Adv. Appl. Prob. **47,** 251–269 (2015)
Printed in Northern Ireland
© *Applied Probability Trust* 2015

DIFFUSION-SCALE TIGHTNESS OF INVARIANT DISTRIBUTIONS OF A LARGE-SCALE FLEXIBLE SERVICE SYSTEM

A. L. STOLYAR,* *Bell Laboratories*

Abstract

A large-scale service system with multiple customer classes and multiple server pools is considered, with the mean service time depending both on the customer class and server pool. The allowed activities (routeing choices) form a tree (in the graph with vertices being both customer classes and server pools). We study the behavior of the system under a *leaf activity priority* (LAP) policy, introduced by Stolyar and Yudovina (2012). An asymptotic regime is considered, where the arrival rate of customers and number of servers in each pool tend to ∞ in proportion to a scaling parameter r, while the overall system load remains strictly subcritical. We prove tightness of diffusion-scaled (centered at the equilibrium point and scaled down by $r^{-1/2}$) invariant distributions. As a consequence, we obtain a limit interchange result: the limit of diffusion-scaled invariant distributions is equal to the invariant distribution of the limiting diffusion process.

Keywords: Many-server model; priority discipline; fluid limit; diffusion limit; tightness of invariant distribution; limit interchange

2010 Mathematics Subject Classification: Primary 60K25
Secondary 60F17

1. Introduction

Large-scale heterogeneous flexible service systems naturally arise as models of large call/contact centers [1], [8], large computer farms (used in network cloud data centers), etc. More specifically, in this paper we consider a service system with multiple customer and server types (or classes), where the arrival rate of class-i customers is Λ_i, the service rate of a class-i customer by a type-j server is μ_{ij}, and the server pool j size (the number of type-j servers) is B_j. It is important that the service rate μ_{ij} in general depends on both the customer type i and server type j. Customers waiting for service are queued, and they cannot leave the system before their service is complete. The system is 'large scale' in the sense that the input rates Λ_i and pool sizes B_j are large. More precisely, we will consider the 'many-server' asymptotic regime, in which the arrival rates Λ_i and pool sizes B_j scale up to ∞ in proportion to a scaling parameter r, i.e. $\Lambda_i = \lambda_i r$ and $B_j = \beta_j r$, while the service rates μ_{ij} remain constant. Furthermore, in this paper we assume that the (appropriately defined) system capacity exceeds the (appropriately defined) traffic load by $O(r)$, i.e. the system is *strictly subcritically* loaded. (This is different from the *Halfin–Whitt* many-server regime, in which the capacity exceeds the load by $O(\sqrt{r})$.)

Received 17 September 2013; revision received 26 March 2014.
* Current address: Lehigh University, Mohler Laboratory, 200 West Packer Avenue, Bethlehem, PA 18015, USA. Email address: stolyar@lehigh.edu

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

If, under a given control policy, the system is stable, i.e. roughly speaking, it has a stationary distribution such that the queues are stochastically bounded, then the average number of occupied servers in a stationary regime is, of course, $O(r)$. A 'good' control policy would keep the steady-state system state within $O(\sqrt{r})$ of its *equilibrium point*, which depends on the system parameters and on the policy itself. More precisely, this means that the sequence (in r) of the system stationary distributions, centered at the equilibrium point and scaled down by $r^{-1/2}$, is tight. We will refer to this property as the *$r^{1/2}$-scale, or diffusion-scale, tightness (of invariant distributions)*.

Typically, it is easy to construct a policy to ensure the diffusion-scale tightness, *if the system parameters λ_i and μ_{ij} are known in advance*. (It is natural to assume that pool sizes *are* available to any control policy.) In this case the equilibrium point can be computed in advance, and then the appropriate fractions of each input flow routed to the appropriate server pools. (See the discussion in [17].) It is much more challenging to establish this property for 'blind' policies, which do not 'know' parameters λ_i and μ_{ij}. In fact, as shown in [17], under a very natural *largest-queue, freest-server load balancing* (LQFSLB) algorithm (which is a special case of the queue-and-idleness ratio (QIR) policy in [9]), the diffusion-scale tightness does not hold in general. The LQFSLB algorithm assumes that the set of allowed 'activities' (ij) (those with $\mu_{ij} > 0$) is known (while the actual μ_{ij} values may not be) and forms a tree in the graph with vertices being customer and server types—let us refer to this as the *tree assumption*; otherwise, the LQFSLB algorithm is blind.

Another example of a blind policy (which also requires the tree assumption) is the *leaf activity priority* (LAP) algorithm, introduced in [16]. (The LAP policy is formally defined in Section 2, and its features and assumptions, including the tree assumption, are discussed in Section 2.4.) It was shown in [16], that the LAP policy ensures $r^{1/2+\varepsilon}$-scale tightness of invariant distributions for any $\varepsilon > 0$.

1.1. Main result and contributions

In this paper we prove that, in fact, the diffusion-scale (i.e. $r^{1/2}$-scale) tightness of invariant distributions holds under the LAP algorithm. We use the weaker, $r^{1/2+\varepsilon}$-scale tightness result given in [16] as a starting point, and make an additional step to obtain the diffusion-scale tightness from it. This additional step is nontrivial and is not a simple extension of the technique in [16]. More specifically, to establish the $r^{1/2+\varepsilon}$-scale tightness in [16], it suffices to work with the process under several *fluid* scalings ('standard' fluid scaling for the many-server regime, as well as *hydrodynamic* and *local-fluid* scalings). In this paper, to prove the diffusion-scale tightness, we also need to work with the process under diffusion scaling. Informally speaking, the major technical challenge here is in showing that the diffusion-scaled process is uniformly close to the corresponding limiting diffusion process on time intervals of the length increasing with r, namely, $O(\log r)$ *long intervals*.

The diffusion-scale tightness under the LAP policy in turn implies a limit interchange property: the limit of (diffusion-scaled) invariant distributions is equal to the invariant distribution of the limit (diffusion) process. Proving this limit interchange in a many-server regime is very challenging, especially for general models with multiple customer and server classes; this is due to the difficulty of establishing the diffusion-scale tightness.

Perhaps more important than establishing the tightness and limit interchange specifically for the LAP policy, is the fact that our technique seems quite generic, and may apply to other policies and/or other many-server models. Speaking very informally, combining the results and proofs in [16] and this paper gives technical 'blocks' which allow one to establish the

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

diffusion-scale tightness, providing the following two properties hold:

(a) global stability on the fluid-scale (r-scale), i.e. convergence of fluid-scaled trajectories to the equilibrium point (plus an additional, related property);

(b) local stability of the linear system in the neighborhood of the equilibrium point, i.e. the drift matrix of the limiting diffusion process has all eigenvalues with negative real parts.

Given properties (a) and (b), our approach is to show tightness in several steps, on the increasingly fine scales: fluid (r), then $r^{1/2+\varepsilon}$, then the diffusion ($r^{1/2}$) scale. We will make this discussion more specific in Section 5.

The distinctive feature of this approach, as opposed to most of the previous results on the diffusion-scale tightness for many-server models (see [6], [7], and [17]) is that it does *not* rely on a single common Lyapunov function. (Finding/constructing a common Lyapunov function is usually a difficult task, especially for the models with multiple server pools, like the model in this paper.) We remind the reader that in this paper we consider a system under strictly subcritical load, and parts of our analysis do use this assumption.

1.2. Brief literature review

A general overview of many-server models, results, and applications to call centers can be found in [1] and [8]. For control policies for general models, with multiple customer and server types, including blind policies, see, e.g. [2], [9], [14], [15], [16], [17], [18], and the references therein. Overviews of diffusion-scale tightness (and limit interchange) results for single-pool models in the many-server Halfin–Whitt regime can be found in, e.g., in [5], [6], and [7]. The diffusion-scale tightness for the LQFSLB policy, with the tree assumption and, additionally, assuming that the service rate (if nonzero) depends only on the server type, was proved in [17]. The results in [6], [7], and [17] use a common Lyapunov function; however, [5] does *not* use a Lyapunov function — it relies instead on a sample-path monotonicity/majorization property for a single-pool system under the first-come–first-served discipline.

1.3. Layout of the rest of the paper

The model and the main result are given in Sections 2 and 3, respectively. Section 4 contains the proofs. In Section 5 we discuss the results and technique.

2. The model

The model we consider is same as that in [16]. To improve the self-containment of this paper, we repeat the necessary definitions in this section.

2.1. The model and the static planning problem

Consider the system in which there are I customer classes, labeled $1, 2, \dots, I$, and J server pools, labeled $1, 2, \dots, J$. (Servers within pool j are referred to as class-j servers. Also, throughout this paper, the terms 'class' and 'type' are used interchangeably.) The sets of customer classes and server pools will be denoted by $\mathcal{I}$ and $\mathcal{J}$, respectively. We will use the indices i and i' to refer to customer classes, and j and j' to refer to server pools.

We are interested in the scaling properties of the system as it grows large. Namely, we consider a sequence of systems indexed by a scaling parameter r. As r grows, the arrival rates and the sizes of the service pools, but not the speed of service, increase. Specifically, in the rth system, customers of type i enter the system as a Poisson process of rate $\lambda_i r$, while the jth server pool has $\beta_j r$ individual servers. (All λ_i and β_j are positive parameters.) Customers may

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

be accepted for service immediately upon arrival, or enter a queue; there is a separate queue for each customer type. Customers do not abandon the system. When a customer of type i is accepted for service by a server in pool j, the service time is exponential of rate μ_{ij}; the service rate depends both on the customer type and the server type, but *not* on the scaling parameter r. If customers of type i cannot be served by servers of class j, the service rate is $\mu_{ij} = 0$.

Remark 2.1. Strictly speaking, the quantity $\beta_j r$ may not be an integer, so we should define the number of servers in pool j as, say, $\lfloor \beta_j r \rfloor$. However, the change is not substantial, and will only unnecessarily complicate the notation.

