Queueing Syst (2010) 66: 1–51
DOI 10.1007/s11134-010-9183-0

Control of systems with flexible multi-server pools: a shadow routing approach

Alexander L. Stolyar · Tolga Tezcan

Received: 1 December 2009 / Revised: 25 May 2010 / Published online: 15 July 2010

Abstract A general model with multiple input flows (classes) and several flexible multi-server pools is considered. We propose a robust, generic scheme for routing new arrivals, which optimally balances server pools' loads, without the knowledge of the flow input rates and without solving any optimization problem. The scheme is based on *Shadow routing* in a virtual queueing system. We study the behavior of our scheme in the Halfin–Whitt (or, QED) asymptotic regime, when server pool sizes and the input rates are scaled up simultaneously by a factor r growing to infinity, while keeping the system load within $O(\sqrt{r})$ of its capacity.

The main results are as follows. (i) We show that, in general, a system in a stationary regime has at least $O(\sqrt{r})$ average queue lengths, even if the so called *null-controllability* (Atar et al., Ann. Appl. Probab. 16, 1764–1804, 2006) on a finite time interval is possible; strategies achieving this $O(\sqrt{r})$ growth rate we call *order-optimal*. (ii) We show that some natural algorithms, such as *MaxWeight*, that guarantee stability, are *not* order-optimal. (iii) Under the *complete resource pooling* condition, we prove the diffusion limit of the arrival processes into server pools, under the Shadow routing. (We conjecture that result (iii) leads to order-optimality of the Shadow routing algorithm; a formal proof of this fact is an important subject of future work.) Simulation results demonstrate good performance and robustness of our scheme.

Keywords Queueing networks · Large flexible server pools · Routing and scheduling · Shadow routing · Many server asymptotics · Halfin–Whitt regime · Diffusion limit · Order-optimality

A.L. Stolyar (✉)
Bell Labs, Alcatel-Lucent, Murray Hill, NJ 07974, USA
e-mail: stolyar@research.bell-labs.com

T. Tezcan
University of Illinois, Urbana-Champaign, Urbana, IL, USA
e-mail: ttezcan@uiuc.edu

Mathematics Subject Classification (2000) 60K25 · 90B15 · 60J70

1 Introduction

In this paper we consider a general problem of devising efficient and robust control strategies for service systems (e.g., call centers) with several input flows (classes) and several flexible multi-server pools. Servers within each pool are homogeneous. The average service time $1/\mu_{ij}$ of a customer depends on both the customer class i and the server pool index j. A control strategy for such a system can generally be thought of as consisting of two parts: (a) a routing algorithm which decides which server pool an arriving customer should go to for service if idle servers are available, or should it wait in a queue; and (b) a scheduling algorithm which determines which customer should a server take for service when becoming idle, if there are customers waiting in the queues, or should it remain idle. It is well known that a "good" routing algorithm is crucial for the efficiency of a multi-server system, see [2, 19, 38], in the (typical) case when service pre-emption is not allowed. This is due to the fact that, under heavy system load, just to be able to handle all input flows while keeping queues stable, the routing should occur, roughly speaking, only along a certain set of *basic activities*. (An activity is a class-pool pair (i, j); basic activities form a subset of them.) If flow input rates are known a priori (along with average service times $1/\mu_{ij}$ for all (i, j)), finding basic activities is a matter of solving a static planning (linear) program (SPP), whose objective is to minimize maximum average utilization ("balance" load) across the server pools.

When basic activities are known in advance, via a priori solution to SPP, efficient control strategies exist (see literature review below). In contrast, in this paper we propose a generic routing algorithm, which optimally balances server pools' loads and "automatically" identifies basic activities, *without the knowledge of the flow input rates and without explicitly solving any optimization problem*. The algorithm is based on *Shadow routing* in a virtual queueing system (and is an instance of a more general *greedy primal-dual* algorithm of [35]). Each routing decision is made via a very simple index rule upon a customer arrival, based on the values and simple updates of virtual queues; "virtual" here means that the "queues" are just variables maintained by the algorithm. When/if flow input rates change, no explicit detection of such event (or any other input rate measurement/estimation) is necessary and no new optimization problem needs to be solved—the virtual queues automatically readjust and the algorithm starts routing along possibly new set of basic activities. We believe that this kind of robustness is a very attractive feature of the algorithm. The Shadow routing algorithm can be used in conjunction with a variety of scheduling algorithms.

An appropriate and commonly used asymptotic regime for the study of many-server systems is the Halfin–Whitt (or, quality and efficiency driven—QED) asymptotic regime, when server pool sizes and the input rates are scaled up simultaneously by a factor r growing to infinity, while keeping the system load within $O(\sqrt{r})$ of its capacity. While our primary goal is to study the behavior of Shadow-routing based control algorithm in the Halfin–Whitt regime, we first establish a general property of this regime, which is of independent interest. We prove that, in general, the average

total queue length in a stationary regime, under any control discipline, scales *at least as* $O(\sqrt{r})$ as $r \to \infty$. While this result is very intuitive and may even seem trivial, it is not, especially in the light of results of [8], where it is shown that under a certain structural *null-controllability* condition, it is possible to control system *on a finite time interval* so that the queues are $o(\sqrt{r})$; our result thus shows that in a steady state such behavior is in general impossible. Motivated by this result, we call a control strategy achieving $O(\sqrt{r})$ steady-state average queue scaling *order-optimal*.

We next address the following question: Is order-optimality achievable by the simple and much studied MaxWeight strategy, which is known to guarantee system stability without a priori knowledge of input rates? (In other words, maybe it does not take much to achieve order-optimality without the knowledge of input rates?) The answer is 'no': we prove that MaxWeight scheduling combined with a very natural fastest-server-first (FSF) routing rule is *not* order-optimal. This is in contrast to the fact that, for systems *with a single customer class* and multiple server pools, FSF rule has been shown to be asymptotically optimal [2].

Finally, we study the behavior of Shadow routing in Halfin–Whitt regime. Namely, under an additional *complete resource pooling* (CRP) condition, we prove the diffusion limit for the arrival processes into server pools under the Shadow routing. This result in turn suggests order-optimality (under CRP) of Shadow routing. (A formal proof of the latter property is a subject of future research.)

Our simulation results demonstrate good performance and robustness of our Shadow-routing based scheme. First, we simulate a relatively complex system with 4 customer classes, 4 server pools and several non-basic activities, in addition to basic ones; Shadow routing very quickly identifies non-basic activities and avoids their usage, thus keeping the system stable and queues short; in addition, we provide intuition and guidance on how to choose the algorithm key parameter. We compare Shadow routing scheme to MaxWeight–FSF scheme by simulating a 2x2-system ("X-system")—our algorithm provides shorter queues and even server utilization; here we also demonstrate how quickly virtual queues of the Shadow routing readjust when a sudden dramatic change of the input rates occurs.

1.1 Review of related previous work

The *many-server heavy-traffic* regime was first formally introduced in [22] although some of the insights are credited to Erlang. This analysis has been mainly used in call center applications. For a detailed literature review on call centers and many-server heavy-traffic analysis we refer the reader to the two survey papers [17] and [1]. See, for example, [25] and [31] for other applications.

The analysis of Halfin and Whitt has been extended in several directions. Paper [18] studied the asymptotic analysis of an $M/M/n$ system with impatient customers and established similar results to those in Halfin and Whitt. Paper [30] established the diffusion limit of a G/PH/n system, where PH stands for a phase type service time distribution, and also established the many-server diffusion limits of a V-model parallel server system under a static priority policy. Paper [40] studied the many-server diffusion limit of a $G/H_2^*/n/m$ system, where H_2^* indicates that the service time

distribution is an extremal distribution among the class of hyperexponential distributions. This analysis is later used in [39] to provide approximations for $G/GI/n/m$ systems. Also see [16, 18, 32, 40], and [26] for other extensions of [22].

Armony and Maglaras studied an $M/M/n$ system with two customer classes in [3, 4]. Paper [21] studied a V-parallel server system with impatient customers; it shows that a static buffer priority policy with a threshold policy is asymptotically optimal. Paper [2] studied an inverted-V-parallel server system and showed that the faster-server-first (FSF) policy is asymptotically optimal.

Most related to our work are papers on so called "skills-based-routing" (SBR) in parallel server systems. As in the current paper, the goal is to devise nearly optimal policies for systems with many servers. Papers [24] and [7] formulated a diffusion control problem to study a V-parallel server system with impatient customers in many-server heavy traffic; [7] proved that the policies obtained from this approach are asymptotically optimal. Works [5, 6] followed a similar approach to that in [7] to find asymptotically optimal policies for tree-like systems. In [38] and [15] it is shown that a greedy policy is asymptotically optimal first for N-systems and then for general systems when the service rates are only server pool dependent. More recently [20] proposed the Fixed Queue Ratio (FQR) routing rule together with corresponding staffing rule for general parallel server systems. Under certain additional conditions, the FQR rule provides an asymptotically optimal solution; in [19] the authors showed that it is asymptotically optimal when the holding costs are convex. The main difference of our approach from this body of literature is that all of the papers above assume that basic and non-basic activities are known a priori. However, if this is not the case, the proposed policies may lead to instability; see [29].

Another line of research considers a different asymptotic regime than the Halfin and Whitt regime; see [9–11]. These papers use a fluid model approach to provide solutions for determining capacity when arrival rates are time varying and uncertain. All of these papers consider staffing and dynamic routing in a setting with uncertainty. Since their analysis is based on a fluid model approach, it is especially appropriate when the uncertainty is significant. Here, our focus is mainly on diffusion limits.

For a review of skills-based-routing problems in the so called conventional heavy-traffic regime see [27, 36] and references therein. In particular, the MaxWeight scheduling algorithm for SBR systems (which is valid and stable regardless of the asymptotic regime) is a special case of the $Gc\mu$-rule in [27].

Finally, our Shadow routing algorithm is a special case of the *greedy primal-dual* (GPD) algorithm [35] for the general problem of maximizing queueing network utility subject to stability of the queues. For another application of GPD algorithm in a (different) shadow routing scheme, in the context of a telecommunication system control, see [37].

1.2 Paper layout

In Sect. 2 we define the model and the asymptotic regime, as well as the CRP condition and related notions. The fact that, in general, the average queue length in stationary regime is at least $O(\sqrt{r})$ is proved in Sect. 3. Section 4 contains formulation and discussion of the result on non-order-optimality of MaxWeight rule. (The proof is in

Sects. 9–12 at the end of paper.) We define Shadow routing algorithm in Sect. 5. In Sect. 6 we prove the diffusion limit for the input processes under the Shadow routing. The conjecture on the order-optimality of the Shadow routing scheme is formulated in Sect. 7. The simulation experiments are presented and discussed in Sect. 8. We conclude and discuss future work in Sect. 13.

1.3 Basic notation

The set of real numbers is denoted $\mathbb{R}$; for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we define a norm by $\|x\| = \sum_i |x_i|$; for a function $f = (f(t), t \geq 0)$, $\|f\|_T \doteq \sup_{0 \leq t \leq T} \|f(t)\|$; $a \vee b \doteq \max\{a, b\}$ and $a \wedge b \doteq \min\{a, b\}$; $\lfloor a \rfloor$ is the largest integer not exceeding a.

For random variables and processes: $\Rightarrow$ and $\stackrel{d}{=}$ denote convergence and equality in distribution in the Skorohod space of RCLL functions on $[0, \infty)$ (or $[0, T]$); $\stackrel{P}{\to}$ is convergence in probability; a.s. and w.p.1 mean *almost surely* and *with probability 1*; unless otherwise specified, convergence $\to$ of random variables means a.s. convergence in the appropriate probability space. Sometimes we just write 0 for an identically zero function.

2 Model and the asymptotic regime

We consider a system with I input flows and J server pools; the sets of flow and pool indices are denoted by $\mathcal{I} = \{1, \ldots, I\}$ and $\mathcal{J} = \{I + 1, \ldots, I + J\}$, respectively. Each server pool $j \in \mathcal{J}$ consists of a number of homogeneous servers; the mean service time of a flow (or type, or class) i customer on a server of pool j is $1/\mu_{ij} > 0$. (If $\mu_{ij} = 0$, this means that class i customers cannot be served in pool j.) We assume that all input flows are independent Poisson and all service time distributions are exponential.

Throughout the paper, we restrict ourselves to the control disciplines that make control decisions only based on the current state (current server occupancy and queue content), and such that customers cannot leave the system before their service is complete; see [14] for more details. (For the Shadow routing scheme, introduced in Sect. 5, the system state is augmented by the state of virtual queues, defined as part of the algorithm.) For some (but not all) results we will require that service is non-pre-emptive. To simplify the definition of stability just below, assume also that at least one customer must be in service as long as there is at least one customer in the system. Under any such control discipline, the process describing system evolution is a continuous-time countable Markov chain. (For the Shadow routing scheme, mild additional conditions are required—see discussion following its definition in Sect. 5.) System stability is understood as positive recurrence of this Markov chain; stability implies the existence of a unique stationary distribution.

We consider a "large number of servers", or Halfin–Whitt, or *quality and efficiency driven* (QED), asymptotic regime, in which the input rates and the numbers of servers in each pool are increased simultaneously with scaling parameter $r \to \infty$. Namely, in a system indexed by r: the number of servers in pool j is $N_j^r = \beta_j r$, with parameter

$\beta_j > 0$; the input rate of flow i is

$$\lambda_i^r = \lambda_i r + \ell_i \sqrt{r} + o\big(\sqrt{r}\big), \tag{2.1}$$

with parameters $\lambda_i > 0$ and real ℓ_i.

For many (but not all!) results in this paper we will assume that a so called *complete resource pooling* (CRP) condition holds. To define it, along with related notions, consider the following ("fluid-scale") *static planning problem* (SPP), first introduced in [23], which is a linear program:

$$\min_{\{\lambda_{ij}\},\rho} \rho, \tag{2.2}$$

subject to

$$\sum_i \lambda_{ij}/(\beta_j \mu_{ij}) \le \rho, \quad \forall j, \tag{2.3}$$

$$\lambda_{ij} \ge 0, \quad \forall i, j, \tag{2.4}$$

$$\sum_j \lambda_{ij} = \lambda_i, \quad \forall i. \tag{2.5}$$

We say that the *complete resource pooling* (CRP) condition (see [12, 23]) holds if (i) problem (2.2)–(2.5) has unique solution, $\{\lambda_{ij}\}$, ρ, and it is such that (ii) $\rho = 1$ and (iii) pairs (edges) (ij) for which $\lambda_{ij} > 0$ form a tree in the graph with the set of vertices being $\mathcal{I} \cup \mathcal{J}$. The pairs (ij) for which $\lambda_{ij} > 0$, are called *basic activities*, and the set of those is denoted by $\mathcal{E}_b$; while the set of all pairs (ij), $i \in \mathcal{I}$, $j \in \mathcal{J}$, called *activities*, is denoted by $\mathcal{E}$.

Now we will introduce some properties and notation, associated with the CRP condition, that we use throughout the paper. If CRP condition holds, then the linear program dual to (2.2)–(2.5) has a unique solution, with duals α_j corresponding to constraints (2.3) being all strictly positive and $\sum_j \alpha_j = 1$. The condition $\rho = 1$ in CRP means that, "up to order $o(r)$" the system is critically loaded, i.e. system load is equal to its capacity. The parameters ℓ_i describe order $O(\sqrt{r})$ deviations of the system load from "exactly critical". Let us denote by $\nu_i \doteq \min_j \alpha_j/(\beta_j \mu_{ij}) > 0$ the *customer workload contribution* [36] of (one customer of) flow i; the minimum in the definition of ν_i is attained if and only if (ij) is a basic activity; also, for each j, $\alpha_j = \max_i \nu_i(\beta_j \mu_{ij}) = \beta_j \max_i \nu_i \mu_{ij}$ with the maximum attained if and only if (ij) is a basic activity (see [36]). Given these definitions, it is easy to verify that $\sum_i \lambda_i \nu_i = \sum_j \alpha_j = 1$, and therefore the rate at which *workload* arrives in the system is

$$\sum_i \lambda_i^r \nu_i = r + \left[\sum_i \ell_i \nu_i\right] \sqrt{r} + o\big(\sqrt{r}\big), \tag{2.6}$$

while the maximum possible rate at which system can serve workload is

$$\sum_j \beta_j r \left[\max_i \nu_i \mu_{ij}\right] = r \sum_j \alpha_j = r. \tag{2.7}$$

Therefore, for a system with index r to be stable, it is necessary that

$$\sum_i \lambda_i^r \nu_i \le r, \tag{2.8}$$

which implies that for systems with all large r to be stable, it is necessary that

$$\sum_i \ell_i \nu_i \le 0. \tag{2.9}$$

Condition

$$\sum_i \ell_i \nu_i < 0 \tag{2.10}$$

is sufficient for the system to be stabilizable for all large r, *under an appropriate control*. (Under condition (2.10), it is possible to route customers so that the load of each pool is strictly below its capacity—see Remark 2 in Sect. 6.)

One more notation we use throughout the paper is: given CRP condition, $\psi_{ij}^* \doteq \lambda_{ij}/\mu_{ij}$; obviously, $\psi_{ij}^* > 0$ if and only if (ij) is a basic activity, and $\sum_i \psi_{ij}^* = \beta_j$ for each j.

3 Order $O(\sqrt{r})$ lower bound on the average queue length

In this section we consider the model and the asymptotic regime as in Sect. 2, and assume CRP condition. If condition (2.10) holds, then for all sufficiently large r there exist control strategies stabilizing the system, and therefore under such strategies a stationary regime of the system exists for each large r. We will show that if parameters ℓ_i are close enough to zero (but are fixed and strictly negative), then no matter what control strategy we choose for each r, as $r \to \infty$ the average queue length (number of waiting customers) grows at least as $\epsilon_2\sqrt{r}$, for some $\epsilon_2 > 0$. Thus we prove that, in general, it is impossible for the *steady-state* average queues to scale as $o(\sqrt{r})$, regardless of the structure of the system. This fact may seem obvious, because it certainly holds for example for any single pool, single class system [22]. However, it is not obvious for a general system with multiple input flows and multiple server pools, especially in view of the results of [8], where authors show that for such general systems satisfying a certain structural condition called *null-controllability*, there exist control strategies making average queue length scale as $o(\sqrt{r})$ *on a finite time interval*. The null-controllability condition says, roughly, that if a system state is such that each server pool allocates a non-zero fraction of servers to each of its basic activities, then it is possible to reassign a positive fraction of customers (being served) between the pools so that the total instantaneous service rate in the system increases. Control disciplines proposed in [8] use this condition to increase service rate, *by employing non-basic activities*, when it is necessary to prevent a queue build-up, thus asymptotically "eliminating" queue on a finite time interval. Our Theorem 3.1 below shows that in steady state such "queue elimination" is impossible.

