TIGHTNESS OF INVARIANT DISTRIBUTIONS OF A LARGE-SCALE FLEXIBLE SERVICE SYSTEM UNDER A PRIORITY DISCIPLINE

By Alexander L. Stolyar and Elena Yudovina[*]

Alcatel-Lucent Bell Labs and University of Michigan

We consider large-scale service systems with multiple customer classes and multiple server pools; interarrival and service times are exponentially distributed, and mean service times depend both on the customer class and server pool. It is assumed that the allowed activities (routing 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, which assigns static priorities to the activities in the order of sequential "elimination" of the tree leaves.

We consider the scaling limit of the system as the arrival rate of customers and number of servers in each pool tend to infinity in proportion to a scaling parameter r, while the overall system load remains strictly subcritical. Indexing the systems by parameter r, we show that (a) the system under LAP discipline is stochastically stable for all sufficiently large r and (b) the family of the invariant distributions is tight on scales $r^{\frac{1}{2}+\epsilon}$ for all $\epsilon > 0$. (More precisely, the sequence of invariant distributions, centered at the equilibrium point and scaled down by $r^{-(\frac{1}{2}+\epsilon)}$, is tight.)

1. Introduction.

Large-scale service systems with heterogeneous customer and server populations bring up the need for efficient dynamic control policies that dynamically match arriving (or waiting) customers and available servers. It is desirable to have algorithms that avoid excessive customer waiting and do not rely on the knowledge of system parameters.

Consider a service system with multiple customer and server types, 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 sizes are B_j. A desirable feature of a dynamic control is insensitivity to parameters Λ_i and μ_{ij}. That is, the assignment of customers to server pools should, to the maximal degree possible, depend only on the current system state (server occupancies, queue sizes), and not on prior knowledge of arrival rates or mean service times,

[*]The research of this author was supported by the NSF Graduate Research Fellowship.

AMS 2000 subject classifications: Primary 60K25, 60F17

Keywords and phrases: many server models, fluid limits, tightness of invariant distributions

because those parameters may not be known in advance and, moreover, they may be changing in time.

If the system objective is to minimize the largest average load of any server pool, a "static" optimal control can be obtained by solving a linear program, called *static planning problem* (SPP), which has B_j's, μ_{ij}'s and Λ_i's as parameters. An optimal solution to the SPP will prescribe optimal average rates Λ_{ij} at which arriving customers should be routed to the server pools. Typically (in a certain sense) the solution to SPP is unique and the *basic activities*, i.e. routing choices (ij) for which $\Lambda_{ij} > 0$, form a tree; let us assume this is the case. Probabilistic routing with static probabilities Λ_{ij}/Λ_i of routing a customer of type i to server pool j, will balance the loads among different server pools and will avoid excessive customer waiting; however, in order to find the routing probabilities, it is necessary to know all of the parameters Λ_i, B_j, and μ_{ij} in advance. The *Shadow Routing* policy in [6] is a dynamic control policy, which achieves the load balancing objective without a priori knowledge of input rates Λ_i; in the process it "automatically identifies" the basic activity tree. Shadow Routing policy, however, does need to "know" the service rates μ_{ij}.

In this paper we assume that the basic activity tree is known, but not the precise rates Λ_i, μ_{ij}; we restrict the routing choices to activities within the basic activity tree. We consider the large-number-of-servers asymptotic regime, in which the arrival rate of customers and number of servers in each pool tend to infinity in proportion to a scaling parameter r; our focus is on the case where the overall system load remains strictly subcritical. In a previous paper [8] we showed that a very natural load balancing policy considered e.g. by [4], [1], [2] may lead to instability at the system equilibrium point: in particular, for certain parameter settings [8, Theorem 7.2] demonstrated the non-tightness (in fact – evanescence to infinity) of invariant measures on the diffusion, $r^{\frac{1}{2}}$- scale. (More precisely, this means that the sequence of invariant distributions, centered at the equilibrium point and scaled down by $r^{-\frac{1}{2}}$, is non-tight, and moreover – escapes to infinity.)