Consider the following *static planning problem* (SPP):

$$\min_{\lambda^\circ_{ij}, \rho} \rho, \tag{2.1a}$$

subject to

$$\lambda^\circ_{ij} \geq 0 \quad \text{for all } i, j, \qquad \sum_j \lambda^\circ_{ij} = \lambda_i \quad \text{for all } i, \quad \text{and} \quad \sum_i \frac{\lambda^\circ_{ij}}{\beta_j \mu_{ij}} \leq \rho \quad \text{for all } j. \tag{2.1b}$$

Throughout this paper, we will suppose that the following two assumptions about the solution to the SPP (2.1) hold.

Assumption 2.1. (Complete resource pooling.) *The SPP (2.1) has a unique optimal solution* $\{\lambda^\circ_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}, \rho$. *Define the* basic activities *to be the pairs, or edges,* (ij)*, for which* $\lambda^\circ_{ij} > 0$. *Let* $\mathcal{E}$ *be the set of basic activities. Furthermore, we assume that the unique optimal solution is such that* $\mathcal{E}$ *forms a tree in the (undirected) graph with vertices set* $\mathcal{I} \cup \mathcal{J}$.

Assumption 2.2. (Strictly subcritical load.) *The optimal solution to (2.1) has* $\rho < 1$.

Remark 2.2. Assumption 2.1 is the *complete resource pooling* (CRP) condition, which holds 'generically' in a certain sense; see [15, Theorem 2.2]. Assumption 2.2 is essential for the main result of the paper.

We assume that the basic activity tree is known in advance, and restrict our attention to the basic activities only. Namely, we assume that a type-i customer service in pool j is allowed only if $(ij) \in \mathcal{E}$. (Equivalently, we can *a priori* assume that $\mathcal{E}$ is the set of *all* possible activities, i.e. $\mu_{ij} = 0$ when $(ij) \notin \mathcal{E}$, and $\mathcal{E}$ is a tree. In this case CRP requires that all feasible activities are basic.) For a customer type i, let $\mathcal{J}(i) = \{j : (ij) \in \mathcal{E}\}$; for a server type j, let $\mathcal{C}(j) = \{i : (ij) \in \mathcal{E}\}$.

2.2. The LAP policy

We analyze the performance of the following policy, which we call the LAP policy. The first step in its definition is the assignment of priorities to customer classes and activities.

Consider the basic activity tree, and assign priorities to the edges as follows. First, we assign priorities to customer classes by iterating the following procedure:

(1) pick a leaf of the tree;

(2) if it is a customer class (rather than a server class), assign to it the highest priority that has not yet been assigned;

(3) remove the leaf from the tree.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Without loss of generality, we assume that the customer classes are numbered in order of priority (with 1 being highest). We now assign priorities to the edges of the basic activity tree by iterating the following procedure:

(1) pick the highest-priority customer class;

(2) if this customer class *is* a leaf, pick the edge going out of it, assign to this edge the highest priority that has not yet been assigned, and remove the edge together with the customer class;

(3) if this customer class is *not* a leaf then pick any edge from it to a server class leaf (such necessarily exists), assign to this edge the highest priority that has not yet been assigned, and remove the edge.

It is not hard to verify that this algorithm will successfully assign priorities to all the edges; it suffices to check that at any time the highest remaining priority customer class will have at most one outgoing edge to a nonleaf server class.

Remark 2.3. This algorithm does *not* produce a unique assignment of priorities, neither for the customer classes nor for the activities, because there may be multiple options for picking a next leaf or edge to remove in the corresponding procedures. This is not a problem, because our results hold for *any* such assignment. Different priority assignments may correspond to different equilibrium points (defined below in Section 2.3); once we have picked a particular priority assignment, there is a (unique) corresponding equilibrium point, and we will be showing steady-state tightness around that point. Furthermore, the flexibility in assigning priorities may be a useful feature in practice. For example, it is easy to specialize the above priority assignment procedure so that the lowest priority is given to any *a priori* picked activity.

We will write $(ij) < (i'j')$ to mean that activity (ij) has higher priority than activity $(i'j')$. It follows from the priority assignment algorithm that $i < i'$ (customer class i has higher priority than i') implies that $(ij) < (i'j')$. In particular, if $j = j'$, we have $(ij) < (i'j)$ if and only if $i < i'$. Without loss of generality, we will assume that the server classes are numbered so that the lowest-priority activity is (IJ).

Now we define the LAP policy itself. The policy consists of two parts: routeing and scheduling. 'Routeing' determines the destination of an arriving customer if it sees available servers of several different types. 'Scheduling' determines which waiting customer a server picks if it sees customers of several different types waiting in the queue.

Routeing. An arriving customer of type i picks an unoccupied server in the pool $j \in \mathcal{S}(i)$ such that $(ij) \leq (ij')$ for all $j' \in \mathcal{S}(i)$ with idle servers. If no server pools in $\mathcal{S}(i)$ have idle servers, the customer queues.

Scheduling. A server of type j upon completing a service picks the customer from the queue of type $i \in \mathcal{C}(j)$ such that $i \leq i'$ for all $i' \in \mathcal{S}(i)$ with $Q_{i'} > 0$. If no customer types in $\mathcal{C}(j)$ have queues, the server remains idle.

We introduce the following notation (for the system with scaling parameter r): $\Psi^r_{ij}(t)$, the number of servers of type j serving customers of type i at time t; $Q^r_i(t)$, the number of customers of type i waiting for service at time t.

Given that the system operates under the LAP policy, the process $((\Psi^r_{ij}(t),\ (ij) \in \mathcal{E}),(Q^r_i(t),\ i \in \mathcal{I}))$, $t \geq 0$, is a Markov process with countable state space.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

There are some obvious relations between system variables, which hold for each process realization: for example, for any $j \in \mathcal{S}(i)$ and any time t, either $Q_i^r(t) = 0$ or $\sum_{i'} \Psi_{i'j}^r(t) = \beta_j r$; and so on.

2.3. LAP equilibrium point

Informally speaking, the equilibrium point $((\psi_{ij}^*,\ (ij) \in \mathcal{E}), (q_i^*,\ i \in \mathcal{I}))$ is the desired operating point for the (fluid-scaled) vector $((\Psi_{ij}^r/r,\ (ij) \in \mathcal{E}), (Q_i^r/r,\ i \in \mathcal{I}))$ of occupancies and queue lengths under the LAP policy. The formal definition is given below.

Let us recursively define the quantities $\lambda_{ij} \geq 0$, which have the meaning of routeing rates, scaled down by a factor $1/r$. (These λ_{ij} are *not* equal to the λ_{ij}° which comprise the optimal solution to the SPP (2.1).) For the activity $(1j)$ with the highest priority, define either $\lambda_{1j} = \lambda_1$ and $\psi_{1j}^* = \lambda_1/\mu_{1j}$, or $\psi_{1j}^* = \beta_j$ and $\lambda_{1j} = \beta_j \mu_{1j}$, according to whichever is smaller. Replace λ_1 by $\lambda_1 - \lambda_{1j}$ and β_j by $\beta_j - \psi_{1j}^*$, and remove the edge $(1j)$ from the tree. We now proceed similarly with the remaining activities.

Formally, set

$$\lambda_{ij} = \min\left(\lambda_i - \sum_{\{j':\,(ij')<(ij)\}} \lambda_{ij'},\ \mu_{ij}\left(\beta_j - \sum_{i'<i} \frac{\lambda_{i'j}}{\mu_{i'j}}\right)\right).$$

Since the definition is in terms of higher-priority activities, this defines the $(\lambda_{ij},\ (ij) \in \mathcal{E})$ uniquely. The LAP equilibrium point is defined to be the vector

$$((\psi_{ij}^*,\ (ij) \in \mathcal{E}), (q_i^*,\ i \in \mathcal{I}))$$

given by

$$\psi_{ij}^* = \frac{\lambda_{ij}}{\mu_{ij}}, \qquad q_i^* = 0 \quad \text{for all } (ij) \in \mathcal{E},\ i \in \mathcal{I}.$$

Clearly, by the above construction, we have

$$\lambda_i = \sum_j \lambda_{ij} = \sum_j \mu_{ij} \psi_{ij}^*, \quad i \in \mathcal{I}, \qquad \sum_i \psi_{ij}^* \leq \beta_j, \quad j \in \mathcal{J}.$$

To avoid trivial complications, throughout this paper we make the following assumption.

Assumption 2.3. *If the* $(\psi_{ij},\ (ij) \in \mathcal{E})$ *are such that* $\psi_{ij} \geq 0$, $\lambda_i = \sum_j \mu_{ij}\psi_{ij}$, *and* $\sum_i \psi_{ij} \leq \beta_j$ *for all* j, *then* $\psi_{ij} > 0$ *for all* $(ij) \in \mathcal{E}$.

This assumption implies, in particular, that, for the equilibrium point, we must have $\psi_{ij}^* > 0$ for all $(ij) \in \mathcal{E}$ and, moreover, $\sum_i \psi_{ij}^* = \beta_j$ for all $j < J$ and $\sum_i \psi_{iJ}^* < \beta_J$.

Assumption 2.3 means that the system needs to employ (on average) all the activities in $\mathcal{E}$ in order to be able to handle the load. It holds, for example, whenever ρ is sufficiently close to 1.