Let us use the following notation, for a system with index r: $A_i^r(t)$ is the number of class i arrivals by time t (i.e., in the interval $[0,t]$), a Poisson process of rate λ_i^r; $S_{ij}^r(s)$ is the number of class i customers which would be served by pool j if it spends the cumulative time s serving class i, a Poisson process of rate μ_{ij}; $X_i^r(t)$ is the number of class i customers present in the system at time t; $Y_i^r(t)$ is the number of class i customers queued (not being served) at time t; $\Psi_{ij}^r(t)$ is the number of pool j servers, serving class i customers at time t; $Z_j^r(t)$ is the number of pool j servers that are idle at time t. We have obvious relations at any time t:

$$X_i^r(t) = Y_i^r(t) + \sum_j \Psi_{ij}^r(t),$$

$$N_j^r = Z_j^r(t) + \sum_i \Psi_{ij}^r(t).$$

For each r, suppose we have a fixed, well-defined control strategy, under which the system is stable. (Such strategies certainly do exist, given (2.10).) For each r, it may be a *different* strategy, and it may be pre-emptive or non-pre-emptive. For each r we can and will consider a stationary version of the process.

Theorem 3.1 *Consider the model and the asymptotic regime, as defined in Sect. 2. Assume CRP condition holds. Then, there exist numbers $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that the following is true. If $-\epsilon_1 \le \ell_i < 0$ for all i, then any sequence of stationary distributions as described above is such that*

$$\liminf_{r\to\infty} E\frac{1}{\sqrt{r}}\sum_i Y_i^r(t) \ge \epsilon_2. \tag{3.1}$$

Theorem 3.1 motivates the following definition.

Definition 3.2 A control strategy is called *order-optimal* in the asymptotic regime of Sect. 2 and under condition (2.10), if system is stable for all large r and the stationary distributions are such that

$$\limsup_{r\to\infty} E\frac{1}{\sqrt{r}}\sum_i Y_i^r(t) < \infty. \tag{3.2}$$

Let us introduce the following "diffusion-scaled" quantities:

$$\hat{Y}_i^r(t) = \frac{1}{\sqrt{r}} Y_i^r(t),$$

$$\hat{Z}_j^r(t) = \frac{1}{\sqrt{r}} Z_j^r(t),$$

$$\hat{\Psi}_{ij}^r(t) = \frac{1}{\sqrt{r}}\big[\Psi_{ij}^r(t) - \psi_{ij}^* r\big].$$

First, we prove the following basic fact that we use throughout the paper.

Lemma 3.3 *Assume the CRP condition and (2.9). Then, as $r \to \infty$, any sequence of system stationary regimes (under any stabilizing control disciplines) is such that*

$$\sum_{(ij)\notin\mathcal{E}_b}\left[\frac{\alpha_j}{\beta_j}-\nu_i\mu_{ij}\right]E\hat{\Psi}^r_{ij}(t)+\sum_j\frac{\alpha_j}{\beta_j}E\hat{Z}^r_j(t)\to-\sum_i\ell_i\nu_i. \tag{3.3}$$

In particular, since $\hat{\Psi}^r_{ij}(t)\geq 0$ and $\alpha_j/\beta_j-\nu_i\mu_{ij}>0$ for each $(ij)\notin\mathcal{E}_b$, and all $\hat{Z}^r_j(t)\geq 0$, and $[-\sum_i\ell_i\nu_i]\geq 0$, we have

$$\limsup_{r\to\infty}E\big[\hat{\Psi}^r_{ij}(t)\big]\leq-\left[\frac{\alpha_j}{\beta_j}-\nu_i\mu_{ij}\right]^{-1}\sum_i\ell_i\nu_i\quad\textit{for each non-basic } (i,j),$$

and

$$\limsup_{r\to\infty}E\big[\hat{Z}^r_j(t)\big]\leq-\left[\frac{\alpha_j}{\beta_j}\right]^{-1}\sum_i\ell_i\nu_i\quad\textit{for all } j.$$

Proof If the system is stable, then in a stationary regime the average arrival and departure rates for each class i are equal. Weighting such equalities by workload contributions ν_i and summing up, we can write

$$\sum_i\lambda^r_i\nu_i=\sum_{ij}\nu_i\mu_{ij}E\big[\Psi^r_{ij}(t)\big], \tag{3.4}$$

meaning that the workload arrival and departure rates are equal. Subtracting (2.6) from (2.7), and then using (3.4), $\beta_j=\sum_i\psi^*_{ij}$ and taking the limit we obtain (3.3). □

Proof of Theorem 3.1 The proof is by contradiction. Suppose the statement of the theorem does not hold; namely, for arbitrarily small $\epsilon_1>0$ and $\epsilon_2>0$ there exists a sequence of systems, with all parameters $\ell_i\in[-\epsilon_1,0)$, such that the $\liminf E(1/\sqrt{r})Y^r_i(t)<\epsilon_2$. Then, by choosing a sequence of vectors $\ell=\{\ell_i\}$, each with strictly negative components, converging to 0 componentwise, and a sequence (on r) of systems for each ℓ, and then choosing an appropriate "diagonal" subsequence (on r), we can construct a sequence (on r) of systems satisfying our specified asymptotic regime with

$$\ell_i=0,\quad\forall i \tag{3.5}$$

and such that

$$E\left\{\sum_i\hat{Y}^r_i(t)\right\}\to 0,\quad r\to\infty,\ \forall i. \tag{3.6}$$

We will show that such a sequence is impossible. First, we observe that given (3.5), by Lemma 3.3 we have:

$$E\left\{\sum_j\hat{Z}^r_j(t)\right\}\to 0,\quad r\to\infty,\ \forall j. \tag{3.7}$$

$$E\{\hat{\Psi}_{ij}^r(t)\} \to 0, \quad \text{for each non-basic activity } (ij). \tag{3.8}$$

The basic evolution of the number of class i customers is given by

$$X_i^r(t) = X_i^r(0) + A_i^r(t) - \sum_j S_{ij}^r\left(\int_0^t \Psi_{ij}^r(s)\,ds\right). \tag{3.9}$$

Consider a stationary version of the process for each r, in the interval $[0,\infty)$. This is equivalent to assuming that the initial state (at $t=0$) has a stationary distribution.

The arrival and service processes are controlled by independent unit rate Poisson processes, $\Pi_i^{(a)}$, $i\in\mathcal{I}$, and $\Pi_{ij}^{(s)}$, $(ij)\in\mathcal{E}$, so that

$$A_i^r(t) \equiv \Pi_i^{(a)}\big(\lambda_i^r t\big), \qquad S_{ij}^r(t) \equiv \Pi_{ij}^{(s)}(\mu_{ij} t), \tag{3.10}$$

and thus the processes for all r are constructed on a common probability space.

We will choose a subsequence of $\{r\}$, such that the following property holds. *With probability 1, for any fixed $t>0$ and $d>0$, uniformly on any sequence of pairs (t_1^r, t_2^r), such that $0\le t_1^r < t_2^r \le rt$ and $t_2^r - t_1^r \ge \sqrt{r}d$,*

$$\lim_{r\to\infty} \frac{\Pi_i^{(a)}(t_2^r) - \Pi_i^{(a)}(t_1^r)}{t_2^r - t_1^r} = 1, \quad \forall i, \tag{3.11}$$

and analogously for each $\Pi_{ij}^{(s)}$. (This can be done in a variety of ways, cf. the proof of analogous property in Sect. 4.2 of [33].)

Consider the processes on a finite time interval $[0,T]$; constant T will be chosen later. The convergence of expected values in (3.6)–(3.8) implies convergence in probability (because the r.v. involved are non-negative). Then, we can choose a further subsequence of r (with r growing sufficiently fast), so that, by Borel–Cantelli, as $r\to\infty$, (3.6)–(3.8) hold w.p.1 at times $t=0$ and $t=T$, namely,

$$\begin{gathered}\sum_i \hat{Y}_i^r(t) \to 0, \quad \forall i, \qquad \sum_j \hat{Z}_j^r(t) \to 0, \quad \forall j,\\ \hat{\Psi}_{ij}^r(t) \to 0, \quad (ij)\notin\mathcal{E}_b, \text{ w.p.1, } t=0,T;\end{gathered} \tag{3.12}$$

moreover, we can choose this subsequence so that, in addition,

$$\int_0^T \left[\sum_i \hat{Y}_i^r(t)\right] dt \to 0, \quad \text{w.p.1}, \tag{3.13}$$

$$\int_0^T \left[\sum_j \hat{Z}_j^r(t)\right] dt \to 0, \quad \text{w.p.1}, \tag{3.14}$$

$$\int_0^T \big[\hat{\Psi}_{ij}^r(t)\big]\,dt \to 0, \quad \text{w.p.1. } (ij)\notin\mathcal{E}_b. \tag{3.15}$$

(We keep notation $\{r\}$ for the subsequence.)

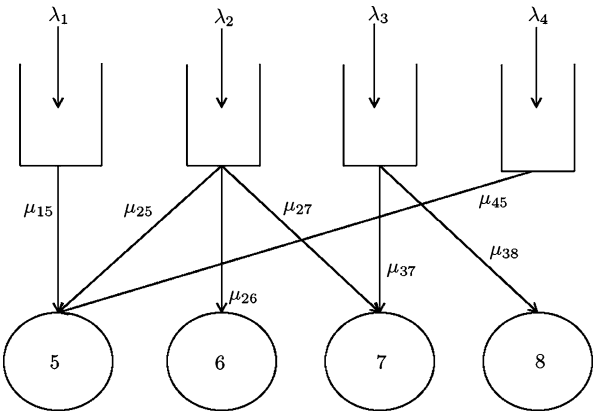


Fig. 1 A sample basic-activity tree (for the proof of Theorem 3.1)

To improve exposition, from this point on in the proof, consider a special system with $\mathcal{I} = \{1, 2, 3, 4\}$, $\mathcal{J} = \{5, 6, 7, 8\}$, and basic activities $\mathcal{E}_b = \{(1, 5), (2, 5), (2, 6), (2, 7), (3, 7), (3, 8), (4, 5)\}$, see Fig. 1. (All other—non-basic—activities are present in the system, but they are not shown in Fig. 1.) It will be clear that the proof applies to the general system (with arbitrary basic-activity tree) as well.

Consider flow (type) 1 which is a leaf of the basic-activity tree, which implies that $\lambda_{1,5} = \lambda_1$ and $\lambda_1/(\beta_1\mu_{1,5}) < 1$. Given (3.13), (3.15), and property (3.11), we obtain from (3.9):

$$X_1^r(t) = X_1^r(0) + \Pi_1^{(a)}(\lambda_1 rt) - \Pi_{1,5}^{(s)}\left(\mu_{1,5}\int_0^t X_1^r(s)\,ds\right) + o\big(\sqrt{r}\big), \quad (3.16)$$

where $o(\sqrt{r})$ holds u.o.c., w.p.1. If we compare the process X_1^r to the process $X_1^{r,0}$ (on the same probability space), solving equation

$$X_1^{r,0}(t) = X_1^r(0) + \Pi_1^{(a)}(\lambda_1 rt) - \Pi_{1,5}^{(s)}\left(\mu_{1,5}\int_0^t X_1^{r,0}(s)\,ds\right), \quad (3.17)$$

it is easy to see that

$$\big|X_1^r(t) - X_1^{r,0}(t)\big| = o\big(\sqrt{r}\big), \quad \text{u.o.c., w.p.1.} \quad (3.18)$$

Indeed, if we denote

$$g^r(t) = \max_{0\le s\le t}\frac{1}{\sqrt{r}}\big|X_1^r(t) - X_1^{r,0}(t)\big|,$$

we have

$$\frac{1}{r}\left|\int_0^t X_1^r(s)\,ds - \int_0^t X_1^{r,0}(s)\,ds\right| \le \frac{1}{\sqrt{r}}g^r(t)t.$$

For some fixed $\epsilon > 0$, define stopping time $\tau^r = T \wedge \min_{t\ge 0}\{g^r(t) \ge \epsilon\}$. Then,

$$g^r\big(\tau^r\big) \le g^r\big(\tau^r\big)\mu_{ij}t + o(1),$$

and thus $g^r(\tau^r) \to 0$ if we choose T small enough so that $\mu_{ij}T < 1$; this in turn implies that for all large r, $\tau^r = T$ and then $g^r(t) \to 0$ uniformly in the interval $[0, T]$. (Repeating the above argument with T as initial time, and so on, we can in fact show that $g^r(t) \to 0$ u.o.c. in $[0, \infty)$.)

What is important about (3.18) is that the process $X_1^{r,0}$ is uniquely determined by the initial state $X_1^r(0)$ and independent primitive processes $\Pi_1^{(a)}$ and $\Pi_{1,5}^{(s)}$; and (3.18) says is that, w.p.1, as $r \to \infty$, the actual trajectories of X_1^r are close to $X_1^{r,0}$, up to a u.o.c. $o(\sqrt{r})$ error.

We can analogously treat another leaf (of the basic-activity tree) type 4, and corresponding activity $(4, 5)$, to obtain an analog of (3.18):

$$\left|X_4^r(t) - X_4^{r,0}(t)\right| = o\left(\sqrt{r}\right), \quad \text{u.o.c., w.p.1,}$$

where $X_1^{r,0}$ is uniquely determined by $X_4^r(0)$ and independent primitive processes $\Pi_4^{(a)}$ and $\Pi_{4,5}^{(s)}$.

Now consider type 2, and denote

$$\Psi_{2,5}^{r,0}(t) \equiv \beta_5 r - X_1^{r,0}(t) - X_4^{r,0}(t), \qquad \Psi_{2,6}^{r,0}(t) \equiv \beta_6 r.$$

We have

$$\left|\int_0^t \Psi_{2,j}^r(s)\,ds - \int_0^t \Psi_{2,j}^{r,0}(s)\,ds\right| = o\left(\sqrt{r}\right), \quad \text{u.o.c., w.p.1, } j = 5, 6.$$

Therefore, there is only one basic activity for type 2, namely $(2, 7)$, for which $\int_0^t \Psi_{2,j}^r(s)\,ds$ is "not yet determined", while for all other basic activities, namely $(2, 5)$ and $(2, 6)$, $\int_0^t \Psi_{2,j}^r(s)\,ds$ is determined, up to a u.o.c. $o(\sqrt{r})$ error, by the independent primitive processes we "used" so far, namely $\Pi_1^{(a)}$, $\Pi_{1,5}^{(s)}$, $\Pi_4^{(a)}$, $\Pi_{4,5}^{(s)}$. (Such a type always exists after we treat all leaf types, which are 1 and 4 in our example.) Then, X_2^r satisfies

$$\begin{aligned} X_2^r(t) = X_2^r(0) + \Pi_2^{(a)}(\lambda_2 rt) - \sum_{j=5,6} \Pi_{2,j}^{(s)}\left(\mu_{2,j}\int_0^t \Psi_{2,j}^{r,0}(s)\,ds\right) \\ - \Pi_{2,7}^{(s)}\left(\mu_{2,7}\int_0^t \left[X_2^r(s) - \sum_{j=5,6}\Psi_{2,j}^{r,0}(s)\right]ds\right) + o\left(\sqrt{r}\right), \end{aligned} \tag{3.19}$$

where $o(\sqrt{r})$ holds u.o.c., w.p.1. From (3.19) we obtain that

$$\left|X_2^r(t) - X_2^{r,0}(t)\right| = o\left(\sqrt{r}\right), \quad \text{u.o.c., w.p.1,}$$

where $X_2^{r,0}$ is the solution to (3.19) with $o(\sqrt{r})$ term dropped, and thus uniquely determined by $X_2^r(0)$, independent primitive processes $\Pi_2^{(a)}$ and $\Pi_{2,7}^{(s)}$, and by other initial conditions and primitive processes we "used" so far.

Finally, the only "untreated" type left is type 3. (In general, we would repeat the above procedure until only one type left.) We denote

$$\Psi_{3,7}^{r,0}(t) \equiv \beta_7 r - \bigl[X_2^{r,0}(t) - \Psi_{2,5}^{r,0}(t) - \Psi_{2,6}^{r,0}(t)\bigr], \qquad \Psi_{3,8}^{r,0}(t) \equiv \beta_8 r,$$

and then have

$$\left| \int_0^t \Psi_{3,j}^r(s)\,ds - \int_0^t \Psi_{3,j}^{r,0}(s)\,ds \right| = o\bigl(\sqrt{r}\bigr), \quad \text{u.o.c., w.p.1,} \ \ j = 7, 8.$$

Then, $X_3^{r,0}$ satisfies:

$$X_3^r(t) = X_3^r(0) + \Pi_3^{(a)}(\lambda_3 r t) - \sum_{j=7,8} \Pi_{3,j}^{(s)}\left(\mu_{3,j} \int_0^t \Psi_{2,j}^{r,0}(s)\,ds\right) + o\bigl(\sqrt{r}\bigr), \tag{3.20}$$

where $o(\sqrt{r})$ holds u.o.c., w.p.1. Consider time $t = T$ we fixed earlier, and recall that (3.12) holds. Then $\sum_i X_i^r(T) = \sum_j \beta_j r + o(\sqrt{r})$, and therefore $X_3^r(T) = X_3^{r,0}(T) + o(\sqrt{r})$, where $X_3^{r,0}(T) = \sum_j \beta_j r - \sum_{i\neq 3} X_i^{r,0}(T)$ is determined by primitive processes "used" so far, along with $X_i^r(0)$ for $i \neq 3$. From (3.20) we have

$$\left[X_3^{r,0}(T) - X_3^r(0) + \sum_{j=7,8} \Pi_{3,j}^{(s)}\left(\mu_{3,j} \int_0^T \Psi_{2,j}^{r,0}(s)\,ds\right)\right] - \bigl[\Pi_3^{(a)}(\lambda_3 r T)\bigr] = o\bigl(\sqrt{r}\bigr), \tag{3.21}$$

where $o(\sqrt{r})$ holds u.o.c., w.p.1. The first term in square brackets in the LHS of (3.21) is uniquely determined by the initial state and all primitive processes except $\Pi_3^{(a)}$, while the second term depends only on $\Pi_3^{(a)}$; therefore these two terms are independent. But then the RHS of (3.21) cannot be $o(\sqrt{r})$ u.o.c. w.p.1, because, clearly, both events $\{\Pi_3^{(a)}(\lambda_3 r T) > \sqrt{r}\}$ and $\{\Pi_3^{(a)}(\lambda_3 r T) < -\sqrt{r}\}$ have probability at least some $\delta > 0$ for all large r. The contradiction completes the proof. □

4 MaxWeight algorithm is not order-optimal

The MaxWeight scheduling policy (in the context of parallel server systems) is defined as follows: (i) When a server in pool j becomes free, it takes for service a customer of a class $i_* \in \arg\max_i Y_i \mu_{ij}$, where Y_i is the current queue length of type i. (If $Y_i = 0$ for all types i with $\mu_{ij} > 0$, the server stays idle.) (ii) No service pre-emptions or interruptions are allowed. (iii) When a new customer of type i arrives when there are idle servers capable of serving it (with $\mu_{ij} > 0$), the customer goes for service in one of them—the specific *routing strategy* for choosing an idle server is arbitrary.

It is well known that the MaxWeight policy is asymptotically optimal in the conventional heavy traffic, see [34] and [27]. The policy is also known to be stable as long as the system load is strictly less than its capacity—this is guaranteed to hold for all large r, under condition (2.10), in the Halfin–Whitt asymptotic regime specified in Sect. 2.

In this section, we would like to explore how the average queue lengths under MaxWeight algorithm scale in the Halfin–Whitt regime.

