For any $\varrho_1 < 1$, we then have*

$$\liminf_{\lambda^{\mathtt{f}}_{tot}\to\infty}\left(-\frac{1}{\lambda^{\mathtt{f}}_{tot}}\log\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty)=\mathbf{0})\right)\geq\frac{1}{\theta}.$$

In particular,

$$\limsup_{\varrho\uparrow 1}\left(\frac{\Lambda_{tot}}{\log(1/(1-\varrho))}\right)\leq\theta.$$

In view of (2.3), the two previous lemmas directly imply Theorem 2.5. In other words, the $\log(1/(1-\varrho))$ scaling of $\mathbf{X}(\infty)$ arises for the two following reasons:

(1) the (at most) linear increase of $\mathbf{X}^{\mathtt{f}}(\infty)\leq\lambda^{\mathtt{f}}_{\mathtt{tot}}\boldsymbol{\xi}^*+o(\lambda^{\mathtt{f}}_{\mathtt{tot}})$ as $\lambda^{\mathtt{f}}_{\mathtt{tot}}\to\infty$;

(2) the exponential decay of $\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty)=\mathbf{0})\leq e^{-\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta+o(\lambda^{\mathtt{f}}_{\mathtt{tot}})}$ as $\lambda^{\mathtt{f}}_{\mathtt{tot}}\to\infty$.

Lemmas 2.7 and 2.8 are proved in Sects. 3 and 4.

Remark 2.9 In Sect. 5, we discuss refinements of these upper bounds: in particular, we show how to prove that $\mathbf{X}^{\mathtt{f}}(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}}\Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}}}\theta^{-1}\boldsymbol{\xi}^*$, and we conjecture that

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty)=\mathbf{0})=\exp(-\kappa\lambda^{\mathtt{f}}_{\mathtt{tot}}+o(\lambda^{\mathtt{f}}_{\mathtt{tot}}))$$

with constant

$$\kappa=\frac{1-\log(1-\varrho_1)}{\theta}.$$

Remark 2.10 The linear increase in $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ of $\mathbf{X}^{\mathtt{f}}(\infty)$ is natural in the setting of single-server queues. Moreover, the refinement $\mathbf{X}^{\mathtt{f}}(\infty)\approx\lambda^{\mathtt{f}}_{\mathtt{tot}}\boldsymbol{\xi}^*$ suggests that $\mathbf{X}^{\mathtt{f}}(\infty)$ is of the order of $\lambda^{\mathtt{f}}_{\mathtt{tot}}$. This makes state 0 far from the typical value of $\mathbf{X}^{\mathtt{f}}(\infty)$ and the exponential decay of the stationary probability of being at 0 is thus expected in view of large deviations theory. The link with large deviations theory is discussed in more detail in Sect. 5.

3 Proof of Lemma 2.7

In the rest of the paper, we use several couplings. We use the notation $\mathbf{X}\prec\mathbf{Y}$ to mean that we can couple $\mathbf{X}$ and $\mathbf{Y}$ such that $\mathbf{X}\leq\mathbf{Y}$. If $\mathbf{X}$ and $\mathbf{Y}$ are random processes, this is to be understood as $\mathbf{X}(t)\leq\mathbf{Y}(t)$ for all t, and vector inequalities are understood component-wise.

In order to prove Lemma 2.7, we first exhibit a family of processes $\mathbf{Y}'$ indexed by some additional parameter $\varepsilon>0$ and with $\mathbf{X}^{\mathtt{f}}\prec\mathbf{Y}'$ for every $\varepsilon>0$, and $\mathbf{Y}'(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}}\Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon}\theta^{-1}\boldsymbol{\xi}^*$. We build this coupling in two steps, and then analyze the process $\mathbf{Y}'$. In order to prove that $\mathbf{Y}'(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}}\Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon}\theta^{-1}\boldsymbol{\xi}^*$, we then exhibit another family of processes $\mathbf{Y}$ with $\mathbf{Y}(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}}\Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon}\theta^{-1}\boldsymbol{\xi}^*$ and such that $(\mathbf{Y}(\infty)-\mathbf{Y}'(\infty))/\lambda^{\mathtt{f}}_{\mathtt{tot}}\Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon}0$.

3.1 First coupling: $\mathbf{X}^{\mathtt{f}} \prec \widetilde{\mathbf{Y}}$

Starting from (1.1), the first step consists of neglecting the term $\mu x_2/(x_1 + x_2)$ in the departure rate of $X_2^{\mathtt{f}}$ by lower bounding it by 0. When we do so, this makes the departure rate smaller for the second coordinate, which makes it larger, which in turn makes the departure rate $\mu y_1/(y_1 + y_2)$ from the first coordinate smaller, and hence the first coordinate larger. Thus if $\widetilde{\mathbf{Y}}$ is the $\mathbb{N}^2$-valued Markov process with non-zero transition rates

$$\mathbf{y} \in \mathbb{N}^2 \longrightarrow \begin{cases} \mathbf{y} + \mathbf{e}_1 & \text{at rate } \lambda_1, \\ \mathbf{y} + \mathbf{e}_2 & \text{at rate } \lambda_{\mathtt{tot}}^{\mathtt{f}}, \\ \mathbf{y} - \mathbf{e}_1 & \text{at rate } \mu y_1/(y_1 + y_2), \\ \mathbf{y} - \mathbf{e}_2 & \text{at rate } \theta y_2, \end{cases}$$

then we have $\mathbf{X}^{\mathtt{f}} \prec \widetilde{\mathbf{Y}}$. For completeness, we provide a proof of this result.

Proof of $\mathbf{X}^{\mathtt{f}} \prec \widetilde{\mathbf{Y}}$. Let the current state of our coupling be $(\mathbf{x}, \widetilde{\mathbf{y}}) \in \mathbb{N}^2 \times \mathbb{N}^2$ with $\widetilde{\mathbf{y}} \geq \mathbf{x}$. We see $\mathbf{x}$ as the "small" system and we index its customers by (i, k) with $i \in \{1, 2\}$ (the user class) and $k = 1, \ldots, x_i$. The "big" system $\widetilde{\mathbf{y}}$ has the same customers and also additional ones which we label $(i, -k)$ with $i \in \{1, 2\}$ and $k = 1, \ldots, \tilde{y}_i - x_i$. The next transition is built as follows:

- at rate λ_1, go to $(\mathbf{x} + \mathbf{e}_1, \widetilde{\mathbf{y}} + \mathbf{e}_1)$;
- at rate $\lambda_{\mathtt{tot}}^{\mathtt{f}}$, go to $(\mathbf{x} + \mathbf{e}_2, \widetilde{\mathbf{y}} + \mathbf{e}_2)$;
- each customer $(2, k)$ of type 2 has an exponential clock with parameter θ and leaves the system if it rings: note that if $k < 0$ this only affects the big system, while if $k > 0$ this affects both systems;
- at rate μ, do the following:

  1. choose a customer $\widetilde{C}$ from the big system uniformly at random, i.e.,

  $$\mathbb{P}(\widetilde{C} = (i, k)) = \frac{1}{\tilde{y}_1 + \tilde{y}_2};$$

  2. if $\widetilde{C}$ is in the small system, let $C = \widetilde{C}$;
  3. otherwise, let C be chosen uniformly at random in the small system independently from everything else;

  Then remove the customer C from the small system and remove the customer $\widetilde{C}$ from the big system if it is of type 1.

This construction is such that

- if a class i customer arrives in the small system it also arrives in the big system;
- if a class i customer leaves the big system and not the small one, then this customer was an "additional" customer which was in the big system but not in the small one.