Remark 2.4. Assumption 2.3 is technical. Our main result, the diffusion-scale tightness in Theorem 3.1, can be proved without it by following the approach presented in this paper. But, it simplifies the statements and proofs of many auxiliary results, and, thus, substantially improves the exposition.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

2.4. Discussion of the LAP policy features and assumptions

The starting point in the definition of the LAP policy is a fixed set of allowed activities $\mathcal{E}$, and the assumption that it forms a tree. How the tree $\mathcal{E}$ is determined is, in a sense, a secondary question. For example, the structure of the system itself may be such that the set of *all* possible activities is a tree $\mathcal{E}$. If not, $\mathcal{E}$ can be computed as a set of basic activities of the SPP (2.1). Solving the SPP (2.1), of course, requires knowledge of the parameters λ_i and μ_{ij}. Note, however, that, typically (in the sense specified in [15, Theorem 2.2]), a small perturbation in the parameters λ_i and μ_{ij}, while changing the SPP solution, will *not* change the set of basic activities. Therefore, computing $\mathcal{E}$ by solving the SSP (2.1) does *not* require *exact knowledge* of the system parameters, and, in many cases, approximate knowledge of the parameters may well be enough to find the 'correct' set $\mathcal{E}$.

A typical solution of the SPP (2.1) is such that the set of basic activities $\mathcal{E}$ forms a forest (graph without cycles), not necessarily a tree (which is a connected forest); moreover, within each tree component of the forest the CRP condition will hold. (Again, see [15, Theorem 2.2].) In this case, the LAP algorithm can be applied to each of the tree components separately.

Finally, we emphasize that while the objective of the SPP (2.1) is load balancing, the LAP algorithm does *not* try to balance the load of the server pools. (Hence, the values of λ_{ij} that define the equilibrium point in Section 2.3 are *not* equal to the values λ°_{ij} solving (2.1).) Instead of balancing the load, the LAP algorithm greedily tries to 'pack' customers into pools according to activity priorities. As a result, the equilibrium point is such that some of the pools are completely 'packed', while other pools (exactly one under the simplifying technical assumption, Assumption 2.3) have a nonzero fraction of idle servers.

2.5. Basic notation

The vector $(\xi_i,\ i \in \mathcal{I})$, where ξ can be any symbol, is often written as (ξ_i); similarly, $(\xi_j,\ j \in \mathcal{J}) = (\xi_j)$ and $(\xi_{ij},\ (ij) \in \mathcal{E}) = (\xi_{ij})$. Furthermore, we often use the notation (η_{ij}, ξ_i) to mean $((\eta_{ij},\ (ij) \in \mathcal{E}), (\xi_i,\ i \in \mathcal{I}))$, and similar notation as well. Unless specified otherwise, $\sum_i \xi_{ij} = \sum_{i \in \mathcal{C}(j)} \xi_{ij}$ and $\sum_j \xi_{ij} = \sum_{j \in \mathcal{S}(i)} \xi_{ij}$. For functions (or random processes) $(\xi(t),\ t \geq 0)$, we often write $\xi(\cdot)$. (And similarly for functions with domains different from $[0, \infty)$.) So, for example, $(\xi_i(\cdot))$ signifies $((\xi_i(t),\ i \in \mathcal{I}),\ t \geq 0)$.

In the Euclidean space $\mathbb{R}^d$ (with appropriate dimension d), $|x|$ denotes the standard Euclidean norm of vector x; the symbol '$\to$' denotes ordinary convergence; we simply write 0 for a zero vector. We always consider the Borel σ-algebra on $\mathbb{R}^d$ when it is viewed as a measurable space. The symbol '$\stackrel{\mathrm{w}}{\to}$' denotes weak convergence of probability distributions. We use 'w.p.1' to mean 'with probability 1'. We will consider a sequence of systems indexed by scaling parameter r increasing to ∞, and will use the abbreviation 'w.p.1-l.r' as shorthand for 'w.p.1 for all sufficiently large r'.

We denote by $\mathrm{dist}[\xi]$ the distribution of a random element ξ, and by $\mathrm{inv}[\xi(\cdot)]$ the stationary distribution of a Markov process $\xi(\cdot)$ (it will be unique in all cases that we consider).

3. Main result

It was shown in [16, Theorem 10] that, if the system under the LAP policy is strictly subcritically loaded, i.e. $\rho < 1$, then, for all large r, the Markov process $(\Psi^r_{ij}(\cdot),\ Q^r_i(\cdot))$ is positive recurrent, has unique stationary distribution $\mathrm{inv}[(\Psi^r_{ij}(\cdot),\ Q^r_i(\cdot))]$ and, moreover, the sequence of stationary distributions is tight on the scale $r^{1/2+\varepsilon}$ with any $\varepsilon > 0$. In this paper we strengthen this result by showing that the invariant distributions are, in fact, tight on the diffusion $(r^{1/2})$ scale. This is, of course, the strongest possible tightness result for the system

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

and the asymptotic regime in this paper. As a consequence, we obtain a limit interchange result: the limit of diffusion-scaled invariant distributions is equal to the invariant distribution of the limiting diffusion process.

Denote by $Z_j^r(t) = \sum_i \Psi_{ij}^r(t) - r\sum_i \psi_{ij}^*$ the 'idleness' of pool j. Recall that, for each $j < J$, $\sum_i \psi_{ij}^* = \beta_j$ and, therefore, $Z_j^r(t) \le 0$. Let L' be the linear mapping (defined in [16, Section 5.2]), which takes a vector (ξ_i) with real components into the vector (η_{ij}), uniquely solving

$$\sum_j \eta_{ij} = \xi_i \quad \text{for all } i, \qquad \sum_i \eta_{ij} = 0, \quad j < J. \tag{3.1}$$

Theorem 3.1. *Consider the sequence of systems under the LAP policy, in the scaling regime and under the assumptions specified in Section 2, with $\rho < 1$. Then the sequence of diffusion-scaled stationary distributions,* $\mathrm{inv}[r^{-1/2}(\Psi_{ij}^r(\cdot) - \psi_{ij}^* r,\ Q_i^r(\cdot))]$, *is tight. Moreover,*

$$\mathrm{inv}[r^{-1/2}(\Psi_{ij}^r(\cdot) - \psi_{ij}^* r)] \xrightarrow{\mathrm{w}} \mathrm{inv}[(\breve{\Psi}_{ij}(\cdot))] \quad \text{as } r \to \infty, \tag{3.2}$$

where $(\breve{\Psi}_{ij}(\cdot))$ is the diffusion process, defined by the stochastic differential equation

$$\mathrm{d}(\breve{\Psi}_{ij}(t)) = L'\,\mathrm{d}(\sqrt{\lambda_i} B_i^{(a)}(t)) - L'\,\mathrm{d}\left(\sum_j \sqrt{\mu_{ij}\psi_{ij}^*}\, B_{ij}^{(s)}(t)\right) - L'\left(\sum_j \mu_{ij}\breve{\Psi}_{ij}(t)\right)\mathrm{d}t, \tag{3.3}$$

with all $B_i^{(a)}(\cdot)$ and $B_{ij}^{(s)}(\cdot)$ being independent standard Brownian motions. For any $\nu > 0$,

$$\mathrm{inv}[r^{-\nu}((Q_i^r(\cdot)), (Z_j^r(\cdot),\ j < J))] \xrightarrow{\mathrm{w}} \mathrm{dist}[0] \quad \text{as } r \to \infty, \tag{3.4}$$

where dist[0] *is the Dirac measure concentrated at the zero vector.*

Remark 3.1. From (3.4) it follows that the distributions of all queue lengths, and of the idlenesses in pools $j \ne J$, are tight on the scale r^ν for any $\nu > 0$. As we will see, this fact is an 'ingredient' of the proof of diffusion-scale tightness and (3.2). Also, it is not surprising, and is a consequence of the priority discipline and (for the queues) of the strict subcriticality, $\rho < 1$. As discussed in Section 2.4, the LAP policy tries to 'pack' server pools according to the activity priority order. As a result, when the idleness in a pool $j \ne J$ is nonzero then, roughly speaking, the arrival rate into the pool exceeds the departure rate by a factor greater than 1; similarly, the departure rate from any nonzero queue exceeds the arrival rate by a factor greater than 1. Therefore, it is natural to expect that an even stronger property than (3.4) holds, namely the sequence of unscaled stationary distributions $\mathrm{inv}[((Q_i^r(\cdot)), (Z_j^r(\cdot),\ j < J))]$ is tight. In this paper we do not pursue the proof of this fact, because establishing diffusion-scale tightness and (3.2) are our main goals.

4. Proof of Theorem 3.1

In the rest of the paper we will use the following additional notation for the system variables. For a system with parameter r, let $X_i^r(t) = \sum_j \Psi_{ij}^r(t) + Q_i^r(t)$ be the total number of type-i customers in the system at time t; let $A_i^r(t)$ be the total number of type-i customer exogenous arrivals into the system in the interval $[0, t]$; let $D_{ij}^r(t)$ be the total number of type-i customers that completed the service in pool j (and departed the system) in the interval $[0, t]$; finally, we will use the short notation $F^r(t) = (\Psi_{ij}^r(t) - \psi_{ij}^* r,\ Q_i^r(t))$.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

We can, and do, assume that a random realization of the system with parameter r is determined by its initial state and realizations of 'driving' unit-rate, mutually independent, Poisson processes $\Pi_i^{(a)}(\cdot)$, $i \in \mathcal{I}$, and $\Pi_{ij}^{(s)}(\cdot)$, $(ij) \in \mathcal{E}$, as follows:

$$A_i^r(t) = \Pi_i^{(a)}(\lambda_i r t), \qquad D_{ij}^r(t) = \Pi_{ij}^{(s)}\left(\mu_{ij} \int_0^t \Psi_{ij}^r(u)\,\mathrm{d}u\right);$$

the driving Poisson processes are common for all r. It is easy to see that, given the LAP policy, with probability 1, the realizations of these driving processes (along with the initial state) uniquely define the system process realization.