In this paper we consider a different algorithm, which we call the *Leaf Activity Priority* (LAP) policy. As specified above, no precise knowledge of the rates Λ_i and μ_{ij} is required, besides the knowledge of the basic activity tree, and routing is restricted to basic activities only. The policy assigns static priorities to the activities in the order of sequential "elimination" of the tree leaves. The precise definition will be given in Section 2.2. Assuming strictly subcritical load, for this policy we first prove that the system is stochastically stable for all sufficiently large values of r. (In contrast to load balancing policies, the stability under LAP is not "automatic".) Next, we demonstrate the r-scale (fluid-scale) tightness of stationary distributions; this fact is closely related to stability – both are"consequences" of the relatively "benign" behavior of the system on the fluid scale. Then, we obtain a much stronger tightness result, namely that the invariant distributions are tight on the $r^{\frac{1}{2}+\epsilon}$-scale, for any $\epsilon > 0$; this is the main contribution of the paper, which involves the analysis of the process under hydrodynamic and local-fluid scaling (in addition to "standard" fluid scaling). We believe that our analysis can be extended to prove still stronger, diffusion scale ($r^{1/2}$) tightness; this is work in progress.

For a general review of literature on the large-number-of-servers asymptotic regime, including design and analysis of efficient control algorithms, see e.g. [4, 6] and references therein.

The rest of the paper is organized as follows. In Section 2 the model, the asymptotic regime, LAP discipline and basic notation and introduced. The main results are stated in Theorem 10 of Section 3, with its statements (i) and (ii) being the stability and tightness results, respectively. Section 4 contains the analysis of the process on the fluid scale, which leads to establishing stability (Theorem 10(i)) and fluid scale (r-scale) tightness of stationary distributions. In Section 5, using the fluid-scale tightness as a starting point, we prove the $r^{1/2+\epsilon}$-scale tightness (Theorem 10(ii)); this is the key part of the paper, which involves the analysis of system dynamics under LAP discipline under hydrodynamic and local-fluid scaling.

2. Model.

2.1. *The model; Static Planning (LP) Problem.* Consider the model in which there are I customer classes, labeled $1, 2, \ldots, I$, and J server pools, labeled $1, 2, \ldots, J$. (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, i' to refer to customer classes, and j, j' to refer to server pools.

We are interested in the scaling properties of the system as it grows large. The meaning of "grows large" is as follows. 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 $\lambda_i^r = r\lambda_i$, while the jth server pool has $r\beta_j$ individual servers. (All λ_i and β_j are positive parameters.) Customers may 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 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, load-balancing, *static planning problem* (SPP):

$$\min_{\lambda_{ij},\rho} \rho, \tag{1a}$$

subject to

$$\lambda_{ij} \geq 0, \quad \forall i, j \tag{1b}$$

$$\sum_j \lambda_{ij} = \lambda_i, \quad \forall i \tag{1c}$$

$$\sum_i \lambda_{ij}/(\beta_j \mu_{ij}) \leq \rho, \quad \forall j. \tag{1d}$$

Throughout this paper we will always make the following two assumptions about the solution to the SPP (1):

Assumption 2 (Complete resource pooling). The SPP (1) has a unique optimal solution $\{\lambda_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\}, \rho$. Define the *basic activities* to be the pairs, or edges, (ij) for which $\lambda_{ij} > 0$. Let $\mathcal{E}$ be the set of basic activities. We further 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 3 (Underload). The optimal solution to (1) has $\rho < 1$.

Remark 4. Assumption 2 is the *complete resource pooling* (CRP) condition, which holds "generically" in a certain sense; see [7, Theorem 2.2]. Assumption 3 is essential for the main results of the paper ($r^{\frac{1}{2}+\epsilon}$-scale tightness), but many of the auxiliary results hold (along with their proofs) for the critically loaded case $\rho = 1$.

Note that under the CRP condition, all ("server pool capacity") constraints (1d) are binding: $\sum_i \lambda_{ij}/(\beta_j \mu_{ij}) = \rho,\ \forall j$. This in particular means that the optimal solution to SPP is such that, if a system with parameter r will route type i customers to pool j at the rate $\lambda_{ij} r$, the server pool average loads will be minimized and "perfectly balanced".