We first note that there are significant differences between conventional and many-server heavy-traffic regimes. Hence the insights provided by the results from the conventional heavy-traffic analysis are insufficient for the analysis in the Halfin–Whitt regime. The differences stem from the fact that in the conventional heavy-traffic regime the servers are busy almost all the time, but in the Halfin–Whitt regime there are available idle servers for a significant fraction of time. Therefore, as is well known, in the Halfin–Whitt regime the routing strategy has a significant impact on the performance. This in particular applies to the MaxWeight policy—the choice of the routing rule, while not affecting system stability, may be crucial for the performance. (In conventional heavy traffic, the routing rule used by MaxWeight makes no difference for the asymptotic behavior of queues.) Here, we show that this indeed is the case. Specifically, we show that the MaxWeight algorithm is not order-optimal when paired with a natural routing strategy, which is known as the fastest-server-first (FSF) rule. Under the FSF routing rule, when a customer arrives to find more than one idle server, it is routed to the fastest available server (largest μ_{ij}) for the customer's class. We haste to note that there may be other routing rules that when matched with MaxWeight can yield better performance. However, if we seek a routing rule that does not know a priori which activities are basic (this requires a priori knowledge of input rates), FSF is a very natural choice.

The main idea behind proving that the MaxWeight–FSF rule is not order-optimal is the following. Consider a sequence of stationary systems in the Halfin–Whitt regime (see (2.1)) indexed by the scaling parameter r. We first show that if the steady-state average queue length under MaxWeight–FSF in the Halfin–Whitt regime is $O(\sqrt{r})$, then with a non-vanishing probability the number of idle servers in the system must be $O(\sqrt{r})$. We next show that if the initial state of the system is such that the number of idle servers is $O(\sqrt{r})$, then under the FSF rule the number of servers in each pool working on each class will reach to a level $M\sqrt{r}$ with arbitrarily large M, including those activities that are non-basic. Hence, the system reaches a state where non-basic activities are used "more than they are supposed to be", which in turn implies that the system cannot be stable—a contradiction.

The result of this section is the following

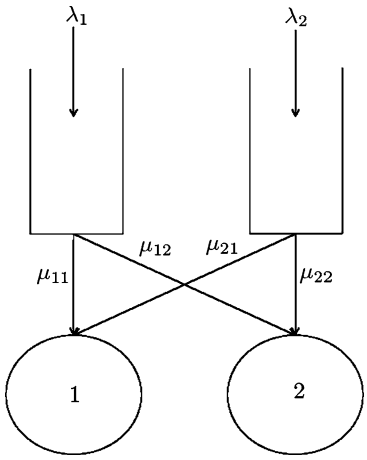
Theorem 4.1 *The MaxWeight–FSF algorithm, with non-pre-emptive service, is not order-optimal for some systems in the asymptotic regime specified in Sect. 2 and satisfying condition* (2.10).

The proof, deferred to Sect. 9 and subsequent sections, uses the "X-model" depicted in Fig. 2 as a counterexample.

5 Shadow routing algorithm, via virtual queues

As demonstrated by Theorem 4.1, even if the subcriticality condition (2.10) holds and a control strategy guarantees system stability, the strategy may be non-order-optimal.

Fig. 2 X-Model



On the other hand, a simple strategy which is *oblivious of the state of the system*, but routes appropriate fractions of the flows along basic activities only, in the way that balances average loads of all pools, will be able to keep stationary average queue length at $O(\sqrt{r})$. (Because the system decomposes into independently operating, subcritically loaded server pools, with independent input flows.) However, this latter strategy would require the problem (2.2)–(2.5) to be solved *in advance*, which in turn would require a priori knowledge (or real-time estimation) of the flow input rates.

We now propose a routing algorithm, which addresses the key problem described above. The algorithm dynamically routes (or, assigns) the arriving customers to server pools in a way that "automatically" (in the sense we elaborate on in the rest of the paper) solves the load-balancing problem (2.2)–(2.5), while *not requiring any a priori knowledge of input rates and without solving any linear program explicitly.* The algorithm is an instance of the *greedy primal-dual* (GPD) algorithm of [35] and, in itself, it does *not* rely on the CRP condition, or even on the subcriticality of the system load. (If system is overloaded, the algorithm still balances the server pool loads. Although the focus of this paper is on the performance of subcritically loaded system.) Thus, we believe that the algorithm addresses a key—routing—component of the control of systems in the many-server regime, in a generic, systematic way and in a variety of contexts.

The Shadow routing algorithm definition

Algorithm maintains virtual ("shadow") queue Q_j for each pool j—these are to "keep track" of the constraints (2.3). Parameter $\eta > 0$ is a small number (we elaborate later on how it should be chosen), which controls the tradeoff between "responsiveness" of the algorithm and its accuracy.

The initial state—the initial values of Q_j—is arbitrary. For example, it can be $\eta Q_j = 1/J$ for all j. Motivation for such initial state is given later.

Begin algorithm

Upon each new (actual) customer arrival, say of class i to be specific, the algorithm does the following (in sequence):

1. A virtual queue

$$m \in \arg\min_j Q_j/(\beta_j \mu_{ij}),$$

is identified; then the customer is routed to pool m, and for this m the following update is done:

$$Q_m := Q_m + 1/(\beta_m \mu_{im}).$$

The latter update has the interpretation of "routing" the amount of $1/(\beta_m \mu_{im})$ work to pool m.

2. If the condition

$$\eta \sum_j Q_j \geq 1$$

holds, the following update ("virtual service") is done:

$$Q_j := [Q_j - c]^+, \quad \text{for each } j, \tag{5.1}$$

where $c > 0$ is a fixed parameter, strictly greater than $\max\{1/(\beta_j \mu_{ij}) \mid \mu_{ij} > 0\}$.

End algorithm

Note that, by the virtual service rule (5.1), the inequality

$$\left| \sum_j Q_j - 1/\eta \right| \leq Jc \tag{5.2}$$

holds at all times (as long as it holds in initial state).

The virtual queueing system fits into the general framework [35] of maximizing queueing network utility subject to stability of the queues. In our case, the utility function of the system is $-U$, where U is the average rate of the virtual service (update (5.1)) activation; and the Shadow algorithm is an instance of GPD algorithm [35], which is asymptotically optimal. More precisely, the Shadow algorithm solves the load-balancing problem (2.2)–(2.5) in the following sense, regardless of whether or not the CRP condition holds. Consider the system with a fixed parameter r (and then input rates λ_i^r), and let $\{\lambda_{ij}^r\}$, ρ^r be an optimal solution of problem (2.2)–(2.5) with λ_i's replaced by λ_i^r's. (Note that λ_{ij}^r's and ρ^r are $O(r)$, because λ_i^r's are.) Suppose all parameters β_j, μ_{ij}, c are rational and the initial values of virtual queues Q_j (fixed for each η) satisfy (5.2). Consequently, for any η, the virtual queueing process $(Q(t), t \geq 0)$ is a finite Markov chain, which always has a stationary distribution. Denote by $\hat{\lambda}_{ij}^r$ the average rate at which the Shadow algorithm routes i-customers to pool j in a stationary regime. Then it follows from results of [35] that any sequence of stationary processes with different values of $\eta > 0$ is such that, as $\eta \to 0$, the maximum utilization of the server pools

$$\max_j \frac{1}{N_j^r} \sum_i \hat{\lambda}_{ij}^r \frac{1}{\mu_{ij}} = \max_j \sum_i \frac{\hat{\lambda}_{ij}^r}{r} \frac{1}{\beta_j \mu_{ij}}$$

converges to the optimal (smallest) max-utilization

$$\max_j \sum_i \frac{\lambda_{ij}^r}{r} \frac{1}{\beta_j \mu_{ij}} = \frac{\rho^r}{r}.$$

Moreover, the steady-state, rescaled virtual queue lengths, $\{\eta Q_j\}$, converge to an optimal set of dual variables $\{\alpha_j\}$ (corresponding to constraints (2.3)) for the problem (2.2)–(2.5). Since CRP condition is not assumed, set $\{\alpha_j\}$ is not necessarily unique, but $\sum_j \alpha_j = 1$ must hold. (This motivates the choice of the initial state, such that $\sum_j \eta Q_j = 1$, in the algorithm description.)

Remark 1 We emphasize again that the Shadow algorithm does not depend on the CRP condition, does not involve any explicit measurement/estimation of input rates, and does not require any optimization problem to be explicitly solved in advance; instead of all that, a simple calculation on a customer arrival assigns this customer to one of the pools.

Remark 2 The Shadow algorithm solves the load-balancing problem *from any initial state* of the virtual queues (after a transient period required for the rescaled queues ηQ_j to get close to optimal duals α_j). This is a very attractive feature of the algorithm: if input rates change (or even if the set of input flows changes), the routing will automatically readjust, to minimize max load of any pool. If input rates change relatively slowly with time, the algorithm will simply "track" such changes, by keeping pools' loads balanced. (This fact is one of the advantages of using an iterative shadow algorithm—we do not need to keep solving the SPP "from scratch," but rather "track" the solution by doing simple updates of variables. See [37] for related application of a shadow algorithm, where this feature is key.) If input rates change suddenly, there is no need to explicitly detect such event - the routing will readjust (after a transient period).

Remark 3 There is a variety of options of how the Shadow algorithm can be used for actual routing in the system. The first, straightforward option is to use the algorithm "literally": if Shadow algorithm assigns a customer to pool j, the customer is actually routed there, and waits for service there if necessary. (We conjecture that with this option, and in the case when CRP and condition (2.10) hold, the Shadow algorithm is order-optimal, see Sect. 7.) Another option is to use Shadow algorithm only to identify basic activities; as we will see in Sect. 6, under CRP, asymptotically, Shadow algorithm assigns customers along all basic activities and only along those. Then, one can apply any algorithm which *assumes* that basic activities are known, e.g. [20].

Remark 4 If one uses the described above "literal" option of the Shadow algorithm and the system is subcritically loaded (2.10), then since the routing itself automatically keeps all server pools subcritically loaded, for the system stability it is irrelevant how scheduling *within each pool* is done. For example, strict priorities, or a $c\mu$-rule, or a "generalized $c\mu$-rule could be used.

Remark 5 We want to point out that the use of virtual queues *cannot* in general be replaced by a direct use of the state of physical system. For example, although each Q_j has the meaning of unfinished work, it is virtual unfinished work in the shadow system, which operates quite differently from the real system, and as a result Q_j does not have a direct relation with the real unfinished work in pool j. To illustrate this, consider the system with $\mathcal{I} = \{1\}$, $\mathcal{J} = \{2, 3\}$, $\beta_2 = \beta_3 = 1$, $\lambda_1 = 3$, $\mu_{12} = 1$, $\mu_{13} = 2$; the servers in pool 3 are twice faster. With these parameters, $\alpha_3 = 2\alpha_2$, and therefore in the shadow system the stationary distribution is such that $Q_3 \approx 2Q_2$. On the other hand, in the real system, *assuming it is controlled so that* $EY_1^r = o(r)$ *in a steady state* (if not, the control is clearly "bad"), the unfinished work in pools 3 and 2 is approximately $(1/\mu_{13})r$ and $(1/\mu_{12})r$, respectively; so it is twice *less* in pool 3.

6 Diffusion limit of the input flows after shadow routing

Since the main focus of this paper is to study the Halfin–Whitt asymptotic regime, to see how Shadow routing performs in it, we need to go beyond the first-order properties, namely the results regarding the average rates of the flows on the algorithm "output", and study the diffusion limit of such processes. In this section we do just that.

Consider the Shadow routing algorithm with parameter η depending on r as $\eta = 1/f(r)$, where the function $f(r)$ is such that $f(r)/\sqrt{r} \to +\infty$. So, for example, it can be $f(r) = r^{3/4}$. (As we discuss later in Sects. 8.1 and 8.2, for a practical use, it is beneficial to *not* have $f(r)$ growing too fast.)

Assume the CRP condition holds. Recall that the input rates λ_i^r scale with r as in (2.1):

$$\lambda_i^r = \lambda_i r + \ell_i \sqrt{r} + o\big(\sqrt{r}\big);$$

that the rates $\{\lambda_{ij}\}$ are the unique solution to the static planning problem (2.2)–(2.5); and that $\{\alpha_j\}$ the unique optimal dual variables corresponding to the constraints (2.3), and $\sum_j \alpha_j = 1$.

Remark 1 For the results in this section, the "$\sqrt{r}$-deviation" parameters ℓ_i can be arbitrary real numbers, so that we do *not* require that the system load is below its capacity. In fact, the actual system load is completely irrelevant to the results, which will only concern with the properties of the processes on the output of the Shadow routing. In particular, for the purposes of this section, we could relax the CRP condition itself by *not* requiring that the optimal load ρ is necessarily 1, as long as it is positive. (We do still require that the SPP solution is unique and the basic activities form a tree.) So, the real system can be overloaded or underloaded, even on the fluid limit scale.

Theorem 6.1 *Consider the sequence of the virtual queueing processes* (*with* $\eta = 1/f(r)$), *using notation* $Q_j^r(t)$ *for the virtual queue lengths. Suppose, the initial conditions are such that*

$$r^{-1/2}\big[Q_j^r(0)/\alpha_j - f(r)\big] \to 0, \quad \forall j.$$

Then

$$\left\{r^{-1/2}\left[Q_j^r(t)/\alpha_j - f(r)\right],\ t \geq 0\right\} \Rightarrow 0, \quad \forall j, \tag{6.1}$$

where 0 *in the RHS signifies the identically zero function.*

Proof It will suffice to show that for any subsequence of indexes $\{r\}$, we can find a further subsequence such that the convergence (6.1) holds w.p.1, i.e.

$$\left\{r^{-1/2}\left[Q_j^r(t)/\alpha_j - f(r)\right],\ t \geq 0\right\} \to 0, \quad \text{u.o.c.} \tag{6.2}$$

Suppose, an arbitrary subsequence of $\{r\}$ is fixed. Assume that, for all r, the input processes A_i^r are constructed on a common probability space, as in (3.10), and then choose a subsequence of $\{r\}$, along which the property (3.11) holds. (We do not need to consider processes S_{ij}^r here, because they are irrelevant for the virtual queues' evolution.) Then, we can prove the following

Assertion 1 *There exists* $\delta > 0$ *such that the following holds w.p.*1. *Suppose an arbitrary sequence* $t^r \to t_0$ *is fixed, where* $t_0 \geq 0$ *is finite*; *and suppose an arbitrary non-negative sequence* $\tau^r \to \tau$ *is fixed, where* $\tau \geq 0$ *is finite. Denote*

$$\hat{q}_j^r(u) = r^{-1/2}\left[Q_j^r\left(t^r + r^{-1/2}u\right)/\alpha_j - f(r)\right], \quad u \geq 0, \tag{6.3}$$

and $\hat{q}^{r,*}(u) = \max_j \hat{q}_j^r(u)$. *Suppose*, $\lim_{r\to\infty} \hat{q}^{r,*}(0) = C$. ($C \geq 0$ *necessarily.*) *Then*,

$$\limsup_{r\to\infty} \hat{q}_j^r\left(\tau^r\right) \leq [C - \delta\tau] \vee 0. \tag{6.4}$$

It then follows that (6.4) *in fact holds uniformly on all values of* C *and all sequences* $\{t^r\}$, $\{\tau^r\}$ *confined to a compact set.*

The proof of Assertion 1 is obtained by looking at the u.o.c. limits of the sequence of trajectories $(\hat{q}^r(u),\ u \geq 0)$. Any such limit $(\hat{q}(u),\ u \geq 0)$ is a *fluid sample path* (FSP). Any FSP is Lipschitz continuous and is such that $\hat{q}^*(u) \equiv \max_j \hat{q}_j(u)$ is (Lipschitz) non-negative; moreover, at any point u where $\hat{q}^*(u) > 0$ and derivative $(d/du)\hat{q}^*(u)$ exists

$$(d/du)\hat{q}^*(u) \leq -\delta, \tag{6.5}$$

for some $\delta > 0$ dependent only on the parameters λ_i, μ_{ij}, and the solution to the SPP. Assertion 1 then follows from these FSP properties, in particular the key property (6.5). In turn, the FSP properties are established analogously to (and in fact much easier than) similar properties of FSPs in papers on "conventional" heavy traffic [27, 36]. In the rest of this paragraph, we describe the key intuition behind (6.5). (For a detailed proof of an analogous fact, cf. Sect. 8 in [27].) When r is large enough all pre-limit virtual queues Q_j^r are "almost exactly proportional" to the corresponding α_j (because $f(r)/\sqrt{r} \to \infty$). Then, all new arrivals are routed along "their" basic activities (because for any i the minimum of α_j/μ_{ij} is attained on the basic activities

and only on them); this means that the average rate at which workload arrives in the queues is constant and equal to the average rate it is removed (served) from the queues. If $\hat{q}_j^r(u)$ for all j happen to be equal "for some time" (which is impossible for pre-limit systems, but is possible for an FSP limit), then the average rates at which workloads arrive into all queues are equal as well (because the service rates of all virtual queues are equal)—let us denote this "nominal" rate γ, which is also the workload service rate from each queue. If maximum of $\hat{q}_j^r(u)$ is attained on the unique queue m, then, as long as this condition holds, the only types i that can be routed into queue m are those for which (im) is the *sole* basic activity; therefore the average rate at which workload arrives into m is *strictly less* than γ, which implies that the queue m is drained at non-zero rate. If maximum of $\hat{q}_j^r(u)$ is attained on more than one queue, the argument is similar but slightly more general.

Given Assertion 1, we can easily show that w.p.1 for any $\epsilon > 0$, and any $T > 0$ the trajectory of

$$\max_j r^{-1/2}\big[Q_j^r(t)/\alpha_j - f(r)\big]$$

can never be above ϵ within $[0, T]$ for all sufficiently large r: otherwise we can construct a contradiction to Assertion 1. Then, $\min_j r^{-1/2}[Q_j^r(t)/\alpha_j - f(r)]$ cannot be below $-\epsilon$ for all large r either, because otherwise (5.2), which is $|\sum_j Q_j^r(t) - f(r)| \le Jc$, would be violated. This proves (6.2) and the theorem. □

Suppose $x = \{x_i,\ i \in \mathcal{I}\}$ is a vector of real numbers. Consider a vector $H(x) = v = \{v_{ij},\ (ij) \in \mathcal{E}_b\}$ such that the following conditions hold:

$$\sum_i v_{ij} \frac{1}{\mu_{ij}\beta_j} \quad \text{are equal for all } j, \tag{6.6}$$

$$\sum_j v_{ij} = x_i, \quad \forall i. \tag{6.7}$$

Lemma 6.2 *Mapping $H(x)$ is a well-defined linear mapping, i.e.* (6.6)–(6.7) *have unique solution v.*

Before we proceed with the proof, let us explain the meaning of operator H. Suppose the components x_i of vector x have the meaning of input rates (or, alternatively, "amounts of customers") of the flows $i \in \mathcal{I}$. (We emphasize, however, that H is defined on vectors x with arbitrary *real* components, *not* necessarily positive or non-negative.) Suppose, the question is: how do we "split" each rate x_i into the input rates v_{ij} to the flow i basic activities j so that the utilizations of all server pools are exactly same, i.e. "perfectly balanced"? (So, this is same problem as (2.2)–(2.5), except we a priori restrict solution to basic activities only.) Obviously, such perfect load balancing may not be possible for some x, if we would require that all v_{ij} are non-negative. Suppose, however, that v_{ij}'s can be any real numbers, as long as the "flow conservation" laws (6.7) holds. Then, the "perfect load-balancing split" is unique and is given by $v = H(x)$. Again, as we already emphasized, not only the components v_{ij} of the image can be negative, but the components x_i are allowed to be negative as well; this

may be the case when x_i are not the absolute flow rates, but rather deviations of input rates from some "nominal" values.