In particular, this construction leads to a state $(\mathbf{x}', \widetilde{\mathbf{y}}')$ with $\widetilde{\mathbf{y}}' \geq \mathbf{x}'$. Moreover, the small system has the same dynamics as $\mathbf{X}^{\mathtt{f}}$ because C is chosen uniformly at random

in the small system, and the big system has the same dynamic as $\widetilde{\mathbf{Y}}$. Thus, this indeed builds a coupling of $\mathbf{X}^{\mathtt{f}}$ and $\widetilde{\mathbf{Y}}$ with $\mathbf{X}^{\mathtt{f}} \preceq \widetilde{\mathbf{Y}}$, as desired. □

3.2 Second coupling: $\widetilde{\mathbf{Y}} \prec \mathbf{Y}'$

Starting from $\widetilde{\mathbf{Y}}$, we build $\mathbf{Y}'$ by lowering the service rate of $\widetilde{Y}_1$: when $\widetilde{Y}_2$ is larger than some threshold ℓ, we put the service to 0, and when $\widetilde{Y}_2 \leq \ell$, we put $\mu y_1/(y_1+\ell)$ instead of $\mu y_1/(y_1+y_2)$, the former being indeed smaller than the latter when $y_2 \leq \ell$. More precisely, we fix $\varepsilon > 0$ (which is omitted from the notation for convenience) and we define $\ell = (1+\varepsilon)\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$ and $\mathbf{Y}'$ the $\mathbb{N}^2$-valued Markov process with non-zero transition rates

$$\mathbf{y} \in \mathbb{N}^2 \longrightarrow \begin{cases} \mathbf{y}+\mathbf{e}_1 & \text{at rate } \lambda_1, \\ \mathbf{y}+\mathbf{e}_2 & \text{at rate } \lambda^{\mathtt{f}}_{\mathtt{tot}}, \\ \mathbf{y}-\mathbf{e}_1 & \text{at rate } \dfrac{\mu\, y_1}{y_1+\ell} \cdot \mathbb{1}\,(y_2 \leq \ell)\,, \\ \mathbf{y}-\mathbf{e}_2 & \text{at rate } \theta y_2, \end{cases}$$

so that $\widetilde{\mathbf{Y}} \prec \mathbf{Y}'$ (in contrast to the inequality $\mathbf{X}^{\mathtt{f}} \prec \widetilde{\mathbf{Y}}$, the proof bears no difficulty and is thus omitted). Since $\mathbf{X}^{\mathtt{f}} \prec \widetilde{\mathbf{Y}}$, this gives $\mathbf{X}^{\mathtt{f}} \prec \mathbf{Y}'$ as desired.

Note that Y'_2 is an $M/M/\infty$ queue, so that $Y'_2(\infty)$ follows a Poisson distribution with parameter $\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$. In particular, we obtain the convergence $Y'_2(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}}} \theta^{-1}\boldsymbol{\xi}^*_2$ and so we only have to prove that $Y'_1(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}^*_1$ in order to prove Lemma 2.7. To do so we resort to another coupling and compare Y'_1 to a birth-and-death process Y_1.

3.3 Third coupling: $\mathbf{Y} \prec \mathbf{Y}'$

As ℓ is larger than the equilibrium point $\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$ of Y'_2, excursions of Y'_2 above level ℓ are rare and so Y'_1 is only rarely turned off. For this reason, it is natural to compare $\mathbf{Y}'$ with the process obtained by putting the indicator function $\mathbb{1}\,(y_2 \leq \ell)$ to 1. To do so, let

$$S = \{(y_1, y'_1, y_2) \in \mathbb{N}^3 : y'_1 \geq y_1\} \subset \mathbb{N}^3;$$

we directly build the coupling that we need and consider (Y_1, Y'_1, Y_2), the S-valued Markov process with the following non-zero transition rates:

$$(y_1, y'_1, y_2) \in S \longrightarrow \begin{cases} (y_1, y'_1, y_2+1) & \text{at rate } \lambda^{\mathtt{f}}_{\mathtt{tot}}, \\ (y_1, y'_1, y_2-1) & \text{at rate } \theta y_2, \\ (y_1+1, y'_1+1, y_2) & \text{at rate } \lambda_1, \\ (y_1-1, y'_1-1, y_2) & \text{at rate } \mu\alpha(y_1)\mathbb{1}\,(y_2 \leq \ell)\,, \\ (y_1-1, y'_1, y_2) & \text{at rate } \mu\alpha(y_1)\mathbb{1}\,(y_2 > \ell)\,, \\ (y_1, y'_1-1, y_2) & \text{at rate } \mu\beta_{y'_1-y_1}(y_1)\mathbb{1}\,(y_2 \leq \ell)\,, \end{cases} \tag{3.1}$$

with

$$\alpha(y) = \frac{y}{y+\ell} \text{ and } \beta_\delta(y) = \alpha(y+\delta) - \alpha(y) = \frac{\ell\delta}{(y+\ell)(y+\ell+\delta)}.$$

In words, what this process does is the following:

- Y_1 and Y_2 are independent Markov processes;
- Y_1 is a state-dependent single-server queue with arrival rate λ_1 and instantaneous service rate $\mu\alpha(y_1)$ when in state y_1;
- Y_2 is an $M/M/\infty$ queue with arrival rate $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ and service rate θ;
- Y_1' has the same arrivals as Y_1, but departures are different: there are additional departures at rate $\beta_{y_1'-y_1}(y_1)$ when $Y_2 \le \ell$, and no departure when $Y_2 > \ell$.

Since the function β has been chosen so that

$$\beta_{y_1'-y_1}(y_1) + \alpha(y_1) = \alpha(y_1'),$$

we see that (Y_1', Y_2) is a Markov process with the same transition matrix as $\mathbf{Y}'$, and so we will actually write $\mathbf{Y}' = (Y_1', Y_2)$ and we have $\mathbf{Y}' \ge \mathbf{Y} := (Y_1, Y_2)$.

In particular, this coupling defines several Markov processes, such as Y_1, Y_2, $Y := (Y_1, Y_2)$, $Y' = (Y_1', Y_2)$ and (Y_1, Y_1', Y_2). For ease of notation, we will use the notation $\mathbb{E}_{\mathbf{x}}$ to denote the law of these Markov processes starting at $\mathbf{x}$, where the dimension of $\mathbf{x}$ depends on the process considered. For instance, if σ is measurable with respect to Y_2 and $\varphi : \mathbb{N} \to \mathbb{R}_+$ is measurable, we will use the notation

$$\mathbb{E}_\ell(\sigma),\ \mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1\right),\ \mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1'\right) \text{ or } \mathbb{E}_{z,z',\ell}\left(\int_0^\sigma \left(\varphi \circ Y_1 - \varphi \circ Y_1'\right)\right)$$

to mean the following:

$$\begin{aligned}
&\mathbb{E}_\ell(\sigma) = \mathbb{E}(\sigma \mid Y_2(0) = \ell),\\
&\mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1\right) = \mathbb{E}\left(\int_0^\sigma \varphi \circ Y_1 \mid Y_2(0) = \ell, Y_1(0) = z\right),\\
&\mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1'\right) = \mathbb{E}\left(\int_0^\sigma \varphi \circ Y_1' \mid Y_2(0) = \ell, Y_1'(0) = z\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{z,z',\ell}\left(\int_0^\sigma \left(\varphi \circ Y_1 - \varphi \circ Y_1'\right)\right)\\
&= \mathbb{E}\left(\int_0^\sigma \left(\varphi \circ Y_1 - \varphi \circ Y_1'\right) \mid Y_1(0) = z, Y_1'(0) = z', Y_2(0) = \ell\right).
\end{aligned}$$

Recall that the goal is to prove that $Y_1'(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}_1^*$: what we will do is first prove this result for Y_1, which is much simpler since Y_1 is a birth-and-death process (whereas Y_1' on its own is not Markov) and then transfer this result to Y_1'.

3.4 Control of Y_1

Let us now prove that $Y_1(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}_1^*$. Let first $y^{\pm} = (1 \pm \varepsilon)\ell\boldsymbol{\xi}_1^*$. Since the function α is increasing, when Y_1 is above level y^+ its departure rate is at least $\mu\alpha(y^+)$. Thus, if L^+ is an $M/M/1$ queue with arrival rate λ_1 and departure rate $\mu\alpha(y^+)$, we have $Y_1 \prec L^+ + y^+$. Likewise, if L^- is an $M/M/1$ queue with arrival rate $\mu\alpha(y^-)$ and departure rate λ_1, we have $y^- - L^- \prec Y_1$.