Finally, the diffusion-scaled variables are defined as

$$(\hat{\Psi}_{ij}^r(t), \hat{Q}_i^r(t)) = r^{-1/2}(\Psi_{ij}^r(t) - \psi_{ij}^* r, Q_i^r(t)),$$
$$\hat{X}_i^r(t) = r^{-1/2}[X_i^r(t) - \sum_j \psi_{ij}^* r], \qquad \hat{Z}_j^r(t) = r^{-1/2} Z_j^r(t).$$

Throughout this section, we will use the following strong approximation of Poisson processes; see, e.g. [3, Chapters 1 and 2].

Proposition 4.1. *A unit-rate Poisson process $\Pi(\cdot)$ and a standard Brownian motion $W(\cdot)$ can be constructed on a common probability space in such a way that the following holds for some fixed positive constants C_1, C_2, C_3: for all $T > 1$ and all $u \geq 0$,*

$$\mathbb{P}\Big(\sup_{0 \leq t \leq T} |\Pi(t) - t - W(t)| \geq C_1 \log T + u\Big) \leq C_2 \mathrm{e}^{-C_3 u}.$$

We will also need the following form of a functional strong law of large numbers for a Poisson process. It is obtained using standard large deviation estimates, e.g. analogously to the approach followed in the proof of [13, Lemma 4.3].

Proposition 4.2. *For a unit-rate Poisson process $\Pi(\cdot)$, the following holds with probability* 1. *For any $\nu \in (0, 1)$ and any $c > 1$, uniformly in $t_1, t_2 \in [0, r^c]$ such that $t_2 - t_1 \geq r^\nu$,*

$$\frac{\Pi(t_2) - \Pi(t_1)}{t_2 - t_1} \to 1 \quad \text{as } r \to \infty.$$

Throughout this paper, we will use Proposition 4.2 with arbitrary fixed $c > 1$: this ensures that, for any fixed $T > 0$, the interval $[0, Tr \log r]$ is contained within $[0, r^c]$ for all large r. Proposition 4.2, in particular, immediately implies the following upper bound on the rate at which system variables can change. There exists $C > 0$ such that, for any $\nu \in (0, 1)$ and any $\alpha > 0$, w.p.1-l.r, uniformly in $t_1, t_2 \in [0, r^{c-1}]$ such that $t_2 - t_1 \geq \alpha r^\nu / r$,

$$\max_{t \in [t_1, t_2]} |Q_i^r(t) - Q_i^r(t_1)| < C(t_2 - t_1)r \quad \text{for all } i, \tag{4.1}$$

and similarly for $\Psi_{ij}^r(\cdot)$ for all (ij), $Z_j^r(\cdot)$ for all j, and $F^r(\cdot)$. Indeed, in a system with parameter r, the customer arrival and departure events occur, 'at most', as

$$\Pi\left(\left[\sum_i \lambda_i + \left(\sum_j \beta_j\right) \max_{(ij)} \mu_{ij}\right] r\right),$$

where $\Pi(\cdot)$ is a unit-rate Poisson process; therefore, the condition $t_2 - t_1 \geq \alpha r^\nu / r$ in the rth system guarantees that the interval $[t_1, t_2]$ corresponds to at least an $O([t_2 - t_1]r) = O(r^\nu)$ long time interval for $\Pi(\cdot)$, and then Proposition 4.2 applies.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Lemma 4.1. *There exists $T > 0$ such that, for any $\varepsilon \in (0, \frac{1}{2})$, the following holds. For any $\delta > 0$, there exists a sufficiently large $C_7 > 0$ such that, uniformly on all sufficiently large r and all $|F^r(0)| \le g(r) = r^{1/2+\varepsilon}$, the probability of $|F^r(t)| \le C_7 r^{1/2}$ occurring within $[0, \varepsilon T \log r]$ is at least $1 - \delta$.*

Proof. The proof is by contradiction. If the lemma does not hold then there exists a function $g_*(r)$ such that $g_*(r)/r^{1/2} \uparrow \infty$ and the probability of starting from $|F^r(0)| \le g(r)$ and not hitting $|F^r(t)| \le g_*(r)$ within time $\varepsilon T \log r$ does not vanish. We will prove that it has to vanish, thus establishing a contradiction.

Define $h(r) = |F^r(0)|$. We now specify the choice of T. We note that all the results in Sections 5.2–5.3 of [16], concerning hydrodynamic and local-fluid limits, hold as is for any function $h(r)$ such that $h(r)/r^{1/2} \to \infty$. (The condition $h(r) \ge r^{1/2+\varepsilon}$ was used in [16] only when the results of Sections 5.2–5.3 therein were *applied*.) Then, by Corollary 25 and Condition (23) of [16], we can, and do, choose a sufficiently large $T > 0$ such that the conditions

$$\max_{t\in[0,T]} |\Pi_i^{(a)}(\lambda_i r t) - \lambda_i r t| \le \delta_2 h(r) \quad \text{for all } i, \tag{4.2}$$

and similarly for $\Pi_{ij}^{(s)}$ for all (ij), with sufficiently small fixed $\delta_2 > 0$, guarantee that condition $g(r) \ge h(r) = |F^r(0)| \ge g_*(r)$ implies that $|F^r|$ decreases at least by a factor $K > 1$ in $[0, T]$. Let us see how the probability of (4.2) depends on $h(r)$, or, more conveniently, on $h_1(r) = h(r)/r^{1/2}$. (Note that $h_1(r) \uparrow \infty$ when $h(r) \ge g_*(r)$.)

Now we will use Proposition 4.1. In its statement let us replace Π with $\Pi_i^{(a)}$, and t with $\lambda_i r t$, and T with $\lambda_i r T$, and make u a function of r, say $u = r^{1/4}$. Then, with probability at least $1 - C_2 \mathrm{e}^{-C_3 r^{1/4}}$,

$$\mathbb{P}\Big\{ \max_{t\in[0,T]} |\Pi_i^{(a)}(\lambda_i r t) - \lambda_i r t| \le \max_{t\in[0,T]} |W(\lambda_i r t)| + C_1 \log(\lambda_i r T) + r^{1/4} \Big\} \ge 1 - C_2 \mathrm{e}^{-C_3 r^{1/4}},$$

where C_1, C_2, and C_3 are universal constants (from the statement of Proposition 4.1). Next, observe that $(W(\lambda_i r t)/h(r),\ t \ge 0)$, where $W(\cdot)$ is a standard Brownian motion, is equal in distribution to $(\sqrt{\lambda_i} W(t)/h_1(r), t \ge 0)$. Therefore,

$$\mathbb{P}\Big\{ \max_{t\in[0,T]} |W(\lambda_i r t)| \le \tfrac{1}{2}\delta_2 h(r) \Big\} \ge 1 - C_4 \mathrm{e}^{-C_5 (h_1(r))^2},$$

where the positive constants C_4 and C_5 depend on δ_2 and T (and the system parameters). We conclude that the probability of (4.2) is lower bounded by

$$1 - C_2 \mathrm{e}^{-C_3 r^{1/4}} - C_4 \mathrm{e}^{-C_5 (h_1(r))^2}.$$

Define

$$p_i = \mathbb{P}\{|F^r(t)| \le g_*(r) \text{ for some } t \in [0, iT] \mid |F^r(0)| \le K^i g_*(r)\}, \qquad i = 0, 1, 2, \ldots.$$

We can write, for any $i \ge 1$,

$$p_i \ge \left[1 - C_2 \mathrm{e}^{-C_3 r^{1/4}} - C_4 \exp\left\{ -C_5 K^{2i} \left(\frac{g_*(r)}{r^{1/2}} \right)^2 \right\} \right] p_{i-1}.$$

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

We are interested in p_k with $k = \varepsilon \log r$, which is lower bounded as

$$p_k \geq \prod_{i=1}^{k}\left[1 - C_2 \mathrm{e}^{-C_3 r^{1/4}} - C_4 \exp\left\{-C_5 K^{2i} \frac{g_*(r)^2}{r}\right\}\right]$$
$$\geq 1 - \sum_{i=1}^{k}\left[C_2 \mathrm{e}^{-C_3 r^{1/4}} + C_4 \exp\left\{-C_5 K^{2i} \frac{g_*(r)^2}{r}\right\}\right].$$

The sum vanishes as $r \to \infty$, and so is $1 - p_k$.

The key part for the remainder of the proof of Theorem 3.1 is to show that, informally speaking, if the process 'hits' the set $\{|F^r| \leq C_7 r^{1/2}\}$ anywhere within $[0, \varepsilon T \log r]$ then it stays 'on the $r^{1/2}$-scale' at time $\varepsilon T \log r$ as well. To achieve this, we will exploit the closeness of the diffusion-scaled process to the diffusion limit on a $\varepsilon T \log r$ long interval (i.e. with length increasing with r), when ε is small enough. This will be formalized in Lemma 4.3 below, but to apply it we require an additional step, given by the following lemma.

Lemma 4.2. *There exist $T_8 > 0$ and $C_8 > 0$ such that the following holds. For any fixed $C_9 > 0$, $\delta_9 > 0$, and $\nu_9 \in (0, \frac{1}{2})$, uniformly on initial states $|F^r(0)| \leq C_9 r^{1/2}$, as $r \to \infty$,*

$$\mathbb{P}\left\{\max_{t\in[0,T_8 C_9 r^{-1/2}]} |F^r(t)| \leq C_8 C_9 r^{1/2}\right\} \to 1, \tag{4.3}$$

$$\mathbb{P}\{\textit{there exists } t \in [0, T_8 C_9 r^{-1/2}]\colon |(Q_i^r(t))| + |(Z_j^r(t),\ j < J)| \leq \delta_9 r^{\nu_9}\} \to 1. \tag{4.4}$$

We will use this lemma (and Lemma 4.4 below) with $0 < \nu_9 < \frac{1}{4}$.

