Control of Many-Server Queueing Systems in Heavy Traffic

Gennady Shaikhet

Control of Many-Server Queueing Systems in Heavy Traffic

Research Thesis
In Partial Fulfillment of the
Requirements for the
Degree of Doctor of Philosophy

Gennady Shaikhet

Submitted to the Senate of
the Technion - Israel Institute of Technology
Nissan 5767 Haifa April 2007

This research thesis was conducted in the Industrial Engineering Faculty in the Technion, under the supervision of Professor Rami Atar from the Electrical Engineering Faculty and Professor Avishai Mandelbaum from the Industrial Engineering Faculty.

Looking back at the whole doctorate period, I see four years of very interesting experience, which was a real joy for me. For that I am deeply grateful to my advisors Rami and Avishai.

I would also like to thank the beautiful people in the Industrial Engineering Faculty, professors and fellow students, for their kindness and support. I was very lucky to meet you on my way!

As always, my love to my family!

Finally, the Generous Financial Help of Technion and Rami Atar and Avishai Mandelbaum is Gratefully Acknowledged.

Contents

List of Figures

Abstract

Consider a queueing model with $I \geq 1$ customer classes and $J \geq 1$ service stations, each consisting of many independent servers with identical capabilities. Customers of different classes can be served at these stations at different rates, that depend on both the class and the station. Service times are exponential while arrival processes are renewal. A system administrator dynamically controls scheduling and routing. The model is studied in the Halfin–Whitt heavy traffic regime: one considers a sequence of models parameterized by $n \in \mathbb{N}$ where the arrival rates and the number of servers are scaled up in such a way that the processes representing the number of class-i customers in the system, $i \in \mathcal{I}$, exhibit diffusive fluctuations about the static fluid model. The static fluid model represents a Law of Large Numbers behavior of the system and is assumed to be critically loaded: the total processing rate devoted to each class's 'material' is equal to its arrival rate and an increase in any of the external arrival rates results with an overloaded system.

We find a new, unusual heavy traffic 'behavior' of a system: under appropriate assumptions there exists a dynamic control policy that maintains a critically loaded system as if it were underloaded. The effect is studied in chapters 2 and 3.

In chapter 2 the above phenomenon is shown to be related to a formulation of the limiting diffusion model as a controlled diffusion with a singular control term. The singular term may be used to constrain the diffusion to lie in certain subsets of $\mathbb{R}^I$ at all times $t > 0$. We say that the diffusion is *null-controllable* if it can be constrained to $\mathbb{X}_-$, the minimal closed subset of $\mathbb{R}^I$ containing all states of the prelimit queueing model for which all queues are empty. We give sufficient conditions for null controllability of the diffusion in terms of the graph that encodes the network's structure. Under these conditions we also show that an analogous, asymptotic result holds for the queueing model, by constructing control policies under which, for any given $0 < \varepsilon < T < \infty$, all queues in the system are kept empty on the time interval $[\varepsilon, T]$, with probability approaching one.

In chapter 3 we introduce and analyze the notion of *throughput sub-optimality* of the underlying static fluid model. Roughly, this means that the servers can be allocated so as to achieve a total processing rate that is greater than the total arrival rate, while, for every $i \in \mathcal{I}$, the mass of servers allocated to serve class i does not exceed the mass of class-i material. Assuming throughput sub-optimality, (which is shown to be a weaker assumption than that of chapter 2), we introduce a dynamic control policy under which, for every finite T, the

measure of the set of times prior to T, at which at least one customer is in the buffer, converges to zero in probability at the scaling limit. The results of chapter 2 allow for both preemptive policies (where service to a customer can be interrupted and resumed at a later time, possibly at a different server) and nonpreemptive ones (where service cannot be interrupted), while chapter 3 only treats preemptive policies. The notion of throughput sub–optimality complements and explains the null controllability.

Chapter 4 deals with limiting diffusion models of queueing systems and their formulation as the controlled diffusions. By imposing conditions on the service rates, significant simplifications of the controlled diffusion model arise, and, in particular, the model process lives in one dimension. We then indicate particular cases when an exact solution is available, and describe how to construct control schemes for the originating queueing model that are conjectured to be asymptotically optimal.

Notations

e_i	A unit coordinate vector.
e	$= (1, 1, ..., 1)$. The dimension of vector e may change from one expression to another.
$v \cdot u$	A scalar product of two vectors v, u of the same dimension.
x_e	$= e \cdot x = x_1 + ... + x_I$ for $x \in \mathbb{R}^I$.
x^+	$= \max\{x, 0\}$, $x \in \mathbb{R}$.
x^-	$= \max\{-x, 0\}$, $x \in \mathbb{R}$.
$\|x\|$	$= \sum_{i=1}^I \|x_i\|$. A norm of $x \in \mathbb{R}^I$.
$\|X\|_t^*$	$= \sup_{0 \le u \le t} \|X(u)\|$. A norm of an $\mathbb{R}^I$-valued function.
$\mathbb{P}$	Probability measure.
$\sigma\{\mathcal{A}\}$	A sigma-field generated by a collection $\mathcal{A}$ of random variables.
$Y_i^n(t)$	= number of class i customers in queue at time t.
$Z_j^n(t)$	= number of idle servers in station j at time t.
$\Psi_{ij}^n(t)$	= number of class i customers being served in station j at time t.
$X_i^n(t)$	= number of class i customers in the system at time t.
$\bar{X}^n(t)$	$= X^n$ in fluid scaling.
$\hat{X}^n(t)$	$= X^n$ in diffusion scaling.
$\mathcal{I}$	$= \{1, 2, ..., I\}$. The set of vertices, corresponding to customer classes.
$\mathcal{J}$	$= \{1, 2, ..., J\}$. The set of vertices, corresponding to server stations.

$\mathcal{V}$ $= \mathcal{I} \cup \mathcal{J}$.

$\mathcal{E}$ $= \{(i,j) \in \mathcal{I} \times \mathcal{J}\}$.

$\mathcal{E}_a$ $= \{(i,j) \in \mathcal{E} \ : \ (i,j)$ is activity$\}$. See Section 2.2.1.

$\mathcal{E}_{ba}$ $= \{(i,j) \in \mathcal{E}_a \ : \ (i,j)$ is basic activity$\}$. See Section 2.2.1.

$\mathcal{E}_{nb}$ $= \{(i,j) \in \mathcal{E}_a \ : \ (i,j)$ is non–basic activity$\} = \mathcal{E}_a \setminus \mathcal{E}_{ba}$.

$\mathcal{G}_a$ $= (\mathcal{V}, \mathcal{E}_a)$. The graph of activities.

$\mathcal{G}_{ba}$ $= (\mathcal{V}, \mathcal{E}_{ba})$. The graph of basic activities.

Chapter 1

Background

1.1. Queues in heavy traffic

1.1.1. Conventional heavy traffic

The vast majority of works on approximation of queueing networks has dealt with *conventional heavy traffic* asymptotics. Consider a sequence of $M/M/N$ models indexed by n. The number of servers $N^n \equiv N$ is kept fixed, but the workload is increased by assuming $\mu^n \to \mu$, $\lambda^n \to N\mu$ and $\rho^n = \lambda^n/(N\mu^n) \to 1$ at the rate of $\sqrt{n}(1-\rho^n) \to \beta, \quad 0 < \beta < \infty$. The scaled queue length corresponds to accelerating time by n and scaling down space coordinates by $\sqrt{n}$ as

$$\hat{Q}^n(t) = \frac{Q^n(nt)}{\sqrt{n}}, \quad t \in [0, \infty).$$

In the sequel we use $\Rightarrow$ to denote weak convergence (see the last paragraph of Section 2.1 for the definition).

Theorem 1.1. *(Iglehart and Whitt, 1970 [39]). Assume $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$. Then $\hat{Q}^n \Rightarrow \hat{Q}$ in $D([0,\infty))$, where $\hat{Q}$ is a reflecting Brownian motion with drift, that is*

$$\hat{Q}(t) = \hat{Q}(0) - \beta N\mu t + \sqrt{2N\mu}\, W(t) + L(t).$$

Here $W(t)$ is a standard Brownian motion, and L is the local time for $\hat{Q}$ at the origin.

We do not attempt to cover all the papers on conventional approximation. For a full account, readers are referred to Whitt [56] and Chen and Yao [17]. The heavy traffic analysis for a single station queueing system was initiated by Kingman [43] in the early 1960's. One should mention Whitt [55] and Iglehart

[39]. Mandelbaum and Pats [48] generalized the M/M/1 treatment for the case of state dependent queues. For multiclass networks, refer to Reiman [51], Harrison [31], Harrison and Williams [36], Chen and Mandelbaum [18]. The stochastic processes that arise there as diffusion approximations are closely related to, or are themselves multidimensional diffusions of the type called reflected Brownian motions (RBM's). For open networks the state space is the nonnegative orthant (see Harrison and Reiman [34], Williams [60], Chen and Mandelbaum [19] on the subject). See also Harrison, Williams and Chen [37] for a treatment of closed queueing networks, i.e, systems where the number of circulating customers is constant in time, with the diffusion approximation resulting in RBM's on the nonnegative simplex.

It is known however that, for some multiclass queueing networks, heavy traffic limit results do not hold (see Dai and Wang [23]). The question then arises on how to determine whether this is true for a particular network. See Williams [61], Bramson [15], where they establish a framework of proving heavy traffic limit theorems for multiclass networks under a variety of queueing disciplines.

1.1.2. Queues in the Halfin–Whitt heavy traffic regime

Theorem 1.1 above represents, in the terminology of Gans, Koole and Mandelbaum [26], an *efficiency* driven operational regime, in the sense that all resources are working extremely close to full capacity.

The Halfin–Whitt regime [30] (also known as the QED regime), in contrast, exhibits a regime where the operation is both Efficiency–Driven (high servers' utilization) and Quality–Driven (high service levels). Consider the M/M/N model consisting of a Poisson arrival process with rate λ, and N independent, statistically identical servers, each offering an exponential service at rate μ. Let $X(t)$ denote the number of customers in the system at time t. The heavy traffic regime, proposed by Halfin and Whitt, arises when taking the number of servers to infinity in a specific manner. Namely, consider a sequence of systems indexed by N, the number of servers. Denote $\rho^N = \frac{\lambda^N}{N\mu^N}$. Take $N \to \infty$ and assume that $\mu^N \to \mu$ and $\frac{\lambda^N}{N} \to \mu$, $0 < \mu < \infty$, at the rate of

$$\sqrt{N}(1 - \rho^N) \to \beta, \quad 0 < \beta < \infty. \tag{1.1}$$

Define the centered, normalized queue length as

$$\hat{X}^N(t) = \frac{X^N(t) - N}{\sqrt{N}}, \quad t \in [0,\ \infty) \tag{1.2}$$

Theorem 1.2. *(Halfin and Whitt, 1981 [30]).*

1. *Assume* $\hat{X}^N(0) \Rightarrow \hat{X}(0)$. *Then* $\hat{X}^N \Rightarrow \hat{X}$ *in* $D([0,\infty))$, *where* $\hat{X}$ *is a diffusion characterized by*

$$\hat{X}(t) = \hat{X}(0) + \int_0^t b(\hat{X}(s))ds + \sqrt{2\mu}\, W(t).$$

Here $W(t)$ *is a standard Brownian motion, and the drift* b *is given as*

$$b(x) = \begin{cases} -\mu\beta & x \geq 0, \\ \\ -\mu(x+\beta) & x < 0. \end{cases}$$

2. *The probability of delay has a nondegenerate limit as follows:*

$$\lim_{N\to\infty} P(X^N(\infty) \geq N) = \alpha, \quad 0 < \alpha < 1, \tag{1.3}$$

where α *is related to* β *from* (1.1) *via* $\alpha = [1 + \beta\Phi(\beta)/\phi(\beta)]^{-1}$. *(Here* Φ *and* ϕ *are, respectively, the distribution and density functions of the standard normal distribution).*

The convergence in (1.1) is in fact necessary and sufficient for the convergence (1.3). Note, however, that the result in Statement 1 remains true without assuming strict positiveness of β in (1.1).

For more information and motivation regarding the QED regime, see [26]. Due to the desirable features of the Halfin–Whitt regime and because it turns out to be a good model for large telephone call centers, it has enjoyed recently considerable attention in the literature. For recent generalizations of [30] see Whitt [58]–[59] and the references therein. Convergence of the scaled queueing process, but for the case when service times have phase-type distribution is discussed in Puhalskii and Reiman [50]. The papers of Jelenkovic, Mandelbaum and Momcilovic [40] and Mandelbaum and Momcilovic [47] study the limiting distribution of appropriately scaled virtual waiting time in queues where the service duration is either deterministic [40], or is a random variable with finite support [47]. Mandelbaum, Massey and Reiman [46] develop limit theorems for queueing models, where the arrival rates, the number of servers and the service rate of each server can depend on time and state. Garnett, Mandelbaum and Reiman [27] extended the results of [30] for queues with possible abandonments. Papers dealing with control aspects are discussed in the next section.

1.2. Scheduling and routing of queueing networks

Optimal scheduling and routing are among the most interesting and difficult challenges in the management of queueing networks. The routing problem is to determine, upon an arrival, *which of the available servers, if any, should we assign to serve a customer*. The scheduling problem is to indicate, upon service completion, *which of the available waiting customers, if any, should be served*.

The earliest control work for single server queues was the exact analysis by Cox and Smith [20]. They looked at a multi–class single–station network (M/G/1) with linear waiting cost, i.e. one pays $c_i\tau$ units for each job of class i that waits for service τ units of time. This is equivalent to looking at $\int_0^t \sum_i c_i Q_i(s)ds$ - the integral over a linear combination of the queue lengths. They proved the classical $c\mu$ rule, which can be described as follows. With each class of jobs we associate an index $c_i\mu_i$ (with μ_i being its service rate) and at a decision point one always serves the highest index. See Walrand [54] for various extensions.

A similar setting was considered in the conventional heavy traffic asymptotic regime by Van Mieghem [53], with his *generalized $c\mu$ rule*. This culminated in the work of Mandelbaum and Stolyar [49]. They treat the parallel server models (J nonidentical servers working in parallel and I customer classes) and convex cost functions $C_i(\cdot)$, $i = 1, ..., I$. Optimal scheduling corresponds to the following: at each time t, when server j becomes idle, it chooses for a service a type i customer with the largest $C_i'(Q_i(t))\mu_{ij}$. Note however that the cost functions $C(\cdot)$ in [49] were restricted to convex functions with $C(0) = 0$ and $C'(0) = 0$. This excludes a direct application to linear costs.

For linear delay costs, one is referred to Williams [62] and to Harrison [32] and Harrison and Lopez [33]. In [33] it is proved that for the parallel server models, the diffusion control problem exhibits a massive state-space collapse and is reduced from multi-dimension to one-dimension, which is much easier to solve. This is done by the striking "equivalent workload formulation". See also Harrison and Van Meighem [35].

The difficulty in [35] is that the asymptotic solution does not have a clear interpretation within the prelimit model. Bell and Williams [11] proved that for the 2-parallel servers model the asymptotically optimal policy is a threshold policy; i.e., the priority of service depends on whether the queue lengths are below or above certain levels - thresholds. The subsequent paper [12] of the same authors deals with extending the threshold strategy to parallel server systems.

Results are less established for networks with many-server stations. By taking the QED diffusion scaling (taking the number of servers N to infinity in an appropriate manner), Armony and Maglaras [2] model and analyze rational customers in equilibrium; they treat jointly the problem of optimal control and staffing. Harrison and Zeevi [38] analyze the diffusion control problem associated with a single pool (multiple customer classes) model with linear costs. Specifically, they show that this control problem has an optimal Markov control policy (cf. [25]) which is characterized in terms of its underlying Hamilton-Jacobi-Bellman (HJB) equation.

The works of Atar, Mandelbaum and Reiman [7] and Atar [3] and [4] established asymptotic optimality of policies in the QED regime, for treelike models (the J nodes, which correspond to the J multi-server stations and the I nodes, corresponding to the I classes, jointly constitute a tree). The scaling then enforces convergence of the prelimit control problem to a diffusion control problem which can be dealt with by stochastic control methods, namely via the HJB equations. Then a method is provided on how to translate the obtained solutions into prelimit policies. The diffusion limit problem arises as a formal weak limit of a preemptive network scheduling problem, i.e. one where a service to a customer can be stopped at any moment and resumed at a later time, possibly in a different station. In the prelimit, the behaviour is clearly different for nonpreemptive networks, but it is proved that this difference vanishes asymptotically.

Related papers, working in the QED asymptotic regime, are those of Dai and Tezcan [21]–[22]. Given a control scheme, [21] gives a recipe for verifying whether this control admits state space collapse, i.e., a reduction of the dimension of the processes involved. Their paper [22] studies a queueing system with 2 customer classes and 2 server stations under some special conditions on service and abandonment rates, where they obtain exact control policies.

Gurvich and Whitt [29] propose a routing control called Fixed–Queue–Ratio. A newly available agent next serves the customer from the head of the queue of the class (from among those he is eligible to serve) whose queue length exceeds the most a prespecified proportion of the total queue length. This results in a dimension reduction, in a sense that vector–valued queue–length process is now asymptotically evolving as a one–dimensional process. The routing problem in [29] is treated simultaneously with the problems of optimal network design and staffing.

Chapter 2

Null controllability in heavy traffic

2.1. Introduction

We consider a multiclass queueing model with heterogeneous service stations, each consisting of many independent servers with identical capabilities. The servers offer service to different classes of customers at rates that may depend on both the station and the class. A system administrator dynamically controls all scheduling and routing in the system. The model is considered in the heavy traffic parametric regime, first proposed by Halfin and Whitt [30], in which the number of servers at each station and the arrival rates grow without bound, while keeping, in an appropriate sense, a critically balanced system. Both the model and the parametric regime have recently received much attention, especially in relation to large telephone call centers (see [26] and references therein). The chapter is based on [8].

When studying queueing models in heavy traffic, one considers a sequence of models parametrized by $n \in \mathbb{N}$ that, under a Law of Large Numbers (LLN) limit, give rise to a *fluid model* which is critically loaded. Typically, an attempt is then made to prove that appropriately scaled fluctuations of the queueing model about the fluid model converge to a diffusion. If, as in the current setting, a control problem is associated to the queueing model, then a similar approach gives rise to a *controlled diffusion model*. In this case, a natural goal is to prove that, given a cost criterion, the (suitably scaled) value of a control problem for the queueing model converges to one for the diffusion model. Moreover, it is often the case that solving the diffusion model for optimal controls helps understand how to construct control schemes for the queueing model that are asymptotically optimal. A similar approach is taken in this chapter. However,

rather than a cost criterion, our formulation will be concerned with a certain property observed for the diffusion model and shown to be inherited, in an asymptotic sense, by the queueing model. The property, that is unusual in heavy traffic formulations, will have to do with the ability to maintain empty queues.

We let I denote the number of customer classes, and let $X^n(t)$, $t \geq 0$ denote an I-dimensional process for which the ith component $X_i^n(t)$ represents the total number of class-i customers in the system at time t, in the nth system. The fluctuations alluded to above are denoted by $\hat{X}^n(t)$, $t \geq 0$. These fluctuations are scaled in such a way that $\hat{X}^n$ gives rise, as weak limits are taken formally, to a (controlled) diffusion process denoted by $X(t)$, $t \geq 0$. Moreover, one has that

$$\hat{X}^n(t) \in \mathbb{X}_- := \left\{ x \in \mathbb{R}^I : \sum_{i=1}^{I} x_i \leq 0 \right\}$$

holds if and only if the total number of customers in the nth system at time t is less than or equal to the total number of servers. Ideally, if one could freely rearrange customers in the system without any constraints, it follows that one could maintain empty queues whenever $\hat{X}^n(t) \in \mathbb{X}_-$. We will thus refer to $\mathbb{X}_-$ as the *null domain*. Although the queueing model considered in this chapter is subject to additional constraints (e.g., a station may offer service to only some of the classes), the null domain will play the same role in the asymptotic regime under consideration.

An important feature of the controlled diffusion model derived in this chapter is that the stochastic differential equation describing it has a *singular control* term, that is a control process with sample paths that are locally of bounded variation, the increments of which take values in a fixed cone $\mathbb{C}$ of $\mathbb{R}^I$. The singular term may be used to constrain the diffusion to lie in certain subsets of $\mathbb{R}^I$ at all times $t > 0$. We say that the diffusion is *null-controllable* if

$$\mathbb{C} \cap \mathbb{X}_-^o \neq \emptyset, \tag{2.1}$$

where $\mathbb{X}_-^o$ denotes the interior of $\mathbb{X}_-$, because under this condition the diffusion can be constrained to $\mathbb{X}_-$. Condition (2.1) will be given in explicit form in terms of the model parameters (see (2.33)). Our main result shows that under (2.1) one can construct control policies for the queueing model in such a way that, for every given $0 < \varepsilon < T < \infty$, all queues are kept empty on the time interval $[\varepsilon, T]$ with probability approaching 1, as $n \to \infty$. We will refer to such behavior as *asymptotic null-controllability*. We will, in fact, consider two versions of the problem: One, referred to as *preemptive scheduling*, in which

service to a customer can be interrupted and resumed at a later time (possibly in a different station). The other, referred to as *non-preemptive scheduling*, where service to customers may not be interrupted before service is completed. The treatment of the non-preemptive case is more complicated than that of the preemptive case. Thus, to keep the exposition simple we have limited ourselves in the non-preemptive case to the simplest possible network structure where null-controllability can show up: two customer classes and two service stations.

Our results on asymptotic null-controllability can be regarded as the demonstration of a new, unusual heavy traffic behavior. On one hand, the system is critically loaded. Indeed, as intuitively expected (and precisely stated in Proposition 2.1), an increase in any of the external arrival rates at the fluid level results with an overloaded system, in the sense that large queues inevitably build up. On the other hand, the system behaves as if it is underloaded as far as the capability of maintaining empty queues is concerned.

Singular control arises in connection with queueing systems in heavy traffic in many references. The singular term is often associated with positivity or finite buffer constraints for the queue length process (see [44], Chapter 8 for discussion and further references), with admission control (ibid.) or with constraints on the so-called workload process to lie in a given cone [16]. The source for the singular term in the current setting is however quite different, and it has to do with the fact that a many-server limit is taken. To explain this point, consider a network in which customers of classes, say 1 and 2, can be served at both station A and station B (these two classes and two stations could be a subset of a larger setup). Assume that the network operates under preemptive scheduling. Suppose that at a certain moment one selects, say r class-1 customers that are in service at station A and r class-2 customers in station B and considers the option of interchanging their position, so that the r class-1 [class-2] customers that were selected are moved to station B [respectively, A]. Since the service rates may depend on both the class and the station, the average rate at which components 1 and 2 of X^n change at that moment may be different depending on whether such an interchange takes place or not. If the interchange is performed instantaneously, the rates alluded to above will change abruptly. Moreover, since in both stations the number of servers is assumed to be large, this effect can be amplified by letting r be large. This, as scaling limit is taken, results with a control term that can have arbitrarily large increments over a given time interval. An appropriate formulation for such a controlled diffusion model will thus have a singular control term. Next, let $\mathcal{G}_a$ denote the graph with classes and stations as vertices, and with (i, j) pairs as edges if and only if station j can serve class i. It is instructive to note that our

explanation above relies on a certain property of the graph $\mathcal{G}_a$, namely that it contains a cycle with vertices 1, 2, A and B. Indeed, this is just another way of saying that customers of each of the two classes can be served in both stations, as we have assumed. A heuristic similar to the one described above will lead to a singular term in the diffusion model whenever $\mathcal{G}_a$ contains cycles, whether with four vertices or more. Thus, in general, cycles contained in $\mathcal{G}_a$ will play an important role in the singular control formulation of the diffusion model.

Note that the phenomenon described above is indeed a result of the many server setting, because it must be possible to occasionally let the number r referred to above take large values, and r is clearly limited by the number of servers. In particular, this phenomenon is not seen in what is sometimes referred to as 'conventional' heavy traffic, where diffusion limits are obtained for systems with a fixed number of servers (and large service rates; cf. [49]).

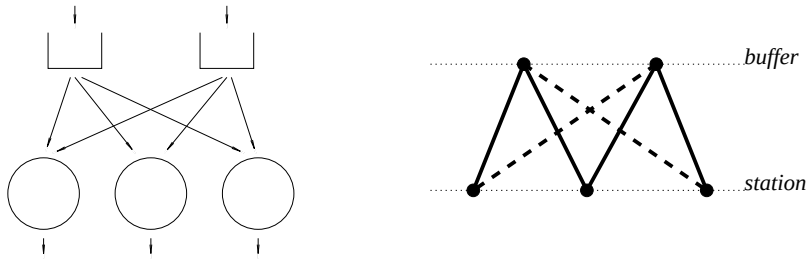
Recall that the so-called fluid model describes the LLN limit of basic quantities of the queueing model. One ingredient of the fluid model is a (deterministic, constant) matrix denoted by ξ^*. The entry ξ^*_{ij} represents the (large n limit) fraction of the number of servers in station j that serve class-i customers. One refers to (i,j) pairs that are edges of $\mathcal{G}_a$ as *activities*, and to activities (i,j) for which $\xi^*_{ij} > 0$ as *basic* activities. In particular, the number of servers that are engaged in non-basic activities ($\xi^*_{ij} = 0$) is of order $o(n)$ (where the total number of servers is proportional to n), in a sense that can be made precise (see, for example, Lemma 2.1). The policy that we shall propose for the preemptive case will basically mimic the construction of a constrained diffusion. Namely, a special rearrangement of customers in the service stations will take place whenever the process $\hat{X}^n$ reaches close to the boundary of $\mathbb{X}_-$ (from the inside). In this rearrangement, the number of servers allocated to work in a certain non-basic activity will be much larger than the typical fluctuations of the process X^n (but still $o(n)$). This rearrangement will have an effect on the dynamics of $\hat{X}^n$ that is a reminiscent of that of a Skorohod mechanism [45] on a constrained diffusion. Namely, it will constrain $\hat{X}^n$ to $\mathbb{X}_-$ by making the term that, in the limit, shows up as singular control, large. Because of (2.1), this can be performed in such a way that the term points into the domain. Although this question is not directly addressed in this chapter, it is expected, in fact, that the restriction to the time interval $[\varepsilon, T]$ of the processes $\hat{X}^n$ converge to a constrained diffusion as $n \to \infty$. The picture is quite different in the non-preemptive setting. The arrangement of customers can only be controlled indirectly (via routing decisions), and the constraining mechanism of the diffusion cannot be faithfully mimicked. A convergence result as above is not to be expected; in fact, the processes $\hat{X}^n$ that we construct in this case are not

even tight. The main idea behind the policy proposed in this case is to assure that a relatively large portion of the servers are engaged in the non-basic activity at all times, rather than at times when the boundary is reached as in the preemptive case.

We reiterate that it is the presence of cycles in $\mathcal{G}_a$ that induces a singular control term in the diffusion. A model similar to the current one is studied in [3] and [4] under structural assumptions that are complementary to those of this chapter, in the sense that the graph $\mathcal{G}_a$ is assumed there to be a tree. Indeed, in these references the diffusion has no singular control term, and a phenomenon as described in this chapter does not occur. Finally, we would like to mention that one can also consider a setting in which the diffusion has a singular term as in the current chapter, but null controllability does not hold, and approach the model from a control theoretic viewpoint so as to minimize costs of interest. This will be subject for future study.

The organization of the chapter is as follows. In Section 2.2 we first introduce the model and describe how its parameters are rescaled. We then present the main step toward the derivation of the diffusion model in Theorem 2.1, where a representation of the prelimit queueing model is provided. The diffusion model, obtained from this representation under the scaling limit, is stated in equations (2.30)–(2.32). We then state that, under (2.1), the diffusion can be constrained to $\mathbb{X}_-$ (Theorem 2.2), and provide the main results regarding asymptotic null-controllability in the preemptive case (Theorem 2.3) and in the non-preemptive case (Theorem 2.4). At the end of Section 2.2 we give numerical examples, and demonstrate that asymptotic null-controllability cannot be expected in overloaded models (Proposition 2.1). Section 2.3 contains the proofs of Theorems 2.2 and 2.3. The proof of Theorem 2.4 appears in Section 2.4. The appendix contains the proofs of Theorem 2.1, Proposition 2.1 and some auxiliary results.

Notation. For E a metric space, we denote by $\mathbb{D}(E)$ the space of all cadlag functions (i.e., right continuous and having left limits) from $\mathbb{R}_+$ to E. We endow $\mathbb{D}(E)$ with the usual Skorohod topology (cf. [13]). If X^n, $n \in \mathbb{N}$ and X are processes with sample paths in $\mathbb{D}(E)$, we write $X^n \Rightarrow X$ to denote weak convergence of the measures induced by X^n (on $\mathbb{D}(E)$) to the measure induced by X.



(a) (b)

FIG 2.1. *(a) A network with 2 classes and 3 stations. (b) Corresponding graph $\mathcal{G}_a$ with basic and non-basic activities (solid and, resp., dashed lines). The subgraph $\mathcal{G}_{ba}$ of $\mathcal{G}_a$ is a tree.*

2.2. Setting and main results

2.2.1. Queueing model

A precise description of the queueing model is as follows. A complete probability space $(\Omega, \mathcal{F}, P)$ is given, supporting all stochastic processes defined below. The processes will all be constructed in such a way that they have cadlag sample paths with probability 1. Expectation with respect to P is denoted by E. The queueing model is parametrized by $n \in \mathbb{N}$. It has $I \geq 2$ customer classes and $J \geq 2$ service stations (see Figure 2.1(a)). Station j has N_j^n identical servers working independently. The classes are labeled as $1, \ldots, I$ and the stations as $I+1, \ldots, I+J$:

$$\mathcal{I} = \{1, \ldots, I\}, \quad \mathcal{J} = \{I+1, \ldots, I+J\}.$$

Arrivals are modeled as renewal processes with finite second moment for the interarrival time. More precisely, we are given parameters $\lambda_i^n > 0$, $i \in \mathcal{I}$, $n \in \mathbb{N}$, and independent sequences of strictly positive i.i.d. random variables $\{\check{U}_i(k), k \in \mathbb{N}\}$, $i \in \mathcal{I}$, with mean $E\check{U}_i(1) = 1$ and variance $C_{U,i}^2 = \mathrm{Var}(\check{U}_i(1)) \in [0, \infty)$. With $\sum_1^0 = 0$, the number of class-i arrivals up to time t for the nth system is given by

$$A_i^n(t) = \sup\left\{ l \geq 0 : \sum_{k=1}^{l} \frac{\check{U}_i(k)}{\lambda_i^n} \leq t \right\}, \quad t \geq 0.$$

For $i \in \mathcal{I}, j \in \mathcal{J}$ and $n \in \mathbb{N}$, we are given parameters $\mu_{ij}^n \geq 0$, representing the service rate of a class-i customer by a server of station j. If $\mu_{ij}^n = 0$, we

say that class-i customers cannot be served at station j. Consider the graph $\mathcal{G}_a$ having a vertex set $\mathcal{I} \cup \mathcal{J}$ and an edge set

$$\mathcal{E}_a = \{(i,j) \in \mathcal{I} \times \mathcal{J} : \mu_{ij}^n > 0\}.$$

We assume that $\mathcal{E}_a$ does not depend on n. We denote $i \sim j$ if $(i,j) \in \mathcal{E}_a$. A class-station pair (i,j) is said to be an *activity* if $i \sim j$, or equivalently, if class-i customers can be served at station j. For every $(i,j) \in \mathcal{I} \times \mathcal{J}$, we denote by $\Psi_{ij}^n(t)$ the number of class-i customers being served in station j at time t. By definition,

$$\Psi_{ij}^n(t) = 0 \text{ for } i \not\sim j. \tag{2.2}$$

Service times are modeled as independent exponential random variables. To this end, let S_{ij}^n, $(i,j) \in \mathcal{I} \times \mathcal{J}$ be Poisson processes with rate μ_{ij}^n (where a Poisson process of zero rate is the zero process), mutually independent and independent of the arrival processes. Note that the time up to t devoted to a class-i customer by a server, summed over all servers of station j, is given as $\int_0^t \Psi_{ij}^n(s)ds$. The number of service completions of class-i customers by all servers of station j up to time t is, by assumption, given by $S_{ij}^n(\int_0^t \Psi_{ij}^n(s)ds)$. The precise description of the processes $\Psi^n = (\Psi_{ij}^n, i \in \mathcal{I}, j \in \mathcal{J})$ is not given at this point. We do emphasize however that they will be constructed in such a way that future service completion times are independent of the current state, which results with independent exponential service times (cf. [4]). We note in passing that whereas renewal arrivals are quite natural in the Halfin–Whitt setting, assumptions on service times that go beyond exponential (not to be dealt with in this chapter) lead to far more complicated diffusion models [50].

The processes A^n and S^n will be referred to as the *primitive processes*.

Denoting by $X_i^n(t)$ the number of class-i customers in the system (meaning: in the queue or being served) at time t, and setting $X_i^{0,n} = X_i^n(0)$, it is clear from the above that:

$$X_i^n(t) = X_i^{0,n} + A_i^n(t) - \sum_{j \in \mathcal{J}} S_{ij}^n \left(\int_0^t \Psi_{ij}^n(s)ds \right), \qquad i \in \mathcal{I}, t \geq 0. \tag{2.3}$$

For simplicity, the initial conditions $X_i^{0,n}$ are assumed to be deterministic. Finally, we introduce the processes $Y_i^n(t)$, representing the number of class-i customers in the queue (not being served) at time t, and $Z_j^n(t)$, representing the number of servers at station j that are idle at time t. Clearly, we have the following relations:

$$Y_i^n(t) + \sum_{j \in \mathcal{J}} \Psi_{ij}^n(t) = X_i^n(t), \qquad i \in \mathcal{I}, \tag{2.4}$$

$$Z_j^n(t) + \sum_{i \in \mathcal{I}} \Psi_{ij}^n(t) = N_j^n, \qquad j \in \mathcal{J}. \tag{2.5}$$

Also, the following holds by definition

$$Y_i^n(t) \geq 0,\ Z_j^n(t) \geq 0,\ \Psi_{ij}^n(t) \geq 0, \qquad i \in \mathcal{I}, j \in \mathcal{J}, t \geq 0. \tag{2.6}$$

We will write X^n for the vector $(X_i^n, i \in \mathcal{I})$ and similarly $Y^n = (Y_i^n, i \in \mathcal{I})$, $Z^n = (Z_j^n, j \in \mathcal{J})$.

2.2.2. Control and rescaling

Equations (2.2)–(2.6) indicate some properties of the processes involved, but they do not characterize these processes, because the control processes Ψ^n have not yet been described. This is the purpose of the following definitions.
Preemptive scheduling. We will regard scheduling as preemptive if service to a customer can be stopped and resumed at a later time, possibly in a different station. Formally, such a scheduling is a scheme according to which one selects $\Psi^n(t)$ at every t. In this chapter we will be concerned only with scheduling of feedback form, in the sense that the selection of $\Psi^n(t)$ depends only on $X^n(t)$, for every t. The precise definition is as follows.

Definition 2.1. *Let n be given. We say that a map $f^n : \mathbb{Z}_+^I \to \mathbb{Z}_+^{IJ}$ is a preemptive resume scheduling control policy (P-SCP) and X^n is the controlled process corresponding to f^n, initial data $X^{0,n}$ and primitive processes A^n and S^n, if $\Psi^n(t) = f^n(X^n(t))$ and equations (2.2)–(2.6) hold.*

Non-preemptive scheduling. By this we mean that it is impossible to interrupt service to customers. Thus, the quantities Ψ_{ij}^n cannot be directly controlled, but they are affected by the routing decisions according to the following equation:

$$\Psi_{ij}^n(t) = \Psi_{ij}^n(0) + B_{ij}^n(t) - S_{ij}^n\left(\int_0^t \Psi_{ij}^n(s)ds\right). \tag{2.7}$$

Above, for each $(i,j) \in \mathcal{I} \times \mathcal{J}$, B_{ij}^n is a nondecreasing $\mathbb{Z}_+$-valued process starting from zero, that increases by k each time k class-i customers are routed to station j. Of course, $B_{ij}^n = 0$ for $i \not\sim j$. In non-preemptive scheduling, a control policy is a scheme for selecting $B_{ij}^n(t)$ for every t. In this chapter we will need a randomized formulation, in which the scheme according to which B_{ij}^n are determined may depend on an auxiliary stochastic process. In addition, in our formulation we will only need the routing decisions to take place at the times when arrivals occur. To this end, for $i \in \mathcal{I}$ and $n \in \mathbb{N}$ we let $\tau_k^{n,i}$ denote the time of the kth jump of the process A_i^n, i.e., the time of the kth

class-i arrival. Finally, it will be assumed that all customers in service at time zero begin their service at that time, and the initial arrangement of these customers in the stations, i.e., $\Psi^n_{ij}(0)$, is determined by the policy. We write $\Xi^n(t) = (X^n(t), Y^n(t), Z^n(t), \Psi^n(t))$.