Recall that $\boldsymbol{\xi}_1^* = \varrho_1/(1-\varrho_1)$ and that $\varrho_1 < 1$: the load of L^+ is

$$\frac{\lambda_1}{\mu\alpha(y^+)} = \frac{\varrho_1((1+\varepsilon)\ell\boldsymbol{\xi}_1^* + \ell)}{(1+\varepsilon)\ell\boldsymbol{\xi}_1^*} = \frac{1+\varepsilon\varrho_1}{1+\varepsilon} < 1$$

and the load of L^- is

$$\frac{\mu\alpha(y^-)}{\lambda_1} = \frac{1-\varepsilon}{1-\varepsilon\varrho_1} < 1.$$

We thus deduce that $L^{\pm}$ are subcritical $M/M/1$ queues (uniformly in $\lambda^{\mathtt{f}}_{\mathtt{tot}}$, with $\varepsilon > 0$ fixed), so that

$$\frac{L^{\pm}(\infty)}{\lambda^{\mathtt{f}}_{\mathtt{tot}}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}}} 0.$$

Since $y^{\pm}/\lambda^{\mathtt{f}}_{\mathtt{tot}} \to_{\lambda^{\mathtt{f}}_{\mathtt{tot}}} (1 \pm \varepsilon)(1+\varepsilon)\theta^{-1}\boldsymbol{\xi}_1^*$, we obtain

$$\frac{y^- - L^-(\infty)}{\lambda^{\mathtt{f}}_{\mathtt{tot}}}, \quad \frac{L^+(\infty) - y^+}{\lambda^{\mathtt{f}}_{\mathtt{tot}}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}_1^*$$

and in view of $y^- - L^- \prec Y_1 \prec L^+ - y^+$, we finally get the desired result for $Y_1(\infty)$, namely $Y_1(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}_1^*$.

3.5 Transfer to Y_1'

We now transfer the result for $Y_1(\infty)$ to $Y_1'(\infty)$ thanks to their coupling (3.1). Recall that Y_1 and Y_1' obey the same dynamics, with the exception that service in Y_1' is interrupted when Y_2 makes excursions above ℓ. To compare their stationary distributions, we consider their trajectories over cycles of Y_2, where a cycle starts when $Y_2 = \ell$ and ends when Y_2 returns to ℓ from above: so there is a long period corresponding to $Y_2 \le \ell$ where Y_1 and Y_1' have the same dynamics, $Y_1' \ge Y_1$ and they get closer (because the departure rate from Y_1' is larger), and then a short period when $Y_2 \ge \ell + 1$ where departures from Y_1' are turned off and Y_1' and Y_1 get further apart (when there is a departure from Y_1). Considering such cycles makes the comparison between Y_1 and Y_1' tractable.

To formalize this idea, define recursively the stopping times $\sigma_0 = 0$ and

$$\tau_k = \inf\{t \geq \sigma_k : Y_2(t) \geq \ell + 1\}, \quad \sigma_{k+1} = \inf\{t \geq \tau_k : Y_2(t) = \ell\}$$

and let $Z_k = Y_1(\sigma_k)$, $Z'_k = Y'_1(\sigma_k)$. Note that Z and Z' are ergodic Markov chains. Let Z_∞ and Z'_∞ be their respective stationary distribution and note, since Y_2 and Y_1 are independent, that $Z_\infty = Y_1(\infty)$ in distribution.

Now let $\sigma = \sigma_1$ and $\tau = \tau_0$; for any function $\varphi : \mathbb{N} \to \mathbb{R}_+$, define the functions Ψ_φ and $\widetilde{\Psi}_\varphi$ by

$$\Psi_\varphi(z) = \mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1\right) \text{ and } \widetilde{\Psi}_\varphi(z) = \mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y'_1\right).$$

The following result then relates the stationary distribution of $Y_1(\infty)$ and $Y'_1(\infty)$ to that of Z_∞ and Z'_∞, respectively. In the sequel, we write $\|f\| = \sup_{t\geq 0}|f(t)|$ for the L_∞-norm of a function $f : \mathbb{R}_+ \to \mathbb{R}$.

Lemma 3.1 *For any bounded function $\varphi : \mathbb{N} \to \mathbb{R}_+$ we have*

$$\mathbb{E}\left[\varphi(Y_1(\infty))\right] = \frac{1}{\mathbb{E}_\ell(\sigma)}\mathbb{E}\left[\Psi_\varphi(Z_\infty)\right] \text{ and } \mathbb{E}\left[\varphi(Y'_1(\infty))\right] = \frac{1}{\mathbb{E}_\ell(\sigma)}\mathbb{E}\left[\Psi_\varphi(Z'_\infty)\right].$$

Proof We present the arguments only for Y'_1, as the same arguments apply to Y_1. In this proof, $\to$ denotes almost sure convergence as $n \to \infty$. Since $(\sigma_k)_{k\geq 0}$ is a (possibly delayed) renewal process, by the strong Markov property we have

$$\frac{1}{n}\int_0^{\sigma_n} \varphi \circ Y'_1 \to \mathbb{E}_\ell(\sigma) \times \mathbb{E}(\varphi(Y'_1(\infty))).$$

The rest of the proof is devoted to showing that we also have

$$\frac{1}{n}\int_0^{\sigma_n} \varphi \circ Y'_1 \to \mathbb{E}(\Psi_\varphi(Z'_\infty)).$$

Recall that $[\sigma_k, \sigma_{k+1}]$ represents the kth cycle of Y_2: $Y_2(\sigma_k) = \ell$, then Y_2 reaches $\ell + 1$ at time τ_k and goes back to ℓ at time σ_{k+1}. For each cycle, Y'_1 starts in a random location $Y'_1(\sigma_k)$: call the i-th z-cycle the i-th cycle of Y_2 such that Y'_1 starts in z, and denote its corresponding time interval by $[\sigma_i(z), \sigma_{i+1}(z)]$. If

$$\Upsilon_i(z) = \int_{\sigma_i(z)}^{\sigma_{i+1}(z)} \varphi \circ Y'_1$$

represents the "reward" accumulated along the i-th z-cycle, then writing

$$N_n(z) = \sum_{k=0}^{n-1} \mathbb{1}\left(Z'_k = z\right)$$

for the number of z-cycles starting before time n, partitioning the cycles depending on their starting point (for Y_1') provides

$$\frac{1}{n}\int_0^{\sigma_n} \varphi \circ Y_1' = \frac{1}{n}\sum_{z\geq 0}\sum_{k=1}^{N_n(z)} \Upsilon_i(z) = \sum_{z\geq 0}\frac{N_n(z)}{n} \times \frac{1}{N_n(z)}\sum_{i=1}^{N_n(z)} \Upsilon_i(z).$$

The ergodic theorem for Z' implies that $N_n(z)/n \to \mathbb{P}(Z'_\infty = z)$. Moreover, for each fixed z, the $(\Upsilon_k(z), k \geq 0)$ are i.i.d. with common distribution that of $\int_0^\sigma \varphi \circ Y_1'$ under $\mathbb{P}_{z,\ell}$, so the strong law of large numbers implies that

$$\frac{1}{N_n(z)}\sum_{i=1}^{N_n(z)} \Upsilon_i(z) \to \mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi \circ Y_1'\right) = \Psi_\varphi(z).$$

Wrapping up, this suggests that

$$\sum_{z\geq 0}\frac{N_n(z)}{n} \times \frac{1}{N_n(z)}\sum_{i=1}^{N_n(z)} \Upsilon_i(z) \to \sum_{z\geq 0}\mathbb{P}(Z'_\infty = n)\Psi_\varphi(z),$$

which is equal to $\mathbb{E}(\Psi_\varphi(Z'_\infty))$, as desired. Let us justify the latter assertion. If we restrict the sum in z to a finite number of terms, then the previous arguments can then be applied and they give, for any $z^* \geq 0$,

$$\frac{1}{n}\sum_{z\leq z^*}\sum_{i=1}^{N_n(z)} \Upsilon_i(z) \to \sum_{z\leq z^*}\mathbb{P}(Z'_\infty = z)\Psi_\varphi(z) = \mathbb{E}(\Psi_\varphi(Z'_\infty);\, Z'_\infty \leq z^*).$$

Since $\Psi_\varphi \geq 0$ (because $\varphi \geq 0$), monotone convergence implies that the above right-hand side converges to $\mathbb{E}(\Psi_\varphi(Z'_\infty))$ as $z^* \to \infty$. Thus, in order to complete the proof, it remains to show that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{z\geq z^*}\sum_{i=1}^{N_n(z)} \Upsilon_i(z) \xrightarrow[z^*\to\infty]{} 0.$$