Proof of Lemma 4.2. Let us first discuss the basic intuition behind the result, which is extremely simple, and will be useful not only for this proof, but also for some other proofs in this paper. Within a fixed $O(r^{-1/2})$ time, $F^r(t)$ can change at most by $O(r^{1/2})$ (see (4.1)) and, therefore, for all (ij), $\Psi_{ij}^r(t)/[\psi_{ij}^* r] \approx 1$ holds. Now, consider the highest-priority activity $(1j)$. Suppose that customer class 1 is a leaf. Then there must exist at least one other activity (ij), associated with the same pool j. The arrival rate of type-1 customers is $\lambda_1 r = \mu_{1j}\psi_{1j}^* r$, while the total service completion rate at pool j is at least $\mu_{1j}\Psi_{1j}^r(t) + \mu_{ij}\Psi_{ij}^r(t) \approx \mu_{1j}\psi_{1j}^* r + \mu_{ij}\psi_{ij}^* r = \lambda_1 r + \mu_{ij}\psi_{ij}^* r$. This means that, since a type-1 customer has the highest priority at pool j, the queue $Q_1^r(t)$, when nonzero, 'drains' at a rate of at least $O(r)$, 'hits' the r^{ν_9}-scale within $O(r^{-1/2})$ time and 'stays there'. Now suppose that customer class 1 is not a leaf. Then pool j must be a leaf, i.e. it serves type-1 customers exclusively, $\psi_{1j}^* = \beta_j$, and there must be at least one other activity $(1m)$, associated with type-1 customers, implying that $\lambda_1 \geq \mu_{1j}\psi_{1j}^* + \mu_{1m}\psi_{1m}^* > \mu_{1j}\beta_j$. The difference between a type-1 arrival rate and the rate at which type-1 customers are served by pool j is at least $[\lambda_1 - \mu_{1j}\beta_j]r = O(r)$. This means that the idleness $|Z_j^r(t)|$, when nonzero, decreases at a rate of at least $O(r)$, 'hits' r^{ν}-scale within $O(r^{-1/2})$ time and 'stays there'. We 'remove' activity $(1j)$ from the activity tree. The argument proceeds by considering all activities (ij) in sequence, from the highest to lowest priority; at each step either $Q_i^r(t)$ or $Z_j^r(t)$ is 'eliminated', depending on i or j, respectively, being a leaf of the current activity tree. The exception is when $j = J$ is the pool serving the lowest-priority activity (IJ): in this case $Z_J^r(t)$ is *not* eliminated. We now proceed with a sketch of a formal argument; details can be easily 'recovered' by the reader.

The proof of (4.3) is an immediate consequence of (4.1). Indeed, for any $T_8 > 0$, w.p.1-l.r, the value of $|F^r(t) - F^r(0)|$ with $t \in [0, T_8 C_9 r^{-1/2}]$ is upper bounded by $C T_8 C_9 r^{1/2}$. So, for any chosen T_8, we can choose $C_8 > 1 + C T_8$.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Property (4.3), in particular, means that, for any fixed $T_8 > 0$, w.p.1, for any $(ij) \in \mathcal{E}$, uniformly in $t \in [0, T_8 C_9 r^{-1/2}]$, we have

$$\frac{\Psi_{ij}^r(t)}{[\psi_{ij}^* r]} \to 1. \tag{4.5}$$

To prove (4.4), we consider and 'eliminate' activities one by one, in the order of their priority. The choice of T_8 will be made later; for now, it is a fixed constant, and we consider the process on the interval $[0, T_8 C_9 r^{-1/2}]$. We start with the highest-priority activity $(1j)$. Suppose first that customer class-1 is a leaf of the activity tree. (In this case, $\mathcal{C}(j)$ necessarily contains at least one customer class in addition to 1.) Consider any $0 < C_1 < \sum_{i \neq 1} \mu_{ij} \psi_{ij}^*$. Then, for any $\delta > 0$, there exists a sufficiently small $\delta_1 > 0$, such that, w.p.1-l.r, uniformly in $t \in [0, T_8 C_9 r^{-1/2}]$, condition $Q_1^r(t) \geq \delta r^{\nu_9}$ implies that $Q_1^r(t + \delta_1 r^{\nu_9}/r) - Q_1^r(t) < -C_1 \delta_1 r^{\nu_9}$ (because *all* departures from pool j are replaced by class 1 customers from the queue), and, for any $Q_1^r(t)$, we have (by (4.1)) $\max_{\tau \in [0, \delta_1 r^{\nu_9}/r]} Q_1^r(t + \tau) < Q_1^r(t) + C\delta_1 r^{\nu_9}$. This means that, w.p.1.,

$$\max_{t \in [T', T_8 C_9 r^{-1/2}]} Q_1^r(t) \leq (\delta + C\delta_1) r^{\nu_9},$$

where $T' = 2(1/C_1) C_9 r^{-1/2}$. Note that this holds for any δ and the corresponding δ_1, both of which can be chosen arbitrarily small. We conclude that, w.p.1.,

$$\max_{t \in [T', T_8 C_9 r^{-1/2}]} \frac{Q_1^r(t)}{r^{\nu_9}} \to 0. \tag{4.6}$$

This means, in particular, that in $[T', T_8 C_9 r^{-1/2}]$ the number of exogenous class-1 arrivals matches the number of class-1 customers entering service, up to $o(r^{\nu_9})$ quantities. Formally, the following holds. Denote by $\Xi_{ij}^r(t_1, t_2)$ the number of type-i customers that enter service in pool j in the time interval $(t_1, t_2]$. For any fixed $\delta_1 > 0$, w.p.1, uniformly in $t_1, t_2 \in [T', T_8 C_9 r^{-1/2}]$ such that $t_2 - t_1 \geq \delta_1 r^{\nu_9}/r$,

$$\frac{\Xi_{1j}^r(t_1, t_2)}{[\lambda_1 r (t_2 - t_1)]} \to 1. \tag{4.7}$$

Finally, note that, again by (4.1), w.p.1-l.r, at time T', $|F^r|$ is at most by a constant factor (depending on C_1) greater than $C_9 r^{1/2}$. Our conclusions about the $(1j)$ activity can be informally summarized as follows: within a time $T' = 2(1/C_1) C_9 r^{-1/2}$, proportional to $C_9 r^{-1/2}$, the value of $Q_1^r(t)/r^{\nu_9}$ 'drains to 0' and 'stays there' (in the sense of (4.6)) until the end of the interval $[0, T_8 C_9 r^{-1/2}]$; moreover, on the interval $[T', T_8 C_9 r^{-1/2}]$, the rate at which server pool j 'takes' type-1 customers is 'equal' (in the sense of (4.7)) to their arrival rate $\lambda_1 r$. Therefore, from time T' on, we can 'eliminate' and 'ignore' activity $(1j)$ in the sense that we know that the rate at which pool j can take for service customers of types other than 1 is 'at least' $[\sum_{i \neq 1} \mu_{ij} \psi_{ij}^*] r$. More precisely, if we denote by $S_{(\neq 1), j}^r(t_1, t_2)$ the number of times on the interval $(t_1, t_2]$ when a service completion by a server in pool j was *not* followed (either immediately or after some idle period) by taking a type-1 customer for service, then the following holds: for any fixed $\delta_1 > 0$, w.p.1, uniformly in $t_1, t_2 \in [T', T_8 C_9 r^{-1/2}]$ such that $t_2 - t_1 \geq \delta_1 r^{\nu_9}/r$,

$$\frac{S_{(\neq 1), j}^r(t_1, t_2)}{[\sum_{i \neq 1} \mu_{ij} \psi_{ij}^*] r (t_2 - t_1)} \to 1. \tag{4.8}$$

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Moreover, $|F^r(T')|$ is at most by a constant factor greater than $C_9 r^{1/2}$, which is the upper bound on $|F^r(0)|$.

Now suppose that customer class 1 is not a leaf. Then necessarily poll j is a leaf and $j < J$. In this case, by looking at the evolution of idleness $Z_j^r(t)$, and using similar arguments, we can show that, again, within a time proportional to $C_9 r^{-1/2}$, let us call it T'', the value of $Z_j^r(t)/r^{\nu_9}$ 'drains to 0' and 'stays there' (in the sense analogous to (4.6)) until the end of the interval $[0, T_8 C_9 r^{-1/2}]$; this in turn means that the rate at which type-1 customers will enter pool j on the interval $[T'', T_8 C_9 r^{-1/2}]$ will be 'equal' (in the sense analogous to (4.7)) to $\mu_{1j}\beta_j r$. And again, w.p.1-l.r, $|F^r(T'')|$ is at most by a constant factor greater than $C_9 r^{1/2}$. Therefore, from time T'' on, we can 'eliminate' activity $(1j)$ in the sense that we can 'ignore' pool j and 'assume' that the arrival rate of type-1 customers in the rest of the system is 'equal' to $\lambda_1 r - \mu_{1j}\beta_j r$. (The latter is in the sense analogous to (4.8), but where we count the type-1 arrivals that were *not* taken for service on the corresponding interval $(t_1, t_2]$.)

We can proceed to 'eliminate' the second highest-priority activity, and so on. The total time for all scaled queues $Q_i^r(t)/r^{\nu_9}$ and all idlenesses $Z_j^r(t)/r^{\nu_9}$, $j < J$, to 'drain to 0' will be proportional to $C_9 r^{-1/2}$, say $T_8' C_9 r^{-1/2}$. We then choose $T_8 > T_8'$. We omit further details, except to emphasize again that property (4.4) does *not* include 'idleness' Z_J^r for the pool J serving the lowest-priority activity (IJ).