Proof of Lemma 6.2 Let us show uniqueness first. Denote

$$y = \sum_i v_{ij} \frac{1}{\mu_{ij}\beta_j}, \tag{6.8}$$

the value is same for all j. Recall from Sect. 2 that, for each flow i, the customer workload contribution $v_i = \alpha_j \frac{1}{\mu_{ij}\beta_j} > 0$ for any basic activities (ij) (for this flow). Then,

$$y = \sum_j \alpha_j y = \sum_j \alpha_j \sum_i v_{ij} \frac{1}{\mu_{ij}\beta_j} = \sum_i x_i v_i. \tag{6.9}$$

Therefore, the value of y is uniquely determined by x. But, given y, we can uniquely determine all v_{ij}. Indeed, if for a given i, the values of v_{ij} are known for all basic activities (ij), except (im), then v_{im} is uniquely determined from condition (6.7); if for a given j, the values of v_{ij} are known for all basic activities (ij), except (mj), then v_{mj} is uniquely determined from (6.8). Since basic activities form a tree, we can uniquely recover all v_{ij}.

The existence essentially follows from the same constructive argument. For given x we set $y = \sum_i x_i v_i$. Then, using tree structure, we determine all v_{ij} one by one. It remains to show that this procedure is well defined: the value v_{ij} for the "last" edge (ij) is same whether it is derived from (6.7) for i or from (6.8) for j, and conditions (6.6)–(6.7) indeed hold. This easily follows from the relation $v_i = \alpha_j \frac{1}{\mu_{ij}\beta_j}$ for each basic activity. □

For the solution $\{\lambda_{ij}, (ij)\}$ to the static planning problem (SPP), we clearly have $\{\lambda_{ij},\ (ij) \in \mathcal{E}_b\} = H(\{\lambda_i,\ i \in \mathcal{I}\})$, because it equalizes server pool loads. Recall that $\lambda_{ij} = 0$ for non-basic (ij).

For each r define $\{\lambda^r_{ij},\ (ij) \in \mathcal{E}_b\} = H(\{\lambda^r_i,\ i \in \mathcal{I}\})$, and set, by convention, $\lambda^r_{ij} = 0$ for non-basic (ij)'s. (Equivalently, $\{\lambda^r_{ij}\}$ is the solution to problem (2.2)–(2.5) with $\{\lambda_i\}$ replaced by $\{\lambda^r_i\}$.) It is easy to verify that, for basic activities (ij),

$$\lambda^r_{ij} = r\lambda_{ij} + \ell_{ij}\sqrt{r} + o\big(\sqrt{r}\big),$$

where, in turn, $\{\ell_{ij},\ (ij) \in \mathcal{E}_b\} = H(\{\ell_i,\ i \in \mathcal{I}\})$. Note that for all sufficiently large r, $\lambda^r_{ij} > 0$ for all basic activities (ij).

Remark 2 If λ^r_{ij} were the rates at which type i customers would be routed to different pools j (e.g., by choosing a pool j with probability $\lambda^r_{ij}/\lambda^r_i$ independently for each type i arrival), then the loads of all pools would be same (by definition of $\{\lambda^r_{ij}\}$) and equal to

$$\sum_i \lambda^r_{ij}/(\mu_{ij}\beta_j r) = \rho + \big(1/\sqrt{r}\big)\sum_i \ell_i v_i + o\big(1/\sqrt{r}\big). \tag{6.10}$$

(To see this, it suffices to sum up the left-hand sides of (6.10), weighted by α_j for each j, and manipulate the sum similarly to (6.9).) In particular, since $\rho = 1$ by CRP, condition (2.10) guarantees that there exists a control strategy stabilizing the system for all large r.

For each r and each flow i, denote by $A_i^r(t)$ an independent Poisson arrival process, of rate λ_i^r. We have a functional central limit theorem (FCLT):

$$\left\{\frac{1}{\sqrt{r}}\big[A_i^r(t) - \lambda_i^r t\big],\ t \geq 0\right\} \Rightarrow \big\{B_i(t),\ t \geq 0\big\} = B_i, \tag{6.11}$$

where B_i is a zero-drift one-dimensional Brownian motion with variance λ_i. Denote $B = \{B_i, i \in \mathcal{I}\}$ the corresponding multi-dimensional Brownian motion with independent components.

Theorem 6.3 *Assume conditions of Theorem 6.1 hold. Denote $A_{ij}^r(t)$ the number of class i customers routed to virtual queue j by time t. Denote*

$$W_{ij}^r(t) \doteq \frac{1}{\sqrt{r}}\big[A_{ij}^r(t) - \lambda_{ij}^r t\big].$$

Then,

$$\big\{\big(W_{ij}^r(t),\ t \geq 0\big),\ (ij)\big\} \Rightarrow \big\{\big(W_{ij}(t),\ t \geq 0\big),\ (ij)\big\}, \tag{6.12}$$

where

$$\big\{W_{ij}(t),\ (ij) \in \mathcal{E}_b\big\} = H\big(B(t)\big),$$

B is the Brownian motion defined earlier, and $W_{ij}(t) \equiv 0$ for $(ij) \notin \mathcal{E}_b$. (Thus, W is also a Brownian motion, but with correlated components.)

Proof Using the Skorohod representation, we can WLOG assume that for each i, all input processes A_i^r, for all r, are constructed on a common probability space so that the convergence (6.11) holds uniformly on compact sets (u.o.c.) w.p.1; and this common probability space is such that processes A_i^r with different i are independent. So, the rest of the argument is sample-path based. Consider a finite interval $[0, t]$. By Theorem 6.1, w.p.1, for all sufficiently large r, the virtual queues in the entire interval $[0, t]$ are strictly positive, and moreover each difference $Q_j^r(t) - \alpha_j f(r)$ is $o(\sqrt{r})$ uniformly in $[0, t]$. (This immediately implies that in $[0, t]$, for all large r, no customer will be routed along any non-basic activity, and so $A_{ij}^r(t) \equiv 0$ for each non-basic (ij). In the rest of the proof we consider only basic (ij).) This means that in $[0, t]$ the "virtual service" operation (5.1) serves exactly equal amount of work from each queue and, consequently, the amounts of work arrived ("routed") to each queue j are equal up to $o(\sqrt{r})$. In other words, we see that

$$\sum_i A_{ij}^r(t)\frac{1}{\mu_{ij}\beta_j} = \epsilon_1^{(r)} + o\big(\sqrt{r}\big) \quad \text{for some } \epsilon_1^{(r)}, \text{ for all } j, \tag{6.13}$$

and the flow conservation holds:

$$\sum_j A_{ij}^r(t) = A_i^r(t), \quad \forall i. \tag{6.14}$$

Now, using the definition of λ_{ij}^r, convergence (6.11), and rescaling by $1/\sqrt{r}$, we can rewrite (6.13)–(6.14) as:

$$\sum_i W_{ij}^r(t)\frac{1}{\mu_{ij}\beta_j} = \epsilon_2^{(r)} + o(1) \quad \text{for some } \epsilon_2^{(r)}, \text{ for all } j, \tag{6.15}$$

$$\sum_j W_{ij}^r(t) = B_i(t) + o(1), \quad \forall i. \tag{6.16}$$

(The $o(1)$ terms go to 0 uniformly on finite time intervals.) For each r and t, we can always slightly change $W_{ij}^r(t)$'s to obtain values $\tilde{W}_{ij}^r(t)$ such that

$$\begin{aligned} &\big|\tilde{W}_{ij}^r(t) - W_{ij}^r(t)\big| = o(1), \quad \forall (ij), \\ &\sum_i \tilde{W}_{ij}^r(t)\frac{1}{\mu_{ij}\beta_j} = \epsilon_3^{(r)} \quad \text{for some } \epsilon_3^{(r)}, \text{ for all } j, \end{aligned} \tag{6.17}$$

and then

$$\sum_j W_{ij}^r(t) = B_i(t) + o(1), \quad \forall i.$$

Therefore (slightly abusing notation),

$$\big\{\tilde{W}_{ij}^r(t),\ (ij) \in \mathcal{E}_b\big\} = H\big(B(t) + o(1)\big).$$

Since mapping H is linear (and then continuous), we see that the finite limits $W_{ij}(t)$ of $\tilde{W}_{ij}^r(t)$ exist and are given by $H(B(t))$; by (6.17), $\{W_{ij}(t)\}$ is also the u.o.c. limit of $\{W_{ij}^r(t)\}$. □

Theorem 6.4 *Suppose all parameters μ_{ij}, β_j, and c are rational. For each r consider a stationary version of the virtual queueing process. (Such always exists—see discussion following the definition of Shadow algorithm in Sect. 5.) Then, as $r \to \infty$,*

$$\big\{r^{-1/2}\big[Q_j^r(0)/\alpha_j - f(r)\big],\ j \in \mathcal{J}\big\} \xrightarrow{P} 0. \tag{6.18}$$

Consequently, the convergence (6.12) holds.

Proof Only (6.18) needs to be proved. Let us consider the discrete-time (sampled) Markov chain $(\{\hat{q}_j^r(u),\ j \in \mathcal{J}\},\ u = 0, 1, 2, \ldots)$, where

$$\hat{q}_j^r(u) = r^{-1/2}\big[Q_j^r\big(r^{-1/2}u\big)/\alpha_j - f(r)\big],$$

and denote $\hat{q}^{r,*}(u) = \max_j \hat{q}_j^r(u)$. (The definition of $\hat{q}_j^r(u)$ is same as in (6.3), but specialized to the case $t^r \equiv 0$.) Clearly, the stationary distribution of this chain is

same as that of the continuous-time chain $(\{r^{-1/2}[Q_j^r(t)/\alpha_j - f(r)],\ j \in \mathcal{J},\ t \geq 0)$, we are interested in. The following, Assertion 2, holds.

Assertion 2 *There exists $\delta > 0$ (same as in Assertion 1) and $\epsilon > 0$ such that, uniformly for all large r, and uniformly on all possible states of the chain at time u*:

$$E\left\{\left[\hat{q}^{r,*}(u+1)\right]^2 - \left[\hat{q}^{r,*}(u)\right]^2 \mid \hat{q}^{r,*}(u)\right\} \leq -\delta \hat{q}^{r,*}(u) + \epsilon. \tag{6.19}$$

Assertion 2 is proved using FSPs, analogously to their use in the proof of Assertion 1 (in the proof of Theorem 6.1). More precisely, FSPs are used to establish (uniformly on states at time u)

$$P\left\{\hat{q}^{r,*}(u+1) - \hat{q}^{r,*}(u) \leq -\delta \hat{q}^{r,*}(u) \mid \hat{q}^{r,*}(u)\right\} \to 1, \quad r \to \infty.$$

This, along with the easily obtained uniform upper bound on $[\hat{q}^{r,*}(u+1) - \hat{q}^{r,*}(u)]^2$, implies (6.19).

Assuming the chain is in stationary regime, and taking expectation of both parts of (6.19) (note that for each r, all involved random variables are uniformly bounded), we obtain $0 \leq -\delta E[\hat{q}^{r,*}(u)] + \epsilon$, and then

$$E\left[\hat{q}^{r,*}(u)\right] \leq \epsilon/\delta. \tag{6.20}$$

Since $E[\hat{q}^{r,*}(u)]$ remains finite as $r \to \infty$, the sequence of stationary distributions of $\hat{q}^{r,*}(u)$ is tight; in other words, for any $\epsilon_1 > 0$ we can pick large enough $C > 0$ such that

$$P\left[\hat{q}^{r,*}(u) \leq C\right] > 1 - \epsilon_1. \tag{6.21}$$

Now, if we fix arbitrary $u' > C/\delta$, then using (6.21) and Assertion 1 we can establish the following

Assertion 3 *For an arbitrary $\epsilon_2 > 0$, for all sufficiently large r and uniformly on $\hat{q}^{r,*}(u) \leq C$*:

$$E\left[\hat{q}^{r,*}\left(u+u'\right) \mid \hat{q}^{r,*}(u)\right] \leq \epsilon_2.$$

Combining (6.21) and Assertion 3, we see that in a stationary regime, $P[\hat{q}^{r,*}(u) > \epsilon_2] < \epsilon_1$ for arbitrary positive ϵ_1 and ϵ_2. This of course means that the stationary distribution of $\hat{q}^{r,*}(u) \vee 0$ convergence to the Dirac measure concentrated at 0. From here, we immediately see that the stationary distribution of $[\min_j \hat{q}_j^r(u)] \wedge 0$ must also converge to the Dirac measure at 0. (Otherwise, the property (5.2), $|\sum_j Q_j^r(t) - f(r)| \leq Jc$, which holds always, would be violated.) The proof is complete. □

7 Conjecture on the order-optimality of shadow routing scheme

Theorem 6.3 shows that, if CRP and (2.10) conditions hold, then for large r, the average load of each server pool is by $O(\sqrt{r})$ less than its capacity (see the definition

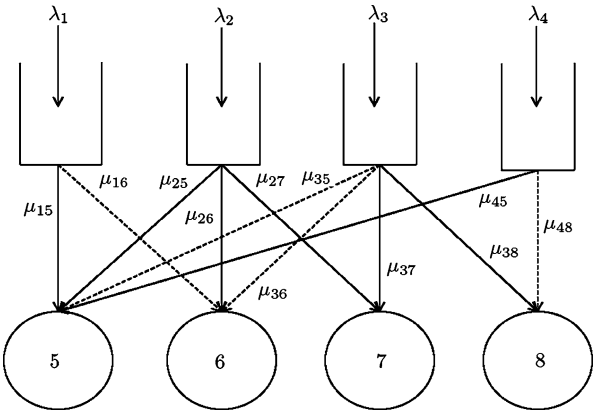


Fig. 3 Simulated 4 × 4 system. *Dashed lines* are feasible non-basic activities

of rates λ_{ij}^r in Sect. 6), and the fluctuations of the arrival process around its mean are of the order $O(\sqrt{r})$ and in fact asymptotically described by a Brownian motion. This naturally suggests the following conjecture.

Conjecture 7.1 *Assume CRP conditions and* (2.10) *hold. Consider the Shadow routing algorithm with parameter* η *depending on* r *as specified in Sect.* 6 (*for example* $\eta = r^{-3/4}$), *and assume that actual routing follows Shadow algorithm assignments "literally." Then, there exist work-conserving, non-pre-emptive scheduling disciplines within each server pool, such that the entire control strategy is order-optimal.*

Even though, given Theorem 6.3, Conjecture 7.1 seems very natural indeed, this fact is by no means "automatic". We plan to address this conjecture in future work.

8 Simulation experiments

8.1 Performance of shadow routing algorithm in a 4 × 4 system

In this section we present the simulation results of the model used in the proof of Theorem 3.1 and depicted in Fig. 3, with four feasible (i.e., those with $\mu_{ij} > 0$) non-basic activities. This system consists of four customer classes and four server pools. The service rates are $\mu_{15} = 4$, $\mu_{16} = 5$, $\mu_{25} = 2$, $\mu_{26} = 4$, $\mu_{27} = 1$, $\mu_{36} = 2$, $\mu_{37} = 1$, $\mu_{38} = 1$, $\mu_{35} = 1.5$, $\mu_{45} = 4$, $\mu_{48} = 1.5$ and all other μ_{ij}'s are zero (for non-feasible activities); $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 1$. The input rate parameters are $\lambda_1 = 2$, $\lambda_2 = 4.8$, $\lambda_3 = 1.6$, and $\lambda_4 = 1.2$, which makes all feasible activities basic, except activities $(3, 5)$, $(3, 6)$, $(1, 6)$, and $(4, 8)$. We use the scaling parameter $r = 100$, so that there are 100 servers in each pool. The actual arrival rates are $(190, 456, 152, 114)$, so that the actual system load is 95%. Our goal is two fold; we would like to validate the performance of the Shadow algorithm in a relatively complicated system and illustrate the effect of parameter η. As discussed in Sect. 5, *after routing is performed according to Shadow algorithm*, for the customers waiting for service in their assigned pools, any non-idling policy would keep the system stable. Then if, for instance, the goal is

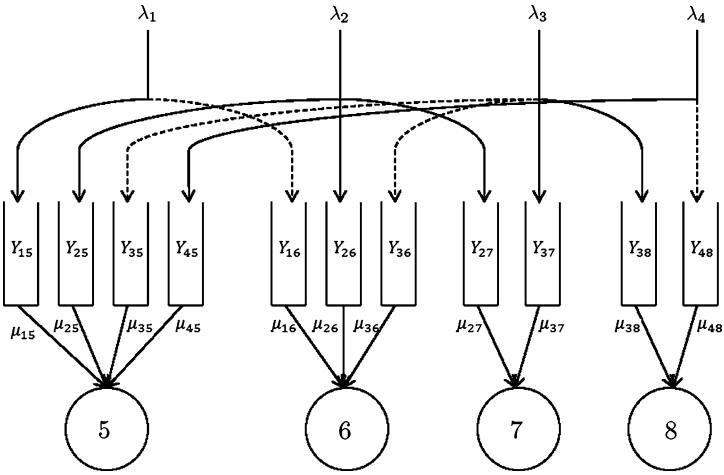


Fig. 4 Simulated 4×4 system under Shadow algorithm

Table 1 Simulation results under the Shadow algorithm

η	Y_{15}	Y_{25}	Y_{26}	Y_{27}	Y_{37}	Y_{38}	Y_{45}	Y_{36}	Y_{16}	Y_{48}	Y_{35}	Total
200^{-1}	3.498	0.605	5.962	1.46	2.525	2.911	2.092	0	0	0	0	19.053
100^{-1}	3.871	0.699	6.154	1.123	2.02	3.13	2.298	0	0	0.006	0	19.297
50^{-1}	4.008	0.774	6.343	1.087	2.139	2.971	2.323	0.001	0	0.083	0	19.729
25^{-1}	4.385	0.947	7.92	1.63	3.142	3.541	2.368	0.042	0.01	0.368	0.005	24.514

to minimize linear holding costs, a strict priority rule may be applied, giving priority to the fastest service customers. For our simulations, we use the following rule in conjunction with Shadow routing. (This is a version of MaxWeight, but applied only to scheduling *within each server pool*. It does not give explicit a priori preference to any class.) Let Y_{ij} denote the number of class i customers waiting for service in pool j. In Fig. 4, we illustrate how the 4×4 system looks like under Shadow routing algorithm (compare to Fig. 3). Customers are routed to one of the servers at the time of their arrival. When a server in pool j becomes available, it picks for service a customer of class i for which $Y_{ij}\mu_{ij}$ is the largest. Table 1 summarizes our results. Column labeled Y_{ij} is the *average* value of Y_{ij}. We simulate the system for 1200 time units (around 1.1 millions arrivals) and we use the first 120 time units as the warm-up period. All the simulations are run using common realizations of input processes, so a fair comparison is possible from our results.