We have

$$\frac{1}{n}\sum_{z\geq z^*}\sum_{i=1}^{N_n(z)} \Upsilon_i(z) \leq \frac{\|\varphi\|}{n}\sum_{z\geq z^*}\sum_{i\leq N_n(z)} (\sigma_{i+1}(z) - \sigma_i(z)) = \frac{\|\varphi\|}{n}\sum_{k=0}^{n-1}\delta_{k+1}\mathbb{1}\left(Z_k \geq z^*\right),$$

where $\delta_{k+1} = \sigma_{k+1} - \sigma_k$. The process (δ_{k+1}, Z'_k) is Markov: given the past until time k, δ_{k+2} is independent and distributed according to σ under $\mathbb{P}_\ell$, and Z'_{k+1} corresponds to the evolution of Y_1' in-between a Z'_k-cycle of Y_2 with length δ_{k+1}. Note that, because each sequence (δ_{k+1}) and (Z'_k) is tight, the sequence (δ_{k+1}, Z_k) is also tight and since

it is also Markov, it converges to $(\delta_\infty, Z'_\infty)$, say. Thus, the ergodic theorem implies that

$$\frac{1}{n}\sum_{k=0}^{n-1}\delta_{k+1}\mathbb{1}\left(Z_k \geq z^*\right) \to \mathbb{E}\left(\delta_\infty; Z'_\infty \geq z^*\right),$$

and since $\mathbb{E}(\delta_\infty) = \mathbb{E}_\ell(\sigma) < \infty$, we obtain the desired result by monotone convergence and letting $z^* \to \infty$. □

By Lemma 3.1, we thus deduce that

$$\mathbb{E}\left[\varphi(Y(\infty))\right] - \mathbb{E}\left[\varphi(Y'(\infty))\right] = \frac{1}{\mathbb{E}_\ell(\sigma)}\left(\mathbb{E}\left[\Psi_\varphi(Z_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z'_\infty)\right]\right). \tag{3.2}$$

To control the right-hand side, we decompose it as

$$\left(\mathbb{E}\left[\Psi_\varphi(Z_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right]\right) + \left(\mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z'_\infty)\right]\right)$$

and control each difference in the next two lemmas.

Lemma 3.2 *We have*

$$\left|\mathbb{E}\left[\Psi_\varphi(Z_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right]\right| \leq 2\|\varphi\|\mathbb{E}_{\ell+1}(\sigma).$$

Proof Using the strong Markov property at time τ, we can write

$$\Psi_\varphi(z) = \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi \circ Y_1\right) + \sum_{z'\geq 0}\mathbb{P}_{z,\ell}\left(Y_1(\tau) = z'\right)\mathbb{E}_{z',\ell+1}\left(\int_0^\sigma \varphi \circ Y_1\right)$$

and

$$\begin{aligned}\widetilde{\Psi}_\varphi(z) &= \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi \circ Y'_1\right) + \sum_{z'\geq 0}\mathbb{P}_{z,\ell}\left(Y'_1(\tau) = z'\right)\mathbb{E}_{z',\ell+1}\left(\int_0^\sigma \varphi \circ Y'_1\right)\\ &= \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi \circ Y_1\right) + \sum_{z'\geq 0}\mathbb{P}_{z,\ell}\left(Y_1(\tau) = z'\right)\mathbb{E}_{z',\ell+1}\left(\int_0^\sigma \varphi \circ Y'_1\right),\end{aligned}$$

where the second equality comes from the fact that Y_1 and Y'_1 coincide on $[0, \tau]$ if they start at the same level. Since Y_1 is independent of Y_2 (and hence of τ) and Y and Z have the same stationary distribution, we have

$$
\begin{aligned}
&\mathbb{E}\left[\Psi_\varphi(Z_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right] \\
&\quad = \sum_{z\geq 0} \mathbb{P}(Z_\infty = z)\mathbb{E}_{z,z,\ell+1}\left(\int_0^\sigma \varphi\circ Y_1\right) \\
&\qquad - \sum_{z\geq 0} \mathbb{P}(Z_\infty = z)\mathbb{E}_{z,z,\ell+1}\left(\int_0^\sigma \varphi\circ Y_1'\right),
\end{aligned}
$$

from which the result follows. □

Lemma 3.3 *If $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is differentiable with derivative φ', then we have*

$$
\left|\mathbb{E}\left[\widetilde{\Psi}_\varphi(Z'_\infty)\right] - \mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right]\right| \leq \|\varphi'\|\mathbb{E}(Z'_\infty - Z_\infty)\mathbb{E}_\ell(\sigma).
$$

Proof Fix temporarily $z' \geq z$. Using the relation

$$
\mathbb{E}_{z,\ell}\left(\int_0^\sigma \varphi\circ Y_1'\right) = \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi\circ Y_1\right) + \mathbb{E}_{z,\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right),
$$

owing to the fact that Y_1 and Y_1' have the same dynamics on $[0, \tau]$, write

$$
\begin{aligned}
\widetilde{\Psi}_\varphi(z') - \widetilde{\Psi}_\varphi(z) \ &= \mathbb{E}_{z',\ell}\left(\int_0^\tau \varphi\circ Y_1\right) - \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi\circ Y_1\right) \\
&\quad + \mathbb{E}_{z',\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right) - \mathbb{E}_{z,\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right).
\end{aligned}
$$

To compute

$$
\mathbb{E}_{z',\ell}\left(\int_0^\tau \varphi\circ Y_1\right) - \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi\circ Y_1\right),
$$

we couple Y_1 starting from two different initial conditions z and z': in fact, when considered on $[0, \tau]$, this is exactly what the coupling between Y_1 and Y_1' does, and so we thus have

$$
\mathbb{E}_{z',\ell}\left(\int_0^\tau \varphi\circ Y_1\right) - \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi\circ Y_1\right) = \mathbb{E}_{z,z',\ell}\left(\int_0^\tau \left(\varphi\circ Y_1' - \varphi\circ Y_1\right)\right)
$$

and so

$$
\begin{aligned}
\left|\mathbb{E}_{z',\ell}\left(\int_0^\tau \varphi\circ Y_1\right) - \mathbb{E}_{z,\ell}\left(\int_0^\tau \varphi\circ Y_1\right)\right| &\leq \|\varphi'\|\mathbb{E}_{z,z',\ell}\left(\int_0^\tau \left(Y_1' - Y_1\right)\right) \\
&\leq \|\varphi'\|(z' - z)\mathbb{E}_\ell(\tau),
\end{aligned}
$$

with the last inequality coming from the fact that $Y_1' - Y_1$ is non-increasing on $[0, \tau]$.

We now control the difference

$$\mathbb{E}_{z',\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right)-\mathbb{E}_{z,\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right).$$

Consider (Υ,Υ') and A such that

- (Υ,Υ'), A and Y_2 are mutually independent;
- (Υ,Υ') is distributed as $(Y_1(\tau),Y_1'(\tau))$ under $\mathbb{P}_{z,z',\ell}$;
- A is a Poisson process with intensity λ_1.

On the interval $[\tau,\sigma]$, $Y_1'-Y_1'(\tau)$ is simply a Poisson process distributed as A: the strong Markov property therefore implies that

$$\begin{aligned}&\mathbb{E}_{z',\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right)-\mathbb{E}_{z,\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right)\\&\quad=\mathbb{E}_{\ell+1}\left(\int_0^\sigma\big(\varphi(\Upsilon'+A(s))-\varphi(\Upsilon+A(s))\big)\,\mathrm{d}s\right)\end{aligned}$$

and so

$$\left|\mathbb{E}_{z',\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right)-\mathbb{E}_{z,\ell}\left(\int_\tau^\sigma \varphi\circ Y_1'\right)\right|\le\|\varphi'\|\mathbb{E}_{z,z',\ell}\left(Y_1'(\tau)-Y_1(\tau)\right)\mathbb{E}_{\ell+1}(\sigma).$$

Finally, using $Y_1'(\tau)\le Y_1'(\sigma)$ and averaging over (Z_∞,Z_∞'), we obtain

$$\left|\mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty')\right]-\mathbb{E}\left[\widetilde{\Psi}_\varphi(Z_\infty)\right]\right|\le\|\varphi'\|\mathbb{E}(Z_\infty'-Z_\infty)(\mathbb{E}_\ell(\tau)+\mathbb{E}_{\ell+1}(\sigma)),$$

from which the result follows since $\mathbb{E}_\ell(\sigma)=\mathbb{E}_\ell(\tau)+\mathbb{E}_{\ell+1}(\sigma)$ by the strong Markov property. □