Definition 2.2. *Let n be given. We say that a triplet $(\Psi^{0,n}, F^n, R^n)$ is a non-preemptive scheduling control policy (N-SCP) and X^n is the controlled process corresponding to $(\Psi^{0,n}, F^n, R^n)$, initial data $X^{0,n}$ and primitive processes A^n, S^n if the following hold:*
i. $\Psi^{0,n} \in \mathbb{Z}_+^{IJ}$ is the initial condition for Ψ^n i.e., $\Psi^n(0) = \Psi^{0,n}$. In particular, it satisfies $\Psi^{0,n}_{ij} = 0$ for $i \not\sim j$, $\sum_j \Psi^{0,n}_{ij} \le X^{0,n}_i$ and $\sum_i \Psi^{0,n}_{ij} \le N^n_j$.
ii. R^n is a collection $\{R^{n,i}, i \in \mathcal{I}\}$, where, for each i, $R^{n,i} = \{R^{n,i}_k : k \in \mathbb{N}\}$ is a sequence of $\mathbb{R}$-valued independent random variables. The sequences $R^{n,i}$ are mutually independent and independent of the primitive processes.
iii. F^n is a collection $\{F^n_{ij}, (i,j) \in \mathcal{I} \times \mathcal{J}\}$ of measurable maps $F^n_{ij} : \mathbb{R}_+ \times \mathbb{Z}_+^I \times \mathbb{Z}_+^I \times \mathbb{Z}_+^J \times \mathbb{Z}_+^{IJ} \times \mathbb{R} \to \mathbb{Z}_+$ and for each $(i,j) \in \mathcal{I} \times \mathcal{J}$, B^n_{ij} is given in the form

$$B^n_{ij}(t) = \sum_{k:\tau^{n,i}_k \le t} F^n_{ij}(\tau^{n,i}_k, \Xi^n(\tau^{n,i}_k-), R^{n,i}_k), \qquad t \ge 0.$$

iv. Equations (2.2)–(2.7) hold.

Remarks. 1. The restriction to policies in which decisions take place only when arrivals occur may not appear to be natural, and one could consider extensions of this definition, say, by allowing decisions to take place upon arrivals or service completions, or even continuously in time. Note however that this restriction does not affect our result (i.e., Theorem 2.4) that is concerned with the existence (and construction) of N-SCP with a certain property: Clearly, the existence of such policies under the preceding definition implies the existence of policies under any extension of it.
2. Existence and uniqueness of the processes X^n and Ψ^n (given the primitive processes) is easily obtained by induction on the jump times of the primitive processes. In addition, one can argue that the future service completion times are independent of all that has occurred up to the current time. For a precise statement and an argument to this effect in a more restricted setup, see Proposition 1 of [7]. This can be adapted to the current setup, but we have omitted the details.
Fluid scaling. We assume that the parameters of the processes involved are scaled in the following way. There are constants $\lambda_i, \nu_j \in (0, \infty)$, $i \in \mathcal{I}$, $j \in \mathcal{J}$

and $\mu_{ij} \in (0,\infty)$, $(i,j) \in \mathcal{E}_a$ such that

$$n^{-1}\lambda_i^n \to \lambda_i, \qquad \mu_{ij}^n \to \mu_{ij}, \qquad n^{-1}N_j^n \to \nu_j. \tag{2.8}$$

We also define $\mu_{ij} = 0$ for $i \not\sim j$. Note that this is consistent with (2.8) because $\mu_{ij}^n = 0$ for $i \not\sim j$. Let

$$\bar{\mu}_{ij} = \nu_j \mu_{ij}, \quad (i,j) \in \mathcal{I} \times \mathcal{J}, \tag{2.9}$$

and consider the following linear program:

$$\begin{aligned} &\textit{Minimize } \rho \in \mathbb{R}_+ \textit{ subject to} \\ &\sum_{j \in \mathcal{J}} \bar{\mu}_{ij}\xi_{ij} = \lambda_i, \quad \sum_{i \in \mathcal{I}} \xi_{ij} \le \rho, \quad \xi_{ij} \ge 0, \qquad i \in \mathcal{I},\ j \in \mathcal{J}. \end{aligned} \tag{2.10}$$

Throughout, we assume that the fluid model is critically loaded. More precisely, we will assume that the *Heavy Traffic Condition* [33] holds: There exists a unique optimal solution (ξ^*, ρ^*) to the linear program (2.10), and moreover, $\sum_{i \in \mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$ (and consequently $\rho^* = 1$). We shall keep the notation ξ_{ij}^* throughout the chapter. We also let

$$x_i^* = \sum_j \xi_{ij}^* \nu_j, \qquad \psi_{ij}^* = \xi_{ij}^* \nu_j, \qquad i \in \mathcal{I}, j \in \mathcal{J}, \tag{2.11}$$

and refer to the quantities ξ^*, ψ^*, x^* as the *static fluid model*, or just *fluid model* for short (see in [4] what these quantities intuitively represent).

Following [33], an activity $(i,j) \in \mathcal{E}_a$ is said to be *basic* if $\xi_{ij}^* > 0$. Define the graph of basic activities $\mathcal{G}_{ba}$ to be the subgraph of $\mathcal{G}_a$ having $\mathcal{I} \cup \mathcal{J}$ as a vertex set, and the collection $\mathcal{E}_{ba}$ of basic activities as an edge set. We will also denote the set of non-basic activities as $\mathcal{E}_{nb} = \mathcal{E}_a \setminus \mathcal{E}_{ba}$ (see Figure 2.1(b)).

Like in some other papers that study a similar fluid model in heavy traffic (e.g., [33]), we will have one more assumption in this chapter about the fluid model, namely, that the *complete resource pooling* condition holds. This condition expresses, in a sense, a strong mode of cooperation between the service stations. More precisely, one of the equivalent formulations of this condition (see [33]) is that all vertices in $\mathcal{J}$ communicate via edges in $\mathcal{G}_{ba}$. It was shown by Williams [62] that this condition is equivalent to the condition that the basic activities form a tree. Thus we assume throughout:

$$\textit{The graph } \mathcal{G}_{ba} \textit{ is a tree.} \tag{2.12}$$

We now introduce a rescaled version of the processes describing the queueing model:

$$\bar{X}_i^n(t) = n^{-1}X_i^n(t), \quad \bar{Y}_i^n(t) = n^{-1}Y_i^n(t),$$

$$\bar{Z}_j^n(t) = n^{-1} Z_j^n(t), \quad \bar{\Psi}_{ij}^n(t) = n^{-1} \Psi_{ij}^n(t).$$

Denote $\bar{X}^n = (\bar{X}_i^n, i \in \mathcal{I})$, and use a similar notation for $\bar{Y}^n$, $\bar{Z}^n$ and $\bar{\Psi}^n$. We will sometimes consider $\bar{X}^n$, $\bar{Y}^n$ and $\bar{Z}^n$ as column vector-valued processes. Heuristically, one expects that $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*)$, and this is indeed the case under appropriate conditions (see, for example, equation (2.57) and Lemma 2.1 for such statements in the preemptive and, respectively, non-preemptive case). For this reason, these processes are referred to as the fluid-level rescaled processes.

Diffusion scaling. We further assume that there are constants $\hat{\lambda}_i, \hat{\mu}_{ij} \in \mathbb{R}$, $i \in \mathcal{I}, j \in \mathcal{J}$, such that

$$\hat{\lambda}_i^n := n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) \to \hat{\lambda}_i, \qquad \hat{\mu}_{ij}^n := n^{1/2}(\mu_{ij}^n - \mu_{ij}) \to \hat{\mu}_{ij}, \tag{2.13}$$

$$\hat{N}_j^n := n^{1/2}(n^{-1}N_j^n - \nu_j) \to 0. \tag{2.14}$$

We introduce a centered, rescaled version of the primitive processes:

$$\hat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t), \quad \hat{S}_{ij}^n(t) = n^{-1/2}(S_{ij}^n(nt) - n\mu_{ij}^n t). \tag{2.15}$$

Similarly, the processes representing the queueing model are centered about the fluid model quantities and rescaled:

$$\hat{X}_i^n(t) = n^{-1/2}(X_i^n(t) - nx_i^*), \tag{2.16}$$

$$\hat{Y}_i^n(t) = n^{-1/2} Y_i^n(t), \qquad \hat{Z}_j^n(t) = n^{-1/2} Z_j^n(t), \tag{2.17}$$

$$\hat{\Psi}_{ij}^n(t) = n^{-1/2}(\Psi_{ij}^n(t) - \psi_{ij}^* n). \tag{2.18}$$

The processes denoted with hats will be referred to as diffusion-level rescaled processes. Similarly to the fluid-level processes, define $\hat{X}^n = (\hat{X}_i^n, i \in \mathcal{I})$, with an analogous definition for $\hat{Y}^n$, $\hat{Z}^n$ and $\hat{\Psi}^n$, and consider $\hat{X}^n$, $\hat{Y}^n$ and $\hat{Z}^n$ as column vector-valued processes.

Scaling of initial conditions. It is assumed that there are constants x_i, $i \in \mathcal{I}$, such that the initial conditions satisfy

$$\hat{X}_i^{0,n} := \hat{X}_i^n(0) \to x_i. \tag{2.19}$$

Throughout, $x = (x_i, i \in \mathcal{I})$.

2.2.3. Main results

Our first result expresses a relation directly between the diffusion-level processes. Although its proof requires only some elementary manipulations of the relations (2.2)–(2.6) and (2.16)–(2.18), it has an important role in this chapter as the basis for deriving the diffusion model. In particular, it will make clear how the singular control formulation arises. To present it we need some notation.

Denote by $\mathcal{C}$ the set of all cycles that are subgraphs of $\mathcal{G}_a$, for which exactly one edge is a non-basic activity. We call an element $c \in \mathcal{C}$ a *simple cycle* (see Figure 2.2(a)). Lemma 2.4 in the appendix shows, using (2.12), that every non-basic activity belongs to a simple cycle (as an edge). Consequently, there is a one-to-one correspondence between $\mathcal{E}_{nb}$ and $\mathcal{C}$, which we denote by σ. More precisely,

$$\sigma(i,j) = c \text{ whenever } (i,j) \in \mathcal{E}_{nb} \text{ and } c \text{ is the simple cycle through } (i,j). \tag{2.20}$$

With an abuse of notation, we will write $(i,j) \in c$ when we mean that a (not necessarily non-basic) activity $(i,j) \in \mathcal{E}_a$ belongs to the edge set of the graph c.

Next, we associate directions with the edges of simple cycles. Let c be a simple cycle with vertices $i_0, j_0, i_1, j_1, \ldots, i_k, j_k$, where for $0 \le l \le k$, $i_l \in \mathcal{I}$ and $j_l \in \mathcal{J}$, and edges $(i_0, j_0) \in \mathcal{E}_{nb}$ and $(j_0, i_1), \ldots, (i_k, j_k), (j_k, i_0) \in \mathcal{E}_{ba}$. The direction that we associate with the non-basic element (i_0, j_0) is $i_0 \to j_0$ (in words: from i_0 to j_0). The direction of the other edges, *when considered as edges of* c, is consistent with that of the non-basic element, namely: $i_0 \to j_0 \to i_1 \to j_1 \to \cdots j_k \to i_0$. Note that an edge (corresponding to a basic activity) may have different directions when considered as an edge of different simple cycles. We signify the directions along the simple cycles by $s(c,i,j)$, defined, for all $c \in \mathcal{C}$ and $(i,j) \in c$, as

$$s(c,i,j) = \begin{cases} -1 & \text{if } (i,j), \text{ considered as an edge of } c, \text{ is directed from } i \text{ to } j \\ 1 & \text{if } (i,j), \text{ considered as an edge of } c, \text{ is directed from } j \text{ to } i. \end{cases}$$

We will denote

$$m^n_{i,c} = \sum_{j:(i,j)\in c} s(c,i,j)\mu^n_{ij}, \qquad i \in \mathcal{I}, \qquad m^n_c = (m^n_{i,c}, i \in \mathcal{I}). \tag{2.21}$$

Next, consider the system of equations in ψ:

$$\begin{cases} \sum_{j\in\mathcal{J}} \psi_{ij} & = a_i, \quad i \in \mathcal{I}, \\ \sum_{i\in\mathcal{I}} \psi_{ij} & = b_j, \quad j \in \mathcal{J}, \\ \psi_{ij} & = 0, \quad (i,j) \in \mathcal{E}_{nb}. \end{cases} \tag{2.22}$$

It is known that (2.22) has a unique solution ψ for every a, b satisfying $\sum_i a_i = \sum_j b_j$ (see [3], Proposition A.2). With

$$D_G = \{(a,b) \in \mathbb{R}^I \times \mathbb{R}^J : \sum_{i\in\mathcal{I}} a_i = \sum_{j\in\mathcal{J}} b_j\}, \tag{2.23}$$

denote by $G : D_G \to \mathbb{R}^{IJ}$ the solution map, namely

$$\psi_{ij} = G_{ij}(a,b), \qquad (i,j) \in \mathcal{E}_a. \tag{2.24}$$

The function G is linear and so Lipschitz (a fact that will be used in the sequel). Denote also

$$H_i^n(a,b) = -\sum_j \mu_{ij}^n G_{ij}(a,b), \qquad i \in \mathcal{I},\ a \in \mathbb{R}^I, b \in \mathbb{R}^J, \qquad H^n = (H_i^n, i \in \mathcal{I}), \tag{2.25}$$

Let $\ell_i^n = \hat{\lambda}_i^n - \sum_{j\in\mathcal{J}} \hat{\mu}_{ij}^n \psi_{ij}^*$ and set

$$\hat{W}_i^n(t) = \hat{A}_i^n(t) - \sum_i \hat{S}_{ij}^n \left(\int_0^t \bar{\Psi}_{ij}^n(s) ds \right) + \ell_i^n t. \tag{2.26}$$

Finally, let the quantities $\{\Psi_{ij}^n\}$, as (i,j) ranges over the non-basic activities, be labeled by simple cycles, namely define for every $c \in \mathcal{C}$, $\Psi_c = \Psi_{ij}$ where $(i,j) = \sigma^{-1}(c)$. Let a diffusion-level version of these processes be defined as $\hat{\Psi}_c^n = n^{-1/2}\Psi_c^n$.

Theorem 2.1. *Let X^n, Y^n, Z^n, Ψ^n satisfy (2.2)–(2.6) and let $\hat{X}^n, \hat{Y}^n, \hat{Z}^n, \hat{\Psi}^n$ be defined by (2.16)–(2.18). Then the following relations hold for all $t \geq 0$:*

$$\begin{aligned} \hat{X}^n(t) = \hat{X}^{0,n} + \hat{W}^n(t) + \int_0^t H^n(\hat{X}^n(s) - \hat{Y}^n(s), \hat{N}^n - \hat{Z}^n(s)) ds \\ + \sum_{c\in\mathcal{C}} m_c^n \int_0^t \hat{\Psi}_c^n(s) ds, \end{aligned} \tag{2.27}$$

$$\hat{Y}_i^n(t) \geq 0,\ i \in \mathcal{I}, \quad \hat{Z}_j^n(t) \geq 0,\ j \in \mathcal{J}, \quad \sum_{i\in\mathcal{I}} [\hat{X}_i^n(t) - \hat{Y}_i^n(t)] = \sum_{j\in\mathcal{J}} [\hat{N}_j^n - \hat{Z}_j^n(t)], \tag{2.28}$$

$$\hat{\Psi}_c^n(t) \geq 0, \quad c \in \mathcal{C}. \tag{2.29}$$

See the appendix for a proof. We now explain how a diffusion model is derived from the above result. First, note that the limit, as $n \to \infty$, in the definition (2.21) of m_c^n exists, and is equal to the expression obtained from (2.21) by replacing μ_{ij}^n by μ_{ij}. We denote the limit by m_c. The vectors m_c will be referred to as *directions of control* because of the role they will play in the singular control term. In a similar fashion, we denote by H the limit of H^n as $n \to \infty$, or equivalently, as the expression obtained in (2.25) by replacing μ_{ij}^n by μ_{ij}. Next, we take a formal limit in (2.26). We let $\ell = (\ell_i, i \in \mathcal{I})$ where $\ell_i = \lim_{n\to\infty} \ell_i^n = \hat{\lambda}_i - \sum_{j\in\mathcal{J}} \hat{\mu}_{ij}\psi_{ij}^*$ and let W denote a Brownian motion for which the mean and the covariance matrix of $W(1)$ are given by ℓ and, respectively, $\Sigma := \mathrm{diag}(\lambda_i C_{U,i}^2 + \lambda_i)$ (for short: an (ℓ, Σ)-Brownian motion). The expression above for the covariance is obtained by calculating the covariance matrix of $\hat{W}^n(1)$ upon formally replacing the fluid-level quantities $\bar{\Psi}_{ij}^n(s)$ in (2.26) by the quantities ψ_{ij}^* from the fluid model, and finally taking limits as $n \to \infty$.

Next, equation (2.29) imposes a constraint on $\hat{\Psi}_c^n$ in the form of a lower bound at zero. It is also not hard to see that an upper bound on $\hat{\Psi}_c^n$ follows from (2.5), (2.14) and (2.18). However, such a bound is of the order of $n^{1/2}$, and as n grows without bound it becomes irrelevant. It therefore makes sense that, in the proposed diffusion model, each integral in the last term of (2.27) is replaced by a process η_c that is required to have increasing sample paths, and no additional limitation.

We thus obtain a diffusion model involving processes $X(t)$, $Y(t)$, $Z(t)$ taking values in $\mathbb{R}_+^I$, $\mathbb{R}_+^I$ and, respectively, $\mathbb{R}_+^J$ as well as $\mathbb{R}_+$-valued processes η_c, $c \in \mathcal{C}$. Equations (2.30)–(2.32) below, that describe the diffusion model, are analogous to equations (2.27)–(2.29), respectively:

$$X(t) = x + W(t) + \int_0^t H(X(s) - Y(s), -Z(s))ds + \sum_{c\in\mathcal{C}} m_c \eta_c(t) \tag{2.30}$$

$$Y_i(t) \geq 0,\ i \in \mathcal{I}, \quad Z_j(t) \geq 0,\ j \in \mathcal{J}, \quad \sum_{i\in\mathcal{I}} Y_i(t) = \sum_{i\in\mathcal{I}} X_i(t) + \sum_{j\in\mathcal{J}} Z_j(t),\ t \geq 0, \tag{2.31}$$

$$\text{For every } c \in \mathcal{C},\ \eta_c \text{ is nondecreasing and } \eta_c(0) \geq 0. \tag{2.32}$$

We will consider Y, Z and $\{\eta_c\}$ as control processes, and regard (2.31), (2.32) as constraints that they must satisfy. To define precisely what controls will be regarded as admissible, recall that W is an (ℓ, Σ)-Brownian motion defined on $(\Omega, \mathcal{F}, P)$. Let $(\mathcal{F}_t)$ be a right-continuous filtration of sub-sigma-fields of $\mathcal{F}$ such that $(W(t) - \ell t, \mathcal{F}_t : t \geq 0)$ is a martingale. Let a deterministic initial

condition $x \in \mathbb{R}^I$ be given. A triplet (Y, Z, η) is said to be an *admissible control* and X a corresponding *controlled process* if Y, Z and η are processes with cadlag sample paths, Y and Z are $(\mathcal{F}_t)$-progressively measurable, η_c is $(\mathcal{F}_t)$-adapted, and (2.30)–(2.32) hold for all $t \in [0, \infty)$ P-a.s.

A principal hypothesis of all the results below is what we refer to as the *null controllability condition:*

$$\text{There exists } c \in \mathcal{C} \text{ such that } e \cdot m_c < 0. \tag{2.33}$$

Note that the cone $\mathbb{C}$ referred to in the introduction, in which the increments of the I-dimensional process $\sum_c m_c \eta_c(t)$ of (2.30) take values, is simply the cone generated by $\{m_c : c \in \mathcal{C}\}$. Also, condition (2.1) of the introduction is given explicitly by (2.33). The following result is proved in Section 2.3.

Theorem 2.2. *Assume* (2.33) *holds. Then there exists an admissible control* (Y, Z, η) *under which* $e \cdot X(t) \leq 0$ *and* $Y(t) = 0$ *on* $[0, \infty)$, P*-a.s.*

The main results of this chapter establish the validity of statements about the queueing model, that are analogous to Theorem 2.2 in an asymptotic sense. The first is concerned with preemptive scheduling (see Section 2.3 for a proof).

Theorem 2.3. *Assume* (2.33) *holds. Then there exist P-SCPs under which, for every* ε *and* T *satisfying* $0 < \varepsilon < T < \infty$,

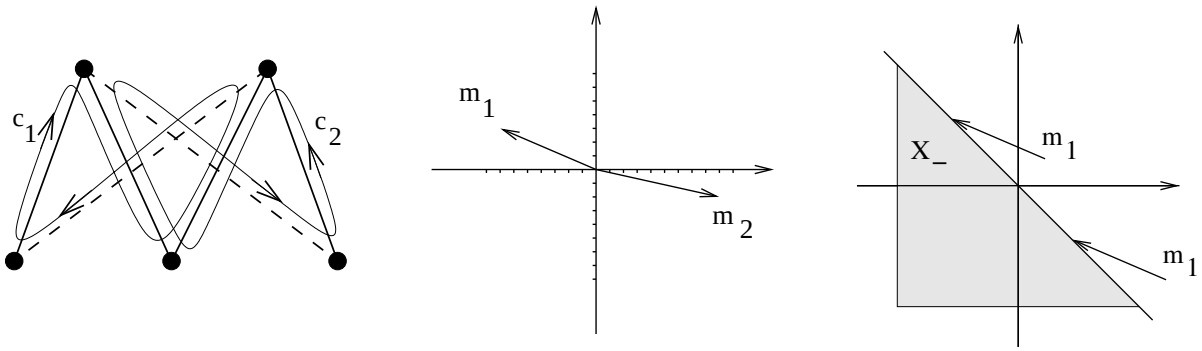
$$\lim_{n \to \infty} \mathbb{P}(Y^n(t) = 0 \text{ for all } t \in [\varepsilon, T]) = 1.$$

The treatment of non-preemptive scheduling is more involved. In order to keep the notation simple we have focused in this chapter on the most simple case where one can expect a null-controllability result: The case $I = J = 2$. Clearly, in this case there is at most one simple cycle. It is expected that the general case can be treated with similar ideas. The result below is proved in Section 2.4.

Theorem 2.4. *Assume* $I = J = 2$ *and let* (2.33) *hold. Then there exist N-SCPs under which, for every* ε *and* T *satisfying* $0 < \varepsilon < T < \infty$,

$$\lim_{n \to \infty} \mathbb{P}(Y^n(t) = 0 \text{ for all } t \in [\varepsilon, T]) = 1.$$

Remark. As can be seen in Sections 2.3 and 2.4, Theorems 2.3 and 2.4 hold with $\varepsilon = 0$ in case that the initial condition x satisfies $e \cdot x < -1$ (or even $e \cdot x < 0$, with an obvious modification of the proofs).



(a) (b) (c)

FIG 2.2. *(a) A graph as in Figure 2.1(b) with two simple cycles. Dashed lines represent non-basic activities. (b) A possible set of directions of control corresponding to the two cycles. (c) The direction m_1 may be used to constrain the diffusion model to the null domain $\mathbb{X}_-$.*

2.2.4. Discussion

Let us consider some numerical examples. Consider first a system with structure as depicted in Figure 2.1(a). Assume $\nu_j = 1$ for $j = 1, 2, 3$, and

$$\lambda = \begin{pmatrix} 8 \\ 4 \end{pmatrix}, \qquad \bar{\mu} = \mu = \begin{pmatrix} 3 & 10 & 1 \\ 1 & 4 & 2 \end{pmatrix},$$

(where in this subsection we abuse notation and label j by $1, \ldots, J$). One checks that the heavy traffic condition holds, and

$$\xi^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix}, \qquad m_1 = \begin{pmatrix} -7 \\ 3 \end{pmatrix}, \qquad m_2 = \begin{pmatrix} 9 \\ -2 \end{pmatrix}.$$

Above, m_1 and m_2 are the directions of control corresponding to the two non-basic activities $(2, 1)$ and $(1, 3)$. In fact, the graph that appears in Figure 2.1(b) precisely describes $\mathcal{G}_a$ and $\mathcal{G}_{ba}$ in the current example. Clearly the null-controllability condition holds, since $e \cdot m_1 < 0$. The simple cycles and directions of control are depicted in Figure 2.2(a) and (b). To demonstrate the geometric aspect we refer to Figure 2.2(c), where the null domain $\mathbb{X}_-$ is shown along with a vector field defined on its boundary assuming the constant value m_1. Under appropriate assumptions, one can construct a diffusion in $\mathbb{R}^2$ with a boundary term according to this vector field, that will be constrained to $\mathbb{X}_-$. In contrast, m_2 can not be used to constrain the diffusion to the same domain.

In general, the diffusion can be constrained to $\mathbb{X}_-$ provided that at least one of the vectors m_c satisfies $e\cdot m_c < 0$. This is the source for condition (2.33).

We next consider examples with two classes and two stations:

$$\mu = \bar{\mu} = \begin{pmatrix} 8 & 10 \\ 3 & 6 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 13 \\ 3 \end{pmatrix}, \quad \xi^* = \begin{pmatrix} 1 & 0.5 \\ 0 & 0.5 \end{pmatrix}, \quad m = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \tag{2.34}$$

$$\mu = \bar{\mu} = \begin{pmatrix} 4 & 7 \\ 2 & 4 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 7.5 \\ 2 \end{pmatrix}, \quad \xi^* = \begin{pmatrix} 1 & 0.5 \\ 0 & 0.5 \end{pmatrix}, \quad m = \begin{pmatrix} -3 \\ 2 \end{pmatrix}. \tag{2.35}$$

Above, we have presented both the data and the solution to the linear program in each case. In both cases the heavy traffic condition holds and the activities are as depicted in Figure 2.3(a). Note that in these examples there is a single cycle. The null controllability condition does not hold in the first example, and it does hold in the second. In fact, one can write the null controllability condition (2.33) for the above examples in a straightforward way as:

$$\mu_{11} + \mu_{22} < \mu_{12} + \mu_{21}. \tag{2.36}$$

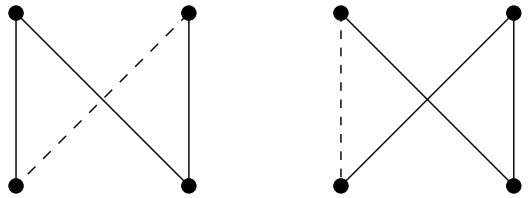
It is instructive to note that there are values of μ for which (2.36) does not express the null controllability condition (2.33). Consider the following example:

$$\mu = \bar{\mu} = \begin{pmatrix} 3 & 7 \\ 6 & 11 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 3.5 \\ 11.5 \end{pmatrix}, \quad \xi^* = \begin{pmatrix} 0 & 0.5 \\ 1 & 0.5 \end{pmatrix}, \quad m = \begin{pmatrix} 4 \\ -5 \end{pmatrix}.$$

Note that the values of both sides of the inequality (2.36) are the same as in example (2.34). Still, as can be verified by calculating $e\cdot m$, the null controllability condition, that did not hold in example (2.34), does hold in this example. The reason is that the structure of the graph has changed and it now corresponds to Figure 2.3(b). In particular, the direction of the cycle is reversed, and condition (2.33) is equivalent to (2.36) with a reversed inequality. We can see that the null controllability condition depends on the values of μ as well as the direction of the cycle, which in turn is determined by the fluid model parameters λ and $\bar{\mu}$.

We now come back to the point referred to in the introduction regarding the unusual heavy traffic behavior. A particular consequence of our main results can be stated as follows. One can find policies (preemptive and non-preemptive) under which, for every $t > 0$,

$$\lim_{n\to\infty} P\left(e\cdot Y^n(t) = 0\right) = 1. \tag{2.37}$$



(a) (b)

FIG 2.3. *Two examples with cycles in opposite directions (dashed lines represent non-basic activities).*

What we referred to as unusual is that a critically loaded system shows a behavior that is typical to underloaded systems: the possibility to maintain empty queues with probability approaching one. We would like to make precise the statement that the system under study is critically loaded. One aspect of this is that the underlying fluid model is critically loaded, in the sense that the linear program (2.10) is solved with $\rho^* = 1$, by assumption. More significant, however, is a statement that can be made regarding the probabilistic model. Namely, one can show that in a probabilistic model associated with an overloaded fluid model, queues inevitably build up.

To this end, let λ'_i, $i \in \mathcal{I}$ be constants satisfying $\lambda'_i \geq \lambda_i$ for all i and $\lambda'_i > \lambda_i$ for at least one $i \in \mathcal{I}$. The parameters $(\lambda, \bar{\mu})$ lead to a fluid model that is critically loaded. In the same sense, the pair $(\lambda', \bar{\mu})$ correspond to an overloaded fluid model. We consider a sequence of processes A^n_{OL} defined analogously to A^n, but with λ'^n replacing the parameters λ^n, where λ'^n is a sequence satisfying $n^{-1}\lambda'^n \to \lambda'$ (compare with (2.8)). Let X^n_{OL} stand for the processes X^n obtained by replacing throughout our probabilistic model A^n by A^n_{OL}. Define analogously all other processes involved, e.g., Y^n_{OL} in place of Y^n. As the following result shows, the model thus obtained is indeed overloaded in the sense that queues (in fact, large queues) necessarily build up. The result shows a sharp contrast with (2.37).

Proposition 2.1. *There exist constants C_1 and $C_2 > 0$ depending only on $(\lambda, \lambda', \bar{\mu})$ (and not depending on n or t) such that under any policy, for every t,*

$$\lim_{n\to\infty} P\left(e \cdot Y^n_{OL}(t) \geq (-C_1 + C_2 t)n\right) = 1. \tag{2.38}$$

A proof is sketched in the appendix.

2.3. Diffusion model and queueing model in the preemptive case

In this section we prove Theorems 2.2 and 2.3.

PROOF OF THEOREM 2.2. Let c_0 be such that $e\cdot m_{c_0} < 0$. Consider the domain

$$D_\alpha = \{\xi \in \mathbb{R}^I : e\cdot \xi < -\alpha\}$$

for some fixed $\alpha \geq 0$. Fix some $j_0 \in \mathcal{J}$. We will construct a control with the following properties:

$$Y(t) = 0, \quad Z_j(t) = 0, \text{ for all } j \neq j_0, \quad \eta_c(t) = 0 \text{ for all } c \neq c_0, \quad t \geq 0. \tag{2.39}$$

The process Z_{j_0} will satisfy

$$Z_{j_0}(t) = -e\cdot X(t), \ t \geq 0. \tag{2.40}$$

As a result, (2.31) will be satisfied. Finally, the process η_{c_0} will serve as a constraining term of a reflected diffusion on the domain D_α with reflection field identically equal to m_{c_0} on the boundary ∂D_α. We therefore consider equation (2.30) in the special form

$$X(t) = x + W(t) + \int_0^t \widetilde{H}(X(s))ds + m_{c_0}\eta_{c_0}(t), \tag{2.41}$$

where $\widetilde{H}(\xi) = H(\xi, (e\cdot \xi)e_{j_0})$. Note that $\widetilde{H}$ is linear. Consider first the case $x \in \overline{D_\alpha}$. The result of [45] regarding existence of strong solutions to stochastic differential equations with oblique reflection, stated for a bounded domain, implies, using a standard localization argument, the existence of a pair (X, η_{c_0}) with the following properties. The process X is progressively measurable, η_{c_0} is continuous nondecreasing, adapted, with values in $\mathbb{R}_+$, equation (2.41) holds for all $t \geq 0$, a.s. and $X(t) \in \overline{D_\alpha}$, $t \geq 0$, a.s. In particular, we have constructed a process X with the property $e\cdot X(t) \leq -\alpha$, $t \geq 0$, a.s. Letting now Y, Z and η_c, $c \neq c_0$ be defined via (2.39) and (2.40), we have constructed a triplet (Y, Z, η) that is an admissible control, and have shown that X and Y satisfy the conclusion of the theorem.

In the case where $e\cdot x > -\alpha$, clearly $x + \beta m_{c_0} \in \overline{D_\alpha}$ for β large. Fix any β as above, and set $y = x + \beta m_{c_0} \in \overline{D_\alpha}$. Denote the control and the controlled process corresponding to starting from y as $(\widetilde{Y}, \widetilde{Z}, \widetilde{\eta})$ and, respectively, $\widetilde{X}$. Now set $Y = \widetilde{Y}$, $Z = \widetilde{Z}$, $\eta_{c_0}(t) = \beta + \widetilde{\eta}_{c_0}(t)$ and $X(t) = \widetilde{X}(t)$. Then $X(0) = y$, and clearly (2.41) still holds for all $t \geq 0$. As a result, (Y, Z, η) is an admissible control and the conclusion of the theorem holds. □

We remark that in the proof above we can simply take $\alpha = 0$. Our results regarding asymptotic null controllability will be inspired by a similar construction but α will be taken to be positive.

Recall that, by definition, a P-SCP is a map that determines $\Psi^n(t)$ for a given value of $X^n(t)$ in such a way that (2.2)–(2.6) hold. The following lemma shows that, under suitable conditions, we can determine the value of Ψ^n by first selecting values for $Y^n(t)$, $Z^n(t)$ and $\{\Psi^n_c(t), c \in \mathcal{C}\}$.

Lemma 2.1. *There is a constant $a_0 > 0$ such that the following holds for all n large. Suppose that, for some t, the following relations hold: $X^n(t) \in \mathbb{Z}^I_+$, $\|X^n(t) - nx^*\| \leq a_0 n$, $Y^n(t) \in \mathbb{Z}^I_+$, $Z^n(t) \in \mathbb{Z}^J_+$,*

$$e\cdot Y^n(t) + e\cdot N^n = e\cdot X^n(t) + e\cdot Z^n(t), \qquad [e\cdot Y^n(t)] \vee [e\cdot Z^n(t)] \leq a_0 In + 1,$$

$\Psi^n_c(t) \in \mathbb{Z}_+$, $c \in \mathcal{C}$, and

$$\Psi^n_c(t) \leq a_0 n, \quad c \in \mathcal{C}.$$

Set

$$\Psi^n_{ij}(t) = G_{ij}(X^n(t) - Y^n(t), N^n - Z^n(t)) - \sum_{c\in\mathcal{C}:(i,j)\in c} s(c,i,j)\Psi^n_c(t). \tag{2.42}$$

Then the quantities

$$\{X^n_i(t), i \in \mathcal{I}\},\ \{Y^n_i(t), i \in \mathcal{I}\},\ \{Z^n_j(t), j \in \mathcal{J}\},\ \{\Psi^n_{ij}(t), (i,j) \in \mathcal{I}\times\mathcal{J}\}$$

satisfy (2.2) *and* (2.4)–(2.6).

The proof of this lemma appears in the appendix.

Proof of Theorem 2.3. We begin by describing the policy. By Definition 2.1 we need to describe a map that determines $\Psi^n(t)$ for a given value of $X^n(t)$. Given the one-to-one relations (2.16) and (2.18), this is equivalent to describing $\hat{\Psi}^n(t)$ for a given value of $\hat{X}^n(t)$. Let a_0 be the constant from Lemma 2.1. If $\|\hat{X}^n(t)\| > a_0 n^{1/2}$ we assume that $\hat{\Psi}^n(t)$ is determined as the image of $\hat{X}^n(t)$ under some fixed map f^n, in a way that is consistent with Definition 2.1 but otherwise arbitrary (there will be no need to describe f^n more precisely). Focusing below on the case $\|\hat{X}^n(t)\| \leq a_0 n^{1/2}$, we fix a sequence K_n, $n \in \mathbb{N}$ so that $n^{1/2}K_n \in \mathbb{Z}_+$ for all n and

$$K_n \to \infty, \qquad n^{-1/2}K_n \to 0, \qquad \text{as } n \to \infty. \tag{2.43}$$

Denote

$$D_1 = \{\xi \in \mathbb{R}^I : e\cdot\xi < -1\}.$$

Fix throughout a simple cycle c_0 for which $e\cdot m_{c_0} < 0$. Also fix throughout $i_0 \in \mathcal{I}$ and $j_0 \in \mathcal{J}$. The proposed policy sets

$$\hat{\Psi}^n_{c_0}(t) = \begin{cases} 0, & \hat{X}^n(t) \in D_1 \\ K_n, & \hat{X}^n(t) \in D_1^c \end{cases} \qquad t \geq 0, \tag{2.44}$$

and $\hat{\Psi}^n_c(t) = 0$ for all $c \neq c_0$, $t \geq 0$. It also sets $\hat{Y}^n_i(t) = 0$ for all $i \neq i_0$, $\hat{Z}^n_j(t) = 0$ for all $j \neq j_0$ and

$$\hat{Y}^n_{i_0}(t) = (e\cdot \hat{X}^n(t) - e\cdot \hat{N}^n)^+, \qquad \hat{Z}^n_{j_0}(t) = (e\cdot \hat{X}^n(t) - e\cdot \hat{N}^n)^-, \qquad t \geq 0. \tag{2.45}$$

By (2.43) and (2.44), $\Psi^n_{c_0}(t) \leq a_0 n$ for all n large. By Lemma 2.1, $\Psi^n(t)$ are well defined given $X^n(t)$, and (2.2), (2.4)–(2.6) hold. This completes the description of the P-SCP. Clearly, this description, along with equation (2.3), uniquely define the processes Ψ^n and X^n given the initial conditions and primitive processes. Although the description applies for any initial condition $\hat{X}^{0,n}$, the treatment with be slightly different for $\hat{X}^{0,n}$ in D_1 and outside.