In this 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 $\mathcal{S}(i) = \{j : (ij) \in \mathcal{E}\}$; for a server type j, let $\mathcal{C}(j) = \{i : (ij) \in \mathcal{E}\}$.

Under the CRP condition, optimal dual variables $\nu_i,\ i \in \mathcal{I}$, and $\alpha_j,\ j \in \mathcal{J}$, corresponding to constraints (1c) and (1d), respectively, are unique and all strictly positive. The dual variable ν_i is interpreted as the "workload" associated with one type i customer, and $\frac{\alpha_j}{\beta_j}$ is interpreted as the (average) rate at which one server in pool j processes workload when it is busy, regardless of the customer type on which it is working, as long as $i \in \mathcal{C}(j)$. The dual variables satisfy the relations $\nu_i \mu_{ij} = \alpha_j/\beta_j$ for any $(ij) \in \mathcal{E}$, and $\sum_j \alpha_j = 1$, which in particular imply that

$$(2) \qquad \sum_i \nu_i \lambda_i = \sum_i \sum_j \nu_i \lambda_{ij} = \sum_i \sum_j \lambda_{ij} \frac{\alpha_j}{\beta_j \mu_{ij}} = \sum_j \alpha_j \sum_i \frac{\lambda_{ij}}{\beta_j \mu_{ij}} = \rho.$$

Given $\rho < 1$, this means, for example, that when all servers in the system are busy, the total rate $\sum_i \nu_i \lambda_i r$ at which new workload arrives in the system is strictly less than the rate $\sum_j \alpha_j r = r$ at which it is served.

REMARK 5. Although (1) is the load-balancing SPP, and the notions introduced in this subsection are defined in terms of this SPP, the policy we consider in this paper (defined in Section 2.2) is *not* a load balancing policy. In particular, the system equilibrium point under the policy, will *not* balance server pool loads, but rather will keep all pools, except one, fully occupied.

2.2. *Leaf activity priority (LAP) policy.* For the rest of the paper, we analyze the performance of the following policy, which we call *leaf activity priority* (LAP). 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 hasn't yet been assigned;
3. Remove the leaf from the tree.

Without loss of generality, we assume 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 this edge the highest priority that hasn't 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 hasn't yet been assigned, and remove the edge.

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

Remark 6. 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 illustrate one such priority assignment in Figure 1.

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 $(ij) < (i'j')$. In particular, if $j = j'$, we have $(ij) < (i'j)$ if and only if $i < i'$. Without loss of generality, we shall assume that the server classes are numbered so that the lowest-priority activity is (IJ). (In Figure 1, this corresponds to assigning the number 3 to server pool C.)

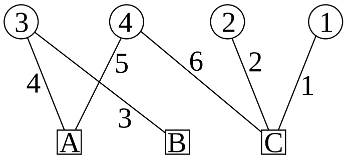

Fig 1. *An example assignment of priorities to customer classes and activities to an example network. Circles represent customer classes, squares represent server pools.*

Now we define the LAP policy itself. It consists of two parts: routing and scheduling. "Routing" determines where an arriving customer goes 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 queue.

Routing: 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.

2.3. *LAP equilibrium point.* Informally speaking, the equilibrium point $(\psi^*_{ij}, q^*_i)_{(ij)\in\mathcal{E},i\in\mathcal{I}}$ is the desired operating point for the (fluid scaled) vector $(\Psi^r_{ij}/r, Q^r_i/r)_{(ij)\in\mathcal{E},i\in\mathcal{I}}$ of occupancies and queue lengths under the LAP policy. Specifically, we will be showing that in steady state the fluid-scaled vector converges in distribution to the equilibrium point, and will then show that the deviations from it are small. We define the equilibrium point below; it will be the stationary point of the fluid models defined in Section 4.