Plugging in the bounds of the two previous lemmas into (3.2), we conclude that for any function $f:\mathbb{R}_+\to\mathbb{R}_+$ bounded, differentiable and with bounded derivative, we have

$$\begin{aligned}&\left|\mathbb{E}\left[f\left(\frac{Y_1(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]-\mathbb{E}\left[f\left(\frac{Y_1'(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]\right|\\&\quad\le\frac{1}{\mathbb{E}_\ell(\sigma)}\left(2\|f\|\mathbb{E}_{\ell+1}(\sigma)+\frac{1}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\|f'\|\mathbb{E}(Z_\infty'-Z_\infty)\mathbb{E}_\ell(\sigma)\right),\end{aligned}$$

hence

$$\left|\mathbb{E}\left[f\left(\frac{Y_1(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]-\mathbb{E}\left[f\left(\frac{Y_1'(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]\right|\le 2\|f\|\frac{\mathbb{E}_{\ell+1}(\sigma)}{\mathbb{E}_\ell(\sigma)}+\frac{1}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\|f'\|\mathbb{E}(Z_\infty'-Z_\infty).$$

The two following lemmas therefore imply that

$$\mathbb{E}\left[f\left(\frac{Y_1(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]-\mathbb{E}\left[f\left(\frac{Y_1'(\infty)}{\lambda_{\mathtt{tot}}^{\mathtt{f}}}\right)\right]\to 0$$

as $\lambda^{\mathtt{f}}_{\mathtt{tot}} \to \infty$ for any differentiable, bounded function f with bounded derivative. Since $Y_1(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}^*_1$, this implies that

$$\mathbb{E}\left[f\left(\frac{Y_1'(\infty)}{\lambda^{\mathtt{f}}_{\mathtt{tot}}}\right)\right] \to_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} f(\theta^{-1}\boldsymbol{\xi}^*_1),$$

which shows that $Y_1'(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}^*_1$, as claimed.

Lemma 3.4 *As $\lambda^{f}_{tot} \to \infty$, we have $\mathbb{E}_{\ell+1}(\sigma)/\mathbb{E}_{\ell}(\sigma) \to 0$.*

Proof For $\gamma > 0$, let T^{γ} be the hitting time of 0 by an $M/M/1$ queue started at 1 and with input rate γ and output rate $(1+\varepsilon)\gamma$. Above level $\ell+1$, Y_2 is upper bounded by an $M/M/1$ queue with input rate $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ and output rate $\theta\ell = (1+\varepsilon)\lambda^{\mathtt{f}}_{\mathtt{tot}}$, so that $\sigma \prec T^{\lambda^{\mathtt{f}}_{\mathtt{tot}}}$, where σ is considered under $\mathbb{P}_{\ell+1}$. Since $T^{\gamma} = T^1/\gamma$ in distribution, this yields $\mathbb{E}_{\ell+1}(\sigma) \le \mathbb{E}(T^1)/\lambda^{\mathtt{f}}_{\mathtt{tot}}$. Since clearly $\mathbb{E}_{\ell}(\sigma) \to \infty$, we obtain the result. □

Lemma 3.5 *As $\lambda^{f}_{tot} \to \infty$, we have $\limsup \mathbb{E}(Z'_{\infty} - Z_{\infty}) < \infty$. In particular, $\mathbb{E}(Z'_{\infty} - Z_{\infty})/\lambda^{f}_{tot} \to 0$.*

Proof Let $\Delta_k = Z'_k - Z_k$. The idea is that when Δ_k is large, then on $[\sigma_k, \tau_k]$ the function $\beta_{Y_1 - Y_1'}$ takes (relatively) large values which brings the processes Y_1' and Y_1 closer and makes Δ_{k+1} smaller. To formalize this idea, we use Theorem 2.3 in [11]: to apply this result, we need to control the exponential moments of Δ_1. To do so, we consider P and P', two Poisson point processes on $\mathbb{R}_+ \times [0,1]$ with intensity $\mu \mathrm{d}t \otimes \mathrm{d}x$, such that P, P', Y_1 and Y_2 are independent, and we write

$$\begin{aligned}\Delta_1 - \Delta_0 = &-\int \mathbb{1}\left(0 \le s \le \tau, \zeta \le \beta_{Y_1'(s-)-Y_1(s-)}(Y_1(s-))\right) P'(\mathrm{d}s\mathrm{d}\zeta)\\ &+\int \mathbb{1}\left(\tau \le s \le \sigma, \zeta \le \alpha(Y_1'(s-))\right) P(\mathrm{d}s\mathrm{d}\zeta). \qquad (3.3)\end{aligned}$$

The first negative term translates the fact that on $[0,\tau]$, Y_1' and Y_1 get closer at rate $\mu\beta_{Y_1'-Y_1}(Y_1)$ (which is the rate at which there is a departure from Y_1' and not from Y_1), while on $[\tau,\sigma]$ they get further apart at rate $\mu\alpha(Y_1)$ (which is the rate at which there is a departure from Y_1). Moreover, as in the previous proof, let T^{γ} be the hitting time of 0 by an $M/M/1$ queue started at 1 and with input rate γ and output rate $(1+\varepsilon)\gamma$, and η be given by

$$\mu(e^{\eta} - 1) = \lambda^{\mathtt{f}}_{\mathtt{tot}}\left(\sqrt{1+\varepsilon} - 1\right)^2.$$

First case: $\Delta_0 = 0$. If $\Delta_0 = 0$, we then have $Y_1 = Y_1'$ on $[0,\tau]$ and so (3.3) reduces to

$$\Delta_1 - \Delta_0 = \int \mathbb{1}\left(\tau \le s \le \sigma, \zeta \le \alpha(Y_1'(s-))\right) P(\mathrm{d}s\mathrm{d}\zeta) \le P^* := P([\tau,\sigma] \times [0,1]),$$

hence

$$\mathbb{E}_{z,z,\ell}\left(e^{\eta\Delta_1}\right) \le \mathbb{E}_{\ell+1}\left(e^{\eta P^*}\right) = \mathbb{E}_{\ell+1}\left(e^{\mu(e^\eta-1)\sigma}\right)$$

after using the strong Markov property for the first inequality and the fact that, under $\mathbb{P}_{\ell+1}$ and conditionally on Y_2, P^* is a Poisson random variable with parameter $\mu\sigma$. Using the same argument as in the previous lemma, namely $\sigma \prec T^1/\lambda^{\mathtt{f}}_{\mathtt{tot}}$, we obtain

$$\mathbb{E}_{z,z,\ell}\left(e^{\eta\Delta_1}\right) \le \mathbb{E}\left(e^{\mu(e^\eta-1)T^1/\lambda^{\mathtt{f}}_{\mathtt{tot}}}\right) \le c := (\varepsilon^{-1}-1)^{1/2},$$

where the last inequality is provided by Proposition 5.4 in [21]. If $\mathbb{E}$ denotes the expectation under the stationary distribution of (Z_k, Z'_k), the strong Markov property then entails

$$\mathbb{E}\left(e^{\eta\Delta_{k+1}} \mid \mathcal{F}_k\right)\mathbb{1}\left(\Delta_k = 0\right) \le c,$$

where $\mathcal{F}_k = \sigma((Y_1(s), Y'_1(s), Y_2(s)), s \le \sigma_k)$.
Second case: $\Delta_0 \ge 1$. Next, consider the case $\Delta_0 \ge 1$. Then

$$\int \mathbb{1}\left(0 \le s \le \tau, \zeta \le \beta_{Y'_1(s-)-Y_1(s-)}(Y_1(s-))\right) P'(\mathrm{d}s\mathrm{d}\zeta)$$

counts the number of points of P' that fall below the curve $\beta_{Y'_1-Y_1}(Y_1)$ before time τ. Each time a point falls below this curve, this makes $Y'_1 - Y_1$ decrease by one, and the β curve lowers until $Y'_1 - Y_1$ possibly hits 0 in which case $\beta_0 = 0$ and no more points can fall below this line. In particular, we have

$$\int \mathbb{1}\left(0 \le s \le \tau, \zeta \le \beta_{Y'_1(s-)-Y_1(s-)}(Y_1(s-))\right) P'(\mathrm{d}s\mathrm{d}\zeta) \ge B,$$

where B is the Bernoulli random variable $B = \mathbb{1}\left(I \ge 1\right)$ with

$$I = \int \mathbb{1}\left(0 \le s \le \tau, \zeta \le \beta_1(Y_1(s-))\right) P'(\mathrm{d}s\mathrm{d}\zeta).$$