In what follows let $0 < \varepsilon < T < \infty$ be fixed. Let

$$\bar{\tau}^n = \inf\{t : \|\hat{X}^n(t)\| > a_0 n^{1/2}\}. \tag{2.46}$$

Let also

$$\widetilde{H}^n(\xi) = H^n(\xi - (e\cdot \xi - e\cdot \hat{N}^n)^+ e_{i_0},\, -(e\cdot \xi - e\cdot \hat{N}^n)^- e_{j_0}).$$

For $t \leq \bar{\tau}^n$, one has by (2.27)

$$\hat{X}^n(t) = \hat{X}^{0,n} + \hat{W}^n(t) + \int_0^t \widetilde{H}^n(\hat{X}^n(s))ds + m^n_{c_0} \int_0^t \hat{\Psi}^n_{c_0}(s)ds. \tag{2.47}$$

Note that $\widetilde{H}^n$ satisfy

$$\|\widetilde{H}^n(\xi)\| \leq C_H(\|\xi\| + \|\hat{N}^n\|), \qquad \xi \in \mathbb{R}^I, n \in \mathbb{N}, \tag{2.48}$$

where C_H is a constant independent of n. Denote $C^n_e = -e\cdot m^n_{c_0}$. Note that by assumption $C^n_e \to -e\cdot m_{c_0} > 0$ as $n \to \infty$. It is assumed in what follows that n is so large that

$$C^n_e > |e\cdot m_{c_0}|/2 =: 2C_0. \tag{2.49}$$

Denote $\hat{X}^n_e = e\cdot \hat{X}^n$ and similarly $\hat{W}^n_e = e\cdot \hat{W}^n$, $\hat{X}^{0,n}_e = e\cdot \hat{X}^{0,n}$ and $\widetilde{H}^n_e = e\cdot \widetilde{H}^n$. Then by (2.44) and (2.47) we have

$$\hat{X}^n_e(t) = \hat{X}^{0,n}_e + \hat{W}^n_e(t) + \int_0^t \widetilde{H}^n_e(\hat{X}^n(s))ds - C^n_e K_n \int_0^t 1_{\hat{X}^e_n(s) \geq -1} ds, \qquad t \leq \bar{\tau}^n. \tag{2.50}$$

The rest of our argument is divided into four steps.
Step 1. We first show that there exists a deterministic constant C_1 independent of x, n and K_n such that

$$\|\hat{X}^n\|_T^* \le C_1(1 + \|\hat{X}^{0,n}\| + \|\hat{W}^n\|_T^*) \qquad \text{on the event } \bar{\tau}^n \ge T. \tag{2.51}$$

To this end, denote $A_n = [4\|e\|(\|\hat{X}^{0,n}\| + \|\hat{W}^n\|_T^*)] \vee 1$ and let

$$\tau_1^n = \inf\{t \in [0,T] : \hat{X}_e^n(t) \le -A_n - 1\}.$$

Note that $\tau_1^n > 0$. Since by assumption $\bar{\tau}^n \ge T$, (2.47) and (2.50) are valid for $t \le T$. By (2.48) and (2.50), noting that $\hat{X}_e^n(t) \ge -A_n - 1$ for $t \le \tau_1^n$, we have

$$\begin{aligned} K_n C_e^n \int_0^t 1_{\hat{X}_e^n(s) \ge -1} ds \le A_n + 1 + \beta_n + \|e\|\|\hat{X}^{0,n}\| + \|e\|\|\hat{W}^n\|_T^* \\ + \|e\| C_H \int_0^t \|\hat{X}^n(s)\| ds, \qquad t \le \tau_1^n \wedge T, \end{aligned}$$

where we denote $\beta_n = \|e\| C_H T \|\hat{N}^n\|$. Note that $\beta_n \to 0$. Hence by (2.47),

$$\|\hat{X}^n(t)\| \le C_2(A_n + 1) + C_2 \int_0^t \|\hat{X}^n(s)\| ds, \qquad t \le \tau_1^n \wedge T,$$

where C_2 does not depend on $\hat{X}^{0,n}$, $\hat{W}^n$, K_n or n. By Gronwall's lemma,

$$\|\hat{X}^n\|_{\tau_1^n \wedge T}^* \le C_2(A_n + 1) e^{C_2 T}. \tag{2.52}$$

Since on the event $\tau_1^n \ge T$ the property (2.51) follows from (2.52), we consider in what follows only the case $\tau_1^n < T$. If $\hat{X}_e^n(t) < -1$ for all $t \in [0, \tau_1^n]$, let $\tau_2^n = 0$; otherwise let

$$\tau_2^n = \sup\{t \in [0, \tau_1^n] : \hat{X}_e^n(t) \ge -1\}.$$

Note that $\hat{X}_e^n(\tau_2^n) \ge -1 - \|e\|\|\hat{X}^{0,n}\|$. Hence by (2.50)

$$-A_n + \|e\|\|\hat{X}^{0,n}\| \ge \hat{X}_e^n(\tau_1^n) - \hat{X}_e^n(\tau_2^n) = \hat{W}_e^n(\tau_1^n) - \hat{W}_e^n(\tau_2^n) + \int_{\tau_2^n}^{\tau_1^n} \widetilde{H}_e^n(\hat{X}^n(s)) ds.$$

As a result,

$$A_n/2 \le \left| \int_{\tau_2^n}^{\tau_1^n} \widetilde{H}_e^n(\hat{X}^n(s)) ds \right| \le \|e\| C_H \tau_1^n \|\hat{X}^n\|_{\tau_1^n}^* + \beta_n.$$

This and (2.52) show $A_n/2 - \beta_n \le \|e\| C_H \tau_1^n C_2 (A_n+1) e^{C_2 T}$. Since $A_n \ge 1$, $A_n/(A_n+1) \ge 1/2$. As a result, there exists a deterministic constant $\delta > 0$ not depending on $\hat{X}^{0,n}, \hat{W}_n, K_n, n$ such that, provided that n is large enough,

$$\tau_1^n \ge \delta \quad \text{on the event } \bar{\tau}^n \ge T. \tag{2.53}$$

Combining (2.52) and (2.53) we have that $\|\hat{X}^n\|_\delta^* \le C_3(1 + \|\hat{X}^{0,n}\| + \|\hat{W}^n\|_T^*)$, where C_3 does not depend on $\hat{X}^{0,n}, \hat{W}^n, K_n, n$. In a similar fashion we have for $i = 1, 2, \ldots$

$$\|\hat{X}^n\|_{i\delta}^* \le C_3(1 + \|\hat{X}^n\|_{(i-1)\delta}^* + \|\hat{W}^n\|_T^*).$$

Assuming without loss that $C_3 \ge 1$ the above implies

$$\|\hat{X}^n\|_T^* \le (2C_3)^{T/\delta+1}(1 + \|\hat{X}^{0,n}\| + \|\hat{W}^n\|_T^*)$$

and proves (2.51).

Step 2. Recall that W denotes an (ℓ, Σ)-Brownian motion. We next show that

$$\lim_{n\to\infty} \mathbb{P}(\bar{\tau}^n \le T) = 0, \tag{2.54}$$

and

$$\hat{W}^n \Rightarrow W \qquad \text{on } [0,T]. \tag{2.55}$$

Let $A_i, i \in \mathcal{I}$ and $S_{ij}, (i,j) \in \mathcal{I}\times\mathcal{J}$ be mutually independent Brownian motions with mean zero and variances given by $EA_i^2(1) = \lambda_i C_{U,i}^2$, $ES_{ij}^2(1) = \mu_{ij}$. By Theorem 14.6 of [13],

$$(\hat{A}^n, \hat{S}^n) \Rightarrow (A, S) \quad \text{locally on compacts.} \tag{2.56}$$

By (2.5) and (2.6), $\Psi_{ij}^n(t) \le N_j^n$ for all i, j and t. Thus by (2.8) and (2.26), $\|\hat{W}^n\|_T^* \le \|\hat{A}^n\|_T^* + \|\hat{S}^n\|_{C_4 T}^* + \|\ell^n\| T$ for a suitable constant C_4. Hence $n^{-1/2}\|\hat{W}^n\|_T^*$ converges to zero in probability. By (2.51), $n^{-1/2}\|\hat{X}^n\|_{T\wedge\bar{\tau}^n}^*$ converges to zero in probability. Using the definition (2.46), this establishes (2.54). In turn, this shows that $n^{-1/2}\|\hat{X}^n\|_T^*$ converges to zero in probability. By (2.45), so do $n^{-1/2}\|\hat{Y}^n\|_T^*$ and $n^{-1/2}\|\hat{Z}^n\|_T^*$. Note that $G(x^*, \nu) = \psi^*$, as follows from (2.11) and (2.24). Using (2.42), linearity of the map G, (2.43) and (2.16)–(2.18) we thus obtain, for a suitable constant C_5,

$$\begin{aligned}
\|\bar{\Psi}^n(t) - \psi^*\|_T^* &\le n^{-1}\|G(X^n - nx^* - Y^n, N^n - n\nu - Z^n)\|_T^* + n^{-1/2}K_n \\
&\le C_5 n^{-1/2}(\|\hat{X}^n\|_T^* + \|\hat{Y}^n\|_T^* + \|\hat{N}^n\| + \|\hat{Z}^n\|_T^*) + n^{-1/2}K_n \\
&\to 0 \text{ in probability.}
\end{aligned} \tag{2.57}$$

Combining (2.26), (2.56) and (2.57), the claim (2.55) follows from the lemma on p. 151 of [13] regarding random change of time.

Step 3. Recall (2.19). We prove the theorem in the case $x \in D_1$. In this case, for all n large, $\hat{X}^{0,n} \in D_1$. Denote

$$\tau^n = \inf\{s \in [0,T] : \hat{X}_e^n(s) \geq -1/2\}.$$

Denote also $\Omega^n = \{\bar{\tau}^n > T\}$. By (2.54), $\lim_n \mathbb{P}(\Omega^n) = 1$. If we show

$$\lim_{n\to\infty} \mathbb{P}(\{\tau^n \leq T\} \cap \Omega^n) = 0, \tag{2.58}$$

it would follow that $\mathbb{P}(\hat{X}_e^n(t) \leq -1/2 \text{ for all } t \in [0,T]) \to 1$, and in turn, by (2.17) and (2.45), that $\mathbb{P}(Y^n(t) = 0 \text{ for all } t \in [0,T]) \to 1$, as $n \to \infty$. Hence in order to prove the theorem it suffices to prove (2.58).

To this end, note that the jumps of the process $\hat{W}^n$ are bounded by $n^{-1/2}$ and write

$$\begin{aligned}\mathbb{P}(\{\tau^n \leq T\} \cap \Omega^n) \leq \mathbb{P}(\{&\text{there exist } 0 \leq s < t \leq T \text{ such that} \\ &-1 \leq \hat{X}_e^n(\theta) \leq -1/2 \text{ for all } \theta \in [s,t], \\ &\hat{X}_e^n(s) \leq -7/8 \text{ and } \hat{X}_e^n(t) \geq -5/8\} \cap \Omega^n).\end{aligned} \tag{2.59}$$

Under the event indicated immediately above, on the (random) interval $[s,t]$, $\hat{X}_e^n \geq -1$, and thus by (2.50),

$$\frac{1}{4} \leq \hat{X}_e^n(t) - \hat{X}_e^n(s) = \int_s^t \widetilde{H}_e^n(\hat{X}^n(\theta))d\theta + \hat{W}_e^n(t) - \hat{W}_e^n(s) - C_e^n K_n(t-s). \tag{2.60}$$

Moreover, by (2.48) and (2.51),

$$\int_s^t \widetilde{H}_e^n(\hat{X}^n(\theta))d\theta \leq C_H\|e\|C_1(t-s)[r + \|W_n\|_T^*] + \beta_n,$$

where $r = 1 + \sup_n \|\hat{X}^{0,n}\|$. Since $K_n \to \infty$, it follows that there is a deterministic n_0 such that for all $n \geq n_0$

$$\begin{aligned}\Delta_n(s,t) &:= \hat{W}_e^n(t) - \hat{W}_e^n(s) + C_H\|e\|C_1(t-s)\|\hat{W}^n\|_T^* + \beta_n \\ &\geq \frac{1}{4} + [C_e^n K_n - C_H\|e\|C_1 r](t-s) \\ &\geq \frac{1}{4} + C_0 K_n(t-s) \\ &\geq \begin{cases} C_0 K_n^{1/2}, & t-s \geq K_n^{-1/2}, \\ 1/4, & t-s < K_n^{-1/2}, \end{cases}\end{aligned} \tag{2.61}$$

where C_0 is as in (2.49), and on the first inequality we used the fact that $C_H\|e\|C_1 r$ does not depend on n. Combining (2.59) and (2.61), we see that there are constants $C_6, C_7 > 0$ not depending on n and K_n such that

$$\begin{aligned}
&\mathbb{P}(\{\tau^n \le T\} \cap \Omega^n) \\
&\quad \le \mathbb{P}(\text{there exist } s < t \le T \text{ such that } \Delta_n(s,t) \ge C_0 K_n^{1/2}) \\
&\qquad + \mathbb{P}(\text{there exist } s < t \le T \text{ such that } t - s < K_n^{-1/2},\ \Delta_n(s,t) \ge 1/4) \\
&\quad \le \mathbb{P}(\|\hat{W}_n\|_T^* \ge C_6 K_n^{1/2}) \\
&\qquad + \mathbb{P}(w_T(\hat{W}_n, K_n^{-1/2}) \ge 1/8 - \beta_n) + \mathbb{P}(C_H\|e\|C_1 K_n^{-1/2}\|\hat{W}_n\|_T^* \ge 1/8) \\
&\quad \le 2\mathbb{P}(\|\hat{W}_n\|_T^* \ge C_7 K_n^{1/2}) + \mathbb{P}(w_T(\hat{W}_n, K_n^{-1/2}) \ge 1/8 - \beta_n),
\end{aligned}$$

where $w_T(f,\delta)$ denotes the modulus of continuity of a function f on $[0,T]$. Since by (2.55) $\hat{W}^n$ are tight, and the weak limit process has continuous sample paths, (2.58) follows (see e.g., Theorem 16.8 of [13]) and hence the result.
Step 4. Finally, we prove the theorem in the case $x \in D_1^c$. Let $\tau_3^n = \inf\{t \in [0,T] : \hat{X}_e^n < -1\}$. Then

$$\mathbb{P}(\{\tau_3^n > \varepsilon\} \cap \Omega^n) \le \mathbb{P}(\{\hat{X}_e^n(\theta) \ge -1 \text{ for all } \theta \in [0,\varepsilon]\} \cap \Omega^n).$$

An argument similar to step 3 (only simpler) shows that under the event indicated on the r.h.s. of the above display,

$$\Delta_n(0,\varepsilon) \ge -1 - r + [C_e^n K_n - C_H\|e\|C_1 r]\varepsilon,$$

and in turn that $\mathbb{P}(\tau_3^n > \varepsilon)$ converges to zero as $n \to \infty$. We can now review the argument of step 3, replacing $\hat{X}^{0,n}$ by $\hat{X}^n(\tau_3^n)$ and τ^n by $\tau_4^n := \inf\{t \in [\tau_3^n, T] : \hat{X}_e^n(t) \ge -1/2\}$ (where $\tau_4^n = \infty$ on the event $\tau_3^n > T$) so as to show that $\mathbb{P}(\tau_4^n \le T) \to 0$ as $n \to \infty$. The only issue that is different now is that the "initial condition" $\hat{X}^n(\tau_3^n)$ is random and cannot be bounded by a deterministic constant r. Instead, let us define the random variable $r^n := 1 + \|\hat{X}^n(\tau_3^n)\|$ and let $\Omega_1^n := \{r^n \le C_0(C_H\|e\|C_1)^{-1}K_n\} \cap \{\tau_3^n \le T\}$. Coming back to (2.61) with r^n in place of r, it is clear that the second inequality will hold on Ω_1^n, and so the remainder of the argument of step 3 is valid once Ω^n is replaced by $\Omega^n \cap \Omega_1^n$. Since $\mathbb{P}(\tau_3^n \le T)$ converges to one, the relations (2.51), (2.54) and (2.55) imply that the random variables r^n are tight, and therefore $\mathbb{P}(\Omega_1^n \cap \Omega^n) \to 1$ as $n \to \infty$. This establishes the theorem. □

2.4. The non-preemptive case

In this section we treat the non-preemptive case and prove Theorem 2.4. Recall that we only consider the case $I = J = 2$. Thus $\mathcal{I} = \{1,2\}$ and $\mathcal{J} = \{3,4\}$.

The heavy traffic and complete resource pooling conditions, that are in force, imply that the graph of basic activities $\mathcal{G}_{ba}$ is a tree with vertex set $\mathcal{I} \cup \mathcal{J} = \{1, 2, 3, 4\}$ (cf. (2.12)). It follows that there are exactly 3 activities that are basic. Had we not had a fourth activity, the graph $\mathcal{G}_a$ would not contain a cycle and it would not be possible to fulfill the null controllability condition (2.33). Thus, the hypotheses of the theorem require that there be a fourth, non-basic activity. We have labeled the classes as '1' and '2'. Although $\mathcal{J} = \{3, 4\}$, it will be more natural in the discussion that follows, and throughout the section, to refer to the stations as 'station 1' and 'station 2', and with an abuse of notation regard the index set for the stations as $\{1, 2\}$. Accordingly, we have four activities, (i, j), $i, j \in \{1, 2\}$, and without loss of generality, we let $(2, 1)$ be the only non-basic activity (see Figure 2.3(a)). As a result, the direction of the only simple cycle, that we denote as c, is: class 2 $\to$ station 1 $\to$ class 1 $\to$ station 2 $\to$ class 2. By (2.21) we get $m_c^n = (\mu_{11}^n - \mu_{12}^n, -\mu_{21}^n + \mu_{22}^n)$ and $m_c = (\mu_{11} - \mu_{12}, -\mu_{21} + \mu_{22})$. In what follows we will write m^n for m_c^n and m for m_c. We also let $C_m^n = -e \cdot m^n$ and $C_m = -e \cdot m$. Note that condition (2.33) can be written as $C_m \equiv -e \cdot m > 0$.

We now specify the N-SCP for which the conclusions of Theorem 2.4 will be shown. According to Definition 2.2, we must specify the initial arrangement and how routing is determined upon each arrival. To this end we need some notation. Note that by (2.23) and (2.25) there exists a constant C'_H such that

$$\|H^n(a, b)\| \leq \frac{C'_H}{2I}(\|a\| + \|b\|), \qquad (a, b) \in D_G, n \in \mathbb{N}. \tag{2.62}$$

Let

$$\kappa = \frac{2 + 16C'_H}{C_m}, \qquad \delta = \frac{1}{8\kappa\|m\|} \wedge \frac{\log 2}{C'_H}, \qquad \gamma = \frac{\log 8}{\delta}. \tag{2.63}$$

The initial arrangement and the routing rules will depend on $\hat{X}^{0,n}$, and in particular on whether $e \cdot \hat{X}^{0,n} < -1$ or not. First consider the case where $e \cdot \hat{X}^{0,n} < -1$.

Initial arrangement. Recall that $X^n(0) = X^{0,n}$ is given and we have to specify $\Psi^n(0)$. We set

$$\Psi_{21}^n(0) = \lceil n^{5/8}\kappa \rceil, \tag{2.64}$$

$$\Psi_{11}^n(0) = \lceil (N_1^n - N_2^n + X_1^{0,n} + X_2^{0,n})/2 \rceil - \Psi_{21}^n(0), \tag{2.65}$$

$$\Psi_{12}^n(0) = \lfloor (N_2^n - N_1^n + X_1^{0,n} - X_2^{0,n})/2 \rfloor + \Psi_{21}^n(0), \tag{2.66}$$

$$\Psi_{22}^n(0) = X_2^{0,n} - \Psi_{21}^n(0). \tag{2.67}$$

Using (2.4), (2.5) one verifies that

$$Y_1^n(0) = Y_2^n(0) = 0, \quad |Z_1^n(0) - Z_2^n(0)| \leq 1. \tag{2.68}$$

Routing. The routing decisions can be based only on n, the value of Ξ^n right before the arrival, and some auxiliary randomness that we have denoted by R^n. The way we use the randomness in the proposed policy is so as to split the customers of class 2 into two sub-classes, that we label as α and β. Upon the kth class-2 arrival, an independent biased coin is tossed according to which it is decided what sub-class the arrival belongs to. The bias of the coin is allowed to depend on k. We denote by $\alpha_n(k)$ the probability that the kth class-2 arrival is classified as a class-α customer. The value of $\alpha_n(k)$ is determined as

$$\alpha_n(k) = \left\{ \frac{\kappa(\gamma + \mu_{21})}{\lambda_2 n^{3/8}} \exp\left[\frac{\gamma(k-1)}{n\lambda_2} \right] \right\} \wedge 1. \tag{2.69}$$

Since we assume that $\hat{X}^{0,n} < -1$ and the difference between $Z_1^n(0)$ and $Z_2^n(0)$ is at most 1, it follows that the initial arrangement is such that there are free servers in both stations. Below, we shall describe the routing policy only as long as there are free servers in both stations. The description of the policy at other times is not important and will be left completely open (in fact, one of the main ingredients of the proof will be to show that the event that there are no free servers in one of the stations some time before T has probability approaching zero as $n \to \infty$). The routing rules are as follows.

1. Class-1 customers are routed to the station with more free servers. More precisely, if a class-1 customer arrives at time t, it is instantaneously routed to station j, where

$$j = \begin{cases} 1 & \text{if } \hat{Z}_1^n(t-) > \hat{Z}_2^n(t-) \\ 2 & \text{otherwise.} \end{cases}$$

2. Class-α customers are routed to station 1 upon arrival.
3. Class-β customers are routed to station 2 upon arrival.

Next, consider the case where the initial condition satisfies $e \cdot \hat{X}^{0,n} \geq -1$. Denote $r_n = e \cdot X^{0,n} - e \cdot N^n + \lceil n^{1/2} \rceil$. For the initial arrangement, we let r_n class-1 customers be left in the queue, and let $\Psi_{ij}^n(0)$ be defined as in (2.64)–(2.67), except that we substitute $X^{0,n} - r_n$ for $X^{0,n}$. As a result, in place of the left part of (2.68) we will have $Y_1^n(0) = r_n$, $Y_2^n(0) = 0$. The routing is determined as follows. The r_n class-1 customers that are initially put in the queue are kept in the queue. Rules 1–3 above apply for all the other customers. Upon the first arrival after the time τ_0^n when the number of free servers in the

system first exceeds $r_n + \lceil n^{1/2} \rceil$, all the r_n customers are moved into service: γ_1^n into station 1 and γ_2^n into station 2. Here, $\gamma_1^n = \lceil (Z_2^n(\tau_0^n-) - Z_1^n(\tau_0^n-) + r_n)/2 \rceil$, $\gamma_2^n = \lfloor (Z_1^n(\tau_0^n-) - Z_2^n(\tau_0^n-) + r_n)/2 \rfloor$. Clearly, $e \cdot \hat{X}^n(\tau_0^n) \leq -1$. Also, by the above choice for γ_i^n, one achieves that $Y_i^n(\tau_0^n) = 0$ and $|Z_1^n(\tau_0^n) - Z_2^n(\tau_0^n)| \leq 1$, a situation similar to (2.68). From the time τ_0^n on, the routing rules 1–3 above apply for all the new arrivals. Note that once again we have ignored the scenario that there are no free servers in one of the stations at some time before T, for reasons as mentioned before.

Here and for most of this section we shall assume that $e \cdot x < -1$. Since $\hat{X}^{0,n} \to x$, we have

$$e \cdot \hat{X}^{0,n} < -1 \tag{2.70}$$

for all large n, and as a result only the first part of the definition of the policy will be relevant until, at the end of the section, we return to the treatment of the case $e \cdot x \geq -1$.

Without loss of generality assume that T is a multiple of δ, and set $\bar{k} = T/\delta$. Divide the time interval $[0, T]$ as follows: $I_0 = \{0\}$, $I_1 = (0, \delta]$, $\ldots$, $I_{\bar{k}} = ((\bar{k}-1)\delta, \bar{k}\delta]$. Let $\bar{h} : [0, T] \to \mathbb{R}$ be defined as

$$\bar{h}(t) = 2^{-k} \text{ for all } t \in I_k, \ k = 0, 1, \ldots, \bar{k}.$$

Define

$$\tau^n = \inf \left\{ t \geq 0 : \hat{X}_e^n(t) \geq -\frac{\bar{h}(t)}{2} \right\} \wedge T. \tag{2.71}$$

Let $h' = \frac{1}{32} 2^{-\bar{k}} \equiv \frac{1}{32} \bar{h}(T)$. Let

$$\sigma^n = \inf\{t \geq 0 : \ \hat{Z}_1^n(t) \wedge \hat{Z}_2^n(t) \leq 4h'\} \tag{2.72}$$

and $\zeta^n = \tau^n \wedge \sigma^n$. By (2.70) and since $\hat{Z}_1^n(0)$ and $\hat{Z}_2^n(0)$ are nearly equal (cf. (2.68)), it is useful to note that

$$\hat{Z}_1^n(0) \wedge \hat{Z}_2^n(0) \geq 1/4. \tag{2.73}$$

Also, it is clear that there are idle servers in both stations up to time σ^n. Denote

$$D(a, p) = \{\xi \in \mathbb{R}^I : \|\xi\| \leq a, \ e \cdot \xi \leq -p\}, \qquad a > 0, \ p > 0.$$

Let $K_n = n^{1/8}$ and

$$D_k^n = D(e^{\gamma \delta k} K_n, 2^{-k}).$$

For $k = 0, 1, \ldots, \bar{k}$, denote

$$\Omega_k^n = \{\hat{X}^n(t) \in D_{k'}^n, \ t \in I_{k'}, \text{ for all } k' \leq k\} \cap \{\zeta^n > k\delta\}. \tag{2.74}$$

Proposition 2.1. *Let the assumptions of Theorem 2.4 hold and assume also that* $e\cdot x < -1$. *Then for* $k = 0, 1, \ldots, \bar{k}$, $\mathbb{P}(\Omega_k^n) \to 1$.

The above result implies that $\mathbb{P}(\sigma^n \leq T) \to 0$. Since by construction of the policy, $Y^n(t) = 0$ for $t \leq \sigma^n$, this establishes Theorem 2.4 in this case.

In preparation for the proof of Proposition 2.1 we need some notation and preliminary results. Recall that we have denoted the only cycle by c, and that according to our notation $\Psi_c^n(t) \equiv \Psi_{21}^n(t)$ and $\hat{\Psi}_c^n(t) \equiv \hat{\Psi}_{21}^n(t)$. Let $A_\alpha^n(t)$ $[A_\beta^n(t)]$ denote the number of class-α [respectively, class-β] customers that have arrived up to time t. Let

$$\hat{A}_\alpha^n(t) = n^{-1/2} A_\alpha^n(t) - K_n \int_0^t (\mu_{21} + \gamma)\psi_0(s)ds, \tag{2.75}$$

where

$$\psi_0(t) = \kappa e^{\gamma t}, \qquad t \geq 0. \tag{2.76}$$

A sequence of processes is said to be C-tight if it is tight and every subsequential limit has continuous sample paths with probability one. The following three lemmas will be proved later in this section.

Lemma 2.1. *Under the assumptions of Proposition 2.1 one has*

$$|\hat{A}_\alpha^n|_T^* \to 0 \quad \textit{in probability}, \tag{2.77}$$

$$\|\bar{X}^n - x^*\|_{T\wedge\sigma^n}^* \equiv n^{-1/2}\|\hat{X}^n\|_{T\wedge\sigma^n}^* \to 0 \quad \textit{in probability.} \tag{2.78}$$

$$\|\bar{\Psi}^n - \psi^*\|_{T\wedge\sigma^n}^* \equiv n^{-1/2}\|\hat{\Psi}^n\|_{T\wedge\sigma^n}^* \to 0 \quad \textit{in probability.} \tag{2.79}$$

In addition, the processes $\hat{A}_i^n(\cdot \wedge \sigma^n)$, $\hat{S}_{ij}^n \left(\int_0^{\cdot\wedge\sigma^n} \bar{\Psi}_{ij}^n(s)ds\right)$ *are* C*-tight. In particular, one has* $|\hat{S}_{21}^n \left(\int_0^{\cdot} \bar{\Psi}_{21}^n(s)ds\right)|_{T\wedge\sigma^n}^* \to 0$ *in probability.*

Lemma 2.2. *Under the assumptions of Proposition 2.1,* $|\hat{\Psi}_c^n - K_n\psi_0|_{T\wedge\sigma^n}^* \to 0$ *in probability, where* ψ_0 *is as in* (2.76).

Lemma 2.3. *Under the assumptions of Proposition 2.1,* $\lim_{n\to\infty} \mathbb{P}(\sigma^n \leq \tau^n) = 0$.

The following will be used several times in the proofs of both Proposition 2.1 and the above lemmas. Recall that for $t \leq \sigma^n$, $Y^n(t) = 0$. Hence have from (2.27)

$$\hat{X}^n(t) = \hat{X}^{0,n} + \hat{W}^n(t) + \int_0^t [H^n(s) + m^n\hat{\Psi}_c^n(s)]ds, \qquad t \leq \sigma^n, \tag{2.80}$$

where we have denoted

$$H^n(s) = H^n(\hat{X}^n(s), -\hat{Z}^n(s)). \tag{2.81}$$

Also, substituting zero for $\hat{Y}^n$ in (2.28), using (2.14) and positivity of $\hat{Z}^n_j$, we have for all n large

$$\|\hat{Z}^n(t)\| \le \|\hat{X}^n(t)\| + 1, \qquad t \le \sigma^n. \tag{2.82}$$

Thus by (2.25), (2.62), (2.81) and (2.82), for all n large,

$$\|H^n(s)\| \vee \|e \cdot H^n(s)\| \le C'_H(1 + \|\hat{X}^n(s)\|), \qquad t \le \sigma^n. \tag{2.83}$$

Proof of Proposition 2.1. We argue by induction on k. As the induction base, consider $k = 0$. By (2.70) and since $\hat{X}^{0,n}$ converge, we have $\hat{X}^n(0) \in D^n_0$ for all large n. By Lemma 2.3, it remains to show that $\tau^n > 0$ with probability approaching 1. This follows from (2.71) and the fact that the jumps of $\hat{X}^n$ are bounded by $n^{-1/2}$.

Next consider the induction step. Assuming that for a given $k < \bar{k}$ one has

$$\lim_{n\to\infty} \mathbb{P}(\Omega^n_k) = 1, \tag{2.84}$$

we shall prove

$$\lim_{n\to\infty} \mathbb{P}(\Omega^n_{k+1}) = 1. \tag{2.85}$$

In view of Lemma 2.3, it suffices to show

$$\lim_{n\to\infty} \mathbb{P}(\tau^n \le (k+1)\delta) = 0, \tag{2.86}$$

and

$$\lim_{n\to\infty} \mathbb{P}(\|\hat{X}^n\|^*_{\delta(k+1)\wedge\sigma^n} \le e^{\gamma\delta(k+1)} K_n) = 1. \tag{2.87}$$

To this end, note that from (2.80) we have

$$\hat{X}^n_e(t) = \hat{X}^{0,n}_e + \hat{W}^n_e(t) + \int_0^t [H^n_e(s) - C^n_m \hat{\Psi}^n_c(s)]ds, \qquad t \le \sigma^n. \tag{2.88}$$

By (2.63), one verifies that the constants κ, δ and γ satisfy

$$e^{C'_H\delta} \le 2, \qquad 1 + C'_H e^{\gamma\delta} \le \frac{1}{2} C_m \kappa, \tag{2.89}$$

$$4(1 + \|m\|\kappa\delta e^{\gamma\delta}) \le e^{\gamma\delta}. \tag{2.90}$$

Denote

$$\widetilde{\Omega}_k^n = \Omega_k^n \cap \{\sigma^n > \tau^n\} \cap \{\tau^n \le (k+1)\delta\}.$$

Let $h = \frac{1}{8}2^{-k}$. On Ω_k^n (hence on $\widetilde{\Omega}_k^n$) the following must hold:

$$\|\hat{X}^n(k\delta)\| \le e^{\gamma\delta k}K_n, \qquad \hat{X}_e^n(k\delta) \le -8h. \tag{2.91}$$

In addition, on $\widetilde{\Omega}_k^n$,

$$\text{there exist } k\delta < s < t \le (k+1)\delta \wedge \sigma^n \text{ such that}$$
$$\hat{X}_e^n(s) \le -7h,\ \hat{X}_e^n(t) \ge -5h.$$

Hence in view of (2.88), the following must hold on $\widetilde{\Omega}_k^n$:

$$2h \le \hat{X}_e^n(t) - \hat{X}_e^n(s) = \hat{W}_e^n(t) - \hat{W}_e^n(s) + \int_s^t [H_e^n(u) - C_m^n\hat{\Psi}_c^n(u)]du. \tag{2.92}$$

Denote $a_n = |\hat{\Psi}_c^n - K_n\psi_0|^*_{T\wedge\sigma^n}$, $b_n = \|\hat{W}^n\|^*_{T\wedge\sigma^n}$ and $d_{u,v}^n = \hat{W}_e^n(v) - \hat{W}_e^n(u)$. Let

$$\hat{\Omega}_k^n = \Omega_k^n \cap \{C_m a_n \le 2\} \cap \{2b_n + 2\delta\|m\|a_n + T \le K_n\}. \tag{2.93}$$

By Lemmas 2.1–2.3 and (2.84),

$$\lim_{n\to\infty} \mathbb{P}(\hat{\Omega}_k^n) = 1. \tag{2.94}$$

On the event $\hat{\Omega}_k^n$, by (2.80), (2.83), denoting

$$\Lambda^n = \|\hat{X}^n(k\delta)\| + 2b_n + \delta\|m^n\|\,|\hat{\Psi}_c^n|^*_{T\wedge\sigma^n} + T$$

we have

$$\|\hat{X}^n(u)\| \le \Lambda^n + \int_{k\delta}^u C'_H\|\hat{X}^n(u')\|du', \quad u \le (k+1)\delta\wedge\sigma^n.$$

By Gronwall's inequality and the first part of (2.89) this shows that $\|\hat{X}^n(u)\| \le 2\Lambda^n$, for u as above. Using this along with (2.93), and assuming in what follows that n is so large that $\|m^n\| \le 2\|m\|$ and $C_m^n \ge C_m/2$, we have on the event $\hat{\Omega}_k^n$

$$\|\hat{X}^n(u)\| \le 4K_n e^{\gamma\delta k}(1 + \|m\|\kappa\delta e^{\gamma\delta}) \le K_n e^{\gamma\delta(k+1)}, \quad u \le (k+1)\delta\wedge\sigma^n, \tag{2.95}$$

where in the last inequality we used (2.90). Equations (2.94) and (2.95) establish (2.87). Next, combining (2.92) and (2.95), and using again (2.93), we have on the event $\widetilde{\Omega}_k^n \cap \hat{\Omega}_k^n$:

$$2h \le d_{s,t}^n + (C'_H + 1)(t-s) + K_n(t-s)e^{\gamma\delta k}[C'_H e^{\gamma\delta} - C_m\kappa/2].$$

Writing $C = C'_H + 1$ and using the second part of (2.89), we conclude that for all large n, on the event $\widetilde{\Omega}^n_k \cap \hat{\Omega}^n_k$, there exist s and t such that $k\delta < s < t \leq (k+1)\delta \wedge \sigma^n$ and

$$d^n_{s,t} \geq 2h + (e^{\gamma\delta k}K_n - C)(t-s) \geq \begin{cases} \frac{1}{2}K_n^{1/2}, & t-s \geq K_n^{-1/2}, \\ 2h, & t-s < K_n^{-1/2}. \end{cases}$$

Hence for all large n

$$\begin{aligned}
&\mathbb{P}(\widetilde{\Omega}^n_k) \\
&\leq \mathbb{P}((\hat{\Omega}^n_k)^c) + \mathbb{P}(\text{ there exist } 0 \leq s < t \leq T \wedge \sigma^n \text{ such that } 2d^n_{s,t} \geq K_n^{1/2}) \\
&\quad + \mathbb{P}(\text{ there exist } 0 \leq s < t \leq T \wedge \sigma^n \text{ such that } t-s < K_n^{-1/2} \text{ and } d^n_{s,t} \geq 2h) \\
&\leq \mathbb{P}((\hat{\Omega}^n_k)^c) + \mathbb{P}(4b_n \geq K_n^{1/2}) + \mathbb{P}(w_{T\wedge\sigma^n}(\hat{W}_n, K_n^{-1/2}) \geq h).
\end{aligned}$$

By (2.94), the tightness of b_n and the C-tightness statement in Lemma 2.1, we see that $\mathbb{P}(\widetilde{\Omega}^n_k) \to 0$ as $n \to \infty$. By (2.84) and Lemma 2.3, this shows (2.86). □

Let a constant $r > 0$ be given and let $\theta^n = \inf\{t \in [0,\delta] \ : \hat{X}^n_e(t) \leq -r\}$. Let ε be as in Theorem 2.4. The following consequence of the above proof will be useful.