The LAP discipline is not designed with load balancing in mind, so its equilibrium point does *not*, of course, achieve load balancing among the server pools. To define it, we recursively define the quantities $\lambda_{ij} \geq 0$, which have the meaning of routing rates, scaled down by factor $1/r$. (These λ_{ij} are *not* the same as those given by the optimal solution to the SPP (1).) For the activity $(1j)$ with the highest priority, define either $\lambda_{1j} = \lambda_1$ and $\psi^*_{1j} = \frac{\lambda_1}{\mu_{1j}}$, 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

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

given by

$$\psi_{ij}^* = \frac{\lambda_{ij}}{\mu_{ij}}, \quad q_i^* = 0 \text{ for all } (ij) \in \mathcal{E},\ i \in \mathcal{I}. \tag{3}$$

(Since we are in the underloaded case $\rho < 1$, all queues should be 0 at equilibrium.) Clearly, by the above construction, we have

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

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

Assumption 7. If $(\psi_{ij})_{(ij)\in\mathcal{E}}$ are such that $\psi_{ij} \ge 0$, $\lambda_i = \sum_j \mu_{ij}\psi_{ij}$, and $\sum_i \psi_{ij} \le \beta_j$ 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 $\sum_i \psi_{iJ}^* < \beta_J$.

The Assumption 7 means that the system needs to employ (on average) all activities in order to be able to handle the load. It holds, for example, whenever ρ is sufficiently close to 1. Indeed, suppose the arrival rates $(\lambda_i)_{i\in\mathcal{I}}$ are such that $\rho = 1-\epsilon$, and consider the system with arrival rates $\hat{\lambda}_i = \frac{1}{1-\epsilon}\lambda_i$. The basic activity tree $\mathcal{E}$ will be the same for $(\hat{\lambda}_i)_{i\in\mathcal{I}}$ as for $(\lambda_i)_{i\in\mathcal{I}}$. Since $\rho = 1$ for $(\hat{\lambda}_i)_{i\in\mathcal{I}}$ and CRP holds, there exists a unique set of $(\hat{\psi}_{ij})_{(ij)\in\mathcal{E}}$ that satisfies the conditions, and it has $\hat{\psi}_{ij} > 0$ for all $(ij) \in \mathcal{E}$. Also, if ϵ is sufficiently small, we must have $\psi_{ij} \approx \hat{\psi}_{ij}$ for all $(ij) \in \mathcal{E}$, and hence $\psi_{ij} > 0$ for all $(ij) \in \mathcal{E}$.

Remark 8. Assumption 7 is technical – our main result, Theorem 10, can be proved without it, by following the approach presented in the paper. But, it simplifies the statements and proofs of many auxiliary results, and thus substantially improves the exposition.

Remark 9. Although the LAP equilibrium point does not achieve load balancing, when system is heavily loaded (i.e. ρ is close to 1), the load balancing is achieved approximately, in the sense that all queues are kept small and all servers are almost fully loaded – which is the best any "load balancer" could do (when ρ is close to 1).

2.4. *Basic notation.* 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 notation (ξ_i, η_{ij}) to mean $((\xi_i,\ i \in \mathcal{I}), (\eta_{ij},\ (ij) \in \mathcal{E}))$, and similar notations 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 domain different from $[0,\infty)$.) So, for example, $(\xi_i(\cdot))$ signifies $((\xi_i(t), i \in \mathcal{I}),\ t \geq 0)$.

The symbol $\Longrightarrow$ denotes convergence in distribution of random variables in the Euclidean space $\mathbb{R}^d$ (with appropriate dimension d). The symbol $\rightarrow$ denotes ordinary convergence in $\mathbb{R}^d$. Standard Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted $|x|$, while $\|x\|$ denotes the L_1-norm (sum of absolute values of the components); *u.o.c.* means *uniform on compact sets* convergence of functions, with the domain defined explicitly or by the context. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

3. Main result.

We are now in position to state our main result.