Lemma 4.3. *Let $T > 0$ be fixed. For a sufficiently small $\varepsilon > 0$, the following holds. For any fixed $C_{11} > 0$, $\delta_9 > 0$, and $\nu_9 \in (0, \frac{1}{4})$, uniformly on initial states satisfying $|F^r(0)| \le C_{11} r^{1/2}$ and $|(Q_i^r(0))| + |(Z_j^r(0),\ j < J)| \le \delta_9 r^{\nu_9}$,*

$$\max_{t \in [0, \varepsilon T \log r]} |(\hat{\Psi}_{ij}^r(t)) - (\breve{\Psi}_{ij}^r(t))| \Longrightarrow 0, \tag{4.9}$$

where $(\breve{\Psi}_{ij}^r(\cdot))$ is a (strongly) unique strong solution of the stochastic integral equation (4.19) (constructed on a common probability space with $(\hat{\Psi}_{ij}^r(\cdot))$), with the initial state $(\breve{\Psi}_{ij}^r(0)) = (\hat{\Psi}_{ij}^r(0))$.

To prove this lemma we will need a series of auxiliary results.

Lemma 4.4. *There exists $C_{10} > 0$ such that the following holds for any $\varepsilon > 0$, $T > 0$, $C_{11} > 0$, $\delta_9 > 0$, and $\nu_9 \in (0, \frac{1}{2})$. As $r \to \infty$, uniformly on all the initial states such that $|F^r(0)| \le C_{11} r^{1/2}$ and $|(Q_i^r(0))| + |(Z_j^r(0),\ j < J)| \le \delta_9 r^{\nu_9}$, we have*

$$\mathbb{P}\left\{ \max_{t \in [0, T \log r]} |F^r(t)| \le r^{1/2+\varepsilon} \right\} \to 1, \tag{4.10}$$

$$\mathbb{P}\left\{ \max_{t \in [0, T \log r]} [|(Q_i^r(t))| + |(Z_j^r(t),\ j < J)|] \le C_{10} \delta_9 r^{\nu_9} \right\} \to 1. \tag{4.11}$$

Proof. The proof of property (4.10) is already contained in the proof of [16, Theorem 10(ii)]. Indeed, that proof considers the process on the interval $[0, T \log r]$ and shows that, starting with $|F^r(0)| = o(r)$, w.p.1-l.r, $|F^r(t)|$ 'hits' the $r^{1/2+\varepsilon}$-scale somewhere within $[0, T \log r]$, and then 'stays' on this scale until the end of the interval. In our case, $|F^r(0)|$ is already on the $r^{1/2+\varepsilon}$-scale, and so the process, w.p.1-l.r, stays on it for the entire interval $[0, T \log r]$.

Given (4.10), to prove (4.11), we can 'reuse' the proof of (4.4) of Lemma 4.2. In that proof we showed that starting with $|F^r(0)| = O(r^{1/2})$, w.p.1-l.r, the quantity $|(Q_i^r(t))| + |(Z_j^r(t),\ j < J)|$ 'hits the r^{ν_9}-scale' within an $O(r^{-1/2})$ long time interval and 'stays there' until the end of that time interval. (See (4.6).) In our case, the initial state is already such that $|(Q_i^r(0))| + |(Z_j^r(0),\ j < J)| = O(r^{\nu_9})$, and, therefore, this quantity stays $O(r^{\nu_9})$ on the entire interval.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

The fact that here we consider a much longer interval, namely, $O(\log r)$ as opposed to $O(r^{-1/2})$, is immaterial, because (4.10), and therefore (4.5), hold on the entire interval and $r \log r = o(r^c)$ (so we can use Proposition 4.2). We omit further details.

Proposition 4.3. *There exists a set of independent standard Brownian motions, $W_i^{(a)}(\cdot)$ and $W_{ij}^{(s)}(\cdot)$, constructed on the same probability space as the set of Poisson processes $\Pi_i^{(a)}(\cdot)$ and $\Pi_{ij}^{(s)}(\cdot)$ such that the following holds. For any fixed $T > 0$, as $r \to \infty$, for each i,*

$$\sup_{0 \le t \le T \log r} r^{-1/4} |\Pi_i^{(a)}(rt) - rt - W_i^{(a)}(rt)| \to 0 \quad w.p.1,$$

and, for each $(ij) \in \mathcal{E}$,

$$\sup_{0 \le t \le T \log r} r^{-1/4} |\Pi_{ij}^{(s)}(rt) - rt - W_{ij}^{(s)}(rt)| \to 0 \quad w.p.1.$$

Proof. This follows from Proposition 4.1: in its statement we replace t with rt, T with $rT \log r$, and u with $r^{1/8}$.

Proposition 4.4. *Consider any sequence of standard Brownian motions, $B_1(\cdot), B_2(\cdot), \ldots$, defined on a common probability space. (They may be dependent.) Let $T > 0$, $C_{12} > 0$, and $\varepsilon \in (0, \frac{1}{4})$ be fixed. Then, w.p.1-l.r, conditions $t_1, t_2 \in [0, T \log r]$ and $|t_2 - t_1| \le C_{12} r^{-1/2+\varepsilon}$ imply that $|B_r(t_2) - B_r(t_1)| < r^{-1/8}$.*

Proof. This proof follows from the basic properties of Brownian motion. Fix $\varepsilon' \in (\frac{1}{8}, \frac{1}{4} - \frac{1}{2}\varepsilon)$. Then, for some fixed $C_{13} > 0$,

$$\mathbb{P}\left\{ \max_{t \in [0, C_{12} r^{-1/2+\varepsilon}]} |B_r(t) - B_r(0)| \ge r^{-\varepsilon'} \right\} \le \exp\left\{ -C_{13} \left[\frac{r^{-\varepsilon'}}{r^{-1/4+\varepsilon/2}} \right]^2 \right\}. \tag{4.12}$$

This probability decays very fast with r. We divide the interval $[0, T \log r]$ into (a polynomial in r number of) $C_{12} r^{-1/2+\varepsilon}$ long subintervals, and use the above probability estimate for each of them; by the Borel–Cantelli lemma, w.p.1-l.r, the event (analogous to the event) in (4.12) will not hold for any of the subintervals. The result follows.

Proof of Lemma 4.3. Suppose that, for each r, the initial state is fixed so that it satisfies the conditions of the lemma. Suppose that the process, for any r, is driven by a common set of Poisson processes, and associated Brownian motions constructed on the same probability space, as specified in Proposition 4.3. It will suffice to show that, for any subsequence of r, there exists a further subsequence, along which the conclusion of the lemma holds. So, let us fix an arbitrary subsequence of r. We fix any $\nu_9 \in (0, \frac{1}{4})$ and choose a further subsequence of r, with r increasing sufficiently fast, so, w.p.1-l.r, the events in (4.10) and (4.11) hold.

Let

$$\hat{A}_i^r(t) = r^{-1/2}[\Pi_i^{(a)}(\lambda_i rt) - \lambda_i rt], \qquad \hat{W}_i^{(a),r}(t) = r^{-1/2} W_i^{(a)}(\lambda_i rt),$$
$$\hat{D}_{ij}^r(t) = r^{-1/2}[\Pi_{ij}^{(s)}(\mu_{ij}\psi_{ij}^* rt) - \mu_{ij}\psi_{ij}^* rt], \qquad \hat{W}_{ij}^{(s),r}(t) = r^{-1/2} W_{ij}^{(s)}(\mu_{ij}\psi_{ij}^* rt).$$

Note that, *for any* r, the *law* of $((\hat{W}_i^{(a),r}(\cdot)), (\hat{W}_{ij}^{(s),r}(\cdot)))$ is equal to that of $((\sqrt{\lambda_i} B_i^{(a)}(\cdot)), (\sqrt{\mu_{ij}\psi_{ij}^*} B_{ij}^{(s)}(\cdot)))$, where all $B_i^{(a)}(\cdot)$ and $B_{ij}^{(s)}(\cdot)$ are independent standard Brownian motions.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Using a standard sample path representation (see, e.g. [12]), we can write, for each i, and all $t \geq 0$,

$$X_i^r(t) = X_i^r(0) + A_i^r(t) - \sum_j D_{ij}^r\left(\mu_{ij}\int_0^t \Psi_{ij}^r(s)\,\mathrm{d}s\right). \tag{4.13}$$

Switching, again in a standard way, to diffusion-scaled variables and to a (I-dimensional) vector form, we rewrite (4.13) as

$$\begin{aligned}(\hat{X}_i^r(t)) = (\hat{X}_i^r(0)) + (\hat{A}_i^r(t)) - \left(\sum_j \hat{D}_{ij}^r\left((\psi_{ij}^* rt)^{-1}\left[\int_0^t \Psi_{ij}^r(s)\,\mathrm{d}s\right]t\right)\right)\\ - \left(\sum_j \mu_{ij}\int_0^t \hat{\Psi}_{ij}^r(s)\,\mathrm{d}s\right).\end{aligned} \tag{4.14}$$

Suppose that $\varepsilon \in (0, \frac{1}{4})$ (so that we can apply Proposition 4.4 later). We will make the choice of ε more specific below.

We claim that, w.p.1-l.r, the following properties hold uniformly for $t \in [0, T\log r]$:

$$|\hat{A}_i^r(t) - \hat{W}_i^{(a),r}(t)| < r^{-1/4} \quad \text{for all } i, \qquad |\hat{D}_{ij}^r(t) - \hat{W}_{ij}^{(s),r}(t)| < r^{-1/4} \quad \text{for all } (ij), \tag{4.15}$$

$$\left|(\psi_{ij}^* rt)^{-1}\left[\int_0^t \Psi_{ij}^r(s)\,\mathrm{d}s\right]t - t\right| \leq r^{-1/2+\varepsilon}\varepsilon T\log r < r^{-1/2+\varepsilon'} \quad \text{for all } (ij), \tag{4.16}$$

$$|L'(\hat{X}_i^r(t)) - (\hat{\Psi}_{ij}^r(t))| < r^{-1/4}. \tag{4.17}$$

Here ε' is a fixed number within $(\varepsilon, \frac{1}{4})$ and the linear mapping L' is defined by (3.1). (L' was defined in [16, Section 5.2]. It maps a vector of *centered* customer quantities onto the vector of *centered* occupancies, assuming all queues and idlenesses in pools $j < J$ are zero.) Indeed, the properties in (4.15) follow from Proposition 4.3; property (4.16) follows from (4.10); property (4.17) follows from (4.11) and the definition of the operator L'.