Indeed, if $B = 0$, then this inequality is true; if $B = 1$, then this means that $I \ge 1$, i.e., a point of P' fell below the curve $\beta_1(Y_1)$, and since $\beta_1 \le \beta_{Y'_1(0)-Y_1(0)}$, this necessarily implies that a point of P' fell below the curve $\beta_{Y'_1-Y_1}(Y_1)$, i.e., the left-hand side of the previous display is also ≥ 1. We thus derive that $\Delta_1 - \Delta_0 \le P([\tau,\sigma]\times[0,1]) - B$ and by independence between P, P', Y_1 and Y_2, the strong Markov property provides

$$\mathbb{E}_{z,z',\ell}\left(e^{\eta(\Delta_1-\Delta_0)}\right) \le c\mathbb{E}_{z,\ell}(e^{-\eta B}).$$

Averaging with respect to (Z_∞, Z'_∞) and using the strong Markov property, we obtain

$$\mathbb{E}\left(e^{\eta(\Delta_{k+1}-\Delta_k)} \mid \mathcal{F}_k\right)\mathbb{1}\left(\Delta_k \ge 1\right) \le c\,\mathbb{E}_\ell(e^{-\eta B}),$$

where here and in the rest of the proof, $\mathbb{E}_\ell$ corresponds to an initial state of (Y_1, Y_2) distributed as $(Y_1(\infty), \ell)$. Assume at this stage that

$$\mathbb{E}_\ell(e^{-\eta B}) \to 0 \tag{3.4}$$

(we will prove this claim at the end of the proof). Then Theorem 2.3 in [11] gives, for large enough $\lambda^{\mathtt{f}}_{\mathtt{tot}}$,

$$\mathbb{P}(\Delta_\infty \geq d) \leq \frac{c\, e^{\eta}}{1 - c\, \mathbb{E}_\ell(e^{-\eta\, B})} e^{-\eta d},$$

from which we get

$$\mathbb{E}(\Delta_\infty) = \sum_{d \geq 1} \mathbb{P}(\Delta_\infty \geq d) \leq \frac{c}{1 - c\, \mathbb{E}_\ell(e^{-\eta\, B})} \frac{1}{1 - e^{-\eta}}.$$

Since $e^\eta \to \infty$ with $\lambda^{\mathtt{f}}_{\mathtt{tot}}$, we obtain the desired result.

In order to conclude the proof, we now prove the claim (3.4). Since B is a Bernoulli random variable, we have $\mathbb{E}_\ell(e^{-\eta B}) = e^{-\eta}\mathbb{P}(B = 1) + \mathbb{P}(B = 0)$ and as $\eta \to \infty$, we only have to show that $\mathbb{P}(B = 0) \to 0$. Let $Y_1^* = \sup_{[0,\ell^2]} Y_1$. Since β_1 is decreasing, when $\tau \leq \ell^4$ and $Y_1^* \leq \ell^2$, we have

$$\begin{aligned}
&\int \mathbb{1}\left(0 \leq s \leq \tau, \zeta \leq \beta_1(Y_1(s-))\right) P'(\mathrm{d}s\mathrm{d}\zeta) \\
&\quad \leq I^* := \int \mathbb{1}\left(0 \leq s \leq \ell^4, \zeta \leq \beta_1(\ell^2)\right) P'(\mathrm{d}s\mathrm{d}\zeta),
\end{aligned}$$

hence

$$\begin{aligned}
\mathbb{P}(B = 0) \leq \mathbb{P}(I^* = 0) + \mathbb{P}_\ell(\tau \leq \ell^4) + \mathbb{P}(Y_1^* \geq \ell^2) \leq \mathbb{P}(I^* = 0) + \mathbb{P}_\ell(\tau \leq \ell^4) \\
+ \mathbb{P}(Y_1^* \geq \ell^2 \mid Y_1(0) \leq \ell^{3/2}) + \mathbb{P}(Y_1(\infty) \geq \ell^{3/2}),
\end{aligned}$$

where $Y_1(0)$ is distributed according to $Y_1(\infty)$. Since I^* is a Poisson random variable with parameter $\mu\beta_1(\ell^2)\ell^4$, we have

$$\mathbb{P}(I^* = 0) = \exp\left(-\mu\ell^4\beta_1(\ell^2)\right),$$

which vanishes as $\lambda^{\mathtt{f}}_{\mathtt{tot}} \to \infty$ since $\beta_1(\ell^2)$ decays like $1/\ell^3$. Moreover, by proceeding as in Sect. 3.4 and comparing Y_1 with a subcritical $M/M/1$, it is easy to see that $\mathbb{P}(Y_1(\infty) \geq \ell^{3/2}) \to 0$. It thus remains to control the two last terms $\mathbb{P}_\ell(\tau \leq \ell^4)$ and $\mathbb{P}(Y_1^* \geq \ell^2 \mid Y_1(0) \leq \ell^{3/2})$, which can be done by comparison with a subcritical $M/M/1$ queue. In fact, it is well known that it takes an exponential time for a subcritical $M/M/1$ queue to reach high values (see for instance Proposition 5.11 in [21]) and we can compare Y_1 and Y_2 to such a queue to transfer this behavior to these two processes:

- for Y_1, we can use the fact that, when in the range $[\ell^{3/2}, \ell^2]$, it is smaller than a subcritical queue $M/M/1$ queue with input rate λ_1 and output rate $\mu\alpha(\ell^{3/2})$;
- for Y_2, we can use the fact that, when in the range $[\ell', \ell]$ with lower bound $\ell' = (\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta + \ell)/2$, it is smaller than a subcritical $M/M/1$ queue with input rate $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ and output rate $\theta\ell'$.

The proof is thus complete. □

4 Proof of Lemma 2.8

Fix an integer $k \geq \mu/\theta$, let $S = \{-k, -k+1, \dots, \}$ and Z be the S-valued Markov process with non-zero transition rates

$$z \in S \longrightarrow \begin{cases} z+1 & \text{at rate } \lambda^{\mathtt{f}}_{\mathtt{tot}}, \\ z-1 & \text{at rate } \theta(k+z). \end{cases}$$

Compared to the transition rates of $X^{\mathtt{f}}_2$, this amounts to upper bounding the rate $\mu x_2/(x_1 + x_2)$ by θk, which makes X_2 smaller. Note also that the downward rate $\theta(k+z)$ is 0 for $z = -k$, so Z indeed lives in S. This implies $Z \prec X^{\mathtt{f}}_2$ and therefore

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}) \leq \mathbb{P}(Z(\infty) = 0).$$

From its transition rates, it is apparent that $Z + k$ is an $M/M/\infty$ queue with input rate $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ and service rate θ, and so $Z(\infty) - k$ follows a Poisson distribution with parameter $\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$. In particular,

$$\mathbb{P}(Z(\infty) = 0) = e^{-\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta} \frac{(\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta)^k}{k!}$$

and so $-\log \mathbb{P}(Z(\infty) = 0)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \to \theta^{-1}$, which gives the desired bound.

5 Possible extensions

In this paper, we aimed to prove the minimal result that shows that mobility makes delay increase like $\log(1/(1-\varrho))$ in heavy traffic, instead of the usual $1/(1-\varrho)$ scaling. However, we can go a bit further than Theorem 2.5 by formulating an interesting conjecture, which we can only partially prove. In this section, we will also discuss the link with large deviations theory, and possible extensions of our model.

5.1 Extension of Theorem 2.5

As explained earlier, Theorem 2.5 is a direct consequence of Lemmas 2.7 and 2.8: informally, these two lemmas state, respectively, that $\mathbf{X}(\infty) \leq \theta^{-1}\xi^*\Lambda_{\mathtt{tot}}$ and

$\Lambda_{\mathtt{tot}} \leq \theta \log(1/(1-\varrho))$, where these inequalities are to be understood in an asymptotic sense. These two results can actually be strengthened.