Corollary 2.1. *Under the assumptions of Proposition 2.1,* $\lim_{n\to\infty} \mathbb{P}(\theta^n > \varepsilon/2) = 0$. *In addition,* $\|\hat{X}^n\|^*_{\theta^n}$ *are tight.*

Proof. With the notation from (2.74) and (2.93), define $\Omega^n = \{\theta^n > \varepsilon/2\} \cap \hat{\Omega}^n_1$. By Lemmas 2.1–2.3 and Proposition 2.1, $\lim_{n\to\infty} \mathbb{P}(\hat{\Omega}^n_1) = 1$ and therefore it suffices to show that $\mathbb{P}(\Omega^n) \to 0$. On Ω^n we have

$$-r \leq \hat{X}^n_e(\varepsilon) - \hat{X}^n_e(0) = \hat{W}^n_e(\varepsilon) - \hat{W}^n_e(0) + \int_0^\varepsilon [H^n_e(u) - C^n_m \hat{\Psi}^n_c(u)]du,$$

and an argument along the lines of the proof of Proposition 2.1 proves the first claim of the result.

Next we prove the tightness statement. Note that by assumption $\|\hat{X}^n(0)\| \leq C_0$ for some deterministic constant C_0 independent on n. Let $\varsigma^n = (2r + \|e\|C_0)e^{-\gamma\delta}K_n^{-1}$ and $\widetilde{\Omega^n} = \{\theta^n > \varsigma^n\} \cap \hat{\Omega}^n_1$. On $\widetilde{\Omega^n}$ we have (just as in the proof of Proposition 2.1)

$$\begin{aligned}
-r \leq \hat{X}^e_n(\varsigma^n) &= \hat{X}^n_e(0) + \hat{W}^n_e(\varsigma^n) + \int_0^{\varsigma^n} [H^n_e(u) - C^n_m \hat{\Psi}^n_c(u)]du \\
&\leq \hat{X}^n_e(0) + \hat{W}^n_e(\varsigma^n) + (C'_H + 1)\varsigma^n - K_n\, \varsigma^n e^{\gamma\delta} \\
&\leq \hat{W}^n_e(\varsigma^n) + (C'_H + 1)\varsigma^n - 2r.
\end{aligned}$$

As a result,

$$\lim_{n\to\infty} \mathbb{P}(\theta^n > \varsigma^n) = 0. \tag{2.96}$$

Let $\widetilde{C} = C_0 + 4r\|m\|\kappa + 2C_H r + 1$. On $\hat{\Omega}_1^n$, for n so large that $\varsigma^n < \delta$, with the notation $a^n = |\Psi_c^n - K_n\psi_0|^*_{T\wedge\sigma^n}$, we have

$$\begin{aligned}\|\hat{X}^n\|^*_{\varsigma^n} &\le \|\hat{X}^n(0)\| + w_T(\hat{W}^n, \varsigma^n) + \|m^n\|\delta a^n + \|m^n\|\kappa K_n e^{\gamma\delta}\varsigma^n + C_H K_n e^{\gamma\delta}\varsigma^n \\ &\le w_T(\hat{W}^n, \varsigma^n) + 2\|m\|\delta a^n + \widetilde{C} - 1.\end{aligned}$$

Since by Lemmas 2.1–2.2, we have $\lim_{n\to\infty}\mathbb{P}(w_T(\hat{W}^n,\varsigma^n)+2\|m\|\delta a^n > 1) = 0$, we conclude that $\lim_{n\to\infty}\mathbb{P}(\|\hat{X}^n\|^*_{\varsigma^n} \le \widetilde{C}) = 1$. In view of (2.96), this shows that $\|\hat{X}^n\|^*_{\theta^n}$ are tight. □

We turn to prove Lemmas 2.1–2.3.

Proof of Lemma 2.1. We first prove (2.77). Let R_k^n denote the indicator of the event that in the nth system, the kth class-2 arrival was classified as an arrival of class α. Then $ER_k^n = \mathbb{P}(R_k^n = 1) = \alpha_n(k)$ given in (2.69). Since $A_\alpha^n(t)$ represents the number of class-2 arrivals up to time t that were classified as class α, we can write $A_\alpha^n(t) = \sum_{k=1}^{A^n(t)} R_k^n$. By (2.75) we have

$$\hat{A}_\alpha^n(t) = n^{-1/2}A_\alpha^n(t) - n^{1/8}\,\frac{\lambda_2 C}{\gamma}\,(e^{\gamma t} - 1),$$

where throughout this proof $C = (\mu_{21} + \gamma)\kappa/\lambda_2$. Fix $\varepsilon > 0$. Then $\hat{A}_\alpha^n(t) \ge \varepsilon$ if and only if $A_\alpha^n(t) \ge n^{1/2}\varepsilon + K_t^n$, where

$$K_t^n = n^{5/8}\,\frac{\lambda_2 C}{\gamma}\,(e^{\gamma t} - 1).$$

Therefore

$$\mathbb{P}(\text{ there exists } t \le T, \text{ such that } \hat{A}_\alpha^n(t) \ge \varepsilon) \le p_1^n + p_2^n, \tag{2.97}$$

where

$$p_1^n = \mathbb{P}(\sup_{t\le T}|A_2^n(t) - \lambda_2^n t| \ge n^{3/4}) \to 0 \text{ as } n \to \infty \tag{2.98}$$

$$\begin{aligned}p_2^n = \mathbb{P}(&\sup_{t\le T}|A_2^n(t) - \lambda_2^n t| < n^{3/4}, \text{ and there exists } t \le T, \\ &\text{such that } \sum_{k=1}^{A^n(t)} R_k^n \ge n^{1/2}\varepsilon + K_t^n) \\ \le \mathbb{P}(&\text{ there exists } t \le T, \text{ such that } \sum_{k=1}^{\beta(n,t)} R_k^n - K_t^n \ge n^{1/2}\varepsilon),\end{aligned}$$

where $\beta(n,t) = \lfloor \lambda_2^n t + n^{3/4} \rfloor$. The convergence statement in (2.98) is due to the tightness of $\hat{A}_2^n$ (cf. (2.56)). A direct calculation based on (2.69) shows

$$\sum_{k=1}^{\beta(n,t)} E(R_k^n) \leq K_t^n + C' n^{3/8},$$

for all large n, where C' is a deterministic constant not depending on n. Hence

$$p_2^n \leq \mathbb{P}(\text{there exists } t \leq T \text{ such that } \sum_{k=1}^{\beta(n,t)} (R_k^n - ER_k^n) \geq n^{1/2}\varepsilon - C' n^{3/8}).$$

Denoting $m(l) = \sum_{k=1}^{l} (R_k^n - ER_k^n)$ we can write

$$p_2^n \leq \mathbb{P}(|m|^*_{2\lambda^n T} \geq n^{1/2}\varepsilon/2) \leq 16 \frac{E[m(2\lambda_2^n T)^2]}{\varepsilon^2 n},$$

where in the last step we used Doob's inequality for the martingale m. In turn, $E[m(l)^2] = \sum_{k=1}^{l} \alpha_n(k)$, and substituting $2\lambda_2^n T$ for l one finds that $p_2^n \to 0$. As a result, the l.h.s. of (2.97) converges to zero. A similar calculation, that we omit, shows that the probability that there exists $t \leq T$ such that $\hat{A}_\alpha^n(t) \leq -\varepsilon$ also converges to zero. Since ε is arbitrary, (2.77) follows.

Exactly as in the proof of Theorem 2.3 we have

$$n^{-1/2} \hat{W}^n \Rightarrow 0. \tag{2.99}$$

Since $\hat{X}^{0,n}$ converges, we have from (2.80), (2.81) and (2.83), for all large n,

$$\|\hat{X}^n(t)\| \leq K_n + \|\hat{W}^n(t)\| + C'_H \int_0^t (1 + \|\hat{X}^n(s)\|) ds + T\|m^n\| \, \|\hat{\Psi}_c^n\|_t^*, \qquad t \leq \sigma^n. \tag{2.100}$$

Using the inequality

$$\hat{\Psi}_c^n(t \wedge \sigma^n) \leq \hat{\Psi}_c^n(0) + n^{-1/2} A_\alpha^n(T) \tag{2.101}$$

and Gronwall's inequality we have

$$\|\hat{X}^n(t)\| \leq C_1(K_n + \|\hat{W}^n(t)\| + n^{-1/2} A_\alpha^n(T) + 1), \qquad t \leq \sigma^n \wedge T, \tag{2.102}$$

for an appropriate constant C_1 not depending on n. It follows from (2.77) that

$$\lim_{n \to \infty} \mathbb{P}(A_\alpha^n(T) \geq n^{3/4}) = 0. \tag{2.103}$$

Combining (2.99), (2.102) and (2.103) we have (2.78). Using (2.4), (2.5), (2.11), (2.16)–(2.18) and the fact that $\hat{Y}^n(t) = 0$ we have, for all $t \le \sigma^n$,

$$\hat{\Psi}^n_{11}(t) + \hat{\Psi}^n_{12}(t) = \hat{X}^n_1(t),\ \hat{\Psi}^n_c(t) + \hat{\Psi}^n_{22}(t) = \hat{X}^n_2(t),\ \hat{Z}^n_1(t) + \hat{\Psi}^n_{11}(t) + \hat{\Psi}^n_c(t) = \hat{N}^n_1.$$

These equations along with (2.82) show that

$$\|\hat{\Psi}^n(t)\| \le C_2(1 + \|\hat{X}^n(t)\| + \hat{\Psi}^n_c(t)), \quad t \le \sigma^n, \tag{2.104}$$

for a constant C_2 not depending on n. Combining (2.101), (2.103) and (2.78) we have (2.79). The result (2.79) implies that the processes $\bar{\Psi}^n_{ij}(\cdot \wedge \sigma^n)$ are tight and that every subsequential limit has continuous sample paths with probability one. Using this and the time change lemma [13] along the lines of the proof of Theorem 2.3 above proves C-tightness as claimed. Finally, consider the last claim in the statement of the lemma. Since $\psi^*_{21} = 0$, (2.79) shows that $|\bar{\Psi}^n|^*_{T\wedge\sigma^n} \to 0$ in probability. Using this and (2.56), the claim follows using again the time change lemma. □

Proof of Lemma 2.2. By construction of the policy, $B^n_{21}(t) = A^n_\alpha(t)$ for all $t \le \sigma^n$. Thus by (2.7),

$$\Psi^n_{21}(t) = \Psi^n_{21}(0) + A^n_\alpha(t) - S^n_{21}\left(\int_0^t \Psi^n_{21}(s)ds\right).$$

Denote

$$\varrho(t) = (\mu_{21} + \gamma)\psi_0(t), \quad t \ge 0. \tag{2.105}$$

Recalling that $\hat{\Psi}^n_c(t) = n^{-1/2}\Psi^n_c(t) = n^{-1/2}\Psi^n_{21}(t)$, using (2.75), one checks by direct calculation that

$$\hat{\Psi}^n_c(t) = \hat{\Psi}^n_c(0) + n^{1/8}\int_0^t \varrho(s)ds - \mu^n_{21}\int_0^t \hat{\Psi}^n_c(s)ds + \widetilde{W}^n_0(t), \qquad t \le \sigma^n,$$

where

$$\widetilde{W}^n_0(t) = \hat{A}^n_\alpha(t) - \hat{S}^n_{21}\left(\int_0^t \bar{\Psi}^n_{21}(s)ds\right).$$

Note that $\psi_0(t) = \kappa + \int_0^t \varrho(s)ds - \mu_{21}\int_0^t \psi_0(s)ds$, and let ψ^n_0 be the unique solution to

$$\psi^n_0(t) = \kappa + \int_0^t \varrho(s)ds - \mu^n_{21}\int_0^t \psi^n_0(s)ds. \tag{2.106}$$

Then, for $t \le \sigma^n$,

$$\hat{\Psi}^n_c(t) - n^{1/8}\psi^n_0(t) = \hat{\Psi}^n_c(0) - n^{1/8}\kappa - \mu^n_{21}\int_0^t [\hat{\Psi}^n_c(s) - n^{1/8}\psi^n_0(s)]ds + \widetilde{W}^n_0(t).$$

By (2.64), $\hat{\Psi}_c^n(0) - n^{1/8}\kappa \to 0$. By Lemma 2.1, $|\widetilde{W}_0^n|^*_{T\wedge\sigma^n} \to 0$ in probability. An application of Gronwall's inequality therefore shows that

$$|\hat{\Psi}_c^n - n^{1/8}\psi_0^n|^*_{T\wedge\sigma^n} \to 0 \quad \text{in probability.} \tag{2.107}$$

Also, by (2.13) and (2.106) it is easy to see that $\delta^n(t) \le C\varepsilon_n + C\int_0^t \delta^n(s)ds$ for $t \le T$, where $\delta^n(t) = |\psi_0(t) - \psi_0^n(t)|$, C is a constant and $n^{1/8}\varepsilon_n \to 0$. Hence $n^{1/8}|\delta^n|^*_T \to 0$. Along with (2.107), this proves the lemma. □

Proof of Lemma 2.3. By (2.73) the initial values for both $\hat{Z}_1^n$ and $\hat{Z}_2^n$ are greater than $8h'$. We thus have $\{\sigma^n \le \tau^n\} \subseteq \Omega_1^n \cup \Omega_2^n$, where for $i = 1, 2$,

$$\begin{aligned}\Omega_i^n = \{&\text{there exist } u_1 < u_2 \le \zeta^n \text{ such that} \\ &\sup_{u\in[u_1,u_2]} \hat{Z}_i^n(u) \le 7h',\ \ \hat{Z}_i^n(u_1) \ge 6h',\ \hat{Z}_i^n(u_2) \le 5h'\}.\end{aligned}$$

Step 1. We show that $\mathbb{P}(\Omega_1^n) \to 0$. Let u_1 and u_2 be as specified in the expression for Ω_1^n. Note that the size of the jumps of $\hat{X}^n$ is bounded by $n^{-1/2}$. Hence for $u \le \tau^n$ we have $\hat{X}_e^n(u) \le -32h' + n^{-1/2}$. This, combined with (2.14), (2.28), and the fact that $Y_i^n(u) = 0$ for $u \le \sigma^n$ imply that $\hat{Z}_1^n(u) < \hat{Z}_2^n(u)$ for $u \in [u_1, u_2]$. According to our definition of the policy, all class-1 customers are routed to station 2 during the period $[u_1, u_2]$ and only class-α customers are routed to station 1 during this period. As a result, $B_{11}^n(u_2) - B_{11}^n(u_1) = 0$ and $B_{21}^n(u_2) - B_{21}^n(u_1) = A_\alpha^n(u_2) - A_\alpha^n(u_1)$. By (2.7), we have

$$\sum_{i=1,2} (\Psi_{i1}^n(u_2) - \Psi_{i1}^n(u_1)) = A_\alpha^n(u_2) - A_\alpha^n(u_1) - D_1^n(u_1, u_2),$$

where

$$D_j^n(t_1, t_2) := \sum_{i=1,2} \left[S_{ij}^n \left(\int_0^{t_2} \Psi_{ij}^n(s)ds \right) - S_{ij}^n \left(\int_0^{t_1} \Psi_{ij}^n(s)ds \right) \right], \qquad j = 1, 2.$$

Note that $D_j^n(t_1, t_2)$ represents the number of departures from station j during $(t_1, t_2]$. Using (2.5) and (2.14), one has on Ω_1^n that

$$h'n^{1/2} \le A_\alpha^n(u_2) - A_\alpha^n(u_1) - D_1^n(u_1, u_2). \tag{2.108}$$

Denoting, for $j = 1, 2$,

$$\widetilde{W}_j^n(t) := \sum_{i=1,2} \hat{S}_{ij}^n \left(\int_0^t \bar{\Psi}_{ij}^n(s)ds \right),$$

$$\widetilde{S}_j^n(t_1,t_2) = \widetilde{W}_j^n(t_2) - \widetilde{W}_j^n(t_1) + \sum_{i=1,2}\left[\mu_{ij}^n \int_{t_1}^{t_2} \hat{\Psi}_{ij}^n(s)ds + n^{1/2}\mu_{ij}^n \psi_{ij}^*(t_2-t_1)\right], \tag{2.109}$$

and

$$\widetilde{A}_1^n(t_1,t_2) = \hat{A}_\alpha^n(t_2) - \hat{A}_\alpha^n(t_1) + n^{5/8}\int_{t_1}^{t_2} \varrho(s)ds,$$

where ϱ is as in (2.105), one checks by direct calculation that (2.108) can be written as $h' \leq \widetilde{A}_1^n(u_1,u_2) - \widetilde{S}_1^n(u_1,u_2)$. Note that ϱ is bounded above. Moreover, $\mu_{11}^n\psi_{11}^*$ is bounded below by a positive constant. Using (2.104) and denoting $\Delta = u_2 - u_1$ we therefore have, for all large n,

$$h' \leq 2|\hat{A}_\alpha^n|_T^* + w_{T\wedge\sigma^n}(\widetilde{W}_1^n, \Delta) - n^{1/2}\Lambda_1^n\Delta, \tag{2.110}$$

where

$$\Lambda_1^n = C_1 - C_2 n^{-1/2}(1 + \|\hat{X}^n\|_{T\wedge\sigma^n}^* + |\hat{\Psi}_c^n|_{T\wedge\sigma^n}^*),$$

and $C_1 > 0$ and C_2 are suitable constants. We conclude that $\mathbb{P}(\Omega_1^n) \leq \mathbb{P}(\Omega_{1,1}^n) + \mathbb{P}(\Omega_{1,2}^n)$, where

$$\Omega_{1,1}^n = \{\text{there exists } \Delta \in (0, n^{-1/4}] \text{ such that (2.110) holds}\},$$

$$\Omega_{1,2}^n = \{\text{there exists } \Delta \in (n^{-1/4}, T] \text{ such that (2.110) holds}\}.$$

By Lemmas 2.1 and 2.2, $\Lambda_1^n \to C_1$ and $|\hat{A}_\alpha^n|_T^* \to 0$ in probability, and $\widetilde{W}_1^n(\cdot\wedge\sigma^n)$ are C-tight. This shows that $\mathbb{P}(\Omega_{1,1}^n) \to 0$. On $\Omega_{1,2}^n$, if $\Lambda_1^n \geq 0$ then $n^{1/4}\Lambda_1^n \leq \varepsilon_n$ must hold, where $\varepsilon_n = 2|\hat{A}_\alpha^n|_T^* + 2|\widetilde{W}_1^n|_{T\wedge\sigma^n}^*$. Hence $\mathbb{P}(\Omega_{1,2}^n) \leq \mathbb{P}(\Lambda_1^n < 0) + \mathbb{P}(\Lambda_1^n \leq n^{-1/4}\varepsilon_n) \to 0$. This shows that $\mathbb{P}(\Omega_1^n) \to 0$.

Step 2. We next show that $\mathbb{P}(\Omega_2^n) \to 0$. Letting u_1 and u_2 be as in the expression for Ω_2^n and arguing as before one obtains that $\hat{Z}_1^n(u) > \hat{Z}_2^n(u)$ for $u \in [u_1,u_2]$ and consequently that all class-1 customers are routed to station 1 during this period. Analogously to (2.108) we find that on Ω_2^n

$$h' n^{1/2} \leq A_\beta^n(u_2) - A_\beta^n(u_1) - D_2^n(u_1,u_2).$$

Using the inequality $A_\beta^n(u_2) - A_\beta^n(u_1) \leq A_2^n(u_2) - A_2^n(u_1)$ and some direct calculation we deduce from the above that

$$h' \leq \widetilde{A}_2^n(u_1,u_2) - \widetilde{S}_2^n(u_1,u_2), \tag{2.111}$$

where $\widetilde{S}_2^n$ is as in (2.109) and

$$\widetilde{A}_2^n(t_1,t_2) = \hat{A}_2^n(t_2) - \hat{A}_2^n(t_1) + \lambda_2^n n^{-1/2}(t_2-t_1).$$

Note that the r.h.s. of (2.111) contains the term $\gamma_n n^{1/2}(u_2 - u_1)$, where $\gamma_n = n^{-1}\lambda_2^n - \sum_{i=1,2} \mu_{i2}^n \psi_{i2}^*$. We claim that γ_n is bounded above by a negative constant for all n large. Indeed, recall that $\xi_{21}^* = 0$ and note that from (2.10) we have $\lambda_2 = \bar{\mu}_{22}\xi_{22}^*$. With (2.9) and (2.11) this shows $\lambda_2 = \mu_{22}\psi_{22}^* < \sum_{i=1,2} \mu_{i2}\psi_{i2}^*$. The claim regarding γ_n thus follows from (2.8).

Along the lines of step 1, we obtain instead of (2.110)

$$h' \leq w_{T\wedge\sigma^n}(\hat{A}_2^n - \widetilde{W}_2^n, \Delta) - n^{1/2}\Lambda_2^n \Delta,$$

where $\Delta = u_2 - u_1$,

$$\Lambda_2^n = C_3 - C_4 n^{-1/2}(1 + \|\hat{X}^n\|_{T\wedge\sigma^n}^* + |\hat{\Psi}_c^n|_{T\wedge\sigma^n}^*),$$

and $C_3 > 0$ and C_4 are constants. The rest of the argument for showing $\mathbb{P}(\Omega_2^n) \to 0$ is similar to that in step 1 and is omitted. We conclude that $\mathbb{P}(\sigma^n \leq \tau^n) \to 0$. □

Proof of Theorem 2.4. In view of Proposition 2.1, it only remains to treat the case where $e\cdot x \geq -1$. We can split the sequence of systems into two subsequences according as $e\cdot \hat{X}^{0,n} < -1$ or not, and since the result is already proved for the subsequence on which $e\cdot \hat{X}^{0,n} < -1$, we will assume without loss of generality that $e\cdot \hat{X}^{0,n} \geq -1$ for all n. As a result, the second part of the definition of the proposed policy applies. In the first part of the proof we used the fact that the system starts with no queues. In the current situation we argue that even if the system starts with a queue, it is brought quickly to zero. Recall the notation τ_0^n from definition of the policy. Note that prior to time τ_0^n there are r_n class-1 customers that are kept in the queue and that as far as all other customers are concerned, the system behaves exactly as in the first part of the definition. As a result, we can use Corollary 2.1 with $r = 1 + \sup_n n^{-1/2} r_n < \infty$, and it follows that $\mathbb{P}(\tau_0^n \leq \varepsilon/2)$ converges to 1. At time τ_0^n the system is in a state very similar to that in which a system satisfying $e\cdot \hat{X}^{0,n} < -1$ is at time zero, in the sense that $e\cdot \hat{X}^n(\tau_0^n) < -1$ and $|Z_1^n(\tau_0^n) - Z_2^n(\tau_0^n)| \leq 1$. A review of the proof of Proposition 2.1, replacing the initial values of all processes by their values at time τ_0^n shows that, with probability approaching 1, $Y^n(t)$ is kept zero for $t \in [\varepsilon, T]$. As in section 2.3, the only remaining issue is that $\hat{X}^n(\tau_0^n)$ must be shown to be tight. This is indeed the case by Corollary 2.1. □

2.5. Appendix to chapter 2

Lemma 2.4. *(i) All nodes $\mathcal{I} \cup \mathcal{J}$ of $\mathcal{G}_{ba}$ are connected through the edges of $\mathcal{G}_{ba}$.*
(ii) Every non-basic activity $(i,j) \in \mathcal{E}_{nb}$ belongs to a simple cycle.

Proof. For (i), consider a node of $\mathcal{G}_a$, $i \in \mathcal{I}$. By (2.10), $\sum_j \bar{\mu}_{ij}\xi^*_{ij} > 0$, and therefore there exists $j \in \mathcal{J}$ such that $\xi^*_{ij} > 0$. This shows that $(i,j) \in \mathcal{E}_{ba}$ and thus i must be a node of $\mathcal{G}_{ba}$. A similar argument holds for a node $j \in \mathcal{J}$ observing that $\sum_i \xi^*_{ij} = 1$ by the heavy traffic condition. Item (ii) follows since $\mathcal{G}_{ba}$ is a tree. □

Proof of Theorem 2.1. We first show that the relations (2.3)–(2.6), when rescaled appropriately, imply the following:

$$\hat{X}^n_i(t) = \hat{X}^{0,n}_i + \hat{W}^n_i(t) - \sum_{j\in\mathcal{J}} \mu^n_{ij} \int_0^t \hat{\Psi}^n_{ij}(s)ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, \tag{2.112}$$

$$\hat{Y}^n_i(t) + \sum_{j\in\mathcal{J}} \hat{\Psi}^n_{ij}(t) = \hat{X}^n_i(t), \qquad i \in \mathcal{I}, \tag{2.113}$$

$$\hat{Z}^n_j(t) + \sum_{i\in\mathcal{I}} \hat{\Psi}^n_{ij}(t) = \hat{N}^n_j, \qquad j \in \mathcal{J}, \tag{2.114}$$

$$\hat{Y}_i(t) \geq 0,\ \hat{Z}_j(t) \geq 0,\ \hat{\Psi}^n_c(t) \geq 0 \qquad i \in \mathcal{I},\ j \in \mathcal{J},\ c \in \mathcal{C}, \quad t \geq 0. \tag{2.115}$$

To this end, note that by (2.15), (2.16) and (2.19),

$$\begin{aligned}\hat{X}^n_i(t) &= n^{-1/2}(X^n_i(t) - nx^*_i)\\ &= \hat{X}^{0,n}_i + \hat{A}^n_i(t) - \sum_{j\in\mathcal{J}} \hat{S}^n_{ij}\left(\int_0^t \bar{\Psi}^n_{ij}(s)ds\right)\\ &\quad + n^{-1/2}\left[\lambda^n_i t - \sum_{j\in\mathcal{J}} n\mu^n_{ij}\int_0^t \bar{\Psi}^n_{ij}(s)ds\right].\end{aligned}$$

By (2.13), (2.17) and (2.18), the last term in the above display is equal

$$-\sum_{j\in\mathcal{J}} \mu^n_{ij}\int_0^t \hat{\Psi}^n_{ij}(s)ds + \hat{\lambda}^n_i t + n^{1/2}\left[\lambda_i - \sum_{j\in\mathcal{J}} \mu^n_{ij}\psi^*_{ij}\right] t.$$

Using (2.10), (2.11) and (2.13),

$$n^{1/2}\left[\lambda_i - \sum_{j\in\mathcal{J}} \mu^n_{ij}\psi^*_{ij}\right] t = -\sum_{j\in\mathcal{J}} \hat{\mu}^n_{ij}\psi^*_{ij} t.$$

Combined with (2.26), this establishes (2.112) above. Equations (2.113) and (2.114) and the first two inequalities in (2.115) directly follow from (2.4), (2.5), (2.6), (2.11), (2.14), (2.16), (2.17) and (2.18). For the last inequality in (2.115) recall that for $c \in \mathcal{C}$, $\hat{\Psi}^n_c = \hat{\Psi}^n_{ij}$ where $(i,j) = \sigma^{-1}(c)$, and that for all (i,j) of this form $\psi^*_{ij} = 0$.

Equations (2.28)–(2.29) follow from (2.113)–(2.115). As for (2.27), it suffices to prove the identity

$$\hat{\Psi}^n_{ij}(t) = G_{ij}(\hat{X}^n(t) - \hat{Y}^n(t), \hat{N}^n - \hat{Z}^n(t)) - \sum_{c \in \mathcal{C}:(i,j)\in c} s(c,i,j)\hat{\Psi}^n_c(t), \quad (2.116)$$

for all $(i,j) \in \mathcal{E}_a,\ t \geq 0$. Indeed, with (2.21), the above implies the identity

$$-\sum_{j\in\mathcal{J}} \mu^n_{ij}\hat{\Psi}^n_{ij}(t) = -\sum_{j\in\mathcal{J}} \mu^n_{ij} G_{ij}(\hat{X}^n(t) - \hat{Y}^n(t), \hat{N}^n - \hat{Z}^n(t)) + \sum_{c\in\mathcal{C}} m^n_{i,c}\hat{\Psi}^n_c(t),$$

which, along with (2.25) and (2.112) establish (2.27). In order to show (2.116), define

$$\Upsilon^n_{ij}(t) = \hat{\Psi}^n_{ij}(t) + \sum_{c\in\mathcal{C}:(i,j)\in c} s(c,i,j)\hat{\Psi}^n_c(t), \qquad (i,j) \in \mathcal{E}_a. \quad (2.117)$$

As follows from the uniqueness statement succeeding (2.22), in order to prove (2.116) it is enough to show that $\{\Upsilon^n_{ij}\}$ satisfy the system of equations (2.22) with $a = \hat{X}^n(t) - \hat{Y}^n(t)$ and $b = \hat{N}^n - \hat{Z}^n(t)$. To this end, note that if (i,j) is a non-basic activity then there is exactly one simple cycle c for which (i,j) is an edge and one has $s(c,i,j) = -1$ and, by definition of $\hat{\Psi}^n_c$, $\Psi^n_c = \Psi^n_{ij}$. This show that $\Upsilon^n_{ij} = 0$ for $(i,j) \in \mathcal{E}_{nb}$. Next, note that for a given $c \in \mathcal{C}$ and a node $i \in \mathcal{I}$ of c, there are exactly two edges of c of the form (i,j). By the construction of directions, for these two values of j, the signifiers $s(c,i,j)$ must have opposite signs. As a result, $\sum_{c\in\mathcal{C}}\sum_{j:(i,j)\in c} s(c,i,j)\hat{\Psi}^n_c(t) = 0$. Changing the order of summation and using an analogous argument for summation over i, we obtain

$$\begin{aligned}
&\sum_{j\in\mathcal{J}}\sum_{c\in\mathcal{C}:(i,j)\in c} s(c,i,j)\hat{\Psi}^n_c(t) = 0, \quad i \in \mathcal{I},\\
&\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}:(i,j)\in c} s(c,i,j)\hat{\Psi}^n_c(t) = 0, \quad j \in \mathcal{J}. \qquad (2.118)
\end{aligned}$$

Equations (2.113), (2.114) and (2.118) now show that $\{\Upsilon^n_{ij}\}$ satisfy (2.22) with the values a and b mentioned above. This proves (2.116), and in turn, the relation (2.27), and completes the proof of the theorem. □

PROOF OF LEMMA 2.1. Let $X^n(t)$, $Y^n(t)$, $Z^n(t)$ and $\{\Psi_c^n(t)\}$ be given, satisfying all requirements in the statement of the lemma. Define $\{\Psi_{ij}^n(t)\}$ according to (2.42). Equations (2.2), (2.4) and (2.5) follow directly from the definition of G and $\{s(c,i,j)\}$. As for (2.6), it remains only to prove that $\Psi_{ij}^n(t) \geq 0$ for every $(i,j) \in \mathcal{E}_{ba}$. Along the lines of Lemma 3 of [4], note that $\psi^* = G(x^*, \nu)$, as follows from (2.11). Using linearity of the map G on the domain D_G and the bound assumed on Ψ_c^n,

$$\Psi_{ij}^n(t) \geq G_{ij}(nx^*, n\nu) + G_{ij}(X^n(t) - nx^* - Y^n(t), N^n - n\nu - Z^n(t)) - C_1 a_0 n,$$

where C_1 is a constant independent of n, t and the choice of a_0, and therefore

$$\Psi_{ij}^n(t) \geq (\psi_{ij}^* - C_1 a_0)n - C_2(\|X^n(t) - nx^* - Y^n(t)\| \vee \|N^n - n\nu - Z^n(t)\|),$$

where C_2 is a constant depending only on the map G. With the assumed bound on $\|X^n - nx^*\| \vee \|Y^n\| \vee \|Z^n\|$, using also (2.14), we obtain $\Psi_{ij}^n(t) \geq (\psi_{ij}^* - C_3 a_0)n - C_4 n^{1/2}$ for appropriate constants C_3, C_4 independent of n, t and a_0. Since $\psi_{ij}^* > 0$ for all $(i,j) \in \mathcal{E}_{ba}$, it follows that $a_0 > 0$ may be chosen so that, for all large n, $\Psi_{ij}^n(t) \geq 0$. □

SKETCH OF PROOF OF PROPOSITION 2.1. We simplify the notation by writing A^n, $\bar{A}^n$, X^n and, respectively, $\bar{X}^n$ for A_{OL}^n, $\bar{A}_{OL}^n$, X_{OL}^n, $\bar{X}_{OL}^n$, etc. (there will be no confusion). Note that an analogous result to (2.56) holds, and in particular, for every t,

$$n^{-1} A_i^n(t) \to t\lambda_i' \qquad \text{in probability} \tag{2.119}$$

as $n \to \infty$. Fix $t \geq 1$. By (2.112) we have

$$\bar{X}_i^n(t) = \bar{X}_i^n(0) + \bar{A}_i^n(t) - \sum_j \mu_{ij} \int_0^t \bar{\Psi}_{ij}^n(s) ds + \Delta_i^n(t),$$

where

$$\Delta_i^n(t) = \sum_j \left[n^{-1/2} \hat{S}_{ij}^n \left(\int_0^t \bar{\Psi}_{ij}^n(s) ds \right) + (\mu_{ij}^n - \mu_{ij}) \int_0^t \bar{\Psi}_{ij}^n(s) ds \right]. \tag{2.120}$$

By (2.5) and (2.6) $\bar{\Psi}_{ij}^n(t)$ is bounded. An argument as in step 2 in the proof of Theorem 2.3 shows that the first term in (2.120) converges to zero in probability. Since $\mu_{ij}^n \to \mu_{ij}$, so does the second term, and it turn $\Delta_i^n(t) \to 0$ in probability. It can be shown that there exists a constant $\delta > 0$ depending only on $(\lambda, \lambda', \bar{\mu})$, such that for any sub-stochastic matrix (ξ_{ij})

$$\max_i [\lambda_i' - \sum_j \bar{\mu}_{ij} \, \xi_{ij}] \geq \delta. \tag{2.121}$$

Let $\Omega^n = \left\{\text{for all } i \in \mathcal{I},\ |\Delta_i^n(t)| + \left|t^{-1}\bar{A}_i^n(t) - \lambda_i'\right| \le \delta/3\right\}$. By (2.119), we have $P(\Omega^n) \to 1$. Define

$$\xi_{ij}^n = \frac{1}{t}\int_0^t \frac{\bar{\Psi}_{ij}^n(s)}{\nu_j} ds\ , \qquad \bar{\xi}_{ij}^n = \frac{\xi_{ij}^n}{\sum_{i'} \xi_{i'j}^n}.$$

It can be shown by (2.5) and (2.8) that $\lambda_i' - \sum_j \bar{\mu}_{ij}\ \xi_{ij}^n \ge \lambda_i' - \sum_j \bar{\mu}_{ij}\ \bar{\xi}_{ij}^n - \delta/3$, for all n large and all i. Since $\bar{\xi}$ is sub-stochastic, (2.121) implies that for some i we must have $\lambda_i' - \sum_j \bar{\mu}_{ij}\ \bar{\xi}_{ij}^n \ge \delta$, hence $\lambda_i' - \sum_j \bar{\mu}_{ij}\ \xi_{ij}^n \ge 2\delta/3$, for some i. This can be used to show $\max_i \bar{X}_i^n(t) \ge \delta t/3$ on Ω^n. Due to the relation $\bar{Y}_i^n(t) + \sum_j \bar{\Psi}_{ij}^n(t) = \bar{X}_i^n(t)$, we have $\bar{Y}_i^n(t) \ge \bar{X}_i^n(t) - C$, for all i, where C does not depend on n or t. Hence $\max_i \bar{Y}_i^n(t) \ge \delta t/3 - C$. We conclude that on Ω^n, for all n large, $\sum_i Y_i^n(t) \ge [\delta t/3 - C]n$, and the result follows. $\square$

Chapter 3

Throughput sub-optimality and heavy traffic

3.1. Introduction

In this chapter we introduce the notion of throughput sub-optimality for an underlying fluid model, and show that it plays a central role in the ability to achieve efficiency in a strong sense, that is usually only seen in systems that are sub-critically loaded. The chapter is based on [10].