Besides the average queue lengths displayed in Table 1, we present the number of customers routed to non-basic activities next. No customers were routed through non-basic activities when $\eta = 200^{-1}$, after the warm-up period. Even during the warm-up period, the number is less than 10 compared to around 100,000 arriving customers in that period. When $\eta = 100^{-1}$, 0.3% of class 4 customers are sent to pool 8, during the warm-up period, but other non-basic activities are never used. When $\eta = 50^{-1}$, around 3% of arriving class 4 customers are routed to server pool 8, less than 0.1% of

class 3 customers are routed to pool 6, and other non-basic activities are almost never used. When $\eta = 25^{-1}$, 10% of arriving class 4 customers are routed to server pool 8 and 1.5% of arriving class 3 customers are routed to pool 6, 4% and 0.3% of arriving class 3 customers are routed to activities 6 and 5 respectively and less than 0.2% of class 1 customers are routed to pool 6. Although these numbers might seem small, they have a degrading effect on performance. In fact, when $\eta = 10^{-1}$, the system becomes unstable.

The key conclusion is this. The Shadow algorithm does a very good job making sure the arriving customers are routed along basic activities only (again, without knowing a priori which activities are basic) and balancing server pool loads, thus keeping system stable and ensuring short queues. The smaller the parameter η the more accurate the routing, because a larger value $1/\eta$ (which is the total length of all virtual queues) results in smaller variations of the *ratios* of the virtual queues—and those ratios determine routing decisions. However, for practical applications it would be incorrect to make parameter η "too small", because this would increase the typical time for the virtual routing mechanism to adapt to changes in input rates. This—robustness—issue is very important for applications. (We discuss it in Sect. 8.2.) For the system in this section, we can conclude that a "good" value of η is about $1/100$.

8.2 Robustness to changes of input rates

How a system responds to arrival rate changes is an important robustness issue. (Cf. [29] for further motivation.) In the simulation experiment here we would like to show the effect on the Shadow routing of setting η at different values when the arrival rates change. We simulate the X-system on Fig. 2, with the following parameters: $\mu_{11} = \mu_{22} = \mu_{12} = 4$, and $\mu_{21} = 1$. The number of servers in each pool is 100. For 300 time units we set the arrival rates to $\lambda_1^r = 370$ and $\lambda_2^r = 410$. At time 300 arrival rates are switched to $\lambda_1^r = 410$ and $\lambda_2^r = 370$. Figure 5 shows the evolution of virtual queue 1, under two different values $\eta = 50^{-1}$ and $\eta = 250^{-1}$. It is clear that when $\eta = 50^{-1}$ the system reaches the new "right" values of the virtual queue a lot quicker than when $\eta = 250^{-1}$. In fact, the transition time is almost exactly 5 times (the factor of η increase) shorter. The latter fact can be rigorously substantiated by looking at the appropriate fluid limit of virtual queues as $\eta \to 0$; the trajectories under $\eta = 50^{-1}$ and $\eta = 250^{-1}$ are close to the same—but differently scaled—fluid limit.

8.3 Comparison to MaxWeight algorithm. Illustration of Theorem 4.1

In this section we compare Shadow Routing to MaxWeight–FSF algorithm. The simulation experiments here also nicely illustrate the key ideas behind the proof of Theorem 4.1 in Sect. 10. We simulate the X-system on Fig. 2 with the following parameters. (The same parameters are used for the system in the proof in Sect. 10.) We set $\mu_{11} = 4, \mu_{12} = 4.1$, $\mu_{21} = 1, \mu_{22} = 1.5$, $\lambda_1 = \lambda_2 = 2$. With these parameters, the only non-basic activity is $(1, 2)$. The scaling parameter is $r = 100$, which means $N_1^r = N_2^r = 100$; input rates $\lambda_1^r = \lambda_2^r = 194$, which makes the system load to be 97%. We would like to observe how often the non-basic activity $(1, 2)$ is used and compare the performance of MaxWeight–FSF with our Shadow algorithm. (As in Sect. 8.1, we

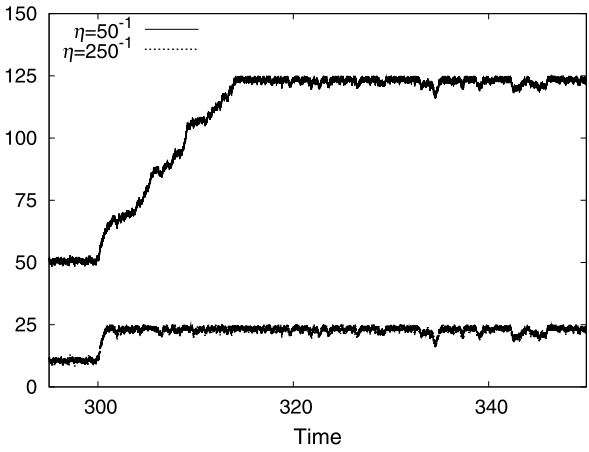


Fig. 5 A virtual queue transition after input rates shift

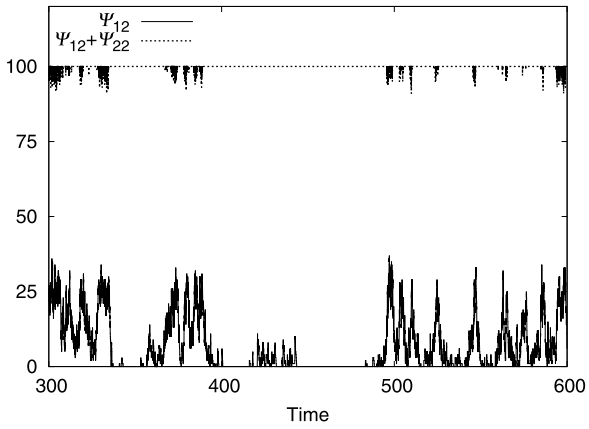


Fig. 6 Occupancy levels in server pool 2

use Shadow routing in conjunction with MaxWeight scheduling *within each server pool, after the routing to the pools is done*.) We simulate the system for 3300 time units, around 1.3 million arrivals, and we use the first 300 time units as the warm-up period. We only collect statistics after the warm-up period. We initially set the system to be empty.

Now we can illustrate the idea behind our proof of Theorem 4.1 in this simulation. In Fig. 6, we display the total number of busy servers in pool 2 and number of class 1 customers being served in pool 2 versus time from time 300 to 600. The plot on top ($\Psi_{12} + \Psi_{22}$) shows the number of busy servers in pool 2. Since servers in pool 2 are faster, under FSF, servers in this pool are busy almost all the time. The plot on the bottom is the number of class 1 customers being served in pool 2. As can be seen from the figure, Ψ_{12} increases when there are idle serves in the second pool. Moreover, it increases very fast and reaches very high levels. This illustrates our proof.

Table 2 summarizes our simulation results for the comparison of MaxWeight–FSF and Shadow algorithm. In Table 2, Columns "Class 1" and "Class 2" are the average number of customers in that class waiting in the queue. Column "Total" is for the

Table 2 Comparison of MaxWeight–FSF and the Shadow algorithm

Policy	Class 1	Class 2	Total	% Activity (1, 2) usage
MaxWeight–FSF	9.19	31.75	40.94	23%
Shadow Algorithm	15.26	17.28	32.54	0%

average total number of customers in system. Column "% Activity (1, 2) usage" is the percentage of class 1 customers that are served by a server in pool 2. We should emphasize that under the Shadow algorithm only one class 1 customer out of around 640,000 class 1 customers has been served by a server in pool 2 and that was very early in the simulation, before the warm-up period expired. This also illustrates that the Shadow algorithm is very fast in "finding" and "avoiding" non-basic activities.

9 Preliminary analysis of X-model (towards the proof of Theorem 4.1)

In this section we start working towards the proof of Theorem 4.1. We consider a sequence of X-models, see Fig. 2, in the Halfin–Whitt regime (of Sect. 2), and assume the following conditions:

$$\text{CRP and condition (2.10) hold,} \tag{9.1}$$

$$\mu_{12} > \mu_{11}, \quad \mu_{22} > \mu_{21}, \quad \mu_{11} > \mu_{21}, \tag{9.2}$$

$$(1,2) \text{ is the only non-basic activity, and then } \psi^*_{12} = 0. \tag{9.3}$$

It is easily seen that (9.3) implies

$$\mu_{11}\mu_{22} > \mu_{21}\mu_{12}.$$

For example, if we set $\lambda_1 = \lambda_2 = 2$, $\mu_{11} = 4$, $\mu_{12} = 4.1$, $\mu_{21} = 1$, $\mu_{22} = 1.5$, $\beta_1 = \beta_2 = 1$, then conditions (9.1)–(9.3) are satisfied, and the unique solution to the SPP is such that $\psi^*_{11} = 0.5$, $\psi^*_{12} = 0$, $\psi^*_{21} = 0.5$, and $\psi^*_{22} = 1$.

Some of the results in this section are not limited to the MaxWeight–FSF control policy, but they always assume non-pre-emptive service.

We first introduce some additional notation and terminology.

9.1 Preliminaries

Recall that $X^r_1(t)$ and $X^r_2(t)$ denote the number of class 1 and 2 customers in the system at time t, and they satisfy:

$$\begin{aligned} X^r_i(t) &= X^r_i(0) + A^r_i(t) - \sum_j S_{ij}\left(\int_0^t \Psi^r_{ij}(s)\,ds\right) \\ &= X^r_i(0) + \left(A^r_i(t) - \lambda^r_i t\right) - \sum_j \left(S_{ij}\left(\int_0^t \Psi^r_{ij}(s)\,ds\right) - \mu_{ij}\int_0^t \Psi^r_{ij}(s)\,ds\right) \end{aligned}$$

$$-\sum_j \mu_{ij} \int_0^t \bigl(\Psi_{ij}^r(s) - r\psi_{ij}^*\bigr)\,ds + \lambda_i^r t - \sum_j r\mu_{ij}\psi_{ij}^* t. \tag{9.4}$$

We use $A_{ij}^r(t)$ to denote the number of class i customers routed to and entered service at pool j by time t in the rth system. We set

$$T_{ij}^r(t) = \int_0^t \Psi_{ij}^r(s)\,ds.$$

Note that

$$\Psi_{ij}^r(t) = \Psi_{ij}^r(0) + A_{ij}^r(t) - S_{ij}\bigl(T_{ij}^r(t)\bigr).$$

We use $\mathbb{X}^r = (X^r, Y^r, \Psi^r, T^r)$ to denote the process describing system evolution, where $X^r = (X_1^r, X_2^r)$, $Y^r = (Y_1^r, Y_2^r)$, $\Psi^r = (\Psi_{ij}^r; i, j = 1, 2)$ and $T^r = (T_{ij}^r; i, j = 1, 2)$. We define the fluid scaling by

$$\bar{\mathbb{X}}^r(t) = r^{-1}\mathbb{X}^r(t).$$

From (9.4), for the diffusion scaled process $\hat{X}^r$ we have the following set of equations.

$$\hat{X}_i^r(t) = \hat{X}_i^r(0) + \ell_i^r t - \sum_{j=1,2} \mu_{ij} \int_0^t \hat{\Psi}_{ij}^r(s)\,ds + B_i^r(t), \tag{9.5}$$

where

$$\begin{aligned} \ell_i^r &\doteq \frac{\lambda_i^r - \lambda_i r}{\sqrt{r}}, \\ B_i^r(t) &= \hat{A}_i^r(t) - \sum_{j=1,2} \hat{D}_{ij}^r(t), \end{aligned} \tag{9.6}$$

and

$$\begin{aligned} \hat{A}_i^r(t) &= \sqrt{r}\left(\frac{A_i^r(t)}{r} - \lambda_i t\right), \\ \hat{S}_{ij}^r(t) &= \sqrt{r}\left(\frac{S_{ij}(rt)}{r} - \mu_{ij} t\right), \\ \hat{D}_{ij}^r(t) &= \hat{S}_{ij}^r\left(\int_0^t \bar{\Psi}_{ij}^r(s)\,ds\right). \end{aligned}$$

Obviously, by (2.1), we have $\ell_i^r \to \ell_i$ as $r \to \infty$. We also define

$$\hat{T}_{ij}^r(t) = \int_0^t \hat{\Psi}_{ij}^r(s)\,ds.$$

9.2 Analysis of Ψ_{12}^r

In the X-model we consider, the only non-basic activity is $(1,2)$ and it plays a crucial role in our analysis. This section is devoted to establishing some of its properties. We first show that if $\hat{\Psi}_{12}^r$ hits a certain level, it will stay above another level for a while, because, essentially, it can decrease at most exponentially fast.

Lemma 9.1 *Consider a sequence of X-models in the Halfin–Whitt regime, operating under a non-pre-emptive rule, and satisfying (9.1)–(9.3). Assume that as* $r \to \infty$

$$\liminf_{r\to\infty} P\Big\{ \sup_{0\le t\le T} \hat{\Psi}_{12}^r(t) > \epsilon_1 \Big\} \ge \epsilon_2,$$

for some $\epsilon_1, \epsilon_2 > 0$ *and* $T > 0$. *Then,*

$$\liminf_{r\to\infty} P\big\{\hat{\Psi}_{12}^r(T) > (\epsilon_1/2)\exp\{-\mu_{12}T\}\big\} \ge \epsilon_2.$$

Proof Consider the process $\hat{\Psi}_{12}^r(t)$ restarted at the first time t_1 when $\hat{\Psi}_{12}^r(t) > \epsilon_1$. Also, from time t_1, we only consider $\epsilon_1\sqrt{r}$ customers present at time t_1. Let $h^r(s)$, $s \ge 0$, denote the number of customers from this subset still present at time $t_1 + s$. Then,

$$\begin{aligned}\frac{h^r(t)}{\sqrt{r}} &= \frac{h^r(0)}{\sqrt{r}} - \frac{1}{\sqrt{r}} S_h\Big(\mu_{12}\int_0^t h^r(s)\,ds\Big)\\ &= \epsilon_1 - \mu_{12}\int_0^t \frac{h^r(s)}{\sqrt{r}}\,ds - \sqrt{r}\Big(S_h\Big(r\mu_{12}\int_0^t \frac{h^r(s)}{r}\,ds\Big) - \mu_{12}\int_0^t \frac{h^r(s)}{r}\,ds\Big),\end{aligned}$$

where the first equality is in the sense of "equal in distribution" and S_h is a Poisson process with rate 1. Since $h^r(t)/r \le \epsilon_1/\sqrt{r}$, we easily see that, as $r \to \infty$, this process converges to a deterministic (fluid) limit $h(t) = \epsilon_1 \exp\{-\mu_{12}t\}$. The result follows. □

Next we prove some properties of the sequences $\{\hat{\Psi}_{12}^r\}$ and $\{\hat{Z}_j^r\}$.

Proposition 9.2 *Consider a sequence of stationary X-models in the Halfin–Whitt regime operating under a non-pre-emptive rule that satisfies (9.1)–(9.3). Let* $T > 0$ *be fixed. Then there exists* $C > 0$ *such that*

$$E\Big[\int_0^T \hat{\Psi}_{12}^r(s)\,ds\Big] < C \quad \textit{and} \tag{9.7}$$

$$E\Big[\int_0^T \hat{Z}_j^r(s)\,ds\Big] < C. \tag{9.8}$$

Consequently, for $M > 0$

$$P\Big(\int_0^T \hat{\Psi}_{12}^r(s)\,ds > M, \hat{\Psi}_{12}^r(0) > M\Big) < C/M, \quad \textit{and} \tag{9.9}$$

$$P\left(\int_0^T \hat{Z}_j^r(s)\,ds > M, \hat{Z}_j^r(0) > M\right) < C/M. \tag{9.10}$$

Also,

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left\{\sup_{0\le t\le T} \hat{\Psi}_{12}^r(t) > M\right\} = 0. \tag{9.11}$$

Proof Properties (9.7) and (9.8) follow from Lemma 3.3 and Fubini's theorem, and then (9.9) and (9.10) follow from Markov inequality. It remains to prove (9.11). For $M > 0$, let

$$\tau^r(M) = \inf\left\{t : \hat{\Psi}_{12}^r(t) > 2M\exp\{\mu_{12}T\}\right\} \wedge T.$$

Then, by Lemma 9.1 and (9.9) for r large enough

$$P\left\{\tau^r(M) < T\right\} < C/M,$$

which gives (9.11). □

9.3 Fluid limit of stationary system

Here we establish that under MaxWeight–FSF, the *fluid limit* of the stationary system must be well behaved if it is order optimal. Specifically, we assume that

$$\limsup_{r\to\infty} E\left[\left|\hat{Y}^r(\infty)\right|\right] < \infty. \tag{9.12}$$

Proposition 9.3 *Consider a sequence of stationary X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies* (9.1)–(9.3). *Assume that* (9.12) *holds. Then for any* $T > 0$, *there exists a subsequence* r_k *of* r *such that, almost surely as* $k \to \infty$

$$\sup_{0\le t\le T} \left\|\bar{Y}^{r_k}(t)\right\| \to 0 \quad \textit{and} \tag{9.13}$$

$$\sup_{0\le t\le T} \left|\bar{\Psi}_{ij}^{r_k}(t) - \psi_{ij}^*\right| \to 0 \tag{9.14}$$

for $i = 1, 2$ *and* $j = 1, 2$.

Proof First of all, we note that all the (fluid) scaled processes $\bar{Y}^r(\cdot)$, $\bar{X}^r(\cdot)$, $\bar{\Psi}^r(\cdot)$, $\bar{Z}^r(\cdot)$, $\bar{A}_{ij}^r(\cdot)$, are easily shown to be asymptotically Lipschitz, in the sense that for some universal $C > 0$, any $t_1 \le t_2$, and any $\epsilon > 0$, we have, for example for $\bar{X}^r(\cdot)$,

$$P\left\{\left\|\bar{X}^r(t_2) - \bar{X}^r(t_1)\right\| < C(t_2 - t_1) + \epsilon\right\} \to 1, \quad \text{as } r \to \infty,$$

and similarly for other processes. This in turn implies that the sequence of processes $\{\bar{X}^r, \bar{\Psi}^r, \bar{Y}^r, \bar{Z}^r\}$ is tight, and moreover, any of its weak limits $\{\bar{X}, \bar{\Psi}, \bar{Y}, \bar{Z}\}$ (along

a subsequence of r) has Lipschitz continuous trajectories. Then, from (9.12) and Proposition 9.2, any weak limit $\{\bar{X}, \bar{\Psi}, \bar{Y}, \bar{Z}\}$ must be such that

$$\bar{Y}(t) \equiv 0, \qquad \bar{Z}(t) \equiv 0, \qquad \bar{\Psi}_{12}(t) \equiv 0.$$

So, we can find a subsequence r_k of r such that (9.13), as well as (9.14) for $(i, j) = (1, 2)$, hold. We drop k from our notation for simplicity.