Concerning the first inequality $\mathbf{X}(\infty) \leq \theta^{-1}\xi^*\Lambda_{\mathtt{tot}}$ of Lemma 2.7, we can prove with similar tools as those used in Sect. 3 that $\mathbf{X}^{\mathtt{f}}(\infty)/\Lambda_{\mathtt{tot}} \Rightarrow_{\varrho} \theta^{-1}\xi^*$. The idea is to prove a matching lower bound to that already proved, by comparing $\mathbf{X}^{\mathtt{f}}$ to a lower bounding process $\mathbf{Y}'$ with non-zero transition rates

$$\mathbf{y} \in \mathbb{N}^2 \longrightarrow \begin{cases} \mathbf{y}+\mathbf{e}_1 & \text{at rate } \lambda_1, \\ \mathbf{y}+\mathbf{e}_2 & \text{at rate } \lambda^{\mathtt{f}}_{\mathtt{tot}}, \\ \mathbf{y}-\mathbf{e}_1 & \text{at rate } \mu\,\dfrac{y_1}{y_1+\ell}\cdot \mathbb{1}\,(y_2 \geq \ell) + \mu \cdot \mathbb{1}\,(y_2 < \ell), \\ \mathbf{y}-\mathbf{e}_2 & \text{at rate } \mu + \theta y_2. \end{cases}$$

We then have $\mathbf{Y}' \prec \mathbf{X}^{\mathtt{f}}$ and the analysis of $\mathbf{Y}'$ proceeds as in Sects. 3.4 and 3.5 and leads to $\mathbf{Y}'(\infty)/\lambda^{\mathtt{f}}_{\mathtt{tot}} \Rightarrow_{\lambda^{\mathtt{f}}_{\mathtt{tot}},\varepsilon} \theta^{-1}\boldsymbol{\xi}^*$. We have here to choose $\ell = (1-\varepsilon)\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$, so that excursions of Y_2 below level ℓ are rare, and thus as for the lower bound, Y'_1 essentially behaves as a birth-and-death process independent of Y'_2.

The second inequality $\Lambda_{\mathtt{tot}} \leq \theta \log(1/(1-\varrho))$ of Lemma 2.8 can also be strengthened, but this is actually much more difficult. More precisely, we can show that

$$\liminf_{\varrho \uparrow 1} \frac{\Lambda_{\mathtt{tot}}}{\log(1/(1-\varrho))} > 0, \tag{5.1}$$

which, using the convergence $\mathbf{X}^{\mathtt{f}}(\infty)/\Lambda_{\mathtt{tot}} \Rightarrow_{\varrho} \theta^{-1}\xi^*$ discussed above, would show that

$$\liminf \frac{\mathbf{X}(\infty)}{\log(1/(1-\varrho))} > 0.$$

Together with the bound

$$\limsup \frac{\mathbf{X}(\infty)}{\log(1/(1-\varrho))} < \infty$$

of Theorem 2.5, this implies that $\log(1/(1-\varrho))$ is indeed the right order of magnitude of $\mathbf{X}(\infty)$.

In order to prove (5.1), since we have

$$\log(1/(1-\varrho)) = -\log \mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty)=0) \geq -\log \mathbb{P}(X^{\mathtt{f}}_1(\infty)=0),$$

we see that it is enough to control $-\log \mathbb{P}(X^{\mathtt{f}}_1(\infty)=0)/\lambda^{\mathtt{f}}_{\mathtt{tot}}$. Using subtle and rather heavy perturbation arguments that we do not detail, we can actually prove the expected result that $\mathbb{P}(X^{\mathtt{f}}_1(\infty)=0)$ has the same exponential order as the corresponding birth-and-death process with death rate $\mu x/(x+\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta)$, obtained by replacing x_2 by its equilibrium value $\lambda^{\mathtt{f}}_{\mathtt{tot}}/\theta$, i.e., that

$$-\frac{1}{\lambda^{\mathtt{f}}_{\mathtt{tot}}} \log \mathbb{P}(X^{\mathtt{f}}_1(\infty) = 0) \to -\theta \log(1 - \varrho_1),$$

which implies the more precise result that $\Lambda_{\mathtt{tot}}/\log(1/(1-\varrho)) \to -\theta \log(1-\varrho_1)$.

We finally conclude this part with a more precise conjecture. Because of time-scale separation arguments explained in Sect. 5.2, we believe that $X^{\mathtt{f}}_1(\infty)$ and $X^{\mathtt{f}}_2(\infty)$ are asymptotically independent and that

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}) \approx \mathbb{P}(X^{\mathtt{f}}_1(\infty) = 0) \times \mathbb{P}(X^{\mathtt{f}}_2(\infty) = 0), \tag{5.2}$$

where the approximation is thought to hold in the logarithmic order. Actually, thanks to perturbation analysis, we know how to prove that

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = \mathbf{0}) \geq \mathbb{P}(X^{\mathtt{f}}_1(\infty) = 0) \times \mathbb{P}(X^{\mathtt{f}}_2(\infty) = 0)$$

and we would need an asymptotically matching upper bound. As argued above, we have $-\log \mathbb{P}(X^{\mathtt{f}}_1(\infty) = 0) \sim -\theta \log(1-\varrho_1)\lambda^{\mathtt{f}}_{\mathtt{tot}}$, while it is easy to show by suitable comparison with an $M/M/\infty$ queue that $-\log \mathbb{P}(X^{\mathtt{f}}_2(\infty) = 0) \sim -\theta\lambda^{\mathtt{f}}_{\mathtt{tot}}$. These various arguments thus lead to the following conjecture.

Conjecture 5.1 *If $\theta > 0$, then*

$$\frac{\mathbf{X}(\infty)}{\log(1/(1-\varrho))} \Longrightarrow \frac{1}{1 - \log(1-\varrho_1)} \cdot \boldsymbol{\xi}^*,$$

as $\varrho \uparrow 1$, with the point $\boldsymbol{\xi}^*$ defined in Theorem 2.5.

If this statement were true, it would have the surprising feature that the heavy traffic limit is independent of the parameter θ: all that matters is that $\theta > 0$, but the precise value is irrelevant in heavy traffic. Moreover, this would give the approximation

$$X_1(\infty) + X_2(\infty) \approx M(\varrho_2) \cdot \log\left(\frac{1}{1-\varrho}\right)$$

for the total number of users with $\varrho_2 \approx 1 - \varrho_1$ and where $M(x) = 1/(x - x\log x)$. As the function $x \in [0,1] \mapsto x - x\log x$ is increasing, this would suggest that for a given load ϱ, the system performance is improved with a larger fraction $1 - \varrho_1$ of mobile users.

5.2 Large deviations for processes undergoing time-scale separation

In order to prove the above conjecture, what we miss is formalizing the approximation (5.2). There is a vast literature on large deviations for Markov processes; we did not find, however, any reference that fits our framework.

What is specific in our problem of controlling the stationary probability in state $\mathbf{0} = (0,0)$ of $\mathbf{X}^{\mathtt{f}}$ is that the two components $X^{\mathtt{f}}_1$ and $X^{\mathtt{f}}_2$ evolve on different time-scales. When $\lambda^{\mathtt{f}}_{\mathtt{tot}}$ is large, Lemma 2.7 shows that $\mathbf{X}^{\mathtt{f}}(\infty)$ is of the order of $\lambda^{\mathtt{f}}_{\mathtt{tot}}$. But

$X_1^{\mathtt{f}}$ is similar to a birth-and-death process with bounded birth-and-death rates, which makes it evolve on the linear time-scale proportional to $\lambda_{\mathtt{tot}}^{\mathtt{f}}$, while $X_2^{\mathtt{f}}$ is similar to an $M/M/\infty$ queue and thus evolves on a constant time-scale. The process $\mathbf{X}^{\mathtt{f}}$ therefore undergoes time-scale separation, or stochastic homogenization: when there are two components with different speeds, the stochastic homogenization principle asserts that the slow one (namely, $X_1^{\mathtt{f}}$) only interacts with the fast one (namely, $X_2^{\mathtt{f}}$) through its stationary distribution. Here the stationary distribution of $X_2^{\mathtt{f}}$ is essentially a Poisson random variable with parameter $\lambda_{\mathtt{tot}}^{\mathtt{f}}/\theta$ and is thus independent of $X_1^{\mathtt{f}}$, which leads to a simpler form of stochastic homogenization.

This stochastic averaging principle is well known, and there is a rich literature on large deviations theory in this case; see for instance [9,12,16,19,26,27]. However, all these works only establish large deviations principles for the empirical measure of the fast process, which is admittedly the most natural question to address. What we presently need, however, is really the probability for the fast process to be exactly in 0 as well.