The number of servers at each pool and the arrival rates of the queueing model in heavy traffic are scaled up at a nearly fixed proportion. When viewed at a scale at which the arrival and service processes exhibit diffusive fluctuations, the processes that represent the number of class-i customers in the system, $i \in \mathcal{I}$, fluctuate about a certain static fluid model. Assume that the fluid model is critically loaded, in a standard sense. In particular, (1) servers can be allocated in such a way that the total processing rate devoted to class-i 'material' is equal to the arrival rate λ_i, for every $i \in \mathcal{I}$; and (2) property (1) does not hold if one of the arrival rates λ_i is replaced by some $\lambda_i' > \lambda_i$ (there are some further assumptions; see Section 3.2). It is possible for such a model to satisfy the following condition: servers can be allocated so as to achieve a total processing rate that is greater than the total arrival rate, while, for every $i \in \mathcal{I}$, the 'mass' of servers allocated to serve class i does not exceed the 'mass' of class-i 'material.' If this condition holds we say that the fluid model is *throughput sub-optimal*. Our main result shows that when the (critically loaded) fluid model is throughput sub-optimal, one can find a dynamic control policy for the queueing model that is efficient in a strong sense: Under this policy, for every finite T, the measure of the set of times prior to T, at which at least one customer is in the buffer, converges to zero in probability

at the scaling limit.

A related analysis appears in Chapter 2, where the same model is proved to satisfy a stronger result under a stronger assumption. While the current chapter addresses the capability to maintain a system with no customer in the buffer 'most of the time', with large probability, the result of Chapter 2 concerns maintaining a system with no customers in the buffer 'at all times' (apart from an initial transition phase), with large probability. More precisely, under appropriate assumptions, it is shown ibid. that there exists a policy under which, for every $0 < \varepsilon < T < \infty$, the probability that at least one customer is present in the buffer any time within $[\varepsilon, T]$ approaches 1 in the scaling limit. This phenomenon is shown to be related, on one hand to a formulation of the limiting diffusion model as a diffusion with singular control (see Section 2.3 of Chapter 2). On the other hand, it is shown to be related to a condition on the graph that encodes the network's structure. This graph has a vertex for each class $i \in \mathcal{I}$, a vertex for each server pool $j \in \mathcal{J}$, and an edge, with an associated weight μ_{ij}, between a class vertex i and a pool vertex j if, and only if $\mu_{ij} > 0$. The assumption of Chapter 2 is the existence of a cycle p in this graph, having a negative total signed weight, $\mu(p)$, where the (p-dependent) signs of the weights μ_{ij} are appropriately defined (as in equations (3.17) in the current chapter; see also equation (3.18) for a definition of $\mu(p)$ as the sum of the signed weights along p). We will show that the algebraic condition alluded to above is a special case of the main assumption of the current chapter, namely throughput sub-optimality. We will also characterize the latter condition in terms of the graph and the signed weights, and show that throughput sub-optimality may occur in one of two ways: The existence of either a cycle or an open path p (appropriately defined), with signed weight $\mu(p) < 0$.

We make two further remarks about the relation to Chapter 2. First, the difference between having no customers in the buffer for a given period of time (as in Chapter 2) and having no customers in the buffer most of the time, may be significant with regard to the queuelength performance measure. In fact, under the policy constructed in the current chapter, there are short time periods in which large queues build. We believe that a result of the type of the previous chapter is not possible under the conditions of the current chapter, but we do not prove this claim. Second, the results of Chapter 2 allow for both preemptive policies (where service to a customer can be interrupted and resumed at a later time, possibly at a different server) and nonpreemptive ones (where service cannot be interrupted), while the current chapter only treats preemptive policies. We leave open the question of whether analogous results

are possible for the nonpreemptive case.

Both chapters 2 and 3 reveal two aspects of a phenomenon, where critically loaded many-server systems behave as sub-critically loaded. As this chapter's main result shows, the notion of throughput sub-optimality captures this phenomenon. It is reasonable to expect that this connection continues to hold in a wider context of critically loaded many server systems, such as ones with similar parametric regime but more general structure than the ones studied here.

The main tool in analyzing the probabilistic model is a related deterministic dynamic fluid model, that, roughly, replaces stochastic fluctuations by deterministic ones. Throughput sub-optimality is shown to have an effect on this model that is similar to the one discussed above. The proof of the result relies on the graph-theoretic characterization alluded to above, and specifically uses the existence of a path p with the property $\mu(p) < 0$. The result for the probabilistic model follows from the deterministic one in a relatively straightforward way.

The organization of the chapter is as follows. Section 3.2 contains the description of the model and assumptions, and the statement of the main result. Some numerical examples are given at the end of this section. Section 3.3 provides an algebraic characterization of throughput (sub) optimality. The dynamic fluid model is introduced in Section 3.4. A property for this model that is analogous to the main result is proved, based on the results of Section 3.3. Relying on the deterministic model results, we provide in Section 3.5 a proof of the main result.

3.2. Setting and main result

3.2.1. Probabilistic queueing model

The queueing model is described in Section 2.2.1. Recall the equations (3.1)–(3.5) below, which indicate some properties of the processes involved.

$$X_i^n(t) = X_i^{0,n} + A_i^n(t) - \sum_{j\in\mathcal{J}} S_{ij}^n\left(\int_0^t \Psi_{ij}^n(s)ds\right), \qquad i\in\mathcal{I}, t\geq 0. \tag{3.1}$$

$$Y_i^n(t) + \sum_{j\in\mathcal{J}} \Psi_{ij}^n(t) = X_i^n(t), \qquad i\in\mathcal{I}, \tag{3.2}$$

$$Z_j^n(t) + \sum_{i\in\mathcal{I}} \Psi_{ij}^n(t) = N_j^n, \qquad j\in\mathcal{J}. \tag{3.3}$$

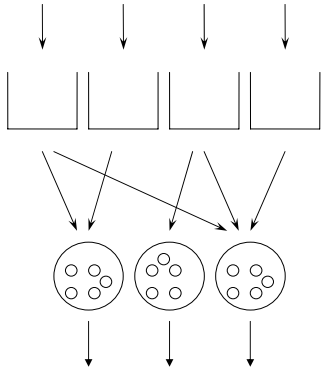


FIG 3.1. *A queueing model with four customer classes and three service pools*

Also, the following holds by definition

$$Y_i^n(t) \geq 0,\ Z_j^n(t) \geq 0,\ \Psi_{ij}^n(t) \geq 0, \qquad i \in \mathcal{I}, j \in \mathcal{J}, t \geq 0. \tag{3.4}$$

$$\Psi_{ij}^n(t) = 0 \text{ for } (i,j) \text{ s.t. } \mu_{ij}^n = 0. \tag{3.5}$$

For simplicity, the initial conditions $X_i^{0,n}$ are assumed to be deterministic. Note that (3.1)–(3.5) do not characterize these processes, because the process Ψ^n has not yet been described. As reflected in the following definition, we regard Ψ^n as a control process, that can be obtained as 'feedback' from the 'state' process X^n and the arrival process A^n.

Definition 3.1. *Fix n. We say that a process Ψ^n with values in $\mathbb{Z}_+^{\mathcal{I}\times\mathcal{J}}$ and cadlag paths is a* scheduling control policy *(SCP) if the following conditions hold:*

i. Given initial data $X^{n,0}$ and primitive processes A^n and S^n, there exist processes X^n, Y^n and Z^n with values in $\mathbb{Z}_+^{\mathcal{I}}$, $\mathbb{Z}_+^{\mathcal{I}}$, and $\mathbb{Z}_+^{\mathcal{J}}$, respectively, such that (3.5)–(3.4) are met;

ii. For every $t \geq 0$, $\Psi^n(t)$ is measurable on $\sigma\{X^n(s), A^n(s) : s \leq t\}$.

Note that uniqueness of the processes X^n, Y^n and Z^n, given A^n, S^n and Ψ^n, is immediate from (3.1)–(3.3). Note also that according to this definition, service to a customer can be stopped and resumed at a later time, possibly in a different station.

We will use some elementary graph theoretic terminology and notation as follows (see e.g., [24] for standard definitions). For a non-empty set V and $E \subseteq V \times V$, we write $G = (V, E)$ for the graph with vertex set V and edge set E. A vertex having exactly one neighbor is called a *leaf vertex*, and an edge

joining a leaf vertex is called a *leaf edge*. A connected graph that does not contain cycles is called a *tree*.

Denote the index set for all customer classes and service stations by $\mathcal{V} := \mathcal{I} \cup \mathcal{J}$, and the set of all class-station pairs by $\mathcal{E} := \mathcal{I} \times \mathcal{J}$. Recall the set of class–station pairs (*activities*), where station j can serve class i

$$\mathcal{E}_a = \{(i,j) \in \mathcal{I} \times \mathcal{J} : \mu_{ij}^n > 0\}, \tag{3.6}$$

We assume that $\mathcal{E}_a$ does not depend on n. Throughout, if $\mathcal{E}_1$ is a subset of $\mathcal{E}$, we write $\mathcal{E}_1^c$ for the complement of $\mathcal{E}_1$ with respect to $\mathcal{E}$. The set of class-station pairs that are not activities is denoted by $\mathcal{E}_a^c \equiv \mathcal{E} \setminus \mathcal{E}_a$.

3.2.2. Static model: heavy traffic and throughput optimality

Heavy traffic condition and related assumptions. Following Chapter 2, recall the constants λ_i, ν_j, $i \in \mathcal{I}$, $j \in \mathcal{J}$, and μ_{ij}, $(i,j) \in \mathcal{E}_a$, such that (2.8) holds.

Consider a fluid model, where the arrivals and service processes are replaced by deterministic flows with corresponding rates λ_i and μ_{ij}. There are I classes of incoming fluid and J processing stations, with capacity ν_j for station j. Let Ξ be the set of $I \times J$ matrices ξ with $\xi_{ij} \geq 0$, $(i,j) \in \mathcal{E}$, and $\sum_i \xi_{ij} \leq 1$, $j \in \mathcal{J}$. For $\xi \in \Xi$, ξ_{ij} will represent the fraction of the service capacity from station j allocated to class i. We call an element of Ξ an *allocation matrix*. The fluid model uses a fixed allocation matrix for all times (hence the term 'static' model). Set $\bar{\mu}_{ij} = \mu_{ij}\nu_j$, $(i,j) \in \mathcal{E}$. Recall the linear program (2.10):

Find $\{\xi_{ij}, (i,j) \in \mathcal{E}\}$ *and* $\rho \in \mathbb{R}_+$ *so as to minimize* ρ *subject to*

$$\begin{cases} \sum_{j \in \mathcal{J}} \bar{\mu}_{ij}\xi_{ij} = \lambda_i, & i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \xi_{ij} \leq \rho, & j \in \mathcal{J}, \\ \xi_{ij} \geq 0, & (i,j) \in \mathcal{E}. \end{cases} \tag{3.7}$$

For $\rho \in [0,1]$, a ξ as above is clearly an allocation matrix. The first line of (3.7) expresses that the system is balanced, in the sense that, for each i, the total processing rate of class-i material equals the arrival rate of this class. We will assume throughout that the system is critically loaded, in the sense of the *Heavy Traffic Condition* [33]. Namely, we assume that *there exists a*

unique optimal solution (ξ^*, ρ^*) *to the linear program* (3.7)*, and moreover,* $\sum_{i\in\mathcal{I}} \xi^*_{ij} = 1$ *for all* $i \in \mathcal{J}$ *(and consequently* $\rho^* = 1$*).* Let

$$\psi^*_{ij} = \xi^*_{ij}\nu_j, \qquad x^*_i = \sum_j \xi^*_{ij}\nu_j, \qquad i \in \mathcal{I}, j \in \mathcal{J}. \tag{3.8}$$

The following simple relations follow directly from the heavy traffic condition

$$\sum_{j\in\mathcal{J}} \psi^*_{ij} = x^*_i, \qquad \sum_{i\in\mathcal{I}} \psi^*_{ij} = \nu_j, \qquad \lambda_i = \sum_{j\in\mathcal{J}} \mu_{ij}\psi^*_{ij}, \qquad i \in \mathcal{I}, j \in \mathcal{J}. \tag{3.9}$$

The quantity ψ^*_{ij} represents the mass of class-i material present at station j under the allocation matrix ξ^*, and x^*_i represents the total mass of class-i material being processed. Recall also the definition of basic and non–basic activities (see Section 2.2.2) and the *complete resource pooling* condition (2.12).

Throughput optimality. The uniqueness statement included in the heavy traffic condition implies that given any allocation matrix ξ other than ξ^*,

$$\text{there is a class } i \in \mathcal{I} \text{ for which } \sum_{j\in\mathcal{J}} \bar{\mu}_{ij}\xi_{ij} < \lambda_i, \tag{3.10}$$

namely, that the processing rate is not sufficient for handling arrivals of this particular class. Note however that there is no analogous limitation concerning the total processing rate. That is, it is possible that there exists an allocation matrix ξ under which

$$\sum_{(i,j)\in\mathcal{E}} \bar{\mu}_{ij}\xi_{ij} > \sum_{i\in\mathcal{I}} \lambda_i. \tag{3.11}$$

The set of allocation matrices $\xi \in \Xi$ that satisfy

$$\sum_{j\in\mathcal{J}} \xi_{ij}\nu_j \le x^*_i \text{ for all } i \in \mathcal{I} \tag{3.12}$$

is of interest. Under these allocation matrices, for each $i \in \mathcal{I}$, the total mass of class-i material being processed does not exceed that under ξ^*. A condition involving simultaneously (3.11) and (3.12) will be key in this chapter. We will say that the static fluid model is *throughput optimal* if the following holds:

$$\textit{Whenever } \xi \in \Xi \textit{ and } \sum_{j\in\mathcal{J}} \xi_{ij}\nu_j \le x^*_i \ \forall\ i \in \mathcal{I}, \textit{ one has } \sum_{(i,j)\in\mathcal{E}} \bar{\mu}_{ij}\xi_{ij} \le \sum_{i\in\mathcal{I}} \lambda_i. \tag{3.13}$$

We will say that the static fluid model is *throughput sub-optimal* if it is not throughput optimal.

When the static fluid model is throughput sub-optimal, one can find $\xi \in \Xi$ meeting (3.11) and (3.12). Recall that x^* represents the mass of material of each class being processed in all service stations. Thus when (3.13) fails to hold, one can keep the same mass of material of each class as under ξ^*, and redistribute it among the stations so as to obtain a greater total processing rate than under ξ^*. Because of (3.10), a use of this allocation matrix will necessarily result with instability. In the probabilistic queueing model, however, one can vary the capacity allocation over time, and the existence of ξ as above turns out to have a crucial impact. This is expressed in Theorem 3.1 below, which is our main result.

The following assumption regards the second order behavior of the parameters and initial condition.

Assumption 3.1. *There exist constants $\hat{x}_i, \hat{\lambda}_i, \hat{\mu}_{ij} \in \mathbb{R}$, $i \in \mathcal{I}, j \in \mathcal{J}$, such that*

$$n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) \to \hat{\lambda}_i, \qquad n^{1/2}(\mu_{ij}^n - \mu_{ij}) \to \hat{\mu}_{ij}.$$
$$n^{1/2}(n^{-1}X_i^{0,n} - x_i^*) \to \hat{x}_i, \qquad n^{1/2}(n^{-1}N_j^n - \nu_j) \to 0. \tag{3.14}$$

Theorem 3.1. *Let the heavy traffic and complete resource pooling conditions hold. Let Assumption 3.1 hold. If the static fluid model is throughput sub-optimal then there exists a sequence of SCPs, under which for any fixed $0 < T < \infty$ and $\varrho > 1/2$,*

$$\int_0^T 1_{\{e\cdot Y^n(s)>0\}} ds \to 0 \quad \textit{in probability, as } n \to \infty, \tag{3.15}$$

$$n^{-\varrho}\|X^n - X^{0,n}\|_T^* \to 0 \quad \textit{in probability, as } n \to \infty. \tag{3.16}$$

3.2.3. Examples

We demonstrate throughput sub-optimality by some numerical examples.

Example 3.1 Consider the following static fluid model in heavy traffic, with 2 classes of customers and 3 stations

$$\nu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 8 \\ 4 \end{pmatrix}, \quad \mu = \bar{\mu} = \begin{pmatrix} 3 & 10 & 1 \\ 1 & 4 & 2 \end{pmatrix},$$

The resulting optimal static allocation is as follows (3.8)

$$\psi^* = \xi^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix} \quad \text{and} \quad x^* = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}.$$

To see that the fluid model is throughput sub-optimal, let $\varepsilon > 0$ be sufficiently small and consider the allocation matrix

$$\widehat{\xi} = \begin{pmatrix} 1-\varepsilon & 0.5+\varepsilon & 0 \\ \varepsilon & 0.5-\varepsilon & 1 \end{pmatrix}.$$

Clearly, we have $\sum_j \widehat{\xi}_{ij}\nu_j = x_i^*$ for every i. However, $\sum_{(i,j)\in\mathcal{E}} \widehat{\xi}_{ij}\bar{\mu}_{ij} > \lambda_1 + \lambda_2$. Thus the condition of throughput optimality (3.13) is not satisfied. The result of Theorem 3.1 holds. We note that the assumptions of [8] are also valid in this example. See more information on this example in the end of Section 3.3.

Example 3.2 In this example, the data is the same as in Example 1 above, except for one entry:

$$\nu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 8 \\ 4 \end{pmatrix}, \quad \mu = \bar{\mu} = \begin{pmatrix} 3 & 10 & 1 \\ 0 & 4 & 2 \end{pmatrix}.$$

The resulting optimal static allocation is as follows

$$\psi^* = \xi^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix} \quad x^* = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}.$$

With $\varepsilon > 0$ sufficiently small, the matrix

$$\widehat{\xi} = \begin{pmatrix} 1-\varepsilon & 0.5+\varepsilon & 0 \\ 0 & 0.5-\varepsilon & 1 \end{pmatrix}$$

is an allocation matrix. Moreover, $\sum_j \widehat{\xi}_{ij}\nu_j = x_i^*$ and $\sum_{(i,j)\in\mathcal{E}} \widehat{\xi}_{ij}\bar{\mu}_{ij} > \lambda_1 + \lambda_2$. Thus the fluid model is throughput sub-optimal. As shown in the end of Section 3.3, the conditions of [8] are not satisfied for this example.

Example 3.3 Consider

$$\nu = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \lambda = \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix}, \quad \mu = \bar{\mu} = \begin{pmatrix} 2 & 4 & 0.5 \\ 0.3 & 1 & 1 \\ 0.1 & 0.5 & 4 \end{pmatrix},$$

The resulting optimal static allocation is as follows

$$\psi^* = \xi^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \end{pmatrix} \quad \text{and} \quad x^* = \begin{pmatrix} 1.5 \\ 1 \\ 0.5 \end{pmatrix}.$$

The fluid model for this example is throughput optimal, as we show in the end of Section 3.3, using the tools we develop in Section 3.3.

3.3. Characterization of throughput optimality

The main result of this section (Theorem 3.2) characterizes throughput optimality in terms of some graph-theoretic properties of the network. To state it we need some definitions.

Recall that by the complete resource pooling condition (2.12) the graph $\mathcal{G}_{ba}$ is a tree, and by construction of it as a subgraph of $\mathcal{G}_a$, all its edges are of the form (i, j) where $i \in \mathcal{I}$ and $j \in \mathcal{J}$. In the definition below and elsewhere in this section it will be convenient to identify (i, j) with (j, i) (where $i \in \mathcal{I}$ and $j \in \mathcal{J}$) when referring to an element of the edge set $\mathcal{E}$. Although the notation is abused, there will be no confusion, since $\mathcal{I}$ and $\mathcal{J}$ do not intersect.

Definition 3.2. *i. A subgraph $q = (\mathcal{V}_q, \mathcal{E}_q)$ of $\mathcal{G}_{ba}$ is called a* basic path *if one has $\mathcal{V}_q = \{i_0, j_0, ..., i_k, j_k\}$, and*

$$\mathcal{E}_q = \{(i_0, j_0), (j_0, i_1), \ldots, (i_k, j_k)\}$$

where $k \geq 1$ and $i_0, \ldots, i_k \in \mathcal{I}$, $j_0, \ldots, j_k \in \mathcal{J}$ are $2k+2$ distinct vertices. Note that every edge of a basic path is a basic activity (i.e., an element of $\mathcal{E}_{ba}$). Denote by BP the set of basic paths. Basic paths are used in this chapter mainly in order to define simple paths, as follows.

ii. Let the leaves i_0 and j_k of a basic path q be denoted by i^q and, respectively, j^q. The pair (i^q, j^q) could be an activity (an element of $\mathcal{E}_a$), in which case it is necessarily a non-basic activity (i.e., an element of $\mathcal{E}_a \setminus \mathcal{E}_{ba}$), and we say that the graph $(\mathcal{V}_q, \mathcal{E}_q \cup \{(i^q, j^q)\})$ is a closed simple path*; otherwise (i^q, j^q) is not an activity (i.e., it is in $\mathcal{E}_a^c$) and we say that q itself is an* open simple path*. We say that p is a* simple path *if it is either a closed or an open simple path. Denote by CSP and OSP the sets of closed and open simple paths, respectively, and by SP the set of simple paths. For a path $p \in SP$, we write $\mathcal{V}_p$ and $\mathcal{E}_p$ for its vertex and edge sets, respectively. Finally, if p is a simple path, let $q^p \in BP$ denote the corresponding basic path q, and let $i^p \in \mathcal{I}$ and $j^p \in \mathcal{J}$ denote the leaves i^q and j^q of q^p.*

Note that if $p = (\mathcal{V}_p, \mathcal{E}_p)$ is a simple path and $q^p = (\mathcal{V}_q, \mathcal{E}_q)$ is its corresponding basic path, then $\mathcal{V}_q = \mathcal{V}_p$, and $\mathcal{E}_q = \mathcal{E}_p \setminus \{(i^p, j^p)\}$.

Next, we associate directions with edges of simple paths. Let p be a simple path and let $q = q^p = (\mathcal{V}_q, \mathcal{E}_q)$ be the corresponding basic path. Write $\mathcal{E}_q = \{(i_0, j_0), \ldots, (i_k, j_k)\}$, where $i_0, \ldots, i_k \in \mathcal{I}$ and $j_0, \ldots, j_k \in \mathcal{J}$. The direction that will be associated with the edges in $\mathcal{E}_q$, when considered as edges of p, is as follows: $j_k \to i_k \to j_{k-1} \to i_{k-1} \to \cdots \to j_0 \to i_0$. In the case of an open simple path, this exhausts all edges of p. In the case of a closed simple path, the

direction of $(i^p, j^p) = (i_0, j_k)$ is $i_0 \to j_k$. We note that an edge (corresponding to a basic activity) may have different directions when considered as an edge of different simple paths. We signify the directions along simple paths by $s(p, i, j)$, defined for $i \in \mathcal{I}$, $j \in \mathcal{J}$, $(i, j) \in \mathcal{E}_p$, $p \in SP$, as

$$s(p,i,j) = \begin{cases} -1 \text{ if } (i,j), \text{ considered as an edge of } p, \text{ is directed from } i \text{ to } j \\ \ \ 1 \text{ if } (i,j), \text{ considered as an edge of } p, \text{ is directed from } j \text{ to } i. \end{cases} \tag{3.17}$$

We will denote

$$\mu(p) = \sum_{(i,j)\in\mathcal{E}_p} s(p,i,j)\mu_{ij}, \quad i \in \mathcal{I} \tag{3.18}$$

Theorem 3.2. *Let the heavy traffic and complete resource pooling conditions hold. Then the following statements are equivalent*

1. *The static fluid model is throughput sub-optimal;*
2. *There exists a simple path $p \in SP$ such that $\mu(p) < 0$.*

Condition (3.13) is stated in terms of the variables $\{\xi_{ij}\}$. It will be convenient to work with the variables $\{\psi_{ij}\}$ in the proof below. To this end, recall that $\nu_j > 0$ for all j and $\psi^*_{ij} = \xi^*_{ij}\nu_j$. Thus the *negation* of (3.13) can be written as follows: *There exists*

$$\psi \in \mathbb{R}^{\mathcal{E}}_+ \tag{3.19}$$

satisfying

$$\begin{aligned} &(a) \quad \textstyle\sum_{i\in\mathcal{I}} \psi_{ij} \le \nu_j \text{ for all } j \in \mathcal{J}, \\ &(b) \quad \textstyle\sum_{j\in\mathcal{J}} \psi_{ij} \le x^*_{ij} \text{ for all } i \in \mathcal{I}, \\ &(c) \quad \textstyle\sum_{(i,j)\in\mathcal{E}} \mu_{ij}\psi_{ij} > \sum_{i\in\mathcal{I}} \lambda_i. \end{aligned} \tag{3.20}$$

Proof that statement 2 of Theorem 3.2 implies statement 1. Assume that statement 2 holds and fix a simple path p with $\mu(p) < 0$. Let $q = q^p$ be the corresponding basic path, and recall that $\psi^*_{ij} > 0$ for $(i, j) \in \mathcal{E}_q$. Denote

$$\alpha = \min_{(i,j)\in\mathcal{E}_q} \psi^*_{ij} > 0. \tag{3.21}$$

For each $(i, j) \in \mathcal{I} \times \mathcal{J}$ we define

$$\sigma_{ij} = -\alpha s(p,i,j) \quad \text{if } (i,j) \in \mathcal{E}_p, \tag{3.22}$$

and $\sigma_{ij} = 0$ otherwise. Let

$$\psi_{ij} = \psi^*_{ij} + \sigma_{ij} \quad \text{for } (i,j) \in \mathcal{I} \times \mathcal{J}. \tag{3.23}$$

To show that statement 1 of the theorem holds, let us show that ψ satisfies (3.19) and (3.20). For $(i,j) \notin \mathcal{E}_p$, $\psi_{ij} = \psi^*_{ij} \geq 0$. For $(i,j) \in \mathcal{E}_q$,

$$\psi_{ij} \geq \psi^*_{ij} - \alpha \geq 0,$$

by (3.21). In the case where p is an open simple path $\mathcal{E}_p = \mathcal{E}_q$, and (3.19) follows. In the case where p is a closed simple path, it is left to show that $\psi_{i^p j^p} \geq 0$. Recall that the direction associated with (i^p, j^p) is $i^p \to j^p$. Thus by (3.17) and (3.22), $\psi_{i^p j^p} \geq \psi^*_{i^p j^p} = 0$, establishing (3.19).

Next, if p is a closed simple path then every vertex v of it has exactly two neighbors along p, say, v' and v'', and the directions of the corresponding edges are $v' \to v$ and $v \to v''$. Hence by (3.17) and (3.22), $\sum_{j \in \mathcal{J}} \sigma_{ij} = 0$ holds for every $i \in \mathcal{I}$, and $\sum_{i \in \mathcal{I}} \sigma_{ij} = 0$ holds for every $j \in \mathcal{J}$. The term $\sigma_{i^p j^p}$, which is positive in the case when p is closed, is in fact zero in the case when p is open, thus yielding $\sum_{j \in \mathcal{J}} \sigma_{ij} \leq 0$ for every $i \in \mathcal{I}$ and $\sum_{i\mathcal{I}} \sigma_{ij} \leq 0$ for every $j \in \mathcal{J}$. Since ψ^* satisfies (3.20)(a) and (b), it follows from(3.23) that so does ψ. Finally, by (3.23) and since (3.20)(c) holds for ψ^* with equality, it suffices to prove

$$\sum_{(i,j) \in \mathcal{E}} \mu_{ij} \sigma_{ij} > 0 \tag{3.24}$$

to establish that ψ satisfies (3.20)(c). By (3.22) and (3.18)

$$\sum_{(i,j) \in \mathcal{E}} \mu_{ij} \sigma_{ij} = -\alpha \sum_{(i,j) \in \mathcal{E}_p} \mu_{ij} s(p,i,j) = -\alpha\, \mu(p) > 0,$$

where the inequality follows from (3.21) and the assumption $\mu(p) < 0$. This establishes (3.24) and completes the proof that statement 2 implies statement 1. □

In the rest of this section we prove that statement 1 of the theorem implies statement 2. Define

$$M(\sigma) := \sum_{(i,j) \in \mathcal{E}} \mu_{ij} \sigma_{ij}, \tag{3.25}$$

for any matrix $\sigma \in \mathbb{R}^{\mathcal{E}}$. Let S denote the set of $\sigma \in \mathbb{R}^{\mathcal{E}}$ satisfying the conditions

$$\sum_{j \in \mathcal{J}} \sigma_{ij} \leq 0 \text{ for all } i \in \mathcal{I}, \quad \sum_{i \in \mathcal{I}} \sigma_{ij} \leq 0 \text{ for all } j \in \mathcal{J}, \tag{3.26}$$

$$\psi^*_{ij} + \sigma_{ij} \geq 0 \quad \text{for all } (i,j) \in \mathcal{E}_a, \tag{3.27}$$

and

$$\sigma_{ij} = 0 \quad \text{for all } (i,j) \in \mathcal{E}_a^c. \tag{3.28}$$

Note that S is non-empty and compact, and let

$$M_{max} = \max\{M(\sigma) : \sigma \in S\}. \tag{3.29}$$

It is easy to see that the existence of a ψ satisfying (3.19) and (3.20) is equivalent to the condition $M_{max} > 0$.

Throughout what follows, let statement 1 hold. By the above discussion, $M_{max} > 0$. Let S_{opt} [resp., S_+] denote the set of $\sigma \in S$ such that $M(\sigma) = M_{max}$ [resp., $M(\sigma) > 0$], and note that S_{opt} and S_+ are non-empty. For $\sigma \in S_+$ consider the graph $G_\sigma = (V_\sigma, E_\sigma)$, where

$$E_\sigma = \{(i,j) \in \mathcal{E}_a : \sigma_{ij} \neq 0\} \tag{3.30}$$

and $V_\sigma = \{i \in \mathcal{I} : (i,j) \in E_\sigma \text{ some } j\} \cup \{j \in \mathcal{J} : (i,j) \in E_\sigma, \text{ some } i\}$ consists of all corresponding vertices. Since $M(\sigma) > 0$, we have:

$$\text{there exists } (i,j) \in E_\sigma \text{ with } \sigma_{ij} > 0. \tag{3.31}$$

By (3.26) and (3.30),

$$\text{if } (i,j) \text{ is a leaf edge of } G_\sigma \text{ then } \sigma_{ij} < 0, \tag{3.32}$$

and

$$\begin{gathered}\text{if } (i,j) \in E_\sigma \text{ and } \sigma_{ij} > 0 \text{ then there exist two edges } (i,j_0), (i_0,j) \in E_\sigma \\ \text{with } \sigma_{i_0,j} < 0 \text{ and } \sigma_{i,j_0} < 0.\end{gathered} \tag{3.33}$$

Definition 3.3. *Let $\sigma \in S_+$ be given. A subgraph $g = (\mathcal{V}_g, \mathcal{E}_g)$ of the graph G_σ is called a* good path for σ*, if it satisfies the following conditions.*

(i) (Connectivity) All vertices in $\mathcal{V}_g$ communicate via the edges in $\mathcal{E}_g$.
(ii) The degree of each vertex is at most 2.
(iii) The number of edges is at least 3.
(iv) (Alternating signs) Whenever $(i_1,j), (i_2,j) \in E_\sigma$, one has $\sigma_{i_1,j}\sigma_{i_2,j} < 0$; whenever $(i,j_1), (i,j_2) \in E_\sigma$, one has $\sigma_{i,j_1}\sigma_{i,j_2} < 0$.
(v) (Maximality) Whenever g is a subgraph of some subgraph g' of G_σ, and g' satisfies properties (i)–(iv) above, one has $g' = g$.

It is not hard to see that observations (3.31) and (3.33) about the graph G_σ imply that, whenever $\sigma \in S_+$, there exists at least one good path for σ.

Let $\sigma \in S_+$ be given. For any edge $(i,j) \in E_\sigma$, define $s_\sigma(i,j) = -\text{sign}(\sigma_{ij})$ and for any good path g for σ set

$$\mu(\sigma, g) := \sum_{(i,j)\in\, \mathcal{E}_g} s_\sigma(i,j)\mu_{ij}. \tag{3.34}$$

We write $\bar{S}_C$ [respectively, $\bar{S}_O$] for the set of all $\sigma \in S_{opt}$ for which there exists a good path [respectively, there exists no good path] g for σ that is a cycle. The letters C and O are mnemonics for closed and open. Note that $S_{opt} = \bar{S}_C \cup \bar{S}_O$. We also set

$$S_C = \{(\sigma, g) : \sigma \in \bar{S}_C \text{ and } g \text{ is a good path for } \sigma \text{ that is a cycle}\},$$
$$S_O = \{(\sigma, g) : \sigma \in \bar{S}_O \text{ and } g \text{ is a good path for } \sigma \text{ that is not a cycle}\}.$$

Lemma 3.1. *Let $(\sigma, g) \in S_O$. Write $g = (\mathcal{V}_g, \mathcal{E}_g)$, where*

$$\mathcal{V}_g = \{v_1, \ldots, v_k\}, \quad \mathcal{E}_g = \{(v_1, v_2), \ldots, (v_{k-1}, v_k)\},$$

and $v_1, \ldots, v_k$ are distinct elements of V_σ. Then $\sigma_{v_1,v} < 0$ for every edge $(v_1, v) \in E_\sigma$, and similarly $\sigma_{v,v_k} < 0$ for $(v, v_k) \in E_\sigma$.

Proof. Argue by contradiction and assume that $\sigma_{v_1,v_2} > 0$. By (3.33) (v_1, v_2) must have a neighbour (v_0, v_1) with $v_0 \neq v_2$, satisfying $\sigma_{v_0,v_1} < 0$. It is easy to see that if we had $v_0 \in \mathcal{V}_g$, there would exist a good path for σ that is a cycle, violating the assumption of the lemma that there exist no such good paths for σ. Define a new graph g' by $\mathcal{V}_{g'} = \mathcal{V}_g \cup \{v_0\}$ and $\mathcal{E}_{g'} = \mathcal{E}_g \cup (v_0, v_1)$, and note that it is a good path (cf. Definition 3.3). This contradicts the maximality property (Definition 3.3(v)), and therefore one must have $\sigma_{v_1,v_2} < 0$. The second leaf edge (v_{k-1}, v_k) is treated similarly.