Theorem 10. *Consider the sequence of systems under LAP policy, in the scaling regime and under the assumptions specified in Section 2, with* $\rho < 1$. *Then:*
(i) For all sufficiently large r, *the system is stable, i.e. the countable Markov chain* $(\Psi^r_{ij}(\cdot), Q^r_i(\cdot))$ *is positive recurrent.*
(ii) For any $\epsilon > 0$, *the stationary distribution of* $r^{-1/2-\epsilon}(\Psi^r_{ij}(\cdot) - \psi^*_{ij} r, Q^r_i(\cdot))$ *weakly converges to the Dirac measure concentrated at* 0.

The proof is given in the rest of the paper, and consists roughly of two stages. First, we study the process under the fluid scaling $r^{-1}(\Psi^r_{ij}(\cdot), Q^r_i(\cdot))$, which allows us to prove stability and statement (ii) for $\epsilon = 1/2$. Then we need a more detailed analysis, involving *hydrodynamic* and *local-fluid* scaling of the process, to prove (ii) for any $\epsilon > 0$.

Throughout the paper, we will use the following additional notation for the system variables. For a system with parameter r, we denote:
$X_i^r(t) = \sum_j \Psi_{ij}^r(t) + Q_i^r(t)$ is the total number of customers of type i in the system at time t;
$A_i^r(t)$ is the total number of customers of type i exogenous arrivals into the system in interval $[0,t]$;
$D_{ij}^r(t)$ is the total number of customers of type i that completed the service in pool j (and departed the system) in interval $[0,t]$;
$\Xi_{ij}^r(t)$ is the total number of customers of type i that entered service in pool j in interval $[0,t]$.
There are some obvious relations between realizations of these processes: $Q_i^r(t) = Q_i^r(0) + A_i^r(t) - \sum_j \Xi_{ij}^r(t)$; $Q_i^r(t) > 0$ implies $\sum_{i'} \Psi_{i'j}^r(t) = \beta_j r$ for each $j \in \mathcal{S}(i)$; and so on.

We can and will 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 process $\Pi_i^{(a)}(\cdot), i \in \mathcal{I}$ and $\Pi_{ij}^{(s)}(\cdot), (ij) \in \mathcal{E}$, as follows:

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

the driving Poisson processes are common for all r.

4. Fluid scaling. We begin by analyzing the LAP discipline on the fluid scale. Namely, consider the scaling

$$\begin{aligned}&\Big(\psi_{ij}^r(t), q_i^r(t), x_i^r(t), a_i^r(t), d_{ij}^r(t), \xi_{ij}^r(t)\Big)\\ &\quad = \frac{1}{r}\Big(\Psi_{ij}^r(t), Q_i^r(t), X_i^r(t), A_i^r(t), D_{ij}^r(t), \Xi_{ij}^r(t)\Big).\end{aligned}$$

PROPOSITION 11. *Suppose $(\psi_{ij}^r(0), q_i^r(0)) \to (\psi_{ij}(0), q_i(0))$. Then, w.p.1, for any subsequence $r \to \infty$ there exists a further subsequence along which $(\psi_{ij}^r(\cdot), q_i^r(\cdot), x_i^r(\cdot), a_i^r(\cdot), d_{ij}^r(\cdot), \xi_{ij}^r(\cdot))$ converges uniformly on compact sets to a set $(\psi_{ij}(\cdot), q_i(\cdot), x_i(\cdot), a_i(\cdot), d_{ij}(\cdot), \xi_{ij}(\cdot))$ of Lipschitz continuous functions satisfying conditions* (4)*. The conditions involving derivatives are to be satisfied at all regular points of the limiting set of functions. (A time point $t \geq 0$ is* regular *if both minimum and maximum of any subset of component functions have derivatives at t. All points $t \geq 0$ are regular, except a subset of zero Lebesgue measure.)*