Beside functional large deviations principles, the analytic singular perturbation theory can provide an alternative approach to derive sharp asymptotics of the distribution of $\mathbf{X}^{\mathtt{f}}(\infty)$. This theory has been applied, in particular, in [28] to obtain asymptotics of the solutions of backward or forward Kolmogorov equations for jump processes with two time-scales; coupled queueing systems ([13,14], [22] - Chap. 9) have been also addressed in this framework. Specifically, an asymptotic expansion for the whole distribution of $\mathbf{X}^{\mathtt{f}}(\infty)$ on $\mathbb{N}^2$ of the form

$$\mathbb{P}(\mathbf{X}^{\mathtt{f}}(\infty) = A\boldsymbol{\xi}) = \frac{1}{A}\exp\left[-A\cdot H(\boldsymbol{\xi}) - h_0(\boldsymbol{\xi}) + O\left(\frac{1}{A}\right)\right], \quad \boldsymbol{\xi} = (x, y) \in \mathbb{R}_+^2,$$

is assumed to exist with large (a-dimensional) scaling parameter $A = \lambda_{\mathtt{net}}^{\mathtt{f}}/\theta$, and where real functions H, h_0 on $\mathbb{R}_+^2$ satisfy $H(\boldsymbol{\xi}^*) = 0$ (with the point $\boldsymbol{\xi}^*$ as in Theorem 2.5) together with smoothness properties. At the present stage, with no claim to formally justify the existence of such an expansion, these singular perturbation methods enable one to determine the functions H and h_0 explicitly, giving, in particular,

$$H(\boldsymbol{\xi}) = \Phi(x) + \Psi(y), \qquad \boldsymbol{\xi} = (x, y),$$

for simple functions Φ and Ψ; the latter relation thus provides another argument for the asymptotic independence (at logarithmic order) of the components of $\mathbf{X}^{\mathtt{f}}(\infty)$ discussed above. This singular perturbation framework for the estimation of the whole distribution of the vector $\mathbf{X}^{\mathtt{f}}(\infty)$ and its justification is the object of current investigations [25].

5.3 Model generalization

In this paper, our goal was to initiate the analysis of a new class of stochastic models for mobile networks. The general idea of these models is to forget about keeping track of all users, but instead to focus on a subset of the whole network and take into account

the rest of the network through a balance equation. Here, we focused as a first step on a single cell of equilibrium but more general situations can be considered.

Specifically, in the case of a single cell, we could for instance consider an "imbalance" parameter $\beta > 0$ and consider the balance equation

$$\lambda^{\mathtt{f}}_{\mathtt{net}} = \beta\theta \cdot \mathbb{E}\left(X^{\mathtt{f}}_2(\infty)\right)$$

instead of (FP). This new fixed-point equation would mean that the ratio of flows from and to the rest of the network is equal to β. Thus, for $\beta > 1$ this would amount to considering a cell where more users enter than exit, and the opposite for $\beta < 1$. Of course, such an imbalance could not be sustained for the whole network but could hold locally. Studying what happens to the constrained model when enforcing this equation instead of (FP) constitutes an interesting research direction.

Finally, another way to generalize the model would be to consider several cells instead of only one. In this case, there are different flows from and to the rest of the network, as well as within the considered cells. The first difficulty to solve would be to find a relevant balance equation generalizing (FP), which would probably be multidimensional. For instance, if one considers n cells, there would now be potentially $2n+n+n(n-1)/2+n$ parameters: one arrival rate per class and per cell, one capacity per cell, a mobility rate between each pair of cells and a mobility rate from each cell to the rest of the network.

References

1. Anton, E., Ayesta, U., Simatos, F.: On the impact of mobility in cellular networks. In: WiOpt 19: Modeling and Optimization in Mobile, Ad-Hoc and Wireless Networks (2019)
2. Baynat, B., Indre, R.-M., Nya, N., Olivier, P., Simonian, A.: Impact of mobility in dense LTE-A networks with small cells. In: IEEE 81st Vehicular Technology Conference (VTC Spring), pp. 1–5 (2015)
3. Bonald, T., Borst, S.C., Proutiere, A: How mobility impacts the flow-level performance of wireless data systems. In: Proc. INFOCOM '04, vol. 3, pp. 1872–1881 (2004)
4. Bonald, T., Proutiere, A.: Wireless downlink data channels: user performance and cell dimensioning. MobilCom'03, pp. 339–352 (2003)
5. Bonald, T., Borst, S., Hegde, N., Jonckheere, M., Proutiere, A.: Flow-level performance and capacity of wireless networks with user mobility. Queueing Syst. **63**(1–4), 131–164 (2009)
6. Borst, S.C., Hegde, N., Proutiere, A.: Mobility-driven scheduling in wireless networks. In: Proc. IEEE INFOCOM '09, pp. 1260–1268 (2009)
7. Borst, S., Proutiere, A., Hegde, N.: Capacity of wireless data networks with intra- and inter-cell mobility. In: Proc. IEEE INFOCOM '06, pp. 58–1069 (006)
8. Borst, S., Simatos, F.: A stochastic network with mobile users in heavy traffic. Queueing Syst. **74**(1), 1–40 (2013)
9. Freidlin, M.I.: The averaging principle and theorems on large deviations. Russ. Math. Surv. **33**(5), 117–176 (1978)
10. Grossglauser, M., Tse, D. Mobility increases the capacity of ad-hoc wireless networks. In:Proc. IEEE INFOCOM '01, vol. 3, pp. 60–1369 (2001)
11. Hajek, B.: Hitting-time and occupation-time bounds implied by drift analysis with applications. Adv. Appl. Probab. **14**(3), 502–525 (1982)
12. Huang, G., Mandjes, M., Spreij, P.: Large deviations for Markov-modulated diffusion processes with rapid switching. Stoch. Process. Appl. **126**(6), 1785–1818 (2016)

13. Knessl, C., Matkowsky, B.J., Schuss, Z., Tier, C.: On the performance of state-dependent single server queues. SIAM J. Appl. Math. **46**(4), 657–697 (1986)
14. Knessl, C., Tier, C.: Applications of singular perturbation methods in queueing, In: Advances in Queueing Theory, Methods and Open Problems, Probability and Stochastics Series, pp. 311–336. CRC Press (1995)
15. Lin, M., Wierman, A., Zwart, B.: The average response time in a heavy-traffic SRPT queue. SIGMETRICS Perform. Eval. Rev. **38**(2), 12–14 (2010)
16. Liptser, Robert: Large deviations for two scaled diffusions. Probab. Theory Relat. Fields **106**(1), 71–104 (1996)
17. Ma, H., Zhao, D., Yuan, P.: Opportunities in mobile crowd sensing. IEEE Commun. Mag. **52**(8), 29–35 (2014)
18. Olivier, P., Simonian, A., Simatos, F.: Performance analysis ofdata traffic in small cells networks with user mobility. In: Puliafio, A., Trivedi, K.S. (eds.) Systems Modeling: Methodologies and Tools, EAI/Springer Innovations in Communicationand Computing, pp. 177–193. Springer, Cham (2019)
19. Puhalskii, A.A.: On large deviations of coupled diffusions with time scale separation. Ann. Probab. **44**(4), 3111–3186 (2016)
20. Rege, K.M., Sengupta, B.: Queue-length distribution for the discriminatory Processor-Sharing queue. Oper. Res. **44**(4), 653–657 (1996)
21. Robert, P.: Stochastic networks and queues. Stochastic modelling and applied probability series, p. xvii+398. Springer, New York (2003)
22. Schuss, Z.: Theory and Applications of Stochastic Processes, An analytical Approach, Applied Mathematical Sciences Series, vol. 170. Springer (2010)
23. Simatos, F., Tibi, D.: Spatial homogenization in a stochastic network with mobility. Ann. Appl. Probab. **20**(1), 312–355 (2010)
24. Simonian, A., Olivier, P.: Performance of data traffic in small cells networks with inter-cell mobility. In: Proceedings of 10th EAI International Conference on Performance Evaluation Methodologies and Tools. ACM (2017). https://doi.org/10.4108/eai.25-10-2016.2266520
25. Simatos, F., Simonian, A.: Heavy load analysis of the multi-class Processor-Sharing queue with impatience, In preparation
26. Veretennikov, AYu.: On large deviations in the averaging principle for SDEs with a "full dependence". Ann. Probab. **27**(1), 284–296 (1999)
27. Veretennikov, AYu.: On large deviations in the averaging principle for SDE's with a "full dependence", revisited. Discrete Contin. Dyn. Syst. Ser. B **18**(2), 523–549 (2013)
28. Yin, G.G., Zhang, A.: Continuous-Time Markov Chains and Applications. A two-time scale approach. Stochastic Modelling and Applied Probability Series. Springer, New York (2013)