To prove the second statement of the lemma, let $v_0 \neq v_2$ be such that $v_0 \in V_\sigma$, $(v_0, v_1) \in E_\sigma$. Argue by contradiction and assume that $\sigma_{v_0,v_1} > 0$. Since we already proved that $\sigma_{v_1,v_2} < 0$, we can again use the assumption that there is no good path for σ that is a cycle to conclude that $v_0 \notin V_\sigma$. Defining g' by $\mathcal{V}_{g'} = \mathcal{V}_g \cup \{v_0\}$ and $\mathcal{E}_{g'} = \mathcal{E}_g \cup (v_0, v_1)$ produces a good path that contains g, contradicting property (v) of Definition 3.3. Hence $\sigma_{v_0,v_1} < 0$. □

Lemma 3.2. *Let $(\sigma, g) \in S_C \cup S_O$. Then there exists a set $SP_g \subset SP$ of simple paths, such that*

$$\mu(\sigma, g) = \sum_{p\in\, SP_g} \mu(p). \tag{3.35}$$

PROOF. Consider first the case where $(\sigma, g) \in S_C$. Write $g = (\mathcal{V}_g, \mathcal{E}_g)$ and let $\gamma^0 \in \mathbb{R}^{\mathcal{E}}$ be defined by

$$\gamma^0_{ij} = \begin{cases} \text{sign}(\sigma_{ij}) \ , & \text{if } (i,j) \in \mathcal{E}_g \\ 0 \ , & \text{otherwise.} \end{cases} \tag{3.36}$$

By (3.34) and (3.36) we have

$$M(\gamma^0) \equiv \sum_{(i,j)\in\mathcal{E}} \gamma^0_{ij}\mu_{ij} = \sum_{(i,j)\in\mathcal{E}_g} \text{sign}(\sigma_{ij})\mu_{ij} = -\mu(\sigma, g). \tag{3.37}$$

The following property is due to (3.36) and the fact that g is a good path for σ that is a cycle:

$$\text{for any } i \in \mathcal{I} \text{ and } j \in \mathcal{J} \text{ we have } \sum_{j\in\mathcal{J}} \gamma^0_{ij} = 0 \text{ and } \sum_{i\in\mathcal{I}} \gamma^0_{ij} = 0. \tag{3.38}$$

Define a finite sequence $\gamma^r \in \mathbb{R}^{\mathcal{E}}$ recursively as follows. Given γ^r, if there are no non-basic activities (i.e., elements of $\mathcal{E}_a \setminus \mathcal{E}_{ba}$) in the set of edges where γ^r is supported then terminate, and set $R = r$. Otherwise, select such a non-basic activity, and let p_r denote the (unique) closed simple path containing it as an edge. Define $\gamma^{r+1} \in \mathbb{R}^{\mathcal{E}}$ by

$$\gamma^{r+1}_{ij} = \begin{cases} \gamma^r_{ij} + s(p_r, i, j) \ , & \text{if } (i,j) \in \mathcal{E}_{p_r} \\ \gamma^r_{ij} \ , & \text{otherwise.} \end{cases} \tag{3.39}$$

For $0 \le r < R$, the selected non-basic activity at step r is (i^{p_r}, j^{p_r}) (using the notation from Definition 3.2). By the discussion following Definition 3.2, the direction for this activity is $i^{p_r} \to j^{p_r}$ and thus by (3.17) we have that

$$\gamma^{r+1}_{i,j} = \gamma^r_{i,j} - 1 \text{ where } (i,j) = (i^{p_r}, j^{p_r}). \tag{3.40}$$

Given a non-basic activity (i,j), let r be the first r' for which (i,j) is the selected non-basic activity at step r' (if such r' exists). Since the transformation (3.39) modifies γ only at basic activities and at the non-basic activity selected at the given step, it follows that $\gamma^r_{i,j} = \gamma^0_{i,j}$. Hence by (3.27), (3.30) and (3.36) that $\gamma^0_{i,j} = 1$. Thus (3.40) shows that $\gamma^{r+1}_{i,j} = 0$. As a result, the support of γ^{r+1} contains one non-basic activity less that that of γ^r. It follows that $R < \infty$. Thus γ^R is well-defined and supported on basic activities.

Next, since by construction, the selected simple paths are closed, it follows by the linearity of the transformation (3.39) that (3.38) holds for each γ^r, and

in particular, for γ^R. It also follows from the linearity of (3.39), using (3.18), that $M(\gamma^{r+1}) = M(\gamma^r) + \mu(p_r)$ for $0 \leq r < R$. Hence

$$M(\gamma^R) = M(\gamma^0) + \sum_{r=0}^{R-1} \mu(p_r). \tag{3.41}$$

The fact that γ^R is supported on basic activities and that these form a tree (cf. (2.12)), combined with the fact that γ^R satisfies (3.38) implies that $\gamma^R = 0$. Hence $M(\gamma^R) = 0$, and using (3.37) and (3.41), we obtain (3.35).

Next, consider $(\sigma, g) \in S_O$. Then $g = (\mathcal{V}_g, \mathcal{E}_g)$, where

$$\mathcal{V}_g = \{v_1, \ldots, v_k\} \quad \mathcal{E}_g = \{(v_1, v_2), \ldots, (v_{k-1}, v_k)\},$$

and $v_1, \ldots, v_k$ are distinct elements of V_σ. First, note that either $(v_1, v_k) \in \mathcal{I} \times \mathcal{J}$ or $(v_k, v_1) \in \mathcal{I} \times \mathcal{J}$. Indeed, by Lemma 3.1 and properties (iii)–(iv) of Definition 3.3, $|\mathcal{E}_g|$ is an odd number, while having both v_1 and v_k belong to either $\mathcal{I}$ or $\mathcal{J}$ would result with an even number for $|\mathcal{E}_g|$.

Also, we claim that $(v_1, v_k) \in \mathcal{E}_a^c$. Argue by contradiction and assume that $\mu_{v_1,v_k} > 0$. If we had $\sigma_{v_1,v_k} > 0$, then by Lemma 3.1, $\sigma_{v_1,v_2} < 0$ and $\sigma_{v_{k-1},v_k} < 0$, and there would exist a good path, which is a cycle. This is prohibited since $(\sigma, g) \in S_O$. Hence $\sigma_{v_1,v_k} \leq 0$. Set $\delta := \min\{|\sigma_{v_1,v_2}|, |\sigma_{v_{k-1},v_k}|\}$ and define a new matrix $\sigma' \in \mathbb{R}^{\mathcal{E}}$ by assigning $\sigma'_{v_1,v_k} = \sigma_{v_1,v_k} + \delta$ and $\sigma'_{ij} = \sigma_{ij}$ for $(i,j) \in \mathcal{E} \setminus \{(v_1, v_k)\}$. By the definition σ' satisfies (3.26)–(3.28) (see also Lemma 3.1), which implies $M(\sigma') = M(\sigma) + \delta\, \mu_{v_1,v_k} > M(\sigma)$. This contradicts the assumption $\sigma \in S_{opt}$. Therefore $\mu_{v_1,v_k} = 0$ meaning $(v_1, v_k) \in \mathcal{E}_a^c$.

The rest of the argument is similar to the treatment of the case where $(\sigma, g) \in S_C$, with some modifications, as follows. Instead of (3.36), consider $\gamma^0 \in \mathbb{R}^{\mathcal{E}}$ defined as

$$\gamma^0_{ij} = \begin{cases} \text{sign}(\sigma_{ij})\,, & \text{if } (i,j) \in \mathcal{E}_g, \\ 1\,, & \text{if } (i,j) = (v_1, v_k), \\ 0\,, & \text{otherwise.} \end{cases} \tag{3.42}$$

Since (v_1, v_k) is not an activity, $\mu_{v_1,v_k} = 0$, and thus (3.37) is still valid. Also, it follows from Lemma 3.1 that $\gamma^0_{v_1,v_2} = \gamma^0_{v_{k-1},v_k} = -1$, and as a result, (3.38) is valid. We can now repeat the construction of $\{\gamma^r\}$, $0 \leq r < R$ and the inductive argument that leads to (3.41). The matrix γ^R, in this case, is supported on basic activities plus the edge (v_1, v_k). Denoting by p the open simple path whose leaves are v_1 and v_k, we apply one last time a transformation of the

form (3.39) as follows:

$$\gamma_{ij}^{R+1} = \begin{cases} \gamma_{ij}^R + s(p,i,j), & \text{if } (i,j) \in \mathcal{E}_p, \\ \gamma_{ij}^R - 1, & \text{if } (i,j) = (i^p, j^p) \equiv (v_1, v_k), \\ \gamma_{ij}^R, & \text{otherwise.} \end{cases}$$

As a result,

$$M(\gamma^{R+1}) = M(\gamma^0) + \sum_{r=0}^{R-1} \mu(p_r) + \mu(p).$$

Arguing as before, we obtain that γ^{R+1} is supported on basic activities, satisfies (3.38) thus vanishes. Hence $M(\gamma^{R+1}) = 0$, and (3.35) follows as before. □

Lemma 3.3. *Let $(\sigma, g) \in S_C \cup S_O$. Then the following statements are true.*

(i) $\mu(\sigma, g) \leq 0$.
(ii) If $\mu(\sigma, g) = 0$ then there exists $\sigma' \in S_{opt}$ satisfying $E_{\sigma'} \subsetneq E_\sigma$.

Proof. We first prove part (i). Consider first the case where $(\sigma, g) \in S_C$. Arguing by contradiction, assume $\mu(\sigma, g) > 0$. Let

$$\alpha = \min_{(i,j) \in \mathcal{E}_g} |\sigma_{ij}| > 0, \tag{3.43}$$

and for each $(i,j) \in E_\sigma$ define

$$\sigma'_{ij} = \sigma_{ij} + s_\sigma(i,j)\alpha \ \text{ if } \ (i,j) \in \mathcal{E}_g, \ \text{ and } \ \sigma'_{ij} = \sigma_{ij} \ \text{ otherwise.} \tag{3.44}$$

We show that σ' satisfies conditions (3.26)–(3.28), and therefore that $\sigma' \in S$. To this end, note that the sums in (3.26) remain unchanged under the transformation from σ to σ', due to the fact that g is a cycle and using the alternating signs property (Definition 3.3(iv)). Thus (3.26) is satisfied by σ'. The relation (3.27) follows from (3.43) and (3.44), since $\sigma'_{ij} > \sigma_{ij}$ for $(i,j) \in \mathcal{E}_g$ with $\sigma_{ij} < 0$ and $\sigma'_{ij} \geq 0$ for $(i,j) \in \mathcal{E}_g$ with $\sigma_{ij} > 0$. The relation (3.28) holds trivially. This shows $\sigma' \in S$. We have

$$\begin{aligned} M(\sigma') &= \sum_{(i,j) \in \mathcal{E}} \sigma'_{ij}\mu_{ij} = \sum_{(i,j) \in \mathcal{E}} \sigma_{ij}\mu_{ij} + \alpha \sum_{(i,j) \in \mathcal{E}_g} s_\sigma(i,j)\mu_{ij} \\ &= M(\sigma) + \alpha\mu(\sigma, g). \end{aligned} \tag{3.45}$$

Since $\mu(\sigma, g) > 0$ by assumption, we have $M(\sigma') > M(\sigma)$, which contradicts the assumption $\sigma \in S_{opt}$. Hence (i) holds.

Consider now the case where $(\sigma, g) \in S_O$. Argue by contradiction and assume that $\mu(\sigma, g) > 0$. Define σ' as in (3.43)–(3.44). Once again, we claim that the

constraints (3.26)–(3.28) are satisfied for σ'. The argument following (3.44) applies, and it remains only to check (3.26) for the vertices v_1 and v_k. The validity of (3.26) in this case follows from (3.43), (3.44) and Lemma 3.1, since we have $\sigma_{v_1,v_2} < 0$, $\sigma_{v_{k-1},v_k} < 0$ and $\sigma'_{ij} \leq 0$ holds for all $(i,j) \in \mathcal{E}_g$ with $\sigma_{ij} < 0$. This shows that $\sigma' \in S$. The rest of the argument is as in the previous case.

Next we prove part (ii). The desired σ' is, in fact, the one constructed in the proof of part (i). Indeed, we have proved that $\sigma' \in S$. Moreover, since $\mu(\sigma, g) = 0$ and $\sigma \in S_{opt}$, we have by (3.45) that $\sigma' \in S_{opt}$. By (3.43) and (3.44),

$$\sigma'_{i'j'} = 0 \text{ for all } (i',j') \in \arg\min_{(i,j)\in\mathcal{E}_g} |\sigma_{ij}|,$$

and therefore statement (ii) holds. □

Proof that statement 1 of Theorem 3.2 implies statement 2. Let $\sigma \in S_{opt}$. Let g be such that $(\sigma, g) \in S_C \cup S_O$. Set $(\sigma^0, g^0) = (\sigma, g)$ and define a finite sequence $(\sigma^r, g^r) \in S_C \cup S_O$ recursively as follows. If $\mu(\sigma^r, g^r) < 0$ then set $R = r$ and terminate. Otherwise, by Lemma 3.3(i), $\mu(\sigma^r, g^r) = 0$. Let σ^{r+1} denote the matrix σ' from Lemma 3.3(ii) corresponding to (σ^r, g^r). Since $\sigma^{r+1} \in S_{opt} = \bar{S}_C \cup \bar{S}_O$, it follows from the definition of S_O and S_C that there exists g such that $(\sigma^{r+1}, g) \in S_C \cup S_O$. Let g^{r+1} be such a good path.

The finiteness of R follows from Lemma 3.3(ii) and the finiteness of the set $\mathcal{E}_\sigma$.

By construction, $(\sigma^R, g^R) \in S_C \cup S_O$ and $\mu(\sigma^R, g^R) < 0$. Lemma 3.2 thus implies that there exists a simple path p such that $\mu(p) < 0$. This concludes the proof of the Theorem. □

We end this section by revisiting the three examples from Section 3.2.3. We can now use Theorem 3.2 to determine throughput sub-optimality for each example.

Example 3.1 The simple path p corresponding to Example 3.1 (see Figure 3.2, left) satisfies $\mu(p) = -4 < 0$. Hence Theorem 3.1 applies. Moreover, p is a *closed* simple path, and one checks that [8, Theorem 2.3] is valid too.

Example 3.2 In the case of Example 3.2, the simple path p is open (see Figure 3.2, right). We have $\mu(p) = -3 < 0$. Theorem 3.1 applies. Since p is open, [8, Theorem 2.3] does not apply.

Example 3.3 To see that the fluid model of Example 3.3 is throughput op-

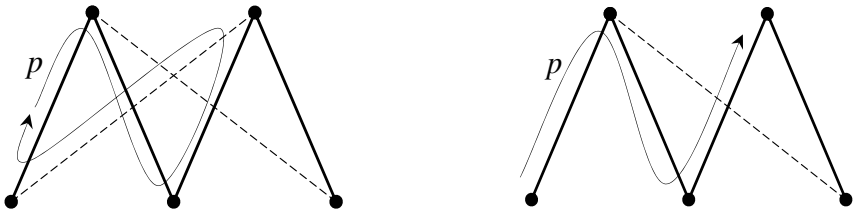


FIG 3.2. *Simple paths for Examples 3.1 and 3.2: On the left p is a closed simple path, while on the right p is open. For Example 3.1,* $\mu_{21} > 0$ *and* $(2,1)$ *is a non-basic activity. For Example 3.2,* $\mu_{21} = 0$ *and* $(2,1)$ *is not an activity*

timal, one could calculate $\mu(p)$ for all simple paths. We can instead use the linear optimization problem described in (3.26)–(3.29), and find that $M_{max} = 0$ (3.29).

3.4. Dynamic fluid model

As a tool for analyzing the probabilistic model, we consider a model with deterministic arrival and service rates. The model is obtained from (3.1)–(3.4) by replacing the primitive processes $A_i^n(t)$ and $S_{ij}^n(t)$ by the deterministic functions $\lambda_i t$, $\mu_{ij} t$, $i \in \mathcal{I}$, $j \in \mathcal{J}$, and adding perturbations. More precisely, the model consists of deterministic cadlag functions X_i, Y_i, Z_j, Ψ_{ij}, W_i, $i \in \mathcal{I}$, $j \in \mathcal{J}$, satisfying the equations below, for all $t \geq 0$.

$$X_i(t) = x_i^* + W_i(t) + \lambda_i t - \sum_{j \in \mathcal{J}} \mu_{ij} \int_0^t \Psi_{ij}(s) ds, \quad i \in \mathcal{I}, \tag{3.46}$$

$$Y_i(t) + \sum_{j \in \mathcal{J}} \Psi_{ij}(t) = X_i(t), \qquad i \in \mathcal{I}, \tag{3.47}$$

$$Z_j(t) + \sum_{i \in \mathcal{I}} \Psi_{ij}(t) = \nu_j + \theta_j, \qquad j \in \mathcal{J}, \tag{3.48}$$

$$\Psi_{ij}(t) = 0, \ (i,j) \in \mathcal{E}_a^c, \tag{3.49}$$

$$Y_i(t) \geq 0, \quad Z_j(t) \geq 0, \quad \Psi_{ij}(t) \geq 0, \quad i \in \mathcal{I}, \ \ j \in \mathcal{J}. \tag{3.50}$$

Above, the constants $\{x_i^*, \lambda_i, \mu_{ij}, \nu_j; i \in \mathcal{I}, j \in \mathcal{J}\}$ are as in Section 3.2, and $\{\theta_j; j \in \mathcal{J}\}$ are additional real constants. We refer to (W, θ) as *data* for the model.

Definition 3.4. *Given* $\sigma > \varepsilon > 0$*, we will say that the data* (W, θ) *for the dynamic fluid model* (3.46)–(3.50) *is an* (ε, σ)-perturbation *if*

$$\|W(t)\| \leq \varepsilon \quad \textit{for all} \quad 0 \leq t < \sigma \ , \ \textit{and} \ \ \|\theta\| \leq \varepsilon. \tag{3.51}$$

Let the heavy traffic and complete resource pooling conditions hold and, in addition, assume that the static fluid model is throughput sub-optimal. Let $\varepsilon > 0$ and $\sigma > 0$ be given and assume that the data (W, θ) is an (ε, σ)-perturbation. Below we will construct functions $\{X_i(t), i \in \mathcal{I}\}$, $\{Y_i(t), i \in \mathcal{I}\}$, $\{Z_j(t), j \in \mathcal{J}\}$ and $\{\Psi_{ij}(t), (i,j) \in \mathcal{I} \times \mathcal{J}\}$ that satisfy (3.46)–(3.50), such that $\int_0^\sigma 1_{\{e \cdot Y(s) > 0\}} ds$ and $\|X(t) - x^*\|_\sigma^*$ are $o(1)$ as ε becomes small. The precise statement will be formulated in Theorem 3.3.

Since the static fluid model is throughput sub-optimal, we have from Theorem 3.2 that there exists a simple path p with $\mu(p) < 0$. Fix such a path p. Set

$$\mathcal{E}_p^+ = \{(i,j) \in \mathcal{E}_p : s(p,i,j) = 1\}, \quad \mathcal{E}_p^- = \{(i,j) \in \mathcal{E}_p : s(p,i,j) = -1\},$$

and note that

$$\mu(p) = \sum_{(i,j) \in \mathcal{E}_p} s(p,i,j)\mu_{ij} = \Sigma_p^+ - \Sigma_p^- < 0, \tag{3.52}$$

where

$$\Sigma_p^\pm := \sum_{(i,j) \in \mathcal{E}_p^\pm} \mu_{ij}.$$

In addition, set

$$\Sigma_p^0 := \sum_{(i,j) \in \mathcal{E}_p^c \cap \mathcal{E}_{ba}} \mu_{ij}. \tag{3.53}$$

Define the constant

$$\alpha = \frac{1}{2}(1 + \Sigma_p^+ / \Sigma_p^-). \tag{3.54}$$

The following inequalities follow from (3.52) and (3.54)

$$\frac{1}{2} < \alpha < 1 \quad , \quad \Sigma_p^+ - \alpha \Sigma_p^- = \frac{1}{2}(\Sigma_p^+ - \Sigma_p^-) < 0. \tag{3.55}$$

We will need the following result from [3, Proposition 7] (that follows from the tree structure of $\mathcal{E}_{ba}$): the system of equations

$$\begin{cases} \sum_{j \in \mathcal{J}} \phi_{ij} & = a_i, \quad i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \phi_{ij} & = b_j, \quad j \in \mathcal{J}, \\ \phi_{ij} & = 0, \quad (i,j) \in \mathcal{E}_{ba}^c, \end{cases} \tag{3.56}$$

in the unknown ϕ has a unique solution, whenever a and b satisfy $\sum_i a_i = \sum_j b_j$. With

$$D_G = \Big\{(a,b) \in \mathbb{R}^I \times \mathbb{R}^J : \sum_{i\in\mathcal{I}} a_i = \sum_{j\in\mathcal{J}} b_j\Big\}, \tag{3.57}$$

denote by $G : D_G \to \mathbb{R}^{I\times J}$ the solution map, namely

$$\phi_{ij} = G_{ij}(a,b), \qquad (i,j) \in \mathcal{E}, \tag{3.58}$$

and note that this map is linear. We need a notion of an operator norm for G, and thus set

$$C_G := \sup\Big\{ \max_{ij} |G_{ij}(a,b)| \ : \ (a,b) \in D_G \ , \ \|a\| \vee \|b\| \le 1\Big\}. \tag{3.59}$$

Recall that, by the definition of basic activities, ψ^* has the property that $\psi^*_{ij} > 0$ for $(i,j) \in \mathcal{E}_{ba}$. Set

$$\delta_1 = \frac{1}{2} \min_{(i,j)\in\mathcal{E}_{ba}} \psi^*_{ij} > 0, \qquad a_0 = (2C_G)^{-1}\delta_1, \tag{3.60}$$

$$\delta_2 = \begin{cases} \delta_1 \min\Big\{ \frac{\alpha\Sigma_p^- - \Sigma_p^+}{2\Sigma_p^0}, (1-\alpha)\Big\} & \text{if } \Sigma_p^0 > 0, \\ \\ 0 & \text{if } \Sigma_p^0 = 0. \end{cases} \tag{3.61}$$

By (3.55) we have $0 \le \delta_2 < \delta_1$.

We can now construct the functions X, Y, Z, Ψ. The construction will be based on a finite sequence of times $0 = \eta_0 < \zeta_1 < \eta_1 < \zeta_2 < \cdots$ that are bounded by τ, where

$$\tau = \widetilde{\tau} \wedge \sigma \quad , \quad \widetilde{\tau} = \inf\{t \ge 0, \ \|X(t) - x^*\| \ge \varepsilon^{1/2}\} \tag{3.62}$$

and where η_k, ζ_k are to be defined recursively. Observe that $\tau > 0$, which follows from (3.46), (3.51) and $\varepsilon < \varepsilon^{1/2}$ for $\varepsilon > 0$ sufficiently small. Let $\beta = \beta(\varepsilon) \in \mathbb{R}^{\mathcal{E}}$ be a constant matrix satisfying

$$|\beta| := \max_{(i,j)\in\mathcal{E}} |\beta_{ij}| \le \varepsilon^2. \tag{3.63}$$

Define the constant matrix $\tilde{\psi} \in \mathbb{R}^{\mathcal{E}}$ as

$$\tilde{\psi}_{ij} = \begin{cases} \psi^*_{ij} + \alpha\delta_1 + \beta_{ij}, & (i,j) \in \mathcal{E}_p^-, \\ \psi^*_{ij} - \delta_1 + \beta_{ij}, & (i,j) \in \mathcal{E}_p^+, \\ \psi^*_{ij} - \delta_2 + \beta_{ij}, & (i,j) \in \mathcal{E}_p^c \cap \mathcal{E}_{ba} \ , \\ 0, & (i,j) \in \mathcal{E}_p^c \cap \mathcal{E}_{ba}^c \ . \end{cases} \tag{3.64}$$

Throughout, we use X_e for $e \cdot X$, and use a similar convention for x^*_e, ν_e and θ_e. Fix $i_0 \in \mathcal{I}$ and $j_0 \in \mathcal{J}$. Set $\eta_0 = 0$. Let $k \geq 1$ and consider the system of equations (3.65)–(3.68):

$$\begin{cases} X_i(t) = X_i(\eta_{k-1}) + W_i(t) - W_i(\eta_{k-1}) \\ \qquad\qquad +\lambda_i(t-\eta_{k-1}) - \sum_{j\in\mathcal{J}} \mu_{ij} \int_{\eta_{k-1}}^t \Psi_{ij}(s)ds, \quad i \in \mathcal{I}, & t \in [\eta_{k-1}, \zeta_k), \\ \Psi(t) = \tilde{\psi}, & t \in [\eta_{k-1}, \zeta_k), \\ Y_i(t) = X_i(t) - \sum_{j\in\mathcal{J}} \Psi_{ij}(t), \quad i \in \mathcal{I}, & t \in [\eta_{k-1}, \zeta_k), \\ Z_j(t) = \nu_j + \theta_j - \sum_{i\in\mathcal{I}} \Psi_{ij}(t), \quad j \in \mathcal{J}, & t \in [\eta_{k-1}, \zeta_k), \end{cases} \tag{3.65}$$

where

$$\zeta_k = \inf\{t \geq \eta_{k-1} \ : \ X_e(t) - X_e(\eta_{k-1}) \leq -7\varepsilon\} \wedge \tau, \tag{3.66}$$

and

$$\begin{cases} X_i(t) = X_i(\zeta_k) + W_i(t) - W_i(\zeta_k) \\ \qquad\qquad +\lambda_i(t-\zeta_k) - \sum_{j\in\mathcal{J}} \mu_{ij} \int_{\zeta_k}^t \Psi_{ij}(s)ds, \quad i \in \mathcal{I}, & t \in [\zeta_k, \eta_k), \\ Y(t) = (X_e(t) - \nu_e - \theta_e)^+ e_{i_0},\ Z(t) = (X_e(t) - \nu_e - \theta_e)^- e_{j_0}, & t \in [\zeta_k, \eta_k), \\ \Psi(t) = G(X(t) - Y(t),\ \nu + \theta - Z(t)), & t \in [\zeta_k, \eta_k), \end{cases} \tag{3.67}$$

where

$$\eta_k = \inf\{t \geq \zeta_k \ : \ \|X(t) - X(\zeta_k)\| \geq 3\varepsilon\} \wedge \tau. \tag{3.68}$$

Lemma 3.4. *Equations* (3.65)–(3.68) *uniquely define a finite sequence* η_k, ζ_k *and functions* X, Y, Z *and* Ψ *on* $t \in [0, \tau)$. *For all* $\varepsilon > 0$ *sufficiently small, these functions satisfy* (3.46)–(3.50) *on* $[0, \tau)$.

Proof. Note first that, if $k \geq 1$ and η_{k-1} and $X(\eta_{k-1})$ are given then the first two equations in (3.65) define X and Ψ uniquely on $[\eta_{k-1}, \zeta_k)$. The last two lines of (3.65) define Y and Z on the same interval. Moreover, it is easy to see that equations (3.46)–(3.49) are satisfied on $[0, \zeta_k)$, provided that they are satisfied on $[0, \eta_{k-1})$. The validity of the constraints (3.50) is argued later.

Next, let ζ_k and $X(\zeta_k)$ be given. Substituting into the first equation in (3.67) the values for Y, Z and Ψ from the last three equations of (3.67) results with an equation of the form

$$X_i(t) = X_i(\zeta_k) + W_i(t) - W_i(\zeta_k) + \int_{\zeta_k}^{t} F_i(X(s))ds, \quad i \in \mathcal{I},$$

where $F_i : \mathbb{R}^I \to \mathbb{R}$ are globally Lipschitz. This uniquely defines X, and in turn, Y, Z and Ψ on $[\zeta_k, \eta_k)$. Thus X, Y, Z and Ψ are uniquely defined on $[0, \eta_k)$, provided that they are on $[0, \zeta_k)$. Let

$$K = \inf\{k \geq 0 \ : \ \eta_k = \tau \ \text{ or } \ \zeta_{k+1} = \tau\}. \tag{3.69}$$

We show that $K < \infty$. To this end, observe that, for $\varepsilon > 0$ sufficiently small, we have $\zeta_1 \geq \varepsilon^2 > 0$. Indeed, if $\zeta_1 = \sigma$, this is clear. Otherwise, (3.51), (3.48) and (3.50) imply the inequality $\Psi_{ij}(t) \leq \nu_j + \varepsilon$, and as a result, from (3.65) and (3.51) it follows that for any $0 \leq t \leq \zeta_1$

$$|X_e(t) - X_e(0)| \leq \|X(t) - X(0)\| \leq 2\varepsilon + (c_1 + c_2\varepsilon)t, \tag{3.70}$$

where $c_1 = \sum_i \lambda_i + \sum_{ij} \mu_{ij}\nu_j$ and $c_2 = \sum_{ij} \mu_{ij}$. By right-continuity of X and (3.66) we have that $|X_e(\zeta_1) - X_e(0)| \geq 7\varepsilon$. Thus $\zeta_1 \geq 5\varepsilon(c_1 + c_2\varepsilon)^{-1} \geq \varepsilon^2$ for ε sufficiently small. For $k \leq K$ denote

$$I_k^1 = [\eta_{k-1}, \ \zeta_k), \quad I_k^2 = [\zeta_k, \ \eta_k), \quad I_k = I_k^1 \cup I_k^2. \tag{3.71}$$

An argument similar to the one for $\zeta_1 \geq \varepsilon^2$ shows that each of the intervals I_k^1 and I_k^2 has a length of at least ε^2. Hence K is finite. The inductive argument given above thus shows that the functions X, Y, Z and Ψ are uniquely defined on $[0, \tau)$ and satisfy (3.46)–(3.49). We now show that the relations (3.50) hold. The two interval types are treated separately.

Intervals I_k^1: Let k and $t \in I_k^1$ be fixed. By (3.65), $\Psi_{ij}(t)$ is given by $\widetilde{\psi}$, defined in (3.64). The nonnegativity of each $\widetilde{\psi}_{ij}$ for all sufficiently small ε follows from (3.60) and $0 \leq \delta_2 < \delta_1$ (cf. (3.61)). To show that Z_j are nonnegative, note that by (3.65) and (3.64)

$$\begin{aligned} Z_j(t) &= \nu_j + \theta_j - \sum_{i \in \mathcal{I}} \Psi_{ij}(t) = \nu_j + \theta_j - \sum_{i \in \mathcal{I}} \widetilde{\psi}_{ij} \\ &\geq \theta_j - C|\beta| + \delta_1 |\{i : (i,j) \in \mathcal{E}_p^+\}| \\ &\quad - \alpha\delta_1 |\{i : (i,j) \in \mathcal{E}_p^-\}| + \delta_2 |\{i : (i,j) \in \mathcal{E}_p^c \cap \mathcal{E}_{ba}\}|, \end{aligned}$$

where $C = |\mathcal{E}_a|$, and we have used (3.9) in the last line. By Definition 3.2 and (3.17), for any vertex $j \in \mathcal{J}$ in $\mathcal{V}_p$ there exists exactly one edge, $(i_1, j) \in \mathcal{E}_p$, with $s(p, i_1, j) = 1$, and there exists at most one edge $(i_2, j) \in \mathcal{E}_p$ with $s(p, i_2, j) = -1$ (in the case where p is an open simple path and j is a leaf, such i_2 does not exist). Thus the positivity of δ_1, nonnegativity of δ_2 and the bounds on β and θ show that $Z_j(t) \geq (1-\alpha)\delta_1 - (C+1)\varepsilon \geq 0$ for all ε sufficiently small and $j \in \mathcal{J}$.

Given $i \in \mathcal{I}$, by similar considerations,

$$\begin{aligned}\sum_{j\in\mathcal{J}} \Psi_{ij}(t) &= \sum_{j\in\mathcal{J}} \widetilde{\psi}_{ij} \\ &\leq x_i^* + C|\beta| - \delta_1|\{j : (i,j) \in \mathcal{E}_p^+\}| \\ &\quad + \alpha\delta_1|\{j : (i,j) \in \mathcal{E}_p^-\}| - \delta_2|\{j : (i,j) \in \mathcal{E}_p^c \cap \mathcal{E}_{ba}\}| \\ &\leq x_i^* + C\varepsilon - (1-\alpha)\delta_1.\end{aligned}$$

Since $t < \tau$, we have by (3.62) that $\|X(t) - x^*\| < \varepsilon^{1/2}$. By (3.55) and (3.60), $(1-\alpha)\delta_1 > 0$. We conclude that $\sum_{j\in\mathcal{J}} \Psi_{ij}(t) \leq X_i(t)$, and in turn by (3.65), $Y_i(t) \geq 0$, provided that ε is sufficiently small.

Intervals I_k^2: Fix k and $t \in I_k^2$. The nonnegativity of $Y_i(t)$ and $Z_j(t)$ is immediate from (3.67). It remains to show that $\Psi_{ij}(t) \geq 0$, $(i,j) \in \mathcal{E}$. This follows from [4, Lemma 3], noting that its assumption $\|X(t) - x^*\| \leq a_0$ is satisfied for $t < \tau$ and ε sufficiently small, and that the special structure assumed in [4, Assumption 3] is not used in the proof of the cited lemma. This concludes the proof of Lemma 3.4. □

Theorem 3.3. *Let the heavy traffic and complete resource pooling conditions hold. Assume that the static fluid model is throughput sub-optimal. Then there exist functions γ_1 and γ_2 from $(0,\infty)$ to itself, satisfying $\lim_{\varepsilon\to 0}\gamma_1(\varepsilon) = 0$ and $\lim_{\varepsilon\to 0}\gamma_2(\varepsilon) = \infty$, such that the following statement holds. If the data (W,θ) for the dynamic fluid model is an (ε,σ)-perturbation then the functions X, Y, Z and Ψ, that are uniquely defined by (3.65)–(3.68), satisfy*

$$\int_0^{\sigma\wedge\gamma_2(\varepsilon)} 1_{\{e\cdot Y(s)>0\}} ds \leq \gamma_1(\varepsilon), \tag{3.72}$$

$$\|X(t) - x^*\| \leq \gamma_1(\varepsilon) \quad \textit{for all} \quad 0 \leq t \leq \sigma \wedge \gamma_2(\varepsilon). \tag{3.73}$$

The proof of the following lemma appears at the end of the section.

Lemma 3.5. *Recall the definitions of K (3.69) and intervals I_k^1 and I_k^2 from (3.71). There exist constants m_1, m_2, $m_3 \in (0,\infty)$, not depending on ε, σ and k, such that for any $k \leq K$*

1. $|I_k^1| \le m_1\varepsilon$;
2. $\|X - x^*\|_{\eta_k}^* \le km_2\varepsilon$;
3. $|I_k^2| \ge m_3/k$.

Proof of Theorem 3.3. We begin by showing that

$$Y(t) = 0 \text{ for all } t \in [\zeta_k, \eta_k), k < K. \tag{3.74}$$

By (3.67), it suffices to show that

$$X_e(t) - \nu_e - \theta_e \le 0, \text{ for all } t \in [\zeta_k, \eta_k), k < K. \tag{3.75}$$

Indeed, from (3.66), (3.51) and using $\nu_e = x_e^*$ (by (3.9)), we have

$$X_e(\zeta_1) - \nu_e - \theta_e \le X_e(\zeta_1) - X_e(0) + (X_e(0) - x_e^*) - \theta_e \le -7\varepsilon + \varepsilon - \theta_e \le -5\varepsilon.$$

Then by (3.68), taking into account the possibility of jumps of at most 2ε for W_e, we have for $t \in [\zeta_1, \eta_1)$:

$$X_e(t) - \nu_e - \theta_e \le X_e(\zeta_1) - \nu_e - \theta_e + \|X(t) - X(\zeta_1)\| \le -5\varepsilon + 5\varepsilon \le 0.$$

A proof by induction that repeats the above argument, using (3.66) and (3.68) shows that (3.75), and in turn (3.74), holds for $k \ge 1$.

Next let us show that

$$\tau \ge \sigma \wedge \gamma_2(\varepsilon), \tag{3.76}$$

where $\gamma_2(\varepsilon) := \frac{m_3}{4}|\log\varepsilon|$, for sufficiently small ε. Consider the number $k_0 = k_0(\varepsilon) := [(2m_2\varepsilon^{1/2})^{-1}] \wedge K$ (where K is as in (3.69)). If $k_0 = K$ and $\tau = \eta_K$ then from (3.62), (3.69) and using Lemma 3.5(2), we have $\tau = \sigma$, since then $\|X - x^*\|_{\eta_K} < \varepsilon^{1/2}$. Otherwise, if $k_0 = K$ and $\tau = \zeta_{K+1}$, or $k_0 = [(2m_2\varepsilon^{1/2})^{-1}]$, one uses (3.62), (3.69) and Lemma 3.5(2),(3) to obtain

$$\tau \ge \eta_{k_0} \ge \sum_{l=1}^{k_0(\varepsilon)} \frac{m_3}{l} \ge \frac{m_3}{4}|\log\varepsilon|.$$

Hence (3.76) follows.

Let $K_0 = K_0(\varepsilon) = \max\{k : \zeta_k \le \sigma \wedge \gamma_2(\varepsilon)\}$. By Lemma 3.5(3), for sufficiently small ε,

$$\sum_{k=1}^{K_0-1} k^{-1} \le m_3^{-1}\gamma_2(\varepsilon) = -\frac{1}{4}\log\varepsilon.$$

This implies that $\frac{1}{2}\log K_0 \le -\frac{1}{4}\log\varepsilon$, hence $K_0 \le \varepsilon^{-1/2}$, provided that ε is small.