The fluid model conditions are:

$$\text{(4a)}\quad q_i(t) \ge 0,\ \ \forall i \in \mathcal{I};\quad \psi_{ij}(t) \ge 0,\ \ \forall (ij) \in \mathcal{E};\quad \sum_i \psi_{ij}(t) \le \beta_j,\ \ \forall j \in \mathcal{J}$$

$$\text{(4b)}\qquad a_i(t) = \lambda_i t,\ \ \forall i \in \mathcal{I};\qquad d_{ij}(t) = \int_0^t \mu_{ij}\psi_{ij}(s)ds,\ \ \forall (ij) \in \mathcal{E}$$

$$\text{(4c)}\qquad \begin{aligned} q_i(t) &= q_i(0) + a_i(t) - \sum_j \xi_{ij}(t),\quad \forall i \in \mathcal{I};\\ \psi_{ij}(t) &= \psi_{ij}(0) + \xi_{ij}(t) - d_{ij}(t),\quad \forall (ij) \in \mathcal{E} \end{aligned}$$

$$\text{(4d)}\quad x_i(t) = q_i(t) + \sum_j \psi_{ij}(t) = x_i(0) + \lambda_i t - \sum_j \int_0^t \mu_{ij}\psi_{ij}(s)ds,\quad \forall i \in \mathcal{I}$$

$$\text{(4e)}\qquad \sum_i \psi_{ij}(t) = \beta_j,\quad \text{whenever } q_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j)$$

$$\text{(4f)}\qquad \frac{d}{dt}\xi_{ij}(t) = 0,\quad \text{whenever } q_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j),\ i' < i$$

$$\text{(4g)}\quad \frac{d}{dt}\xi_{ij}(t) = 0,\quad \text{whenever } \sum_k \psi_{kj'}(t) < \beta_{j'} \text{ for at least one } (ij') < (ij)$$

$$\text{(4h)}\quad \frac{d}{dt}\xi_{ij}(t) = \sum_{i'} \mu_{i'j}\psi_{i'j}(t) - \sum_{(i'j)<(ij)} \frac{d}{dt}\xi_{i'j}(t),$$

for any $(ij) \in \mathcal{E}$ such that $q_i(t) > 0$ (and then necessarily $\sum_k \psi_{kj}(t) = \beta_j$)

$$\text{(4i)}\quad \frac{d}{dt}\xi_{ij}(t) = \lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\xi_{ij'}(t),$$

for any $(ij) \in \mathcal{E}$ such that $\sum_k \psi_{kj}(t) < \beta_j$ (and then necessarily $q_i(t) = 0$)

$$\text{(4j)}\quad \frac{d}{dt}\xi_{ij}(t) = \min\left(\lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\xi_{ij'}(t), \right.$$
$$\left. \sum_{i'} \mu_{i'j}\psi_{i'j}(t) - \sum_{(i'j)<(ij)} \frac{d}{dt}\xi_{i'j}(t)\right)$$
$$\text{for any } (ij) \in \mathcal{E} \text{ such that } q_i(t) = 0 \text{ and } \sum_k \psi_{kj}(t) = \beta_j.$$

Proof of Proposition 11. The proof of convergence fact and of the basic conditions (4a)-(4d) of the limit, is very standard. Indeed, it follows from the Functional Strong Law of Large Numbers (FSLLN) for the driving processes, and the scaling applied, that w.p.1 each component function is asymptotically Lipschitz. For example, for each scaled departure process we have: w.p.1, for a fixed large $C > 0$ and any $0 \le t_1 < t_2 < \infty$,

$$\limsup_{r\to\infty} d^r_{ij}(t_2) - d^r_{ij}(t_1) < C(t_2 - t_1).$$

This implies that, w.p.1 any subsequence of r has a further subsequence along which a u.o.c. convergence $d^r_{ij}(\cdot) \to d_{ij}(\cdot)$ holds, where $d_{ij}(\cdot)$ is Lipschitz. Similar convergence property holds for each arrival process. From here we obtain the convergence (along a subsequence) for all other components. Then, relations (4a)-(4d) are inherited from the corresponding conservation laws for the pre-limit trajectories.

Properties (4e)-(4i) easily follow from the priority rule of LAP; it suffices to consider the behavior of pre-limit trajectories in a small time interval $[t, t+\delta]$ when r sufficiently large. (See e.g. [5, Theorem 1] for this type of argument.)