For the prelimit processes we have (for $i = 1, 2$) by (9.5):

$$\bar{X}_i^r(t) = \bar{X}_i^r(0) + \bar{A}_i^r(t) - \sum_{j=1,2} \mu_{ij} \int_0^t \bar{\Psi}_{ij}^r(s)\, ds + \sum_{j=1,2} \bar{S}_{ij}^r(t),$$

where

$$\bar{S}_{ij}^r(t) = r^{-1} S_{ij}\left(r \int_0^t \bar{\Psi}_{ij}^r(s)\, ds\right) - \mu_{ij} \int_0^t \bar{\Psi}_{ij}^r(s)\, ds.$$

Taking the usual (fluid) limit on $r \to \infty$, using the functional strong law of large numbers for the input and service processes, we obtain the following a.s. relations for the limit process:

$$\bar{X}_i(t) = \bar{X}_i(0) + \lambda_i t - \sum_{j=1,2} \mu_{ij} \int_0^t \bar{\Psi}_{ij}(s)\, ds,$$

$$\bar{\Psi}_{12}(t) = 0,$$

$$\bar{X}_i(t) = \sum_j \bar{\Psi}_{ij}(t),$$

$$\sum_i \bar{\Psi}_{ij}(t) = 1.$$

Note also that the limit process is stationary because each pre-limit process is stationary. From here we immediately have $\bar{\Psi}_{22}(t) \equiv 1 = \psi_{22}^*$. Also,

$$\bar{X}_1(t) = \bar{X}_i(0) + \lambda_1 t - \mu_{11} \int_0^t \bar{X}_1(s)\, ds.$$

Given that $\bar{X}_1(0) \le 1$, the sample paths of $\bar{X}_1(\cdot)$ uniformly converge to ψ_{11}^*. But, $\bar{X}_1(\cdot)$ is stationary. Then, we must have $\bar{X}_1(t) \equiv \bar{\Psi}_{11}(t) \equiv \psi_{11}^*$ a.s. Finally, $\bar{\Psi}_{21}(t) \equiv 1 - \bar{\Psi}_{11}(t) = 1 - \psi_{11}^* = \psi_{21}^*$. The convergence in probability in (9.13) and (9.14) follows; we can choose a further subsequence to obtain a.s. convergence. □

9.4 State space collapse results for X-systems under MaxWeight–FSF

In this section we prove a state space collapse (SSC) result for X-systems under MaxWeight–FSF.

Lemma 9.4 (SSC in X-model) *Consider a sequence of X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies* (9.1)–(9.3). *Assume that for* $T > 0$ *as* $r \to \infty$

$$\sup_{0\le t\le T} \left|\bar{\Psi}^r_{ij}(t) - \psi^*_{ij}\right| \to 0, \tag{9.15}$$

$$\sup_{0\le t\le T} \left\|\bar{Y}^r(t)\right\| \to 0, \quad \textit{and} \tag{9.16}$$

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left(\left\|\hat{\Psi}^r(0)\right\| > M, \left\|\hat{Y}^r(0)\right\| > M\right) = 0. \tag{9.17}$$

Then, for any $0 < s < T$

$$\sup_{s\le t\le T} \left|\mu_{11}\hat{Y}^r_1(t) - \mu_{22}\hat{Y}^r_2(t)\right| \Rightarrow 0 \quad \textit{and} \tag{9.18}$$

$$\sup_{s\le t\le T} \left|\hat{\Psi}^r_{12}(t) + \hat{\Psi}^r_{22}(t)\right| \Rightarrow 0, \tag{9.19}$$

as $r \to \infty$. *If in addition*

$$\left|\mu_{11}\hat{Y}^r_1(0) - \mu_{21}\hat{Y}^r_2(0)\right| \Rightarrow 0 \tag{9.20}$$

and

$$\left|\hat{\Psi}^r_{12}(0) + \hat{\Psi}^r_{22}(0)\right| \Rightarrow 0,$$

as $r \to \infty$, *then* (9.18) *and* (9.19) *hold for* $s = 0$.

Proof The idea of the proof is quite standard for SSC-type results: if we look at the evolution of the process on a finer ("local-fluid") time scale, in our case—over $O(1/\sqrt{r})$-long intervals, then we can observe that $|\mu_{11}\hat{Y}^r_1(t) > \mu_{21}\hat{Y}^r_2(t)|$ has a negative drift when it deviates from 0.

Assume that (9.15)–(9.17) hold for some $T > 0$. Let us fix $\delta > 0$ and for each r cover $[0, T]$ interval with $K = K(r) = \lfloor T\sqrt{r}/\delta \rfloor + 1$ "subintervals," each of length $\delta/\sqrt{r}$. Namely, we consider subintervals $[t^r_k, t^r_{k+1}]$, $k = 0, 1, \ldots, K-1$, where $t^r_k = k\delta/\sqrt{r}$. We will use notation $f(s:t) = f(t) - f(s)$ for increments and denote by $D^r_{ij}(t)$ the number of class i customers who complete service in pool j by time t. Using (9.15) and the law of large numbers in the form (3.11), we can establish the following facts:

$$\max_{k\le K-1} \max_{u\in[0,\delta]} \left|\frac{1}{\sqrt{r}} A^r_i\left(t^r_k : t^r_k + u/\sqrt{r}\right) - \lambda_i u\right| \xrightarrow{P} 0, \quad \forall i, \tag{9.21}$$

$$\max_{k\le K-1} \max_{u\in[0,\delta]} \left|\frac{1}{\sqrt{r}} D^r_{ij}\left(t^r_k : t^r_k + u/\sqrt{r}\right) - \mu_{ij}\psi^*_{ij}u\right| \xrightarrow{P} 0, \quad \forall (ij). \tag{9.22}$$

Now, let us fix $\epsilon = 3[\max_{(ij)} \mu_{ij}][2\max_{(ij)} \mu_{ij} + \sum_i \lambda_i]\delta$ and denote $\Delta^r(t) = r^{-1/2}[\mu_{11}\hat{Y}^r_1(t) - \mu_{21}\hat{Y}^r_2(t)]$. Then, it follows from (9.21) and (9.22) that, with prob-

ability approaching 1 as $r \to \infty$, a process trajectory must satisfy the following properties:

$$\begin{aligned} &\max_{u\in[0,\delta]} \left|\Delta^r\left(t_k^r : t_k^r + u/\sqrt{r}\right)\right| < \epsilon, \quad \forall k, \\ \forall k: \quad &\Delta^r\left(t_k^r\right) > \epsilon \quad \text{implies} \\ &0 < \Delta^r\left(t_k^r : t_k^r + t_{k+1}^r\right) < \epsilon - (1/2)(\mu_{11}+\mu_{21})\mu_{21}\psi_{21}^*\delta, \end{aligned} \tag{9.23}$$

$$\begin{aligned} \forall k: \quad &\Delta^r\left(t_k^r\right) < -\epsilon \quad \text{implies} \\ &0 > \Delta^r\left(t_k^r : t_k^r + t_{k+1}^r\right) > -\epsilon + (1/2)(\mu_{11}+\mu_{21})\mu_{11}\psi_{11}^*\delta. \end{aligned} \tag{9.24}$$

Property (9.23) [resp. (9.24)] holds because if $\Delta^r(t_k^r) > \epsilon$ [resp. $\Delta^r(t_k^r) < -\epsilon$] for some k, then according to the MaxWeight rule, in the interval $[t_k^r, t_{k+1}^r]$ all servers in pool 1 that become free will take for service class 1 [resp. class 2] customers only. Since δ, and the corresponding ϵ, can be chosen arbitrarily small, and given (9.17), we obtain (9.18); and if additionally (9.20) holds, (9.18) holds with $s = 0$. Property (9.19) is proved similarly. □

9.5 Weak limits of X-model under MaxWeight–FSF

We next establish the weak limit of X-model systems under MaxWeight–FSF.

Proposition 9.5 *Consider a sequence of X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies* (9.1)–(9.3). *Assume that for any* $T > 0$

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left\{\sup_{t\le T}\left(\left\|\hat{\Psi}^r(t)\right\| + \left\|\hat{Y}^r(t)\right\|\right) > M\right\} = 0, \tag{9.25}$$

and that

$$\left(\hat{\Psi}^r(0), \hat{Y}^r(0)\right) \Rightarrow \left(\hat{\Psi}(0), \hat{Y}(0)\right) \tag{9.26}$$

as $r \to \infty$. *Then, there exists a subsequence* r_k *of* r *such that*

$$\left(\hat{X}^{r_k}, \hat{T}^{r_k}, B^{r_k}\right) \Rightarrow (\hat{X}, \hat{T}, B) \tag{9.27}$$

as $k \to \infty$, *where* $B^{r_k} = (B_1^{r_k}, B_2^{r_k})$ (*see* (9.6)) *and* B_i*'s are independent driftless Brownian motions with variances*

$$\left(\lambda_i + \sum_j \mu_{ij}\psi_{ij}^*\right), \quad i = 1, 2. \tag{9.28}$$

The limiting process $\hat{X} = (\hat{X}_1, \hat{X}_2)$ *satisfies the following equations*

$$\hat{X}_1(t) = \hat{X}_1(0) - B_1(t) - \mu_{11}\hat{T}_{11}(t) - \mu_{12}\hat{T}_{12}(t) + \ell_1 t, \tag{9.29}$$

$$\hat{X}_2(t) = \hat{X}_2(0) - B_2(t) - \mu_{21}\hat{T}_{21}(t) - \mu_{22}\hat{T}_{22}(t) + \ell_2 t. \tag{9.30}$$

Also

$$\hat{T}_{12}(t) \geq 0 \quad \textit{and} \tag{9.31}$$

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left\{\left\|\hat{T}(\cdot)\right\|_T > M\right\} = 0. \tag{9.32}$$

Proof of Proposition 9.5 Consider a sequence of X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies (9.1)–(9.3). Assume that (9.25) and (9.26) hold. Note that by (9.25) there exists a subsequence r_k of r such that for any $T > 0$

$$\sup_{0\leq t\leq T} \left|\bar{\Psi}_{ij}^{r_k}(t) - \psi_{ij}^*\right| \to 0 \quad \text{and}$$

$$\sup_{0\leq t\leq T} \left\|\bar{Y}^{r_k}(t)\right\| \to 0$$

a.s. $r \to \infty$. Hence, $B_i^{r_k} \Rightarrow B_i$ (recall (9.6)) as $r \to \infty$, where B_i is a Brownian motion with variance given by (9.28); with B_i being mutually independent.

Consider

$$\hat{T}_{ij}^r(t) = \int_0^t \hat{\Psi}_{ij}^r(s)\, ds.$$

By (9.25), $\hat{T}_{ij}^r$ is tight in sup norm over compact sets. By (9.5), $\hat{X}_i^r$ is also tight in sup norm over compact sets. Because $\{B_i^{r_k}\}$ is also tight, there exists a further subsequence, denoted again by r_k for notational simplicity, such that

$$\left(\hat{X}^{r_k}, \hat{T}^{r_k}, B^{r_k}\right) \Rightarrow \left(\hat{X}, \hat{T}, B\right)$$

as $k \to \infty$, with $(\hat{X}, \hat{T}, B)$ having continuous sample paths. Also $\hat{X}$ satisfies (9.29) and (9.30) by (9.5).

Property (9.31) follows from the fact that $\hat{T}_{12}^r(t) \geq 0$ for all $t \geq 0$. Also,

$$\hat{T}_{ij}^r(t) \leq t \left\|\hat{\Psi}_{ij}^r(\cdot)\right\|_t,$$

hence (9.32) follows from (9.25). □

Proposition 9.6 *Assume that the conditions of Proposition 9.5 hold. Let r_k denote the subsequence along which (9.27) holds. Also assume that*

$$\hat{\Psi}_{12}^r(0) + \hat{\Psi}_{22}^r(0) \Rightarrow 0, \quad \textit{as } r \to \infty, \tag{9.33}$$

$$\left|\mu_{11}\hat{Y}_1^r(0) - \mu_{21}\hat{Y}_2^r(0)\right| \Rightarrow 0, \quad \textit{as } r \to \infty, \tag{9.34}$$

and

$$\sup_{0\leq t\leq T} \left|\bar{\Psi}_{ij}^{r_k}(t) - \psi_{ij}^*\right| \to 0 \quad \textit{and} \tag{9.35}$$

$$\sup_{0\leq t\leq T} \left\|\bar{Y}^{r_k}(t)\right\| \to 0. \tag{9.36}$$

If for each $T > 0$

$$\inf_{0\le t\le T}\big(\hat{X}_1(t)+\hat{X}_2(t)\big) > 0 \quad a.s. \tag{9.37}$$

then

$$\big(T^{-1}\hat{T}_{12}(T),\ T^{-1}\hat{T}_{22}(T)\big) \Rightarrow (0,0), \tag{9.38}$$

as $T \to \infty$.

Proof Assume that the conditions of Proposition 9.5 and (9.33)–(9.36) hold. Let

$$\big(\hat{X}^{r_k}, \hat{T}^{r_k}, B^{r_k}\big) \Rightarrow \big(\hat{X}, \hat{T}, B\big) \tag{9.39}$$

as $k \to \infty$. We drop the subscript k from our notation for notational simplicity.

Fix $\epsilon > 0$. We show that there exists $T > 0$ such that

$$P\big\{T^{-1}\hat{T}_{12}(T) > \epsilon, T^{-1}\hat{T}_{22}(T) > \epsilon\big\} < \epsilon.$$

Note that by (9.33), (9.35), (9.36) and Lemma 9.4, there exists a sequence $\delta^r \to 0$ as $r \to \infty$, see Theorem 4.2.3 [13], such that

$$P\big\{\big\|\mu_{11}\hat{Y}_1^r(\cdot) - \mu_{21}\hat{Y}_2^r(\cdot)\big\|_T > \delta^r\big\} < \delta^r.$$

For a sequence of positive numbers $\{\delta_0^r\}$ define

$$\tau_2\big(\delta_0^r\big) = \inf\big\{t : \hat{Y}_1^r(t) \le \delta_0^r \text{ or } \hat{Y}_2^r(t) \le \delta_0^r\big\} \wedge T. \tag{9.40}$$

By (9.35) there exists $c > 0$ such that for $\delta_0^r = c\delta^r$

$$A_{12}^r\big(\tau_2\big(\delta_0^r\big)\big) = 0 \tag{9.41}$$

outside the set

$$\big\{\big\|\mu_{11}\hat{Y}_1^r(\cdot) - \mu_{21}\hat{Y}_2^r(\cdot)\big\|_T > \delta^r\big\}.$$

This follows from the facts that activity $(1,2)$ is non-basic, both queues are non-zero until time $\tau_2(\delta_0^r)$ (see (9.40)), and under the MaxWeight rule, a class 1 customer from the queue cannot be sent to server pool 2 unless $\mu_{12}\hat{Y}_{21}^r(t) > \mu_{22}\hat{Y}_{22}^r(t)$. (Here it is enough to take $c \ge \frac{\mu_{12}}{\mu_{22}\mu_{11}-\mu_{12}\mu_{22}}$. In this case $c > 0$ because of our assumption that $(1,2)$ is non-basic. The exact value of c is not needed in the proof.)

Also, for any $c_2 > 0$

$$\liminf_{r\to\infty} P\Big\{\inf_{0\le t\le T}\hat{X}_1^r(t)+\hat{X}_2^r(t) > c_2\delta^r\Big\} \ge P\Big\{\inf_{0\le t\le T}\hat{X}_1(t)+\hat{X}_2(t) > \xi\Big\}$$

for any $\xi > 0$. Since ξ is arbitrary, (9.37) implies that

$$\liminf_{r\to\infty} P\Big\{\inf_{0\le t\le T}\hat{X}_1^r(t)+\hat{X}_2^r(t) > c_2\delta^r\Big\} = 1. \tag{9.42}$$

Note that by (9.41) this implies, by choosing c_2 large enough, that

$$A^r_{12}(T) \Rightarrow 0$$

as $r \to \infty$, i.e. with probability approaching 1 there is no class 1 arrivals into pool 2. Using this fact and the argument analogous to that in the proof of Lemma 9.1, we see that for any $\epsilon_2 > 0$

$$P\big[\hat{\Psi}^r_{12}(t) \le \hat{\Psi}^r_{12}(0)\exp(-\mu_{12}t) + \epsilon_2,\ \forall t \le T\big] \to 1, \quad \text{as } r \to \infty. \tag{9.43}$$

Fix $\epsilon > 0$. By (9.26) and (9.43), there exist $t > 0$ such that for any $T > 0$ for all sufficiently large r,

$$P\Big\{ \sup_{t \le u \le T} \hat{\Psi}^r_{12}(u) > \epsilon,\ \sup_{0 \le u \le t} \hat{\Psi}^r_{12}(u) > \hat{\Psi}^r_{12}(0) + \epsilon \Big\} < \epsilon.$$

Hence, for T large enough, this with (9.26) implies that

$$P\big\{(T)^{-1}\hat{T}^r_{12}(T) > \epsilon\big\} < \epsilon. \tag{9.44}$$

In addition, (9.42) implies that the probability of all servers remaining busy in $[0, T]$ approaches 1:

$$\liminf_{r\to\infty} P\big\{\hat{T}^r_{12}(T) + \hat{T}^r_{22}(T) = 0\big\} = 1. \tag{9.45}$$

Therefore, (9.38) follows by (9.39), (9.44) and (9.45). □

10 Proof of Theorem 4.1

We consider a sequence of X-models, see Fig. 2, in the Halfin–Whitt regime (of Sect. 2), and assume conditions (9.1)–(9.3). For this sequence we prove Theorem 4.1 by contradiction. Recall that we consider MaxWeight–FSF with non-pre-emptive service. Assume that the MaxWeight–FSF algorithm is order-optimal. This implies that (9.12) holds and so

$$\limsup_{r\to\infty} E\big[\big|\bar{Y}^r(\infty)\big|\big] = 0.$$

We first note that each system with large r has a stationary distribution since under the MaxWeight rule a parallel server system is stable when its load is strictly less than capacity [27].

We prove Theorem 4.1 in two key steps. In the first step, we show that, as r increases, under MaxWeight–FSF rule in stationary regime, the first pool (and then the entire system) must get to a state with $O(\sqrt{r})$ idle servers within finite time and with non-vanishing probability. In the second step, we prove that if the system starts with $O(\sqrt{r})$ idle servers then the number of servers in pool 2 serving class 1 customers gets to the level $M\sqrt{r}$ with arbitrarily large positive M under the FSF routing algorithm. The two steps combined mean that in steady state, for any $M > 0$, the probability of having $M\sqrt{r}$ class 1 customers being served in pool 2 is non-vanishing; this, however, is inconsistent with the system stability, see Lemma 3.3.