Now, using (3.74), Lemma 3.5(1) and the estimate on K_0, we have

$$\int_0^{\sigma\wedge\gamma_2(\varepsilon)} 1_{\{e\cdot Y(s)>0\}} ds \le (K_0(\varepsilon)+1)m_1\varepsilon \le 2m_1\varepsilon^{1/2} \le \gamma_1(\varepsilon),$$

where $\gamma_1(\varepsilon) := 2(m_1 \vee 1)\varepsilon^{1/2}$, establishing (3.72). As a result of (3.62) and (3.76) we obtain that

$$\|X(t) - x^*\| < \varepsilon^{1/2} \text{ for all } t < \sigma \wedge \gamma_2(\varepsilon).$$

As a result, $\|X - x^*\|^*_{\sigma\wedge\gamma_2(\varepsilon)} \le \varepsilon^{1/2} + 2\varepsilon \le \gamma_1(\varepsilon)$. This shows (3.73) and completes the proof of Theorem 3.3. □

Proof of Lemma 3.5. By (3.64) and (3.65), the dynamics of X on the intervals I_k^1 is given by

$$X(t) = X(\eta_{k-1}) + W(t) - W(\eta_{k-1}) + (t-\eta_{k-1})(r+b), \tag{3.77}$$

where

$$r_i := \lambda_i - \sum_{j\in\mathcal{J}} \mu_{ij}\psi^*_{ij} - \alpha\delta_1 \sum_{j:(i,j)\in\mathcal{E}_p^-} \mu_{ij} + \delta_1 \sum_{j:(i,j)\in\mathcal{E}_p^+} \mu_{ij} + \delta_2 \sum_{j:(i,j)\in\mathcal{E}_p^c\cap\mathcal{E}_{ba}} \mu_{ij}, \tag{3.78}$$

and

$$b_i = b_i(\mu,\beta) := -\sum_{j:(i,j)\in\mathcal{E}_p} \mu_{ij}\beta_{ij} - \sum_{j:(i,j)\in\mathcal{E}_p^c\cap\mathcal{E}_{ba}} \mu_{ij}\beta_{ij}.$$

By (3.9), (3.55), (3.60)–(3.61) and (3.78) we have

$$\sum_{i\in\mathcal{I}} r_i = \delta_2\Sigma_p^0 + \delta_1\Sigma_p^+ - \alpha\delta_1\Sigma_p^- \le \frac{1}{2}(\delta_1\Sigma_p^+ - \alpha\delta_1\Sigma_p^-) < 0. \tag{3.79}$$

Note that by (3.51) and (3.46), $|\Delta X_e| \le 2\varepsilon$. From (3.65) and (3.66), and using (3.79), we thus obtain for I_k^1, $k \ge 1$,

$$\begin{aligned} -10\varepsilon &\le X_e(\zeta_k) - X_e(\eta_{k-1}) \\ &\le W_e(\zeta_k) - W_e(\eta_{k-1}) + (e\cdot r + \|b\|)(\zeta_k - \eta_{k-1}) \\ &\le 2\varepsilon + (e\cdot r + c_1\varepsilon^2)(\zeta_k - \eta_{k-1}), \end{aligned}$$

for $c_1 = \sum_{ij} \mu_{ij}$, and where we also used (3.51) and (3.63). Therefore for ε sufficiently small,

$$|I_k^1| = \zeta_k - \eta_{k-1} \le m_1 \varepsilon, \tag{3.80}$$

where $m_1 = 24/|e \cdot r|$. This proves part 1 of the lemma.

From (3.77) and (3.80), for $t \in \overline{I_k^1}$ and sufficiently small ε,

$$\|X(t) - X(\eta_{k-1})\| \le 2\varepsilon + c_2(t - \eta_{k-1}) \le (2 + c_2 m_1)\varepsilon,$$

where $c_2 = 2\|r\|$. We therefore have

$$\sup_{t \in [\eta_{k-1}, \zeta_k]} \|X(t) - X(\eta_{k-1})\| \le (2 + c_2 m_1)\varepsilon. \tag{3.81}$$

By (3.68), $\|X(t) - X(\zeta_k)\| \le 3\varepsilon$ for all $t \in I_k^2$, and taking into account a possible jump at η_k, we have

$$\sup_{t \in [\zeta_k, \eta_k]} \|X(t) - X(\zeta_k)\| \le 5\varepsilon. \tag{3.82}$$

Since by (3.46) and (3.51), $\|X(0) - x^*\| \le \varepsilon$, part 2 of the lemma follows from (3.81) and (3.82).

In view of (3.67), (3.74) and (3.75), we have on I_k^2

$$\begin{cases} Y(t) = 0, \\ Z(t) = -(X_e(t) - \nu_e - \theta_e)\, e_{j_0}, \\ \Psi_{ij}(t) = G_{ij}(X(t),\ \nu + \theta - Z(t)). \end{cases} \tag{3.83}$$

Define $\widetilde{X}(t) = X(t) - x^*$. From the definition of map G (3.56)–(3.58) we have

$$\begin{aligned} &G_{ij}(\widetilde{X}(t) + x^*,\ \nu + \theta - Z(t)) \\ &\qquad = G_{ij}(x^*,\ \nu) + G_{ij}(\widetilde{X}(t) - \theta_e e_{i_0}\ ,\ -Z(t)) + G_{ij}(\theta_e e_{i_0}\ ,\ \theta). \end{aligned} \tag{3.84}$$

Due to (3.9) we have $G_{ij}(x^*, \nu) = \psi_{ij}^*$. Now consider the second term in (3.84). Using (3.83)

$$\begin{aligned} G_{ij}(\widetilde{X}(t) - \theta_e e_{i_0}\ ,\ -Z(t)) &= G_{ij}(\widetilde{X}(t) - \theta_e e_{i_0}\ ,\ (X_e(t) - \nu_e - \theta_e)\, e_{j_0}) \\ &= G_{ij}(\widetilde{X}(t),\ \widetilde{X}_e(t)\, e_{j_0}) \\ &\quad +\ G_{ij}(-\theta_e e_{i_0}\ ,\ -\theta_e e_{j_0}), \end{aligned} \tag{3.85}$$

where we used $\widetilde{X}_e = X_e - \nu_e$ due to $x_e^* = \nu_e$ (3.9). Finally, from (3.83)–(3.85), we have

$$\begin{aligned}\Psi_{ij}(t) = \psi_{ij}^* + G_{ij}(\widetilde{X}(t),\ \widetilde{X}_e(t)\ e_{j_0}) \\ +\ G_{ij}(-\theta_e e_{i_0}\ ,\ -\theta_e e_{j_0}) + G_{ij}(\theta_e e_{i_0}\ ,\ \theta), \qquad t \in I_k^2.\end{aligned} \tag{3.86}$$

Define the map H by

$$H_i(x) := -\sum_j \mu_{ij} G_{ij}(x,\ x_e e_{j_0}), \qquad x \in \mathbb{R}^I\ , \quad i \in \mathcal{I}, \tag{3.87}$$

and the constant H^θ by

$$H_i^\theta := -\sum_j \mu_{ij}\Big[G_{ij}(-\theta_e e_{i_0}\ ,\ -\theta_e e_{j_0}) + G_{ij}(\theta_e e_{i_0}\ ,\ \theta)\Big], \qquad i \in \mathcal{I}. \tag{3.88}$$

By the heavy traffic conditions, $\sum_{j\in\mathcal{J}} \mu_{ij}\psi_{ij}^* = \lambda_i$. Hence using (3.46), (3.83)–(3.88), we have

$$\widetilde{X}(t) = \widetilde{X}(\zeta_k) + W(t) - W(\zeta_k) + \int_{\zeta_k}^t H(\widetilde{X}(u))du + H^\theta(t - \zeta^k), \qquad t \in I_k^2, \tag{3.89}$$

By (3.56)–(3.58) and (3.51), there exist constants $c_H > 0$ and $l_H > 0$, such that

$$\|H(x)\| \le c_H\|x\|, \qquad \|H^\theta\| \le l_H\varepsilon, \tag{3.90}$$

for ε sufficiently small. Therefore, applying (3.51), (3.90) and Lemma 3.5(2) to (3.89), we have from (3.68)

$$3\varepsilon \le \|X(\eta_k) - X(\zeta_k)\| = \|\widetilde{X}(\eta_k) - \widetilde{X}(\zeta_k)\| \le 2\varepsilon + (c_H k m_2 + l_H)(\eta_k - \zeta_k)\varepsilon.$$

The above implies $\varepsilon \le (c_H k m_2 + l_H)(\eta_k - \zeta_k)\varepsilon$. Therefore, for $k \ge 1$

$$|I_k^2| = \eta_k - \zeta_k \ge \frac{1}{c_H k m_2 + l_H} \ge \frac{m_3}{k}, \qquad m_3 := \frac{1}{c_3 c_H m_2} < 1, \tag{3.91}$$

where the constant c_3 satisfies $(c_3 - 1)c_H m_2 \ge l_H$. This concludes the proof of Lemma 3.5. □

3.5. Estimates on the probabilistic model

In this section we prove Theorem 3.1. We begin by introducing a rescaled version of the processes defined in Section 3.2 as follows. For $n \in \mathbb{N}$ and $t \geq 0$, let

$$\bar{X}_i^n(t) = n^{-1} X_i^n(t), \quad \bar{Y}_i^n(t) = n^{-1} Y_i^n(t), \quad i \in \mathcal{I} \tag{3.92}$$

$$\bar{Z}_j^n(t) = n^{-1} Z_j^n(t), \quad \bar{\Psi}_{ij}^n(t) = n^{-1} \Psi_{ij}^n(t), \quad i \in \mathcal{I}, \ \ j \in \mathcal{J}. \tag{3.93}$$

Denote $\bar{X}^n = (\bar{X}_i^n, i \in \mathcal{I})$, and use a similar convention for $\bar{Y}^n$, $\bar{Z}^n$ and $\bar{\Psi}^n$. Following a straightforward calculation, relations (3.5)–(3.4) can be rewritten in terms of the rescaled processes, as equations (3.94)–(3.98) below, holding for $n \in \mathbb{N}$ and $t \geq 0$:

$$\bar{\Psi}_{ij}^n(t) = 0, \ (i,j) \in \mathcal{E}_a^c, \tag{3.94}$$

$$\bar{X}_i^n(t) = x_i^* + \bar{W}_i^n(t) + \lambda_i t - \sum_{j \in \mathcal{J}} \mu_{ij} \int_0^t \bar{\Psi}_{ij}^n(s) ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, \tag{3.95}$$

$$\bar{Y}_i^n(t) + \sum_{j \in \mathcal{J}} \bar{\Psi}_{ij}^n(t) = \bar{X}_i^n(t), \qquad i \in \mathcal{I}, \tag{3.96}$$

$$\bar{Z}_j^n(t) + \sum_{i \in \mathcal{I}} \bar{\Psi}_{ij}^n(t) = \nu_j + \theta_j^n, \qquad j \in \mathcal{J}, \tag{3.97}$$

$$\bar{Y}_i(t) \geq 0, \ \bar{Z}_j(t) \geq 0, \ \bar{\Psi}_{ij}^n(t) \geq 0 \qquad i \in \mathcal{I}, \ j \in \mathcal{J}, \tag{3.98}$$

where we set

$$\begin{aligned} \bar{W}_i^n(t) : = \ & n^{-1}[A_i^n(t) - \lambda_i^n t] \\ & - n^{-1} \sum_{j \in \mathcal{J}} \left[S_{ij}^n \left(n \int_0^t \bar{\Psi}_{ij}^n(s) ds \right) - n \mu_{ij}^n \int_0^t \bar{\Psi}_{ij}^n(s) ds \right] \\ & + (n^{-1} X_i^{0,n} - x_i^*) + (n^{-1} \lambda_i^n - \lambda_i) t - \sum_{j \in \mathcal{J}} (\mu_{ij}^n - \mu_{ij}) \int_0^t \bar{\Psi}_{ij}^n(s) ds \end{aligned} \tag{3.99}$$

and

$$\theta_j^n = n^{-1} N_j^n - \nu_j. \tag{3.100}$$

The above equations resemble the dynamic fluid model studied in Section 3.4, and the proof of Theorem 3.1 will rely on the results of this section.

Lemma 3.6. *Under any SCP, for any given $T \in (0,\infty)$, $\{n^{1/2}\|\bar{W}^n\|^*_T, n \in \mathbb{N}\}$ are tight random variables.*

PROOF. Relations (3.3), (3.4) and (2.8) imply that $0 \leq \bar{\Psi}^n_{ij}(t) \leq c_1$, where c_1 is a constant independent of i, j, n and t. Hence by (3.14), the last three terms in (3.99) are bounded by $c_2(T+1)n^{-1/2}$, where c_2 is a constant independent of i, j, n and t. Denote $\hat{A}^n_i(t) := n^{-1/2}(A^n_i(t) - \lambda^n_i t)$ and $\hat{S}^n_{ij}(t) := n^{-1/2}(S^n_{ij}(nt) - n\mu^n_{ij}t)$. Theorem 14.6 of [13] shows that $\{\hat{A}^n_i, n \in \mathbb{N}\}$ converges weakly to a Brownian motion (with zero mean and variance that depends on i), and that a similar statement holds for $\{\hat{S}^n_{ij}, n \in \mathbb{N}\}$. It follows that $\{|\hat{A}^n_i|^*_T, n \in \mathbb{N}\}$ and $\{|\hat{S}^n_{ij}|^*_{cT}, n \in \mathbb{N}\}$ are tight random variables, for each i, j, whenever c is a constant that is independent of n. By (3.99) we obtain that

$$n^{1/2}|\bar{W}^n_i|^*_T \leq |\hat{A}^n_i|^*_T + |\hat{S}^n_{ij}|^*_{c_1 T} + c_2(T+1). \tag{3.101}$$

As a result, $\{n^{1/2}|\bar{W}^n_i|^*_T, n \in \mathbb{N}\}$ are tight random variables, and the lemma follows. □

PROOF OF THEOREM 3.1. For $n \in \mathbb{N}$, let $\varepsilon^n = n^{-1/2}\log n$. By (3.14), for sufficiently large n,

$$\|\theta^n\| \leq \varepsilon^n. \tag{3.102}$$

For $n \in \mathbb{N}$, let

$$\tilde{\psi}^n_{ij} = \begin{cases} \psi^*_{ij} + \alpha\delta_1 + \beta^n_{ij}, & (i,j) \in \mathcal{E}^-_p, \\ \psi^*_{ij} - \delta_1 + \beta^n_{ij}, & (i,j) \in \mathcal{E}^+_p, \\ \psi^*_{ij} - \delta_2 + \beta^n_{ij}, & (i,j) \in \mathcal{E}^c_p \cap \mathcal{E}_{ba}\ , \\ 0, & (i,j) \in \mathcal{E}^c_p \cap \mathcal{E}^c_{ba}, \end{cases} \tag{3.103}$$

where β^n_{ij} are constants chosen in such a way that, for all sufficiently large n one has $|\beta^n_{ij}|^2 \leq (\varepsilon^n)^2$, and $n\tilde{\psi}^n_{ij}$ has integer values, for each i, j and n. Below, we write a system of equations for the processes $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n)$ that uniquely defines them. We then let the processes (X^n, Y^n, Z^n, Ψ^n) be defined through (3.92), (3.93). These processes will then be shown to form a SCP, and to satisfy the statement of the theorem.

Fix $i_0 \in \mathcal{I}, j_0 \in \mathcal{J}$. Set $\eta_0^n = 0$, and consider the system of equations:

$$\begin{cases} \bar{X}_i^n(t) = \bar{X}_i^n(\eta_{k-1}^n) + \bar{W}_i^n(t) - \bar{W}_i^n(\eta_{k-1}^n) \\ \qquad\qquad +\lambda_i(t-\eta_{k-1}^n) - \sum_{j\in\mathcal{J}} \mu_{ij} \int_{\eta_{k-1}^n}^t \bar{\Psi}_{ij}^n(s)ds, \ i \in \mathcal{I}, & t \in [\eta_{k-1}^n, \zeta_k^n), \\ \bar{\Psi}^n(t) = \tilde{\psi}^n, & t \in [\eta_{k-1}^n, \zeta_k^n), \\ \bar{Y}_i^n(t) = \bar{X}_i^n(t) - \sum_{j\in\mathcal{J}} \bar{\Psi}_{ij}^n(t), \quad i \in \mathcal{I}, & t \in [\eta_{k-1}^n, \zeta_k^n), \\ \bar{Z}_j^n(t) = \nu_j + \theta_j^n - \sum_{i\in\mathcal{I}} \bar{\Psi}_{ij}^n(t), \quad j \in \mathcal{J}, & t \in [\eta_{k-1}^n, \zeta_k^n), \end{cases} \tag{3.104}$$

where $\bar{W}^n$ is given by (3.99),

$$\zeta_k^n = \inf\{t \geq \eta_{k-1}^n \ : \ \bar{X}_e^n(t) - \bar{X}_e^n(\eta_{k-1}^n) \leq -7\varepsilon^n\} \wedge \tau^n, \tag{3.105}$$

and

$$\begin{cases} \bar{X}_i^n(t) = \bar{X}_i^n(\zeta_k^n) + \bar{W}_i^n(t) - \bar{W}_i^n(\zeta_k^n) \\ \qquad\qquad +\lambda_i(t-\zeta_k^n) - \sum_{j\in\mathcal{J}} \mu_{ij} \int_{\zeta_k^n}^t \bar{\Psi}_{ij}^n(s)ds, \quad i \in \mathcal{I}, & t \in [\zeta_k^n, \eta_k^n), \\ \bar{Y}^n(t) = (\bar{X}_e^n(t) - \nu_e - \theta_e^n)^+ e_{i_0}, & t \in [\zeta_k^n, \eta_k^n), \\ \bar{Z}^n(t) = (\bar{X}_e^n(t) - \nu_e - \theta_e^n)^- e_{j_0}, & t \in [\zeta_k^n, \eta_k^n), \\ \bar{\Psi}^n(t) = G(\bar{X}^n(t) - \bar{Y}^n(t), \ \nu + \theta^n - \bar{Z}^n(t)), & t \in [\zeta_k^n, \eta_k^n), \end{cases} \tag{3.106}$$

where

$$\eta_k^n = \inf\{t \geq \zeta_k^n \ : \ \|\bar{X}^n(t) - \bar{X}^n(\zeta_k^n)\| \geq 3\varepsilon^n\} \wedge \tau^n, \tag{3.107}$$

$$\tau^n = \widetilde{\tau}^n \wedge \sigma^n, \quad \widetilde{\tau}^n = \inf\{t \geq 0 : \|\bar{X}^n(t) - x^*\| \geq (\varepsilon^n)^{1/2}\}, \tag{3.108}$$

$$\sigma^n = \inf\{t \geq 0 : \|\bar{W}^n(t)\| \geq \varepsilon^n\}, \tag{3.109}$$

and finally,

$$\begin{cases} \bar{X}_i^n(t) = \bar{X}_i^n(\tau^n) + n^{-1}(A_i^n(t) - A_i^n(\tau^n)), \quad t \geq \tau^n, \\ \bar{Y}_i^n(t) = \bar{X}_i^n(t), \ \bar{Z}_j^n(t) = n^{-1}N_j^n, \ \bar{\Psi}_{ij}^n(t) = 0, \quad i \in \mathcal{I}, j \in \mathcal{J}, t \geq \tau^n. \end{cases} \tag{3.110}$$

The above equations mimic the deterministic model (3.62)–(3.68), and therefore what is established in Section 3.4 can be used here. Notable difference is the definition of the processes for times $t \geq \tau^n$. Note that, by definition of τ^n, $\|\bar{W}^n(t)\| \leq \varepsilon^n$ holds for all $t < \tau^n$. This and (3.102) provide a bound that is similar to (3.51) over $[0, \tau^n)$. As a result, Lemma 3.4 implies that, given the primitive processes A_i^n and S_{ij}^n, the processes $\bar{X}^n, \bar{Y}^n, \bar{Z}^n$ and $\bar{\Psi}^n$ and the random times η_k^n, ζ_k^n and τ^n are uniquely defined by equations (3.104)–(3.110). It also follows from Lemma 3.4 that the processes $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n)$ satisfy (3.94)–(3.98) over $[0, \tau^n)$. In turn, the processes (X^n, Y^n, Z^n, Ψ^n) satisfy (3.5)–(3.4) on this interval. It is also easy to check that these equations are satisfied for $t \geq \tau^n$. We show that the processes $(X_i^n, Y_j^n, Z_j^n, \Psi_{ij}^n)$ take values in $\mathbb{Z}_+$. Since we have proved that (3.5)–(3.4) are satisfied, it suffices to show that Ψ_{ij}^n take integer values. By construction of $\widetilde{\psi}$, Ψ_{ij}^n take integer values for $t \in [\eta_{k-1}^n, \zeta_k^n)$. On the intervals $[\zeta_k^n, \eta_k^n)$, by (3.67), Ψ_{ij}^n will be the solution of the system of equations (3.56) with integer right hand sides. In this case, a simple argument that uses the tree structure of $\mathcal{E}_{ba}$ shows that Ψ_{ij}^n are all integer valued.

To show that the constructed processes form a SCP, it remains to prove that, for every t, $\Psi^n(t)$ is measurable on $\sigma\{X^n(s), A^n(s) : s \leq t\}$ (cf. Definition 3.1). Fix t. We will show in steps (a)–(d) below that the value of $\Psi^n(t)$ is uniquely determined by the sample path $\Lambda[0,t] := \{X^n(s), A^n(s) : s \in [0,t]\}$.

(a) By (3.1), the sample path $\Lambda[0,t]$ uniquely determines the sample paths $\sum_{j \in \mathcal{J}} S_{ij}^n \left(\int_0 \Psi_{ij}^n(u) du \right)$, $i \in \mathcal{I}$ on $[0,t]$.

(b) By (3.105), (3.107), $\Lambda[0,t]$ along with the value $\tau^n \wedge t$ uniquely determine the values $\eta_k^n \wedge t$, $\zeta_k^n \wedge t$, $k = 1, \ldots, K$. Thus by (3.104), (3.106) and (3.110), $\Lambda[0,t]$ and $\tau^n \wedge t$ uniquely determine Ψ^n on $[0,t]$. Equation (3.99), along with (a) above, shows that the same data, $\Lambda[0,t]$ and $\tau^n \wedge t$, uniquely determine $\bar{W}^n$ on $[0,t]$.

(c) We next show that $\Lambda[0,t]$ determines Ψ^n and $\bar{W}^n$ on $[0,t)$. Let $(\Psi_1^n, \bar{W}_1^n)$, $(\Psi_2^n, \bar{W}_2^n)$ be two sample paths that correspond to the same data $\Lambda[0,t]$. Argue by contradiction and assume that on $[0,t)$ they do not agree. It follows from (b) that the corresponding values of τ^n, that we denote by τ_1^n and τ_2^n, do not agree, and that $\tau_1^n \wedge \tau_2^n < t$. Without loss of generality, assume that $\tau_1^n < \tau_2^n \wedge t$. Using (b) again, we have that $(\Psi_1^n, \bar{W}_1^n) = (\Psi_2^n, \bar{W}_2^n)$ on $[0, \tau_1^n]$. In particular,

$$\bar{W}_1^n(\tau_1^n) = \bar{W}_2^n(\tau_1^n). \tag{3.111}$$

Since $\widetilde{\tau}^n$ is defined in terms of X^n,

$$\widetilde{\tau}_1^n \wedge t = \widetilde{\tau}_2^n \wedge t. \tag{3.112}$$

Recall that σ^n is the time when $\bar{W}^n$ leaves an open set of $\mathbb{R}^I$ (3.109). Thus $\bar{W}_1^n(\tau_1^n)$ is either inside the open set, in which case $\tau_1^n = \widetilde{\tau}_1^n < \sigma_1^n \wedge \sigma_2^n$ and therefore by (3.112) $\tau_1^n = \tau_2^n$, or it is outside the open set, in which case $\tau_1^n = \sigma_1^n$, and by (3.111), we have that $\sigma_1^n = \sigma_2^n$ and $\tau_1^n = \tau_2^n$. In both cases we obtain a contradiction to $\tau_1^n < \tau_2^n$. We conclude that statement (c) holds.

(d) By (3.99) and (a) above, $\bar{W}^n(t)$ is uniquely determined by $\mathit{\Lambda}[0,t]$ along with the values of $\mathit{\Psi}^n$ over $[0,t)$. Hence in view of (c), $\bar{W}^n(t)$ is determined by $\mathit{\Lambda}[0,t]$. Thus the right continuity of $\bar{W}^n$ and the definition of σ^n imply that $\sigma^n \wedge t$ is determined by $\mathit{\Lambda}[0,t]$. Hence so is $\tau^n \wedge t$ and, by (a), so is $\mathit{\Psi}^n(t)$.

Finally, we show that (3.15) and (3.16) hold for any fixed $T \in (0,\infty)$ and $\varrho > 1/2$. To this end, note that

$$\mathbb{P}(\sigma^n < T) \le \mathbb{P}(\|\bar{W}^n\|_T^* \ge \varepsilon^n) = \mathbb{P}(n^{1/2}\|\bar{W}^n\|_T^* \ge \log n) \to 0,$$

by Lemma 3.6. On the event $\sigma^n \ge T$, we can use Theorem 3.3. On this event, for all n sufficiently large, we have $T \le \gamma_2(\varepsilon_n) \wedge \sigma^n$, thus Theorem 3.3 implies that $\int_0^T 1_{\{e\cdot Y^n(s)>0\}} ds \le \gamma_1(\varepsilon^n)$. Since $\varepsilon^n \to 0$ and $\gamma_1(0+) = 0$, (3.15) follows.

The second and third parts of Lemma 3.5 imply that there is a constant $C(T) < \infty$, independent of n, such that $\|\bar{X}^n - x^*\|_T^* \le C(T)\varepsilon^n$ on the event $\sigma^n \ge T$. On this event we therefore have

$$n^{-\varrho}\|X^n - X^{0,n}\|_T^* \le C(T)n^{\frac{1}{2}-\varrho}\log n + n^{-\varrho}\|n^{-1}X^{0,n} - x^*\|,$$

where the last term on the above display converges to zero by (3.14). Since $\mathbb{P}(\sigma^n \ge T) \to 1$, (3.16) follows. This completes the proof of Theorem 3.1. $\square$

Chapter 4

Diffusion models: a reduction to one dimension

4.1. Introduction

In this chapter we deal with diffusion models under the assumption of pool–dependent service rates. We find several significant reductions of the correspondent diffusion model, which yield a one–dimensional controlled diffusion. Given some cost, we formulate a one–dimensional stochastic control problem with a compact control space. We then identify particular cases, for which an exact solution is available, and describe how to construct control policies for the prelimit model, that are conjectured to be asymptotically optimal. The chapter is based on the paper [9].

There have been other instances when the reduction to a one–dimensional diffusion model has been discovered in queueing systems with pool–dependent service rates. Armony [1] treated the model with one class of customers and several server stations and showed the optimality of a property, similar to our Statement *(ii)* of Theorem 4.1. The recent paper of Dai and Tezcan [22] treats a model with 2 classes and 2 stations. They show a similar reduction and also, due to a certain condition on abandonment rates, get an exact solution of the control problem. The working paper of Gurvich and Whitt also treats pool–dependent queueing models without the abandonments, adapting the approach from [49].

Our model is more general than the ones treated by the above works, and as a result the 1-dimensional diffusion control problem is sometimes more complicated. In particular, it does not in general admit a pathwise solution. However, a standard control theoretic approach yields a solution via the HJB equation which, due to the one–dimensionality, can be easily solved, at least numerically.

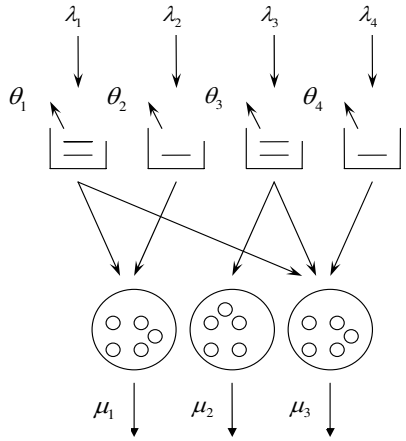


FIG 4.1. *A queueing model with abandonments and pool–dependent service rates*

4.2. Reduction of a pool–dependent diffusion model

Consider the queueing model described in Section 2.2.1 where, in addition, abandonments are allowed, i.e., customers may abandon the system while waiting to be served, and abandonments arise according to exponential clocks. For each $i \in \mathcal{I}$, let $\theta_i^n \equiv \theta_i$ be the abandonment rate of class–i. The queueing system is assumed to be in *heavy traffic* and satisfy the *complete resource pooling* condition (see also Section 2.2.2 and equations (2.8)–(2.12)). In addition, the service rates are assumed to be (asymptotically) *pool–dependent*, i.e., the quantities $\{\mu_{ij}\}$ from (2.8) satisfy

$$\mu_{ij} = \mu_j \ , \qquad i \in \mathcal{I}, \ \ j \in \mathcal{J}.$$

The corresponding *pool–dependent* diffusion model (see e.g., [3]) is given in terms of the following equations:

$$X_i(t) = x_i + W_i(t) - \sum_{j \in \mathcal{J}} \mu_j \int_0^t \Psi_{ij}(s)ds - \theta_i \int_0^t Y_i(s)ds, \quad i \in \mathcal{I}. \tag{4.1}$$

$$\sum_{j \in \mathcal{J}} \Psi_{ij}(t) = X_i(t) - Y_i(t), \qquad i \in \mathcal{I}, \tag{4.2}$$

$$\sum_{i \in \mathcal{I}} \Psi_{ij}(t) = -Z_j(t), \qquad j \in \mathcal{J}. \tag{4.3}$$

$$\Psi_{ij}(t) \equiv 0 \qquad \text{if} \qquad i \not\sim j. \tag{4.4}$$

$$Y_i(t) \geq 0, \quad Z_j(t) \geq 0, \qquad i \in \mathcal{I},\ j \in \mathcal{J}, \quad t \geq 0. \tag{4.5}$$

The last term in (4.1) is due to the abandonments (see [3]–[7] for a rigorous introduction of abandonments into the diffusion model). Set $\mu = (\mu_1, .., \mu_J)$. In what follows we will impose the assumption that the minimal service rate is greater than the maximal abandonment rate, i.e.,

$$\mu_{min} \geq \theta_{max} \qquad \text{for} \qquad \mu_{min} := \min_{j \in \mathcal{J}} \mu_j, \qquad \theta_{max} := \max_{i \in \mathcal{I}} \theta_i. \tag{4.6}$$

By allowing the abandonment rates to exceed the service rates in the prelimit queueing system, it may be optimal not to route customers to service for some periods of time. This will result in having queue length of order $O(n)$ and not just $O(\sqrt{n})$. Therefore, the static fluid model, about which the centering is performed, is no longer relevant. From that point of view it is natural to assume (4.6).

Let a complete filtered probability space $\{\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P}\}$, an $(\mathcal{F}_t)$–Brownian motion W and a deterministic $x \in \mathbb{R}^I$ be given. A set $\mathcal{M}$ of processes (X, Y, Z, Ψ) is said to be a *diffusion model* if the following conditions hold:

- X, Y, Z and Ψ are $(\mathcal{F}_t)$– progressively measurable,
- equations (4.1)–(4.5) are satisfied $\mathbb{P}$–a.s..

A subset $\widetilde{\mathcal{M}}$ of $\mathcal{M}$ is called a *reduction* of $\mathcal{M}$, if for any $(X, Y, Z, \Psi) \in \mathcal{M}$ there exists $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \widetilde{\mathcal{M}}$, such that for any constant $c \in \mathbb{R}^I_+$, $\mathbb{P}$–a.s. $c \cdot \widetilde{Y}(t) \leq c \cdot Y(t)$ for all $t \geq 0$. We emphasize that the above statement regards a pathwise property of Y and $\widetilde{Y}$, meaning $\mathbb{P}\{c \cdot \widetilde{Y}(t) \leq c \cdot Y(t), \forall t \geq 0\} = 1$.

Fix some j_0, satisfying $\mu_{j_0} = \mu_{min}$.

Theorem 4.1. *(i) Let $\mathcal{M}$ be a diffusion model and let $\mathcal{M}_1$ be the subset of $\mathcal{M}$, that, in addition, satisfies 1.–2. below $\mathbb{P}$–a.s. for all $t \geq 0$:*

1. *$Y_e(t) \wedge Z_e(t) = 0$,*
2. *$\Psi_{ij}(t) = 0$ for $(i, j) \in \mathcal{E}_{nb}$.*

Then $\mathcal{M}_1$ is a reduction of $\mathcal{M}$.

(ii) Let $\mathcal{M}_2$ be the subset of $\mathcal{M}_1$ that, in addition, satisfies $Z(t) = Z_e(t) e_{j_0}$ $\mathbb{P}$–a.s. for all $t \geq 0$. Then $\mathcal{M}_2$ is a reduction of $\mathcal{M}_1$.

The proof appears in Section 4.4. We comment that Statement 1 of Theorem 4.1 (i) corresponds to *Joint Work Conservation* (JWC), while Statement 2 says that the system uses only basic activities. Both were introduced in [3]–[4] as

central assumptions. Theorem 4.1 justifies those assumptions in the case of pool–dependent service rates, at least when we are interested in minimizing queue lengths (see also [7], pp. 1103–1104, for a discussion about how some other types of cost, e.g., delay or abandonment costs, can be treated within the framework of queue–length costs). Therefore, as a consequence of Theorem 4.1, we may restrict the diffusion model $\mathcal{M}$ to $\mathcal{M}_2$.

Consider the family $\mathcal{O}$ of $(\mathcal{F}_t)$–progressively measurable processes $(\breve{X}, u) \in \mathbb{R} \times \mathbb{U}$, where

$$\breve{X}(t) = x_e + W_e(t) + \mu_{min} \int_0^t \breve{X}^-(s)ds - \int_0^t [\theta \cdot u(s)]\breve{X}^+(s)ds, \quad t \geq 0; \quad (4.7)$$

$$\mathbb{U} = \{u \in \ R^I \ : \ u_i \geq 0 \ , \ u_e = 1\}. \quad (4.8)$$

Recall (2.22)–(2.24) and set

$$H_i(x, u) = -\sum_j \mu_j G_{ij}(x - x_e^+ u, -x_e^- e_{j_0}) - \theta_i x_e^+ u, \quad i \in \mathcal{I}.$$

The main result of this chapter is the following

Theorem 4.2. *Let $\mathcal{M}^{\mathcal{O}}$ be a set of processes (X, Y, Z, Ψ) such that $\mathbb{P}$–a.s. for all $t \geq 0$ one has*

$$X(t) = x + W(t) + \int_0^t H(X(s), u(s))ds. \quad (4.9)$$

$$Y(t) = X_e^+(t)u(t) \quad (4.10)$$

$$Z(t) = X_e^-(t)e_{j_0} \quad (4.11)$$

$$\Psi_{ij}(t) = G_{ij}(X(t) - X_e^+(t)u(t), -X_e^-(t)e_{j_0}), \qquad i \in \mathcal{I}, \ j \in \mathcal{J}, \quad (4.12)$$

where u is such that $(\breve{X}, u) \in \mathcal{O}$. Then

1. *$\mathcal{M}^{\mathcal{O}} = \mathcal{M}_2$.*
2. *For any $(\breve{X}, u) \in \mathcal{O}$ and a corresponding $(X, Y, Z, \Psi) \in \mathcal{M}^{\mathcal{O}}$ we have $\mathbb{P}$–a.s. $X_e = \breve{X}$.*

See Section 4.4 for a proof. This establishes a reduction of the diffusion model to a one–dimensional model. The steps of reduction are summarized in Fig. 4.2.

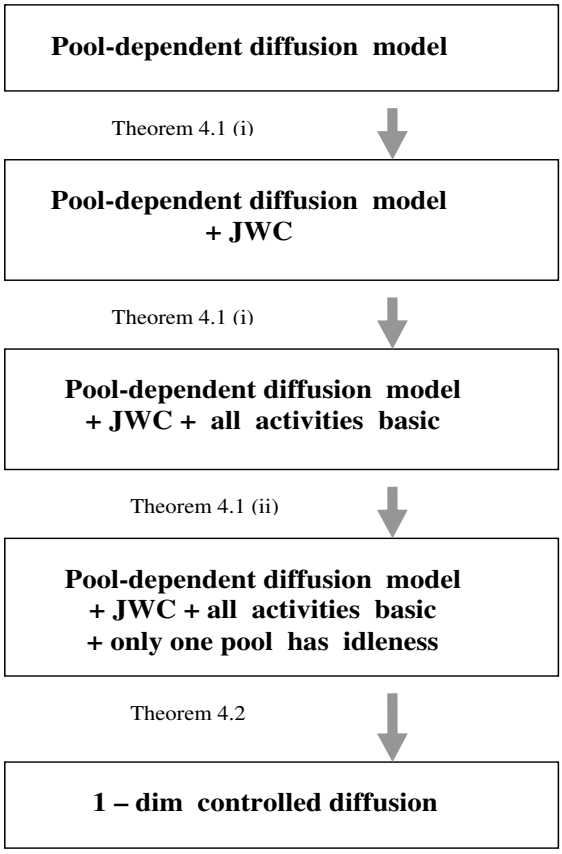


FIG 4.2. *A scheme of dimensionality reduction for the diffusion model*

4.3. Diffusion control problem

As a demonstration of dimension reduction consider the following stochastic control problem. Let $\breve{X}$ be a controlled diffusion with a control u, satisfying (4.7)–(4.8). Assume that we are given deterministic constants $\gamma > 0$, $c \in \mathbb{R}^I_+$. Let $\pi := (\Omega, \mathcal{F}, (\mathcal{F}_t), u, W)$ be an admissible control system system (see [25], Chapter III.8) and consider the cost functional

$$C(x, \pi) = E^\pi_x \int_0^\infty e^{-\gamma t}[c \cdot u(t)]\breve{X}^+(t)dt, \qquad x \in \mathbb{R}. \tag{4.13}$$

Define a diffusion control problem by optimizing $C(x, \pi)$ over the set Π of all admissible control systems:

$$V(x) = \inf_{\Pi} C(x, \pi), \qquad x \in \mathbb{R}. \tag{4.14}$$

As a consequence of Theorem 4.2, the one–dimensional problem (4.7), (4.13), (4.14) is equivalent to the problem of minimizing the queue lengths cost

$$E_y \int_0^\infty e^{-\gamma t}[c \cdot Y(t)]dt$$

for the original diffusion model, whenever $x^+ = e \cdot y$. Regarding the stochastic control problems, such a reduction of dimensionality certainly simplifies the general solution methods and even enables one to get exact solution in particular cases. We will address those in the subsection below.