Finally, to show (4j) we recall that, by the priority assignment procedure, for the activity (ij): either (ij) has the lowest priority among activities associated with customer class i or (ij) has the lowest priority among activities associated with server pool j (or both). (The former case happens if, at the time when (ij) was being assigned a priority, it was the only activity associated with customer class i; analogously, the latter case is when (ij) was the only activity associated with j.) Taking into account that point t is regular (which in particular implies $q_i'(t) = 0$ and $(d/dt)\sum_k \psi_{kj}(t) = 0$), we easily see that in the former case the only possibility is that

$$\frac{d}{dt}\xi_{ij}(t) = \lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\xi_{ij'}(t) \le \sum_{i'} \mu_{i'j}\psi_{i'j}(t) - \sum_{(i'j)<(ij)} \frac{d}{dt}\xi_{i'j}(t),$$

and in the latter case we must have

$$\frac{d}{dt}\xi_{ij}(t) = \sum_{i'} \mu_{i'j}\psi_{i'j}(t) - \sum_{(i'j)<(ij)} \frac{d}{dt}\xi_{i'j}(t) \le \lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\xi_{ij'}(t).$$

This implies (4j). We omit further details of the proof which are, again, rather standard. □

We call any Lipschitz solution $(\psi_{ij}(\cdot), q_i(\cdot), x_i(\cdot), a_i(\cdot), d_{ij}(\cdot), \xi_{ij}(\cdot))$ of (4) a *fluid model* of the system with initial state $(\psi_{ij}(0), q_i(0))$; a set $(\psi_{ij}(\cdot), q_i(\cdot))$, which is a projection of a fluid model we often call a fluid model as well.

REMARK 12. It will not be important for the results in the paper whether the fluid model with given initial conditions is unique; all that will matter is the long-term behavior of all fluid models with given initial conditions.

PROPOSITION 13. *For any $\epsilon' > 0$ and any $K > 0$ there exists a finite time $T = T(K)$ such that all fluid models whose starting state satisfies $|(\psi_{ij}(0), q_i(0))| \le K$ have $\sum_i \psi_{ij}(t) = \beta_j$, $\forall j < J$, $q_i(t) = 0$, $\forall i \in \mathcal{I}$, and $\left|\psi_{ij}(t) - \psi^*_{ij}\right| < \epsilon'$ for all $(ij) \in \mathcal{E}$, for all $t \ge T(K)$.*

SKETCH OF PROOF. For the highest priority activity $(1j)$ there are two cases.
Case a: Type 1 is a leaf. In this case j is the unique server to which type 1 jobs are allowed to go, and they have the highest priority there. Pick a small $\delta > 0$. After a finite time (uniformly bounded above, across all starting states as in the proposition statement), the condition $\psi_{1j}(t) \ge \psi^*_{1j} - \delta$ must hold, because $\psi_{1j}(t) < \psi^*_{1j} - \delta$ implies that $(d/dt)\psi_{1j}(t)$ is positive and bounded away from 0. After such time, $q_1(t) > 0$ implies $\sum'_i \psi_{i'j}(t) = \beta_j$ and (recall that δ is small) $\lambda_1 \le \mu_{1j}\psi_{1j}(t) - \delta_1$ for some $\delta_1 > 0$; and therefore $(d/dt)q_1(t) \le -\delta_1$. We conclude that after a finite time (uniformly bounded above) we must have $q_1(t) = 0$. This in turn implies that $(d/dt)\psi_{1j}(t)$ is negative and bounded away from 0 as long as $\psi_{1j}(t) > \psi^*_{1j} + \delta$. Thus, $|\psi_{1j}(t) - \psi^*_{1j}| \le \delta$ and $q_1(t) = 0$ after a bounded time.
Case b: Pool j is a leaf. Then Assumption 7 implies $\psi^*_{1j} = \beta_j$ and $\lambda_1 > \mu_{1j}\beta_j$. In this case, $\psi_{1j}(t) = \psi^*_{1j}$ starting at some time (that is bounded uniformly on initial states), simply because $(d/dt)\psi_{1j}(t) \ge \lambda_1 - \mu_{1j}\beta_j > 0$ as long as $\psi_{1j}(t) < \beta_j$.
We see that, in either case a or b, for arbitrarily small $\delta > 0$, there exists $T_1 = T_1(\delta)$ such that $|\psi_{1j}(t) - \psi^*_{1j}| < \delta$.
We proceed by induction on the activity priorities and, using Assumption 7,

easily establish analogous properties for every activity $(i'j')$. This implies the result. We omit details. □