Proposition 10.1 *Consider a sequence of stationary X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies (9.1)–(9.3). Assume that (9.12) holds. Let subsequence r_k of r be such that (9.13) and (9.14) hold. Then*

$$\lim_{M\to\infty} \limsup_{k\to\infty} P\left\{\left|\hat{\psi}_{ij}^{r_k}(0)\right| > M\right\} = 0, \tag{10.1}$$

$$\lim_{M\to\infty} \limsup_{k\to\infty} P\left\{\left\|\hat{Y}^{r_k}(0)\right\| > M\right\} = 0, \tag{10.2}$$

$$\hat{\Psi}_{12}^{r_k}(0) + \hat{\Psi}_{22}^{r_k}(0) \Rightarrow 0, \tag{10.3}$$

$$\left|\mu_{11}\hat{Y}_1^{r_k}(0) - \mu_{21}\hat{Y}_2^{r_k}(0)\right| \Rightarrow 0, \tag{10.4}$$

as $k \to \infty$. In addition, for any $T > 0$, there exists a further subsequence $\{r_k'\}$ of $\{r_k\}$ such that

$$\lim_{M\to\infty} \limsup_{k\to\infty} P\left\{\sup_{t\le T}\left(\left\|\hat{\Psi}^{r_k'}(t)\right\| + \left\|\hat{Y}^{r_k'}(t)\right\|\right) > M\right\} = 0, \tag{10.5}$$

and there exist $C > 0$ and $\epsilon > 0$ such that

$$\liminf_{k'\to\infty} P\left(\inf_{0\le t\le T} \hat{X}_1^{r_k'}(t) + \hat{X}_2^{r_k'}(t) < -C\right) > \epsilon. \tag{10.6}$$

Proposition 10.2 *Consider a sequence of X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies (9.1)–(9.3). Assume that*

$$\limsup_{r\to\infty} E\left[\left|\bar{Y}^r(0)\right|\right] = 0,$$

and that for any $T > 0$

$$\sup_{0\le t\le T} \left\|\bar{Y}^r(t)\right\| \to 0, \tag{10.7}$$

$$\sup_{0\le t\le T} \left\|\bar{\Psi}_{ij}^r(t) - \psi_{ij}^*\right\| \to 0, \tag{10.8}$$

a.s. as $r \to \infty$. Also assume that for a constant $C > 0$

$$\liminf_{r\to\infty} P\left\{\hat{X}_1^r(0) + \hat{X}_2^{r_k}(0) < -C\right\} > 0, \tag{10.9}$$

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left\{\left|\hat{\psi}_{ij}^r(0)\right| > M\right\} = 0, \tag{10.10}$$

$$\left|\hat{\Psi}_{12}^r(0) + \hat{\Psi}_{22}^r(0)\right| \Rightarrow 0, \quad \text{as } r \to \infty, \tag{10.11}$$

$$\left|\mu_{11}\hat{Y}_1^r(0) - \mu_{21}\hat{Y}_2^r(0)\right| \Rightarrow 0, \quad \text{as } r \to \infty, \tag{10.12}$$

and

$$\hat{X}^r(0) \Rightarrow \hat{X}(0) \quad \text{as } r \to \infty. \tag{10.13}$$

Then, there exists a subsequence r_k of r and $S > 0$ such that

$$\lim_{M\to\infty} \liminf_{k\to\infty} P\big[\hat{\Psi}_{12}^{r_k}(S) > M\big] > 0. \tag{10.14}$$

Proofs of Propositions 10.1 and 10.2 will be presented in Sects. 11 and 12, respectively. (Auxiliary results for the X-model system, presented in Sect. 9, may be of independent interest, but in this paper they are building blocks for the proofs of Propositions 10.1 and 10.2.)

Proof of Theorem 4.1 Consider a sequence of stationary X-models in the Halfin–Whitt regime operating under the MaxWeight–FSF rule that satisfies (9.1)–(9.3). Assume that (9.12) holds. Let $\{r_k\}$ denote the subsequence such that (9.13)–(9.14) hold. We focus on his subsequence for the rest of the proof and for notational simplicity we drop the subscript k from our notation. Note that by Proposition 10.1 there exists $C > 0$ such that (10.6) holds. Fix C and let

$$\tau^r = \inf\big\{t : \hat{X}_1^r(t) + \hat{X}_2^r(t) < -C\big\} \wedge T.$$

Since X^r is a Markov process, we can analyze the system by restarting the process at τ^r and setting the initial state of the system to be $X^r(\tau^r)$. We denote this process by $X^r_{\tau^r}$ and next show that it satisfies the conditions of Proposition 10.2.

By (9.13) and (9.14), $X^r_{\tau^r}$ satisfies (10.7) and (10.8), respectively. By (10.5) it satisfies (10.10). Note that for the original process whose initial state is its steady state, by (10.3)–(10.4) and Lemma 9.4

$$\sup_{0\le t\le 2T} \hat{\Psi}_{12}^r(t) + \hat{\Psi}_{22}^r(t) \Rightarrow 0, \quad \text{as } r\to\infty,$$

$$\sup_{0\le t\le 2T} \big|\mu_{11}\hat{Y}_1^r(t) - \mu_{21}\hat{Y}_2^r(t)\big| \Rightarrow 0, \quad \text{as } r\to\infty.$$

Therefore, $X^r_{\tau^r}$ satisfies (10.11) and (10.12) since $\tau^r \le T$. In addition, by (10.5), $\hat{X}^r(\tau^r)$ is tight, hence there exists a subsequence r_k such that (10.13) also holds. Finally, by (10.6), $X^r_{\tau^r}$ also satisfies (10.9).

Thus, by Proposition 10.2, there exists a further subsequence r_ℓ of r_k such that

$$\lim_{M\to\infty} \liminf_{\ell\to\infty} P\big(\hat{\Psi}_{12}^{r_\ell}\big(T + \tau^{r_\ell}\big) > M\big) > 0.$$

Since $\tau^{r_\ell} \le T$, by Lemma 9.1

$$\lim_{M\to\infty} \liminf_{\ell\to\infty} P\big(\hat{\Psi}_{12}^{r_\ell}(2T) > \exp\{-\mu_{12}T\}M\big) > 0.$$

Since the original system started at its steady state

$$\lim_{M\to\infty} \liminf_{\ell\to\infty} P\big(\hat{\Psi}_{12}^{r_\ell}(0) > M\big) = \lim_{M\to\infty} \liminf_{\ell\to\infty} P\big(\hat{\Psi}_{12}^{r_\ell}(2T) > M\big) > 0.$$

But this contradicts with (10.1), hence (9.12) cannot hold. □

11 Proof of Proposition 10.1

Observe that (10.2) follows immediately from (9.12), Fubini's Theorem and the fact that $\hat{Y}_i \geq 0$. Hence, for each $M > 0$ and r large enough, this gives (by reselecting C if necessary)

$$P\left(\int_0^t \left\|\hat{Y}^r(s)\right\| ds > M, \left\|\hat{Y}^r(0)\right\| > M\right) < C/M. \tag{11.1}$$

Note also that conditions of Propositions 9.2 and 9.3 are satisfied. For $\hat{\Psi}^r_{12}$, (9.11) gives (10.1). Note also that, because

$$\left|\hat{\Psi}^r_{22}(t)\right| \leq \left|\hat{Z}^r_2(t)\right| + \left|\hat{\Psi}^r_{12}(t)\right|,$$

we have by (9.9) and (9.10) that

$$P\left(\int_0^t \left|\hat{\Psi}^r_{22}(s)\right| ds > M, \left|\hat{\Psi}^r_{22}(0)\right| > M\right) < C/M. \tag{11.2}$$

Below we prove that

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left(\left\|\hat{X}^r(0)\right\| > M\right) = 0. \tag{11.3}$$

Then, because for all $t \geq 0$

$$\hat{\Psi}^r_{11}(t) = \hat{X}^r_1(t) - \hat{Y}^r_1(t) - \hat{\Psi}^r_{12}(t) \quad \text{and} \tag{11.4}$$

$$\hat{\Psi}^r_{21}(t) = \hat{X}^r_2(t) - \hat{Y}^r_2(t) - \hat{\Psi}^r_{22}(t), \tag{11.5}$$

we have (10.1) by (10.2), (9.9), (11.2) and (11.3).

Next we prove (11.3). Assume on the contrary that

$$\lim_{M\to\infty} \limsup_{r\to\infty} P\left(\left\|\hat{X}^r(0)\right\| > M\right) = \epsilon > 0 \tag{11.6}$$

for some $\epsilon > 0$.

Fix $0 < t < 1$. Choose M large enough in (9.9), (9.10), (11.1) and (11.2) such that $C/M < \epsilon/8$. We define the set $\mathcal{B}^r$ as the intersection of the "good" sets in (9.9), (9.10), (11.1), (11.2), and the set $\{|\hat{X}^r_1(0)| \vee |\hat{X}^r_2(0)| > N\}$, for large N, i.e.,

$$\begin{aligned}
\mathcal{B}^r = &\left\{\int_0^t \left\|\hat{Y}^r(s)\right\| ds < M, \left\|\hat{Y}^r(0)\right\| < M\right\} \\
&\cap \left\{\int_0^t \hat{\Psi}^r_{12}(s)\, ds < M, \hat{\Psi}^r_{12}(0) < M\right\} \\
&\cap \left\{\int_0^t \left|\hat{\Psi}^r_{22}(s)\right| ds < M, \left|\hat{\Psi}^r_{22}(0)\right| < M\right\} \\
&\cap \left\{\int_0^t \hat{Z}^r_j(s)\, ds < M, \hat{Z}^r_j(0) < M\right\} \\
&\cap \left\{\left|\hat{X}^r_1(0)\right| \vee \left|\hat{X}^r_2(0)\right| > N\right\}.
\end{aligned} \tag{11.7}$$

For r large enough, we can assume $P(\mathcal{B}^r) > \epsilon/2$. Note also that since B_i^r in (9.6) converges weakly to a Brownian motion by (9.14), in the set $\mathcal{B}^r$ we can assume that $\|B_i^r(\cdot)\|_1 < M$. By (11.1) in this set $\hat{X}_1^r(0) + \hat{X}_2^r(0) < M$, but for any given $N > M$ (by resetting the value of N in (11.7)) and r large enough $|\hat{X}_1^r(0)| \vee |\hat{X}_2^r(0)| > N$.

Assume that $\hat{X}_1^r(0) > N$, then $-\hat{X}_1^r(0) + M \geq \hat{X}_2^r(0) \geq -\hat{X}_1^r(0) - M$. Note that

$$\hat{X}_1^r(t) = \hat{X}_1^r(0) + \ell_1^r t + B_1^r(t) - \mu_{11} \int_0^t \hat{X}_1^r(s)\,ds + \mu_{11} \int_0^t \hat{Y}_1^r(s)\,ds - (\mu_{12} - \mu_{11}) \int_0^t \hat{\Psi}_{12}^r(s)\,ds, \tag{11.8}$$

$$\hat{X}_2^r(t) = \hat{X}_2^r(0) + \ell_2^r t + B_2^r(t) - \mu_{21} \int_0^t \hat{X}_2^r(s)\,ds + \mu_{21} \int_0^t \hat{Y}_2^r(s)\,ds - (\mu_{22} - \mu_{21}) \int_0^t \hat{\Psi}_{22}^r(s)\,ds. \tag{11.9}$$

Next we claim that there exists $\delta > 0$ and $c > 0$ such that, on $\mathcal{B}^r$

$$\hat{X}_1^r(\delta) \leq \big(\hat{X}_1^r(0)\big) \exp\{-\mu_{11}\delta\} + cM, \tag{11.10}$$

$$\hat{X}_2^r(\delta) \leq \big(\hat{X}_2^r(0)\big) \exp\{-\mu_{21}\delta\} + cM. \tag{11.11}$$

To prove (11.10), let

$$x(t) = \hat{X}_1^r(0) - \mu_{11} \int_0^t x(s)\,ds.$$

Then,

$$x(t) = \hat{X}_1^r(0) \exp\{-\mu_{11} t\}.$$

Also,

$$\big|\hat{X}_1^r(t) - x(t)\big| \leq |\ell_1 t| + \big|B_1^r(t)\big| + \mu_{11} \left| \int_0^t \hat{Y}_1^r(s)\,ds \right| + \mu_{21} \left| \int_0^t \hat{\Psi}_{21}^r(s)\,ds \right| + \mu_{11} \int_0^t \big|\hat{X}_1^r(s) - x(s)\big|\,ds.$$

Hence, by Corollary 11.2 in [28], on $\mathcal{B}^r$

$$\big|\hat{X}_1^r(t) - x(t)\big| \leq cM \exp\{\mu_{11} t\}$$

for some $c > 0$. By selecting $\delta > 0$ small enough we have (11.10). The second inequality (11.11) follows similarly.

By (11.10) and (11.11)

$$\hat{X}_1^r(\delta) + \hat{X}_2^r(\delta) \leq \hat{X}_1^r(0) \exp\{-\mu_{11}\delta\} + \hat{X}_2^r(0) \exp\{-\mu_{21}\delta\} + 2cM.$$

Because $\mu_{11} > \mu_{21}$ and N is arbitrary, we have that

$$\hat{Z}_1^r(\delta) + \hat{Z}_2^r(\delta) > N/2.$$

Also, the initial state is stationary, therefore

$$P\{\|\hat{Z}^r(0)\| > N/4\} > \epsilon/2,$$

which contradicts (9.10) since we can select N arbitrarily large. Hence, $\hat{X}_1^r(0) < N$. If $\hat{X}_2^r(0) > N$, we get a similar contradiction for $\hat{Y}_1^r(0) + \hat{Y}_2^r(0)$. Thus (11.6) cannot hold and so (11.3) must hold.

Result (10.3) follows from (10.1), (9.13), and (9.14), because Lemma 9.4 implies that if

$$\limsup_{M\to\infty} \limsup_{r\to\infty} P\{|\hat{\Psi}_{ij}^r(0)| > M\} = 0,$$

then for any $T > s > 0$

$$\limsup_{r\to\infty} P\Big\{\inf_{s\le t\le T} \hat{\Psi}_{12}^r(t) + \hat{\Psi}_{22}^r(t) < -\varepsilon\Big\} = 0, \tag{11.12}$$

for every $\varepsilon > 0$ and $T > 0$. Result (10.4) follows similarly.

Next we prove (10.5). By (11.12) and (9.11), we have

$$\limsup_{M\to\infty} \limsup_{r\to\infty} P\Big\{\sup_{0\le t\le T} |\hat{\Psi}_{22}^r(t)| > M\Big\} = 0. \tag{11.13}$$

Note that by (11.8)

$$\begin{aligned}
|\hat{X}_1^r(t)| &\le |\hat{X}_1^r(0)| + \mu_{11}\int_0^t |\hat{X}_1^r(s)|\,ds + \mu_{11}\int_0^t |\hat{Y}_1^r(s)|\,ds \\
&\quad + |\mu_{12} - \mu_{11}|\int_0^t |\hat{\Psi}_{12}^r(s)|\,ds + |B_1^r(t)| + |\ell_1^r|t \\
&\le |\hat{X}_1^r(0)| + \mu_{11}\int_0^t |\hat{X}_1^r(s)|\,ds + \mu_{11}\int_0^t \big(|\hat{X}_1^r(s)| + |\hat{X}_2^r(s)|\big)\,ds \\
&\quad + |\mu_{12} - \mu_{11}|\int_0^t |\hat{\Psi}_{12}^r(s)|\,ds + |B_1^r(t)| + |\ell_1^r|t.
\end{aligned}$$

The inequality above follows from the fact that

$$\hat{Y}_1^r(t) + \hat{Y}_2^r(t) \le |\hat{X}_1^r(t)| + |\hat{X}_2^r(t)|,\ t \ge 0. \tag{11.14}$$

Similarly, by (11.9)

$$\begin{aligned}
|\hat{X}_2^r(t)| &\le |\hat{X}_2^r(0)| + \mu_{21}\int_0^t |\hat{X}_2^r(s)|\,ds + |\mu_{21} - \mu_{22}|\int_0^t |\hat{\Psi}_{22}^r(s)|\,ds \\
&\quad + \mu_{21}\int_0^t \big(|\hat{X}_1^r(s)| + |\hat{X}_2^r(s)|\big)\,ds + |B_2^r(t)| + |\ell_2^r|t.
\end{aligned}$$

Note that by (9.14), $B_i^r \Rightarrow B_i$ as $r \to \infty$. Therefore, similar to (3.31) in [30], by (9.11) and (11.13), we have that

$$\lim_{M\to\infty} \limsup_{k\to\infty} P\left\{\sup_{t\le T}\left|\hat{X}^r(t)\right| > M\right\} = 0.$$

This gives (10.5) by (9.11), (11.4), (11.5), (11.13), and (11.14).

Next we prove (10.6). Note that the sequences $\hat{Y}^r(0)$ and $\hat{\Psi}^r(0)$ are tight by (10.1) and (10.2), hence we can find a subsequence r_k such that

$$\hat{Y}^{r_k}(0) \Rightarrow \hat{Y}(0) \quad \text{and} \quad \hat{\Psi}^{r_k}(0) \Rightarrow \hat{\Psi}(0),$$

as $k \to \infty$. By Proposition 9.5, this implies,

$$\hat{X}^{r_k} \Rightarrow \hat{X}$$

as $k \to \infty$. We drop the subscript "k" for notational simplicity.

Below, we prove that for some $T > 0$

$$P\left\{\inf_{0\le t\le T} \hat{X}_1(t) + \hat{X}_2(t) \le 0\right\} > 0. \tag{11.15}$$

Note that this implies that there exists $\eta > 0$ and $T > 0$ such that

$$P\left\{\inf_{0\le t\le T} \hat{X}_1(t) + \hat{X}_2(t) \le 0\right\} > \eta.$$

Then, for any $\delta > 0$, if we denote

$$\tau^r(\delta) = \inf\left\{t \ge 0 : \hat{X}_1^r(t) + \hat{X}_2^r(t) \le \delta\right\} \wedge T,$$

we have

$$\liminf_{r\to\infty} P\left\{\tau^r(\delta) < T\right\} > \eta.$$

This in turn means that we can choose a subsequence of r and corresponding sequence of $\delta^r \to 0$, such that, along this subsequence,

$$\liminf_{r\to\infty} P\left\{\tau^r\left(\delta^r\right) < T\right\} > \eta. \tag{11.16}$$

Consider the process restarted at the stopping time $\tau^r(\delta^r)$. This process, denoted by

$$\left\{\left(\check{X}^r(\xi), \check{\Psi}^r(\xi)\right), \xi \ge 0\right\},$$

has the initial distribution of $(\hat{X}^r(\tau^r(\delta^r)), \hat{\Psi}^r(\tau^r(\delta^r)))$. We can find a further subsequence of $\{r\}$, along which this process has a weak limit, $\{\check{X}(\xi), \check{\Psi}(\xi), \xi \ge 0\}$; this follows from (10.5). Also,

$$P\left\{\check{X}_1(0) + \check{X}_2(0) = 0\right\} > \eta, \tag{11.17}$$

since $P\{\tau^r(\delta^r) < T\}$ remains lower bounded by $\eta > 0$, see (11.16).

By Proposition 9.5 and (10.5), we know that the weak limit of the restarted process is such that, w.p.1 $\check{X}(\xi)$ is continuous and $\check{\Psi}(\xi)$ is bounded on bounded time intervals. Now, if we assume that (10.6) does *not* hold, then we also must have

$$\check{X}_1(\xi) + \check{X}_2(\xi) \geq 0, \quad \forall \xi \geq 0.$$

This is, however, impossible. Indeed by Proposition 9.5, we have that w.p.1,

$$\check{X}_1(\xi) + \check{X}_2(\xi) = \check{X}_1(0) + \check{X}_2(0) + \int_0^\xi h(s)\,ds + B(\xi),$$

where $h(s)$ is some (random) function that is finite on finite intervals and $B(\xi)$ is a zero-drift Brownian motion. But, w.p.1., for any $M > 0$ and any $\epsilon_1 > 0$, the value of $B(\xi) + M\xi$ "drops" below 0 in $(0, \epsilon_1)$. This proves (10.6) by (11.17).