Using properties (4.15)–(4.17), the sample path relation (4.14) implies the following relation (written in vector form, with components indexed by (ij)), which holds, w.p.1-l.r, uniformly for $t \in [0, T\log r]$:

$$\begin{aligned}(\hat{\Psi}_{ij}^r(t)) = (\hat{\Psi}_{ij}^r(0)) + L'(\hat{W}_i^{(a),r}(t)) - L'\left(\sum_j \hat{W}_{ij}^{(s),r}(t)\right) - L'\left(\sum_j \mu_{ij}\int_0^t \hat{\Psi}_{ij}^r(s)\,\mathrm{d}s\right)\\ + (\Delta_i^r(t)).\end{aligned} \tag{4.18}$$

Here $|(\Delta_i^r(t))| < r^{-1/9}$. (Instead of $\frac{1}{9}$ we could use any fixed number in $(0, \frac{1}{8})$.) Indeed, in (4.14) we can replace $\hat{A}_i^r$ and $\hat{D}_{ij}^r$ with $\hat{W}_i^{(a),r}$ and $\hat{W}_{ij}^{(s),r}$, respectively, which introduces an $o(r^{1/4})$ error by (4.15); then, we apply the operator L' to both sides and replace $L'(\hat{X}_i^r)$ with $(\hat{\Psi}_{ij}^r)$, which introduces an $o(r^{1/4})$ error by (4.17); finally, we replace time $(\psi_{ij}^* rt)^{-1}[\int_0^t \Psi_{ij}^r(s)\,\mathrm{d}s]t$ with t in the argument of $\hat{W}_{ij}^{(s),r}$, which introduces an $O(r^{1/8})$ error by (4.16) and Proposition 4.4.

For each r and each initial condition $(\hat{\Psi}_{ij}^r(0))$, in addition to (4.18) consider the (strongly) unique strong solution (see Theorems 5.2.9 and 5.2.5 of [10]) $(\breve{\Psi}_{ij}^r(\cdot))$ of the stochastic

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

integral equation

$$(\breve{\Psi}^r_{ij}(t)) = (\breve{\Psi}^r_{ij}(0)) + L'(\hat{W}^{(a),r}_i(t)) - L'\Bigg(\sum_j \hat{W}^{(s),r}_{ij}(t)\Bigg) - L'\Bigg(\sum_j \int_0^t \mu_{ij}\breve{\Psi}^r_{ij}(s)\,\mathrm{d}s\Bigg), \tag{4.19}$$

driven by the same set of Brownian motions $(\hat{W}^{(a),r}_i(\cdot), \hat{W}^{(s),r}_{ij}(\cdot))$ and with the same initial condition $(\breve{\Psi}^r_{ij}(0)) = (\hat{\Psi}^r_{ij}(0))$. Thus, solutions to both (4.18) and (4.19) for all r are constructed on the same probability space associated with the underlying set of independent Brownian motions (and the corresponding Poisson processes coupled with them). It follows that, w.p.1-l.r, we have, for $t \in [0, T\log r]$,

$$|(\hat{\Psi}^r_{ij}(t)) - (\breve{\Psi}^r_{ij}(t))| \le |(\Delta^r_i(t))| + \int_0^t C'|(\hat{\Psi}^r_{ij}(s)) - (\breve{\Psi}^r_{ij}(s))|\,\mathrm{d}s,$$

with some constant $C' > 0$. By the Gronwall inequality (see, e.g. Theorem 5.1 in Appendix 5 of [4]), for $t \in [0, \varepsilon T\log r]$,

$$|(\hat{\Psi}^r_{ij}(t)) - (\breve{\Psi}^r_{ij}(t))| \le r^{-1/9}\mathrm{e}^{C'\varepsilon T\log r} = r^{-1/9+\varepsilon C'T}.$$

Now we specify the choice of ε: it is such that both $-\frac{1}{8} + \varepsilon C'T < 0$ and (for the reasons given earlier) $\varepsilon < \frac{1}{4}$ hold. In other words, $0 < \varepsilon < \min\{\frac{1}{4}, 1/(9C'T)\}$.

Recall that for any r the law of the multidimensional Brownian motion $(\hat{W}^{(a),r}_i(\cdot), \hat{W}^{(s),r}_{ij}(\cdot))$, driving (4.19), is same as that of $(\sqrt{\lambda_i}B^{(a)}_i(\cdot), \sqrt{\mu_{ij}\psi^*_{ij}}B^{(s)}_{ij}(\cdot))$, where all $B^{(a)}_i(\cdot)|$ and $B^{(s)}_{ij}(\cdot)$ are independent standard Brownian motions. Therefore, for any r, the law of the solution to (4.19) is equal to that of the solution to the stochastic differential equation

$$\mathrm{d}(\breve{\Psi}_{ij}(t)) = L'\mathrm{d}(\sqrt{\lambda_i}B^{(a)}_i(t)) - L'\,\mathrm{d}\Bigg(\sum_j \sqrt{\mu_{ij}\psi^*_{ij}}B^{(s)}_{ij}(t)\Bigg) - L'\Bigg(\sum_j \mu_{ij}\breve{\Psi}_{ij}(t)\Bigg)\mathrm{d}t, \tag{4.20}$$

with the same initial state $(\breve{\Psi}_{ij}(0)) = (\breve{\Psi}^r_{ij}(0))$. This is (3.3). Moreover, the drift term in (4.20) can be written as

$$-L'\Bigg(\sum_j \mu_{ij}\breve{\Psi}_{ij}(t)\Bigg)\mathrm{d}t = L(\breve{\Psi}_{ij}(t))\,\mathrm{d}t,$$

where the matrix L is easily checked to be exactly the same matrix in the ordinary differential equation (ODE) $\mathrm{d}(\tilde{\psi}_{ij}(t)) = L(\tilde{\psi}_{ij}(t))\,\mathrm{d}t$ for the local-fluid model, which follows from conditions (24) of [16]. From [16, Theorem 23] we know that all eigenvalues of L have negative real parts.

Proposition 4.5. *Uniformly on all fixed initial conditions* $(\breve{\Psi}_{ij}(0))$ *from any fixed bounded set, the corresponding solutions to the stochastic differential equation (4.20) have the following properties. Uniformly on all* $t \ge 0$, *the random vector* $(\breve{\Psi}_{ij}(t))$ *is Gaussian, with bounded mean and covariance matrix. Moreover, as* $t \to \infty$, *the mean vector and the covariance matrix of* $(\breve{\Psi}_{ij}(t))$ *converge to those of the unique stationary distribution,* $\mathrm{inv}[(\breve{\Psi}_{ij}(\cdot))]$, *which is Gaussian with zero mean.*

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

Proof. This follows from the fact that all eigenvalues of the drift matrix L have negative real parts; see (5.6.12), (5.6.13)′, (5.6.14)′, Problem 5.6.6, and Theorem 5.6.7 of [10].

Conclusion of the proof of Theorem 3.1. Consider the Markov process $F^r(\cdot)$ in a stationary regime. We choose T as in Lemma 4.1, then ε as in Lemma 4.3, and consider the process on the interval $[0, \varepsilon T \log r]$. Fix an arbitrary $\nu_9 \in (0, \frac{1}{4})$. The combination of [16, Theorem 10(ii)], Lemma 4.1, and Lemma 4.2 proves the following fact: uniformly on all sufficiently large r, the process will 'hit' a state, satisfying the conditions of Lemma 4.3, with probability that can be made arbitrarily close to 1 by choosing sufficiently large fixed $C_{11} > 0$.

Now, suppose at some time point within $[0, \varepsilon T \log r]$ the process is in a state satisfying the conditions of Lemma 4.3. First, we obtain a bound on $|F^r(\varepsilon T \log r)|$. Namely, uniformly on all sufficiently large r, $|F^r(\varepsilon T \log r)| \le C_{14} r^{1/2}$ with a probability that can be made arbitrarily close to 1 by choosing a sufficiently large fixed $C_{14} > 0$. This follows from Lemma 4.3 and Proposition 4.5. This establishes the tightness of the sequence of $\mathrm{inv}[(\hat{\Psi}^r_{ij}(\cdot))] \equiv \mathrm{inv}[r^{-1/2}(\Psi^r_{ij}(\cdot) - \psi^*_{ij} r)]$. Secondly, we obtain a bound on $|(Q^r_i(\varepsilon T \log r))| + |(Z^r_j(\varepsilon T \log r),\ j < J)|$. This is even easier; by (4.11),

$$\mathbb{P}\{|(Q^r_i(\varepsilon T \log r))| + |(Z^r_j(\varepsilon T \log r),\ j < J)| \le C_{10} \delta_9 r^{\nu_9}\} \to 1.$$

But, since ν_9 can be chosen arbitrarily small, we obtain property (3.4).