Exact solutions and connection to the prelimit queueing system

Solving (4.7), (4.13), (4.14) requires standard methods of stochastic control (see also our discussion in directions for future work, Section 5.2). However, there are particular cases in which the optimal control policy can be obtained explicitly.

Lemma 4.1. *Assume there exists a pair of indices $i, j \in \mathcal{I}$, such that $c_i \geq c_j$ and $\theta_i < \theta_j$. Then, necessarily the optimal control will satisfy $u_i \equiv 0$.*

The proof appears at the end of Section 4.4. As a consequence, it enables one to get an optimal control policy in the special cases below, which provide a generalization of the models without abandonments treated in [22].

Corollary 4.1. *(i) Assume that $\theta_i = \theta$ for all i. Then following is an optimal control: $u_{opt}(t) \equiv e_{i_0}$, for some fixed i_0, satisfying $c_{i_0} = \min_i c_i$.*

(ii) Assume that

$$\begin{cases} \theta_1 \leq \theta_2 \leq ... \leq \theta_I \\ c_1 \geq c_2 \geq ... \geq c_I \end{cases}$$

Then the control $u_{opt}(t) \equiv e_I$ is optimal.

Remark 4.1. We introduce a problem of minimizing convex queueing costs, for which an exact solution is also obtainable. Assume that $\theta_i = \theta$ for all i. Let the optimization goal be given as

$$V(x) = \inf_{\Pi} E_x^\pi \int_0^\infty e^{-\gamma t} \sum_i C_i\Big(u_i(t) X_e^+(t)\Big)dt, \quad x \in \mathbb{R}^I. \tag{4.15}$$

Note that $u_i X_e^+$ represents queue length of class i. Here $C_i(\cdot)$ are strictly increasing continuously differentiable convex functions. Then optimal controls

are characterized by

$$C_1'(u_1^o(t)X_e^+(t)) = C_2'(u_2^o(t)X_e^+(t)) = ... = C_I'(u_I^o(t)X_e^+(t)), \quad t \geq 0. \quad (4.16)$$

Indeed, when all $\theta' s$ are equal, the scalar product $\theta \cdot u$ in (4.7) equals to constant θ, hence the control u does not influence the dynamics of the system in (4.7). Therefore, one just needs to solve the following minimization problem:

$$\min \sum_i C_i(y_i), \quad s.t. \quad y_1 + .. + y_I = x,$$

The solution y^* satisfies $C_1'(y_1^*) = ... = C_I'(y_I^*)$ and $e \cdot y^* = x$. □

Although the focus of this chapter is on diffusion models, we would like to comment on their relation with routing and scheduling policies for the prelimit queueing models. Given the optimal policy of the diffusion control problem, [4] provides a scheme of constructing an asymptotically optimal policy for the prelimit model. Due to the simplicity of the diffusion optimal policies in Corollary 4.1 and Remark 4.1, the construction of analogous policies is straightforward. Indeed, we conjecture that the following are asymptotically optimal for the prelimit model. We consider here the harder case of non-preemptive policies.

For Corollary 4.1 (case (ii) follows from (i), by taking $i_0 = I$):

- ROUTING: each arriving customer, if not queued, is routed to the fastest server available, otherwise stays in queue. The system "tends" to have idle servers only in station j_0 (recall $Z_{j_0} = Z_e$)
- SCHEDULING: a newly available agent, that can serve class i_0, will accept a waiting i_0–class customer, only if no other classes are waiting for him. All other situations are resolved arbitrarily. In other words, class i_0 always has the lowest priority, and thus the system will seek to have waiting customers only in class i_0 (recall $Y_{i_0} = Y_e$). Note that this setting generalizes that of Dai and Tezcan [22] to multi–dimensional models.

Consider the scheduling policy that satisfies the above. A newly available agent, among all the waiting customers, will accept a customer with the largest holding cost c. Since the service rates do not depend on the class, this scheduling policy is related to the $c\mu$ rule (see e.g., [54]).

Regarding Remark 4.1:

- ROUTING: the same as in Corollary 4.1.
- SCHEDULING: at any time t, among all the waiting customers that are available for him, a newly available agent at station j will serve a customer from class $i^* = \arg\max_{i \sim j}\{C_i'(Y_i^n(t))\}$. The system will seek to achieve approximate equality in (4.16).

Similarly to our $c\mu$ comment from above, this scheduling policy is related to the *generalized* $c\mu$ rule (see [53] and [49]). Note, however that we do not claim that a similarity to the $c\mu$ rule holds in a more general setting, where a customer is to be chosen from a number of classes having class–dependent service rates, as the results of [4] show. In fact, the relation to the $c\mu$ rule is very limited to the current setting.

4.4. Proofs

PROOF OF THEOREM 4.1 (i): We show that each $(X, Y, Z, \Psi) \in \mathcal{M}$ defines $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \mathcal{M}_1$ such that for any $c \in \mathbb{R}^I_+$ and all $t \geq 0$ $\mathbb{P}$–a.s. holds $c \cdot \widetilde{Y}(t) \leq c \cdot Y(t)$.

Consider an arbitrary $(X, Y, Z, \Psi) \in \mathcal{M}$. By summing equations (4.1) over all i's, we obtain

$$X_e(t) = x_e + W_e(t) + \sum_{j \in \mathcal{J}} \mu_j \int_0^t Z_j(s)ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t Y_i(s)ds. \tag{4.17}$$

Define a process M as

$$M(t) = Y_e(t) \ \wedge \ Z_e(t), \qquad t \geq 0. \tag{4.18}$$

Observe that M is $\mathcal{F}_t$–progressively measurable, since both Y and Z are. Also $M(t) \geq 0$, which follows from (4.5). From (4.2), (4.3) we have $X_e = Y_e - Z_e$. Now rewrite (4.18) as $(Y_e - M) \wedge \ (Z_e - M) = 0$ to get $Y_e = X_e^+ + M$ and $Z_e = X_e^- + M$. Therefore, we can define an $\mathcal{F}_t$–measurable process $U = (u, v)$ taking values in

$$\mathcal{A} = \{(u, v) \in \mathbb{R}^I \times \mathbb{R}^J \ : \ u_i, v_j \geq 0 \ , \ u_e = v_e = 1\}, \tag{4.19}$$

such that for all $t \geq 0$, $Y_i(t) = u_i(t) \ (X_e^+(t) + M(t))$, $Z_j(t) = v_j(t) \ (X_e^-(t) + M(t))$. Recall $\mu = (\mu_1, ..., \mu_J)$ and rewrite (4.17) as

$$X_e(t) = x_e + W_e(t) + \int_0^t [\mu \cdot v(s)] X_e^-(s)ds - \int_0^t [\theta \cdot u(s)] X_e^+(s)ds \tag{4.20}$$

$$+ \int_0^t \Big(\mu \cdot v(s) - \theta \cdot u(s) \Big) M(s)ds \tag{4.21}$$

The new processes $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi})$ will be constructed as follows. For any $t \geq 0$ set

$$\widetilde{Y}_i(t) = u_i(t) \widetilde{X}_e^+(t), \qquad i \in \mathcal{I} \tag{4.22}$$

$$\widetilde{Z}_j(t) = v_j(t) \widetilde{X}_e^-(t), \qquad j \in \mathcal{J} \tag{4.23}$$

$$\widetilde{\Psi}_{ij}(t) = G_{ij}(\widetilde{X}(t) - \widetilde{Y}(t), -\widetilde{Z}(t)). \tag{4.24}$$

and $\widetilde{X}$ is constructed by substituting (4.22)–(4.24) into (4.1). By the construction, $(\widetilde{X},\widetilde{Y},\widetilde{Z},\widetilde{\Psi})$ are $\mathcal{F}_t$–progressively measurable and satisfy (4.1)–(4.5) for all $t \geq 0$ $\mathbb{P}$–a.s.. By (4.22)–(4.23) $\widetilde{Y}_e \wedge \widetilde{Z}_e = 0$ and by the properties of G (see (2.22)–(2.24)), we also have $\Psi_{ij} \equiv 0$ for $(i,j) \in \mathcal{E}_{nb}$. Hence $(\widetilde{X},\widetilde{Y},\widetilde{Z},\widetilde{\Psi}) \in \mathcal{M}_1$. Consequently,

$$\widetilde{X}_e(t) = x_e + W_e(t) + \int_0^t [\mu \cdot v(s)]\widetilde{X}_e^-(s)ds - \int_0^t [\theta \cdot u(s)]\widetilde{X}_e^+(s)ds. \qquad (4.25)$$

By (4.6) we have $\mu \cdot v(s) \geq \theta \cdot u(s)$. Therefore, from (4.20) and (4.25),

$$\frac{d}{dt}\Big(\widetilde{X}_e(t) - X_e(t)\Big) \leq [\mu \cdot v(t)]\Big(\widetilde{X}_e^-(t) - X_e^-(t)\Big) + [\theta \cdot u(t)]\Big(X_e^+(t) - \widetilde{X}_e^+(t)\Big)$$

$$\leq [\mu \cdot v(t)]\Big(\widetilde{X}_e(t) - X_e(t)\Big)^- + [\theta \cdot u(t)]\Big(X_e(t) - \widetilde{X}_e(t)\Big)^+,$$

where we used the simple inequalities $a^- - b^- \leq (a-b)^-$ and $a^+ - b^+ \leq (a+b)^+$. By the comparison principle for one dimensional ordinary differential equation (Theorem 7, p. 29, [14]), we get $\widetilde{X}_e(t) \leq X_e(t)$ and, as a result, for any $c \in \mathbb{R}_+^I$ and all $t \geq 0$ $\mathbb{P}$–a.s. we have $\widetilde{Y}_i(t) = u_i(t)\widetilde{X}_e^+(t) \leq u_i(t)(X_e^+(t) + M(t)) = Y_i(t)$, $i \in \mathcal{I}$. This proves part (i) of the theorem.

Proof of Theorem 4.1 (ii): Again, we show that each $(X,Y,Z,\Psi) \in \mathcal{M}_1$ defines $(\widetilde{X},\widetilde{Y},\widetilde{Z},\widetilde{\Psi}) \in \mathcal{M}_2$ such that for any $c \in \mathbb{R}_+^I$ and all $t \geq 0$ $\mathbb{P}$–a.s. holds $c \cdot \widetilde{Y}(t) \leq c \cdot Y(t)$.

Consider an arbitrary $(X,Y,Z,\Psi) \in \mathcal{M}_1$. By summing equations (4.1) over all i's, and using $U(t)$ from the proof of part (i),

$$X_e(t) = x_e + W_e(t) + \int_0^t [\mu \cdot v(s)]X_e^-(s)ds - \int_0^t [\theta \cdot u(s)]X_e^+(s)ds. \qquad (4.26)$$

Fix some arbitrary j_0 satisfying $\mu_{j_0} = \mu_{min}$ and consider a new $\mathcal{F}_t$–measurable process $\widetilde{U} := (\widetilde{u},\widetilde{v})$, defined for all $t \geq 0$ as $\widetilde{u}(t) = u(t)$ and $\widetilde{v}(t) \equiv e_{j_0}$. Using a similar argument as in the proof of part (i), by substituting $\widetilde{U}$ into (4.22)–(4.24), we get a new $(\widetilde{X},\widetilde{Y},\widetilde{Z},\widetilde{\Psi}) \in \mathcal{M}_1$. By the choice of $\widetilde{v}$ and using (4.23), we also have that $(\widetilde{X},\widetilde{Y},\widetilde{Z},\widetilde{\Psi}) \in \mathcal{M}_2$. Since $\mu \cdot v(t) \geq \mu \cdot \widetilde{v}(t) = \mu_{min}$,

$$\frac{d}{dt}\Big(\widetilde{X}_e(t) - X_e(t)\Big) \leq [\mu \cdot v(s)]\Big(\widetilde{X}_e^-(t) - X_e^-(t)\Big) + [\theta \cdot u(t)]\Big(\widetilde{X}_e^+(t) - x_e^+(t)\Big)$$

$$\leq [\mu \cdot v(t)]\Big(\widetilde{X}_e(t) - X_e(t)\Big)^- + [\theta \cdot u(t)]\Big(\widetilde{X}_e(t) - X_e(t)\Big)^+,$$

and, again, by using the standard comparison principle, we get $\widetilde{X}_e(t) \le X_e(t)$. As a result, similarly to the proof of part (i), for any $c \in \mathbb{R}^I_+$ and all $t \ge 0$ $\mathbb{P}$–a.s. holds $\widetilde{Y}_i(t) = u_i(t)\widetilde{X}^+_e(t) \le u_i(t)X^+_e(t) = Y_i(t)$, $i \in \mathcal{I}$. This proves the theorem. □

Proof of Theorem 4.2(1). We first prove the inclusion $\mathcal{M}_2 \subseteq \mathcal{M}^{\mathcal{O}}$. Consider an arbitrary $(X, Y, Z, \Psi) \in \mathcal{M}_2$ The proof of Theorem 4.1 suggests the existence of an $\mathcal{F}_t$–measurable process $u = u(t)$ with values in $\mathbb{U}$, such that (X, Y, Z, Ψ) satisfy (4.10)–(4.12) and the inclusion $(X, Y, Z, \Psi) \in \mathcal{M}^{\mathcal{O}}$ follows by substituting (4.10)–(4.12) into (4.1). The opposite inclusion is obvious. □

Proof of Theorem 4.2(2). We need to show that, after substituting (4.10)–(4.12) into (4.1), the resulting X for all $t \ge 0$ $\mathbb{P}$–a.s. satisfies $X_e(t) = \breve{X}(t)$. Indeed, by the properties of G, we have $\sum_{i\in\mathcal{I}} \Psi_{ij}(t) = Z_j(t)$ and, therefore, by (4.9)–(4.12),

$$X_e(t) = x_e + W_e(t) + \mu_{min}\int_0^t X^-_e(s)ds - \int_0^t [\theta \cdot u(s)]X^+_e(s)ds.$$

Since the process u above is the same as in (4.7), the statement of the theorem follows. □

Proof of lemma 4.1. Assume there exists a control $u(t)$, such that $u_i(t)$ is not identically equal to zero. The corresponding controlled process will be X_e. Now consider another control $\widetilde{u}$, such that $\widetilde{u}_i(t) \equiv 0$ and $\widetilde{u}_j(t) = u_i(t) + u_j(t)$. Denote the corresponding controlled process $\widetilde{X}_e$. Using the relation $\theta \cdot \widetilde{u}(t) \ge \theta \cdot u(t)$ and repeating the comparison argument from the proof of Theorem 4.1 (ii), we get that $\widetilde{X}_e(t) \le X_e(t)$. Moreover, since $c \cdot \widetilde{u}(t) \le c \cdot u(t)$, trivially for all $t \ge 0$ $\mathbb{P}$–a.s. we have $[c \cdot \widetilde{u}(t)]\widetilde{X}^+_e(t) \le [c \cdot u(t)]X^+_e(t)$. □

Chapter 5

Future work

5.1. Extending the existing control–theoretic framework

Adding a singular component. In addition to the *heavy traffic* and *complete resource pooling* conditions (see Section 2.2.2), the theory developed in Atar, Mandelbaum and Reiman [7] and Atar [3, 4] relied on two key assumptions: (a) the absence of non–basic activities and (b) joint work conservation. As a result, the diffusion model for these formulations was the following controlled Markov diffusion:

$$X(t) = X(0) + \sigma W(t) + \int_0^t b(X(s), u(s))ds, \tag{5.1}$$

where $u \in \mathbb{U}$ is a control term and the control space $\mathbb{U}$ is compact. One of the key points discovered in Chapter 2, is that when one drops either one of the two assumptions, a term is added to the diffusion model:

$$X(t) = X(0) + \sigma W(t) + \int_0^t b(X(s), u(s))ds + \eta_t. \tag{5.2}$$

The term η is of bounded variation, but not necessarily absolutely continuous with respect to the Lebesgue measure and, in the literature, is sometimes referred to as "singular". Theorem 2.2 of Chapter 2 states that, under some conditions, the singular term can result in the null–controllability of the controlled diffusion. However, in general, it is natural to consider a stochastic control problem, with both drift and singular controls, for minimizing expected discounted costs, associated with queue lengths, delays or abandonments. Extending the theory from [3, 4, 7], regarding the HJB equations and its use, to such problems seems an interesting and challenging research direction.

Queueing systems with relatively "small" service stations. Besides the systematic study of a diffusion control problem, [4, 7] construct scheduling policies for the prelimit control problem. The policies are based on diffusion optimal controls and are shown to be asymptotically optimal for both the preemptive and non–preemptive regimes. The asymptotical equivalence was possible due to the following property of the heavy traffic regime: the population in each station and the arrival rates are of order $O(n)$, with possible fluctuations of $O(\sqrt{n})$. The control is applied only to diffusive fluctuations of order $O(\sqrt{n})$. As n grows to infinity, service completion or arrival of $O(\sqrt{n})$ customers takes time of $O(1/\sqrt{n})$, which is short.

The cases when some service station has $O(\sqrt{n})$ servers, or an arrival rate of some class is of order $O(\sqrt{n})$ are not covered by the existing theory. As a simple demonstration, consider a queueing system with a single class and two different pools, where the arrival rate and stations' populations are assumed to satisfy

$$\lambda^n = \lambda n + \hat{\lambda}\sqrt{n}, \quad N_1^n = n, \quad N_2^n = \sqrt{n}, \tag{5.3}$$

and the service rates μ_1 and μ_2 are constant (see Fig 5.1).

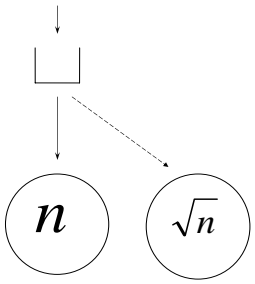


FIG 5.1. *A queueing model with a relatively small service station*

Let $\Psi_i^n(t)$, $i = 1, 2$, represent the number of customers being served in each station at time $t \geq 0$. For simplicity, we assume $\Psi_1^n = (X^n - \Psi_2^n) \wedge N_1^n$, namely, each customer, if not routed to the "small" (with $\sqrt{n}$ servers) station, joins the queue only if the "large" station is fully occupied. As n increases, by applying formal weak limits to appropriate centered and rescaled processes, one can derive the following diffusion model, which is different from (4.1)–(4.5):

$$X(t) = X(0) + W(t) + \mu_1 \int_0^t (X(s) - \Psi_2(s))^- ds - \mu_2 \int_0^t \Psi_2(s) ds \tag{5.4}$$

$$0 \leq \Psi_2(t) \leq 1, \qquad t \geq 0. \tag{5.5}$$

The last condition is due to the fact that no centering is applied to Ψ_2^n (recall the last condition in (5.3)). The relations (5.4)–(5.5) suggest to consider the process Ψ_2 as a control process, and X as a controlled diffusion. Given a cost, we can then formulate a one–dimensional control problem, the study of which is not expected to be complex, especially given the theory from [3, 7]. However, we believe that this diffusion control problem is not the *right problem* to generate asymptotically optimal policies for the prelimit system in the non–preemptive scheduling regime! A rough explanation is as follows. Assume that the diffusion policy suggests to immediately decrease Ψ_2. Translated to the prelimit model, this suggests to decrease Ψ_2^n by an amount of $O(\sqrt{n})$ in a short time, which is impossible due to $N_2^n = \sqrt{n}$.

On the other hand, due to a much higher arrival rate of the order $O(n)$, it is indeed possible to increase Ψ_2^n by an amount of $O(\sqrt{n})$ almost immediately. Therefore, the right approach seems to introduce a non-decreasing process B, with $B(0) = 0$, as a control process, and to consider (X, Ψ_2) as a two–dimensional controlled diffusion which, in addition to (5.4)–(5.5), satisfies

$$\Psi_2(t) = \Psi_2(0) - \mu_2 \int_0^t \Psi_2(s)ds + B(t). \tag{5.6}$$

Here $B(t)$ represents the "input flow" to station 2. Since B is not necessarily absolutely continuous, one should treat it as a singular control. Finally, given some cost, we have a singular control problem with B as control.

Unfortunately, unlike the one–dimensional singular control problems, the structure of the optimal solution for such problems in high dimensions is generally unknown. Shreve and Soner [52], for example, studied a very particular 2–dimensional singular control problem and showed that it is optimal to keep a diffusion inside a certain region, that is given in terms of an HJB equation (note that singular controls make it possible to do so). The extension of their result to higher dimension or to weaker conditions on the diffusion model (in dimension 2 and higher) are hard open problems. However, see Atar and Budhiraja [5] and Atar, Budhiraja and Williams [6] for recent developments.

It is possible, however, that there is some structural property in the current problem that makes it treatable. Thus a better understanding of this setting will be the subject of future study.

Non–exponential service times. Exponential distribution of service times is a key assumption in most of the works (including ours) that deal with many-server queues in heavy traffic. The reason for this is the memoryless property,

which results in some remarkable simplifications in representation of processes. As of today, there are only few papers dealing with convergence of scaled queueing processes in the QED regime, where the service duration is not exponential. [40] and [47] study the limiting behaviour of the virtual waiting time in queues where the service duration is deterministic or has a finite support; [50] studies single pool model with service times distributed as phase–type, which can be thought of as a relative of the exponential one (see Section 1.1.2 for the literature background). For general service time distributions, the queueing models are believed to be infinite dimensional. The recent paper of Kaspi and Ramanan [42] studies measure-valued processes arising as fluid limits for general service time distribution.

To the best of our knowledge, there is no research dealing with stochastic control of queueing systems in heavy traffic in the non-exponential setting. Extending the existing "exponential" control framework to the case of generally distributed service time is an interesting challenge. To start with, one may consider phase–type distributions. Regarding the current work, we believe that the null–controllability phenomenon can be extended to the non–exponential setting.

5.2. Further study of pool-dependent diffusion models

In Chapter 4 we studied pool–dependent diffusion models and showed that they can be reduced to one–dimensional controlled diffusions. We also identified some cases when the control problem (4.7), (4.13), (4.14) has explicit solution. However, the most interesting cases are yet to be solved analytically. Consider, for example the simplest case with two customer classes, where the service and abandonment rates satisfy

$$c_1 > c_2 \quad \text{and} \quad \theta_1 > \theta_2.$$

Translated into the language of queues, the administrator must make a trade–off: either to maintain the waiting population in class 1 and get a higher total abandonment rate, but also higher cost rate; or to move waiting customers into queue 2 and get lower cost rate, but also low abandonment rate. The results of Chapter 4 show that this 2–dimensional model can be reduced to a 1–dimensional one, with value and dynamics as follows (assume $\mu_{min} = 1$):

$$V(x) = \inf_{\pi} E_x^{\pi} \int_0^{\infty} e^{-\gamma t} \Big(c_1 u + c_2 (1-u) \Big) X^+ dt, \tag{5.7}$$

$$dX = dW + X^- dt - \Big(\theta_1 u + \theta_2(1-u)\Big)X^+ dt. \tag{5.8}$$

The corresponding HJB equation is

$$\frac{1}{2}V_{xx} + \inf_{u\in[0,1]}\Big[(c_1-c_2)ux^+ - (\theta_1-\theta_2)ux^+V_x\Big] + (x^- - \theta_2 x^+)V_x + c_2x^+ - \gamma V = 0. \tag{5.9}$$

A significant advantage of the reduction to a one–dimensional model is that the corresponding HJB equation is one–dimensional as well, and thus amenable to numerical schemes. Moreover, initial calculations suggest that the optimal policy is "bang–bang": the control $u(t)$ takes only two different values: 0 or 1, depending on the relation between $V_x(X(t))$ and $\frac{c_1-c_2}{\theta_1-\theta_2}$. To be precise, $u(t) = 1$ if $V_x(X(t)) \geq \frac{c_1-c_2}{\theta_1-\theta_2}$ and $u(t) = 0$ otherwise. Trying to solve this problem and then to generalize it to models with more customer classes will be the subject of future work.

To begin with, one may be interested in the properties of V from (5.7). Computer simulations suggest that V is concave, which implies the existence of a point $x^* \geq 0$, such that $u(X(t)) = 0$ once $X(t) \geq x^*$ and $u(X(t)) = 1$ otherwise.

In the future work we will also address diffusion models for queueing systems with class–dependent service rates:

$$\mu_{ij} = \mu_i\,, \qquad i \in \mathcal{I},\ \ j \in \mathcal{J}.$$

Several reductions are also available in this setting, although one does not expect a reduction to one–dimensional controlled diffusion.

5.3. Staffing in throughput sub–optimal systems

This is a question that is important for further understanding of the throughput sub–optimality property. As a motivating example consider a static fluid model with the following arrival and service rates:

$$\lambda = \begin{pmatrix} 8 \\ 4 \end{pmatrix}, \quad \mu = \begin{pmatrix} 3 & 10 & 1 \\ 1 & 4 & 2 \end{pmatrix}. \tag{5.10}$$

Assume that the staffing vector is given by $\nu = (0.3, 0.3, 6.1)'$. The resulting optimal static allocation matrix is as follows

$$\xi^* = \begin{pmatrix} 1 & 1 & 0.6721 \\ 0 & 0 & 0.3279 \end{pmatrix}.$$

The resulting graph of activities is depicted in Fig 5.2(left). The fluid model is throughput optimal, as can be easily checked.

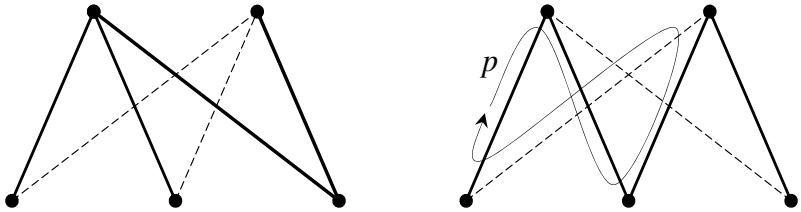


FIG 5.2. *The resulting graph of activities. The left graph corresponds to throughput optimal fluid model, the right graph - to sub–optimal, with the corresponding closed simple path*

Now assume that the staffing vector equals to $\widetilde{\nu} = (1, 1, 1)'$ and the arrival and service rates remain the same (5.10). The resulting optimal static allocation is as follows

$$\widetilde{\xi}^* = \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix}.$$

The fluid model now appears to be throughput sub–optimal, though the total staffing was decreased from $\nu_e = 6.7$ to $\widetilde{\nu}_e = 3$. In Figure 5.2(right) we see the graph of activities together with a corresponding closed simple path. An interesting question is to identify the set of $\nu's$ that result in throughput sub–optimality.

Bibliography

[1] Armony M. (2005). Dynamic routing in large–scale service systems with heterogeneous servers, *Queueing Systems*, 51(3-4), 287-329.

[2] Armony M. and Maglaras C. (2004). On customer contact centers with a callback option: customer decisions, routing rules and system design. *Operatins Research*, 52(2), 271-292.

[3] Atar R. (2005). A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* 15, No. 1B, 820–852.

[4] Atar R. (2005). Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab.*, 15, No. 4, 2606-2650.

[5] Atar R. and Budhiraja A. (2006). Singular control with state constraints on unbounded domain. *Ann. Probab.*, 34, No. 5, 1864–1909.

[6] Atar R., Budhiraja A. and Williams R.J. (2007). HJB equations for certain singularly controlled diffusions. *Preprint.*

[7] Atar R., Mandelbaum A. and Reiman M. (2004). Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.* 14, No. 3, 1084–1134.

[8] Atar R., Mandelbaum A. and Shaikhet G. (2006). Queueing systems with many servers: null controllability in heavy traffic. *Ann. Appl. Probab.*, 16, No. 4, 1764-1804.

[9] Atar R., Mandelbaum A. and Shaikhet G. (2007). Diffusion models: a reduction to one dimension. *In preparation.*

[10] Atar R. and Shaikhet G. (2007). Critically loaded queueing models that are throughput sub-optimal. Submitted to *Ann. Appl. Probab.*

[11] Bell S. L. and Williams R. J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* 11, No. 3, 608–649.

[12] Bell S. L. and Williams R. J. (2005). Dynamic Scheduling of a Parallel Server System in Heavy Traffic with Complete Resource Pooling:

Asymptotic Optimality of a Threshold Policy. *Electronic J. of Probability*, 10, 1044-1115.

[13] Billingsley P. (1999). *Convergence of Probability Measures,* 2nd ed. Wiley, New York.

[14] Birkhoff G. and Rota G.-C. (1989). *Ordinary Differential Equations,* 4th ed. Wiley, New York.

[15] Bramson M. (1998). State–space collapse with applications to heavy traffic limits for queueing networks. *Queueing Systems*, 30, pp. 89-148.

[16] Bramson M. and Williams R. J. (2003). Two workload properties for Brownian networks. *Queueing Systems* 45, 191–221.

[17] Chen H. and Yao D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization,* Springer, New-York.

[18] Chen H. and Mandelbaum A. (1991). Open heterogeneous fluid networks, with applications to multiclass queues. Preprint

[19] Chen H. and Mandelbaum A. (1991). Stochastic discrete flow networks: biffusion and bottlenecks, *Ann. of Probab.*, 19, No. 4, pp. 1453-1519.

[20] Cox D.R. and Smith W.L. (1961). *Queues,* Methuen (London) and Wiley (New-York).

[21] Dai J. G. and Tezcan T. (2006). State space collapse in many server diusion limits of parallel server systems. Preprint.

[22] Dai J. G. and Tezcan T. (2006). Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Preprint

[23] Dai J.G. and Wang Y. (1993). Nonexistence of Brownian models of certain multiclass queueing networks, *Queueing Systems*, Vol 13, 41-46.

[24] Diestel R. (2000). *Graph Theory,* Electronic Edition 2000, Springer–Verlag New–York.

[25] Fleming W. H. and Soner H. M. (1993). *Controlled Markov Processes and Viscosity Solutions,* Springer–Verlag, New York.

[26] Gans N., Koole G. and Mandelbaum A. (2003). Telephone call centers: tutorial, review and research prospects. *Commissioned paper, MSOM* 5(2).

[27] Garnett O., Mandelbaum A. and Reiman M. (2002). Designing a call center with impatient customers. *MSOM* 4(3), 208-227.

[28] Gurvich I., (2004). *Design and control of the M/M/N queue with multi-type customers and many servers*, MSc. Thesis, Technion.

[29] Gurvich I. and Whitt W. (2006). Service–level differentiation in many–server service systems: a solution based on fixed–queue–ratio rout-

ing. *Submitted to Oper. Res.*

[30] Halfin S. and Whitt W. (1981). Heavy traffic limits for queues with many exponential servers. *Oper. Res.* 29, No. 3. 567-588.

[31] Harrison J. M. (1988). *Brownian models of queueing networks with heterogeneous customer populations.* Stochastic differential systems, stochastic control theory and applications (Minneapolis, Minn., 1986), 147–186, IMA Vol. Math. Appl., 10, Springer, New York.

[32] Harrison J.M. (2000). Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.* V.10, No. 1, 75-103.

[33] Harrison J. M. and López M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33: 339-368.

[34] Harrison J. M. and Reiman M. (1981). Reflected Brownian motion on an orthant. *Ann. Probab.* 9, 302-308.

[35] Harrison J. M. and Van Mieghem J. A. (1997). Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Ann. Appl. Probab.* 7, No. 3, 747–771.

[36] Harrison J. M. and Williams R.J. (1987). Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* 22, 77-15.

[37] Harrison J. M., Williams R.J. and Chen H. (1990). Brownian models of closed queueing networks. *Stochastics and Stochastics Reports* 29, 37-74.

[38] Harrison J. M. and Zeevi A. (2004). Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* 52, 243–257.

[39] Iglehart D.L. and Whitt W. (1970). Multiple channel queues in heavy traffic. *Adv. Appl. Probab.* 2, 150-177.

[40] Jelenkovic P., Mandelbaum A. and Momcilovic P. (2004). Heavyt traffic limits for queues with many deterministic servers. *Queueing Systems,* 47, pp. 53-69.

[41] Karatzas I. and Shreve S. E. (1991). *Brownian Motion and Stochastic Calculus.* 2nd ed. Springer-Verlag, New York.

[42] Kaspi H. and Ramanan K. (2007). Fluid limits for the GI/GI/N queueu. *Working paper.*

[43] Kingman J.F.C. (1961). The single server queue in heavy traffic. *Proc. of Cambridge Philosophical Society,* 57, 902-904.

[44] Kushner H. J. and Dupuis P. (2001). *Numerical Methods for Stochastic Control Problems in Continuous Time.* 2nd ed. Springer-Verlag,

New York.

[45] Lions P. L. and Sznitman A. S. (1984). Stochastic differential equations with reflecting boundary conditions. *Comm. Pure appl. Math.* 37, 511–537.

[46] Mandelbaum A., Massey W. A. and Reiman M.I. (1998). Strong approximations for markovian service networks. *Queueing Systems*, 30, 149-201.

[47] Mandelbaum, A. and Momcilovic P. (2007). Queues with Many Servers: The Virtual Waiting-Time Process in the QED Regime. *Preprint*

[48] Mandelbaum A. and Pats G. (1998). State–dependent stochastic networks: approximations and applications with continuous diffusion limits. *Ann. Appl. Probab.* 8, No. 2, 569-646.

[49] Mandelbaum A. and Stolyar A. (2004). Scheduling flexible servers with convex delay costs: heavy–traffic optimality of the generalized c-rule. *Oper. Res. 52*, No. 6, 836–855.

[50] Puhalskii A. A. and Reiman M. I. (2000). The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. in Appl. Probab.* 32, No. 2, 564–595.

[51] Reiman M. I. (1984). Some diffusion approximations with state space collapse. *Proc. of the Internat. Seminar on Modeling and Performance Evaluation Methodology,* Lecture Notes in Control and Informational Science, eds. Baccelli, F. and Fayolle G., Springer NY., pp 209-240.

[52] Shreve, S., and Soner M. (1994). Optimal investment and consumption with transaction costs. *Ann. Appl. Probab.* 4, 609–692.

[53] Van Mieghem J.A. (1995). Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Ann. Appl. Probab.* 5, 809–833.

[54] Walrand J. (1988). *An Introduction to Queueing Networks*, Imprint, Englewood Cliffs, N.J. : Prentice-Hall.

[55] Whitt W. (1974). Heavy traffic limit theorems for queues: a survey. *Math. Methods in Queueing Theory,* Proceedings of a Conference at Western Michigan University, Lecture Notes in Economica and Mathematical Systems, No 98, Springer-Verlag, New-York, pp. 307-350.

[56] Whitt W. (2002). *Stochastic–Process Limits: an Introduction to Stochastic–Process Limits and Their Application to Queues,* Springer, 2002.

[57] Whitt W. (2004). A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* Vol. 52, No. 6, pp. 922-941.

[58] Whitt W. (2005). Heavy-traffic limits for the $G/H2/n/m$ queue. *Math. Oper. Res.*, Vol. 30, No. 1, pp. 1-27.

[59] Whitt W. (2007). Martingale proofs of many–server heavy–traffic limits for markovian queues. Working paper.
[60] Williams R. J. (1987). Reflected Brownian motion with skew symmetric data in a polyhedral domain, *Prob. Theory Related Fields,* 75, 459-485.
[61] Williams R. J. (1998). Diffusion appproximations for open multiclass queueing networks: sufficient conditions involving state space collapse, *Queueing Systems*, 30, 27-88.
[62] Williams R. J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. *Analysis of communication networks: call centres, traffic and performance* (Toronto, ON, 1998), 49–71, Fields Inst. Commun., 28, Amer. Math. Soc., Providence, RI.