Next we prove (11.15) to conclude the proof. Assume on the contrary that for any $T > 0$

$$P\Big\{\inf_{0\leq t\leq T} \hat{X}_1(t) + \hat{X}_2(t) > 0\Big\} = 1. \tag{11.18}$$

Consider the process

$$\hat{W}(t) = \frac{\hat{X}_1(t)}{\mu_{11}} + \frac{\hat{X}_2(t)}{\mu_{21}}.$$

(For the X-system we consider, the vector of workload contributions (ν_1, ν_2) is proportional to $(1/\mu_{11}, 1/\mu_{21})$. So, $\hat{W}(t)$ is the "workload" associated with $(\hat{X}_1(t), \hat{X}_2(t))$.) Note that by (10.1), (10.2), and the fact that the process is stationary, for any $T > 0$

$$\lim_{M\to\infty} P\big\{\big\|\hat{W}(T)\big\| > M\big\} = 0. \tag{11.19}$$

By (9.29), (9.30), and (11.18)

$$\hat{W}(t) = \hat{W}(0) + \hat{\ell}t + B(t) - (\mu_{12}/\mu_{11})\hat{T}_{12}(t) - (\mu_{22}/\mu_{21})\hat{T}_{22}(t),$$

where $\hat{\ell} = \mu_{11}^{-1}\ell_1 + \mu_{21}^{-1}\ell_2 < 0$ (see (2.10)) and B is a driftless Brownian motion.

By Proposition 9.6, (11.18) implies that for any $\epsilon > 0$ fixed, we can find T large enough such that

$$P\big\{T^{-1}\hat{T}_{12}(T) > \epsilon\big\} < \epsilon \quad \text{and} \quad P\big\{T^{-1}\hat{T}_{22}(T) < -\epsilon\big\} < \epsilon. \tag{11.20}$$

Therefore, for any $\epsilon > 0$, because $B(t)/t \to 0$ and $\hat{W}(0)/t \to 0$ (by (11.19)) a.s. as $r \to \infty$, and by (11.20), there exists $T > 0$ large enough such that

$$P\big\{\hat{W}(T) \leq \hat{\ell}T/2\big\} > 1 - \epsilon.$$

Since T is arbitrary, this contradicts (11.19), hence (11.18) cannot hold.

12 Proof of Proposition 10.2

The key part of the proof is the following proposition, which in particular provides a lower bound for the time until all servers become busy if at time zero the number of idle servers is in the order of $\sqrt{r}$.

Proposition 12.1 *Suppose we are in the conditions of Proposition 10.2. Consider the following ("local-fluid") scaled processes $\tilde{X}_i^r$ for $t \geq 0$:*

$$\tilde{X}_i^r(t) = \frac{X_i^r(t/\sqrt{r})}{\sqrt{r}} = \hat{X}_i^r\big(t/\sqrt{r}\big).$$

Then, there exists a subsequence r_k of r, along which

$$\big(\tilde{X}_1^r(\cdot), \tilde{X}_2^r(\cdot)\big) \Rightarrow \big(\tilde{X}_1(\cdot), \tilde{X}_2(\cdot)\big),$$

and the limit process is such that

$$\tilde{X}_1(t) + \tilde{X}_2(t) \equiv \tilde{X}_1(0) + \tilde{X}_2(0), \quad t \geq 0. \tag{12.1}$$

Consequently, along the subsequence, for any $T > 0$,

$$\liminf_{k\to\infty} P\big\{\tau^{r_k} > T/\sqrt{r}, \mathcal{A}^{r_k}\big\} > \liminf_{k\to\infty} P\big(\mathcal{A}^{r_k}\big)/2, \tag{12.2}$$

where

$$\mathcal{A}^r = \big\{\hat{X}_1^r(0) + \hat{X}_2^r(0) < -C\big\} \tag{12.3}$$

and

$$\tau^r = \inf\big\{t : \hat{X}_1^r(t) + \hat{X}_2^r(t) \geq 0\big\}.$$

Proof We can write

$$\tilde{X}_i^r(t) = \tilde{X}_i^r(0) + \frac{A_i^r(t/\sqrt{r})}{\sqrt{r}} - \sum_j \frac{1}{\sqrt{r}} S_{ij}\left(\int_0^{t/\sqrt{r}} \Psi_{ij}^r(s)\,ds\right). \tag{12.4}$$

Note that

$$\frac{1}{\sqrt{r}} S_{ij}\left(\int_0^{t/\sqrt{r}} \Psi_{ij}^r(s)\,ds\right) = \frac{1}{\sqrt{r}} S_{ij}\left(r\int_0^{t/\sqrt{r}} \bar{\Psi}_{ij}^r(s)\,ds\right).$$

Hence, for any fixed $T > 0$, we have

$$\sup_{0\leq t\leq T} \big|\tilde{S}_{ij}^r(t) - \mu_{ij}\psi_{ij}^* t\big| \to 0 \tag{12.5}$$

a.s. as $r \to \infty$ by (10.8), where

$$\tilde{S}_{ij}^r(t) = \frac{1}{\sqrt{r}} S_{ij}\left(\int_0^{t/\sqrt{r}} \Psi_{ij}^r(s)\,ds\right).$$

Similarly, for

$$\tilde{A}_i^r(t) = \frac{1}{\sqrt{r}} A_i^r\left(\frac{t}{\sqrt{r}}\right)$$

we have

$$\sup_{0\le t\le T} \left|\tilde{A}_i^r - \lambda_i t\right| \to 0$$

a.s. as $r \to \infty$. Summing up (12.4) over $i = 1, 2$, and taking limit on $r \to \infty$, we obtain (12.1) because $\sum_i \lambda_i = \sum_i \sum_j \mu_{ij} \psi_{ij}^*$.

Therefore, for any $T > 0$,

$$\left\| \tilde{X}_1^r(\cdot) + \tilde{X}_2(\cdot) - \tilde{X}_1^r(0) + \tilde{X}_2^r(0) \right\|_T \Rightarrow 0 \tag{12.6}$$

as $r \to \infty$. Note that for $\omega \in \mathcal{A}^r$, $\tau^r < T/\sqrt{r}$ implies

$$\left\| \tilde{X}_1^r(\cdot) + \tilde{X}_2(\cdot) - \tilde{X}_1^r(0) + \tilde{X}_2^r(0) \right\|_T > C.$$

The result (12.2) follows from this and (12.6). □

Proof of Proposition 10.2 We have

$$\liminf_{r\to\infty} P\left(\mathcal{A}^r\right) = \eta,$$

where $\mathcal{A}^r$ is given by (12.3). By (10.10), we can choose $M_0 > 0$ large enough such that

$$\liminf_{r\to\infty} P\left\{\hat{\Psi}_{12}^r(0) < M_0, \mathcal{A}^r\right\} \ge \eta/2.$$

Fix $M > M_0$. Below, we prove that, for any $S > 0$, there exists $T > 0$ large enough such that

$$\begin{aligned}\liminf_{r\to\infty} P\left\{\hat{\Psi}_{12}^r(S) > M \exp\{-\mu_{12} S\}\right\} &\ge \liminf_{r\to\infty} P\left\{\hat{\Psi}_{12}^r\left(r^{-1/2}T\right) > M\right\} \\ &> \eta/2. \end{aligned} \tag{12.7}$$

Since $M > M_0$ can be selected arbitrarily large, this yields (10.14).

Let us prove (12.7). Choose $T > 0$ large enough so that

$$p\mu_{22} T > 2M \tag{12.8}$$

for $p = \lambda_1/(\lambda_1 + \lambda_2)$. By Proposition 12.1 and (10.9),

$$\liminf_{r\to\infty} P\left\{\tau^r > T/\sqrt{r}, \mathcal{A}^r\right\} > \eta/2. \tag{12.9}$$

Let $D_{ij}^r(t)$ denote the number of class i customers completed service in pool j by time t, and $D_j^r(t) = D_{1j}^r(t) + D_{2j}^r(t)$. We define

$$\tilde{D}_{ij}^r(t) = \frac{1}{\sqrt{r}} D_{ij}^r\left(t/\sqrt{r}\right), \qquad \tilde{D}_j^r(t) = \frac{1}{\sqrt{r}} D_j^r\left(t/\sqrt{r}\right).$$

Note that by (9.14) as in (12.5)

$$\tilde{D}^r_{ij} \to \tilde{D}_{ij} \quad \text{a.s. u.o.c. as } r \to \infty,$$

where

$$\tilde{D}_{ij}(t) = \mu_{ij}\psi^*_{ij}t. \tag{12.10}$$

Hence,

$$\tilde{D}^r_2 \to \tilde{D}_2 \quad \text{a.s. u.o.c. as } r \to \infty,$$

where

$$\tilde{D}_2(t) = \mu_{22}t.$$

By Lemma 9.4, (10.8) and (10.11)

$$\sup_{0\le t\le T} \left|\tilde{A}^r_{12}(t) + \tilde{A}^r_{22}(t) - \tilde{D}^r_2(t)\right| \Rightarrow 0, \quad \text{as } r \to \infty, \tag{12.11}$$

where $A^r_{ij}(t)$ is the number of class i customers who entered service in pool j before time t and $\tilde{A}^r_{ij}(t) = r^{-1/2}A^r_{ij}(t/\sqrt{r})$. We can choose a further subsequence of r such that (12.11) hold a.s.

Now, in the time interval $[0, \tau^r)$ (in original, unscaled time) the behavior of pool 2 is very simple: when there is at least one idle server in it, the pool receives a Poisson input flow of total rate $\lambda^r_1 + \lambda^r_2$, where each arrival is class 1 with probability $p^r = \lambda^r_1/(\lambda^r_1 + \lambda^r_2)$ independently of the previous history of the system, and is class 2 with probability $1 - p^r$; when all servers in pool 2 are busy, there are no new arrivals in it. Using this observation, we easily obtain the following fact (on the local-fluid time scale):

$$\tilde{A}^r_{12}\left(t \wedge \tau^r\sqrt{r}\right) - p\left(\tilde{A}^r_{12}\left(t \wedge \tau^r\sqrt{r}\right) + \tilde{A}^r_{22}\left(t \wedge \tau^r\sqrt{r}\right)\right) \xrightarrow{P} 0, \quad \forall t \ge 0. \tag{12.12}$$

Since all functions $\tilde{A}^r_{ij}$ in (12.12) are non-decreasing, we can choose a further subsequence of r along which the convergence in (12.12) is u.o.c., w.p.1.

Let

$$\tilde{\Psi}_{ij}(t) = \hat{\Psi}^r_{ij}\left(t/\sqrt{r}\right).$$

Note that

$$\tilde{\Psi}^r_{12}(t) = \tilde{\Psi}^r_{12}(0) + \tilde{A}^r_{12}(t) - \tilde{D}^r_{12}(t),$$

where, as we already established, $\tilde{D}^r_{12}(T) \to 0$. Moreover, by (12.9), $\liminf P(\tau^r\sqrt{r} > T) > \eta/2$, which along with (12.8), (12.10), (12.11) and (12.12) gives

$$\liminf P\left(\tilde{\Psi}^r_{12}(T) > M\right) > \eta/2,$$

thus proving the right inequality of (12.7). The left inequality of (12.7) then follows from Lemma 9.1. □

13 Conclusions and directions of further research

In this paper we proposed a generic *Shadow routing* scheme for flexible many-server pools, which ensures good performance of the system, *without requiring a priori knowledge of the input flow rates*. This is achieved via a very simple virtual queueing mechanism, whose underlying objective is to balance server pool loads; in particular, the mechanism automatically identifies basic activities in the system, and automatically re-identifies them when/if input rates change. We proved, under the additional complete resource pooling (CRP) condition, that the mechanism not only balances pools' loads, but "produces" (on its output) flows with very well-behaved diffusion limits, which naturally suggests *order-optimality* of the proposed scheme—the property that the steady-state average queue lengths grow as $O(\sqrt{r})$ with the scaling parameter r. (Proving this fact is an important subject for future work.) Our simulations confirm good performance of the scheme. We believe that the Shadow routing algorithm is very attractive for practical applications.

Our results suggest many interesting directions of future research, in addition to formally proving the order-optimality of Shadow routing. First, it will be interesting to study the behavior of Shadow routing when the CRP condition does not necessarily hold. (We note however that the basic—load-balancing—property of the Shadow routing holds regardless of CRP; however, the diffusion limits of the flows it forms, will be more complex without CRP.) Secondly, for the application purposes it is very interesting to investigate different "opportunistic" versions of our scheme; namely schemes that do not necessarily take Shadow routing decision "literally", but, for example, use it only for the dynamic automatic identification of basic activities and then may opportunistically re-route customers along their basic activities. Many other options exist and may be useful in practice. Finally, the Shadow algorithm in this paper has the underlying objective of balancing pool loads, which is appropriate for scenarios when the system is not overloaded; however, versions of the algorithm can be designed for other objectives as well, for example for system reward maximization in the case of overload—this is another venue of future work.

We have proved that a simple MaxWeight–FSF discipline is not order-optimal, i.e. it makes average steady-state queues grow faster than $O(\sqrt{r})$. However, the question of what the growth rate actually is, remains open. We also would like to investigate whether one can find a sufficiently simple routing rule, which, when paired with MaxWeight, produces an order-optimal control.

References

1. Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: A multi-disciplinary perspective on operations management research. Prod. Oper. Manag. **16**, 655–688 (2007)
2. Armony, M.: Dynamic routing in large-scale service systems with heterogenous servers. Queueing Syst., Theory Appl. **51**, 287–329 (2005)
3. Armony, M., Maglaras, C.: Contact centers with a call-back option and real-time delay information. Oper. Res. **52**, 527–545 (2004)
4. Armony, M., Maglaras, C.: On customer contact centers with a call-back option: Customer decisions, routing rules and system design. Oper. Res. **52**, 271–292 (2004)
5. Atar, R.: A diffusion model of scheduling control in queueing systems with many servers. Ann. Appl. Probab. **15**, 820–852 (2005)

6. Atar, R.: Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. Ann. Appl. Probab. **15**, 2606–2650 (2005)
7. Atar, R., Mandelbaum, A., Reiman, M.: Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy-traffic. Ann. Appl. Probab. **14**, 1084–1134 (2004)
8. Atar, R., Mandelbaum, A., Shaikhet, G.: Queueing systems with many servers: Null controllability in heavy traffic. Ann. Appl. Probab. **16**, 1764–1804 (2006)
9. Bassamboo, A., Zeevi, A.: On a data-driven method for staffing large call centers. Oper. Res. **57**, 714–726 (2009)
10. Bassamboo, A., Harrison, J.M., Zeevi, A.: Design and control of a large call center: Asymptotic analysis of an LP-based method. Oper. Res. **54**, 419–435 (2006)
11. Bassamboo, A., Harrison, J.M., Zeevi, A.: Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. Queueing Syst., Theory Appl. **51**, 249–285 (2006)
12. Bell, S.L., Williams, R.J.: 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**, 608–649 (2001)
13. Chung, K.L.: A Course in Probability Theory, 3rd edn. Academic Press, New York (2001)
14. Dai, J.G., Tezcan, T.: State space collapse in many server limits of parallel server systems. Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology (2005)
15. Dai, J.G., Tezcan, T.: Dynamic control of parallel server systems in many server heavy traffic. Queueing Syst., Theory Appl. **59**, 95–134 (2008)
16. Gamarnik, D., Momcilovic, P.: Steady-state analysis of a multi-server queue in the Halfin–Whitt regime. Adv. Appl. Probab. **40**, 548–577 (2008)
17. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review and research prospects. Manuf. Serv. Oper. Manag. **5**, 79–141 (2003)
18. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **48**, 566–583 (2002)
19. Gurvich, I., Whitt, W.: Scheduling flexible servers with convex delay costs in many-server service systems. Manuf. Serv. Oper. Manag. **11**, 237–253 (2007)
20. Gurvich, I., Whitt, W.: Service-level differentiation in many-server service systems via queue-ratio routing. Oper. Res. **58**, 316–328 (2010)
21. Gurvich, I., Armony, M., Mandelbaum, A.: Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers. Manag. Sci. **54**, 279–294 (2008)
22. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
23. Harrison, J.M., López, M.J.: Heavy traffic resource pooling in parallel-server systems. Queueing Syst., Theory Appl. **33**, 339–368 (1999)
24. Harrison, J.M., Zeevi, A.: Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. Oper. Res. **52**, 243–257 (2004)
25. Jennings, O., de Vericourt, F.: Nurse staffing and bed capacity: A queueing perspective. Technical report, Duke University, The Fuqua School of Business (2008)
26. Kaspi, H., Ramanan, K.: Law of large numbers limits for many server queues. Working paper (2007)
27. Mandelbaum, A., Stolyar, A.L.: Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. Oper. Res. **52**, 836–855 (2004)
28. Mandelbaum, A., Massey, W.A., Reiman, M.: Strong approximations for Markovian service networks. Queueing Syst., Theory Appl. **30**, 149–201 (1998)
29. Perry, O., Whitt, W.: Responding to unexpected overloads in large-scale service systems. Manag. Sci. **55**, 1353–1367 (2009)
30. Puhalskii, A., Reiman, M.: The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. Adv. Appl. Probab. **32**, 564–595 (2000)
31. Randhawa, R.S., Kumar, S.: Usage restriction and subscription services: Operational benefits with rational users. Manuf. Serv. Oper. Manag. **10**, 429–447 (2008)
32. Reed, J.E.: The $G/GI/N$ queue in the Halfin–Whitt regime I: Infinite server queue system equations. Ann. Appl. Probab. **19**, 2211–2269 (2009)
33. Shakkottai, S., Stolyar, A.L.: Scheduling for multiple flows sharing a time-varying channel: The exponential rule. Anal. Methods Appl. Probab., Am. Math. Soc. Transl. Ser. 2 **207**, 185–202 (2002)
34. Stolyar, A.L.: Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. Ann. Appl. Probab. **14**, 1–53 (2004)

35. Stolyar, A.L.: Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. Queueing Syst., Theory Appl. **50**, 401–457 (2005)
36. Stolyar, A.L.: Optimal routing in output-queued flexible server systems. Probab. Eng. Inf. Sci. **19**, 141–189 (2005)
37. Stolyar, A.L., Viswanathan, H.: Self-organizing dynamic fractional frequency reuse in ofdma systems. In: Proceeding of INFOCOM'2008 (2008)
38. Tezcan, T., Dai, J.G.: Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Oper. Res. **58**, 94–110 (2010)
39. Whitt, W.: A diffusion approximation for the $G/GI/n/m$ queue. Oper. Res. **52**, 922–941 (2004)
40. Whitt, W.: Heavy-traffic limits for the $G/H_2^*/n/m$ queue. Math. Oper. Res. **30**, 1–27 (2005)