Given the tightness of the sequence of $\mathrm{inv}[(\hat{\Psi}^r_{ij}(\cdot))]$ and property (3.4), it is straightforward to prove the remaining property (3.2). (The argument is essentially the same as that used in the proof of [11, Theorem 8.5.1], although that result does not directly apply to our setting.) Consider the Markov process $F^r(\cdot)$ in a stationary regime. We fix an arbitrary $T > 0$, $\delta_9 > 0$, and $\nu_9 \in (0, \frac{1}{4})$, and then a large enough parameter $C_{11} > 0$, so that, with probability arbitrarily close to 1, the conditions on $F^r(0)$ in Lemma 4.3 are satisfied for all large r. We then pick a sufficiently small fixed $\varepsilon > 0$ so that property (4.9) holds. Finally, using Proposition 4.5, we pick a sufficiently large $T' > 0$ so that $\mathrm{dist}[(\breve{\Psi}_{ij}(T'))]$ is close to $\mathrm{inv}[(\breve{\Psi}_{ij}(\cdot))]$, uniformly on the initial states $|(\breve{\Psi}_{ij}(0))| \le C_{11}$. (Here 'close' is in the sense of close Gaussian distribution parameters, namely, means and covariances; or, more generally, it can be in the sense of the Prohorov metric [4].) Note that, for all large r, $T' < \varepsilon T \log r$. Applying Lemma 4.3, we see that, for all large r, $\mathrm{dist}[(\hat{\Psi}^r_{ij}(T'))]$ is close to $\mathrm{dist}[(\breve{\Psi}^r_{ij}(T'))]$, which in turn is close to $\mathrm{inv}[(\breve{\Psi}^r_{ij}(\cdot))] = \mathrm{inv}[(\breve{\Psi}_{ij}(\cdot))]$; and we can make it arbitrarily close by rechoosing parameters. This implies (3.2). We omit further details.

5. Discussion

As already mentioned in the introduction, we believe that the approach developed in [16] and this paper provides a quite generic scheme for establishing the diffusion-scale tightness of invariant distributions under the strictly subcritical load $\rho < 1$. The approach shows that, for the diffusion-scale tightness to hold, it is essentially sufficient to verify the two key stability properties, global stability and local stability, which we (at a high level and informally) describe next. Let $F^r(\cdot)$ be a process describing the system-state deviation from the equilibrium point. (For the LAP policy, $F^r(t) = (\Psi^r_{ij}(t) - \psi^*_{ij} r,\ Q^r_i(t))$ as defined in this paper.)

(a) *Global stability.* The fluid limit $f(t),\ t \ge 0$, is defined as $\lim_r r^{-1} F^r(t),\ t \ge 0$. By global stability we mean the following property:

(a.1) the trajectories $f(t)$ converge to 0, uniformly in the initial states from a bounded set. Moreover, we also require the following related property to hold:

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

(a.2) uniformly on all infinite initial states, $|f(0)| = \infty$, each trajectory $f(t)$ reaches a state, where all server pools are fully occupied, and then stays in such a state forever. (For the LAP policy, the formal statements are [16, Propositions 13 and 16].)

(b) *Local stability.* Suppose that $h(r)$ is a function of r such that $h(r)/r \to 0$ and $h(r)/\sqrt{r} \to \infty$. The local fluid limit $\tilde{f}(t)$, $t \geq 0$, is defined as $\lim_r h(r)^{-1} F^r(t)$, $t \geq 0$. Suppose that the trajectories $\tilde{f}(\cdot)$ satisfy a linear ODE $(\mathrm{d}/\mathrm{d}t)\tilde{f}(t) = L\tilde{f}(t)$. By local stability we mean the property that all eigenvalues of L have negative real parts. (For the LAP policy, the formal statement is [16, Theorem 23]. For the LQFSLB policy of [17], the local stability does *not* hold.)

Properties (a) and (b) may or may not be easy to verify for a given control policy; but the task of proving or disproving them is typically much easier than the full task of verifying the diffusion-scale tightness. We also note that showing local stability may require working with the process under additional space and/or time scalings, such as hydrodynamic scaling for the LAP policy (see [16, Section 5.2]).

If the global and local stability properties hold, the steps needed to establish diffusion-scale tightness of invariant distributions are as follows.

Step 1. Existence and $o(r)$-scale tightness of invariant distributions. Using the global stability property (a.2) and employing the total (appropriately defined) workload in the system as a Lyapunov function, we can prove the positive recurrence (stochastic stability) of the process, and, therefore, existence of a stationary distribution. The proof is fairly standard, using the Lyapunov function average drift argument, which additionally shows that $\mathbb{E}|r^{-1}F^r|$ is bounded, which in turn applies the tightness of distributions of $r^{-1}F^r$. We then employ the global stability property (a.1) to show that, in fact, the invariant distributions of $r^{-1}F^r$ asymptotically concentrate at 0. This can be referred to as $o(r)$-scale tightness. (The formal result for the LAP policy is given in [16, Theorem 14].)

Step 2. The $r^{1/2+\varepsilon}$-scale tightness. Local stability implies the exponentially fast convergence of fluid limit trajectories $\tilde{f}(\cdot)$ to 0. In particular, for a sufficiently large fixed T, the norm $|\tilde{f}(t+T)| \leq \delta|\tilde{f}(t)|$, where $\delta < 1$. We use this, and the probability estimates for deviations of $h(r)^{-1}F^r(t)$ from a corresponding local-fluid limit $\tilde{f}(t)$, to show that if $F^r(0) = h(r) = o(r)$ then, with high probability, $|F^r(T)| \leq \delta|F^r(0)|$. Now, it takes $O(\log r)$ intervals of length T for $|F^r|$ to 'descend' from $o(r)$ to $r^{1/2+\varepsilon}$, and we show that this does in fact happen with high probability. (So, the key technical issue here is that we have to obtain probability estimates not on a finite, but on an $O(\log r)$ interval.) This implies $r^{1/2+\varepsilon}$-scale tightness for any $\varepsilon > 0$; namely, the invariant distributions of $r^{-1/2-\varepsilon}F^r$ asymptotically concentrate at 0. (The formal argument for the LAP policy is given in [16, Section 5.2].) Note that this property is *weaker* than, for example, $\mathbb{E}|r^{-1/2-\varepsilon}F^r| \to 0$.

Step 3. Diffusion-scale ($r^{1/2}$-scale) tightness. Here we start with the $r^{1/2+\varepsilon}$-scale tightness, with $\varepsilon > 0$ being sufficiently small. We show that if $|F^r(0)| = O(r^{1/2+\varepsilon})$ then, with high probability, $|F^r(t)|$ 'hits the diffusion scale' $O(r^{1/2})$ within $\varepsilon \log r$. Again, this is achieved by considering $O(\log r)$ consecutive T long intervals, in each of which $|F^r|$ must decrease by a factor with high probability, unless $|F^r(t)|$ does hit $O(r^{1/2})$. (The formal result for the LAP policy is Lemma 4.1.) Given that, it remains to show that if $|F^r(0)| = O(r^{1/2})$ and ε is small enough, then, for any $t \in [0, \varepsilon \log r]$, we also have $|F^r(t)| = O(r^{1/2})$ with high probability. This is achieved by showing the closeness of process $r^{-1/2}F^r(\cdot)$ to the corresponding limiting diffusion process on the $\varepsilon \log r$ long interval, and the fact that the drift matrix of the diffusion process is exactly the L matrix from the definition of local stability. (For the LAP policy, this takes the bulk of this paper, from Lemma 4.2 on. It involves, in particular, showing that all

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press

queues and all pool idlenesses, except for pool J serving the lowest-priority activity, are in fact $o(r^{\nu})$ for any $\nu > 0$.) Again, we note that the diffusion-scale tightness is *weaker* than, for example, the boundedness of $\mathbb{E}|r^{-1/2}F^r|$.

In conclusion, we remark again that many (although not all) parts of the above scheme do rely on the strict subcriticality condition $\rho < 1$. It would be of interest to explore whether the approach can be extended to establishing diffusion-scale tightness in the Halfin–Whitt regime.

References

[1] Aksin, Z., Armony, M. and Mehrotra, V. (2007). The modern call center: a multi-disciplinary perspective on operations management research. *Production Operat. Manag.* **16,** 655–688.

[2] Atar, R., Shaki, Y. Y. and Shwartz, A. (2011). A blind policy for equalizing cumulative idleness. *Queueing Systems* **67,** 275–293.

[3] Csörgő, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. John Wiley, Chichester.

[4] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes. Characterization and Convergence*. John Wiley, Chichester.

[5] Gamarnik, D. and Goldberg, D. A. (2013). Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Ann. Appl. Prob.* **23,** 2382–2419.

[6] Gamarnik, D. and Momčilović, P. (2008). Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* **40,** 548–577.

[7] Gamarnik, D. and Stolyar, A. L. (2012). Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems* **71,** 25–51.

[8] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: tutorial, review, and research prospects. *Manufacturing Service Operat. Manag.* **5,** 79–141.

[9] Gurvich, I. and Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Math. Operat. Res.* **34,** 363–396.

[10] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus,* 2nd edn. Springer, New York.

[11] Liptser, R. Sh. and Shiryayev, A. N. (1989). *Theory of Martingales*. Kluwer, Dordrecht.

[12] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surveys* **4,** 193–267.

[13] Shakkottai, S. and Stolyar, A. L. (2002). Scheduling for multiple flows sharing a time-varying channel: the exponential rule. In *Analytic Methods in Applied Probability* (Amer. Math. Soc. Trans. Ser. 2 **207**), American Mathematical Society, Providence, RI, pp. 185–201.

[14] Stolyar, A. L. and Tezcan, T. (2010). Control of systems with flexible multi-server pools: a shadow routing approach. *Queueing Systems* **66,** 1–51.

[15] Stolyar, A. L. and Tezcan, T. (2011). Shadow-routing based control of flexible multiserver pools in overload. *Operat. Res.* **59,** 1427–1444.

[16] Stolyar, A. L. and Yudovina, E. (2012). Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. *Stoch. Systems* **2,** 381–408.

[17] Stolyar, A. L. and Yudovina, E. (2013). Systems with large flexible server pools: instability of 'natural' load balancing. *Ann. Appl. Prob.* **23,** 2099–2138.

[18] Ward, A. R. and Armony, M. (2013). Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operat. Res.* **61,** 228–243.

https://doi.org/10.1239/aap/1427814590 Published online by Cambridge University Press