THEOREM 14. *For all sufficiently large r, the LAP discipline stabilizes the network (in the sense of positive recurrence of the underlying Markov process). Moreover, the sequence of invariant distributions of (ψ^r_{ij}, q^r_i) is tight, and the invariant distributions converge weakly to the point mass at the equilibrium point.*

Before we proceed with the proof, we need the following lemma.

LEMMA 15. *There exists $T_1 > 0$ such that for any $T_2 > T_1$ there exists a sufficiently large $C = C(T_2)$ for which the following holds. For any $\epsilon > 0$,*

$$\mathbb{P}\left\{ \big|\sum_{(ij)} \nu_i(d^r_{ij}(T_2) - d^r_{ij}(T_1)) - (T_2 - T_1)\big| \geq \epsilon \right\} \to 0,$$

as $r \to \infty$, uniformly on initial states with $\max_{i\in\mathcal{I}} q^r_i(0) \geq C$.

In turn, to prove this lemma, we will need to use fluid models with infinite initial states. Note that we cannot appeal directly to the properties of "standard" fluid models defined earlier, because we require convergence that is uniform in all large initial states. So, we need the following version of a fluid limit result. We will use notation $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ for the the one-point compactification of $\mathbb{R}$.

PROPOSITION 16. *Consider a sequence of fluid-scaled processes $(\psi^r_{ij}(\cdot), q^r_i(\cdot))$ with deterministic initial states such that $\Big|(\psi^r_{ij}(0), q^r_i(0))\Big| = C'(r) \to \infty$ and*

$$(\psi^r_{ij}(0), q^r_i(0)) \to (\psi_{ij}(0), q_i(0)),$$

where each $q^r_i(0)$ and $q_i(0)$ is viewed as an element of $\bar{\mathbb{R}}$. Partition the customer classes as $\mathcal{I} = \mathcal{I}^\infty \cup \mathcal{I}^0$, where $q_i(0) = \infty$ for $i \in \mathcal{I}^\infty$, and $q_i(0) < \infty$ for $i \in \mathcal{I}^0$. (Necessarily, $\mathcal{I}^\infty$ is non-empty.) Then, with probability 1, any subsequence of trajectories has a further subsequence which converges u.o.c. to a fluid model, satisfying same conditions as (4), except that for all $i \in \mathcal{I}^\infty$ the queue length $q_i(t) = \infty, \forall t \geq 0$. Moreover, all such fluid models are such that, uniformly on all of them, starting at some finite time T'_1, all server pools are fully occupied: $\sum_i \psi_{ij}(t) = \beta_j$, $t \geq T'_1$, $\forall j$.

Proof of this result is very similar to that of Proposition 13 (and in fact simpler), so it is not spelled out here. We just note that Assumption 7 is essential in showing that *all* server pools are occupied after a finite time. Without the assumption, we could still show that the occupancy becomes *strictly greater* than at the equilibrium point, and that would be enough for our purposes; however, it would make Proposition 16 statement and proof more cumbersome.

Proof of Lemma 15. Let us choose $T_1 = 2T_1'$, where T_1' is as in Proposition 16. Now, if the lemma statement would not hold, then for some fixed $\epsilon' > 0$ we could find a sequence of systems with $\left|(\psi^r_{ij}(0), q^r_i(0))\right| = C'(r) \to \infty$, such that

$$\limsup_r \mathbb{P}\left\{ \big|\sum_{(ij)} \nu_i(d^r_{ij}(T_2) - d^r_{ij}(T_1)) - (T_2 - T_1)\big| \ge \epsilon' \right\} > 0.$$

This, however, is impossible because, by Proposition 16, w.p.1 from any subsequence of r we can find a further subsequence such that:

$$\sum_i \psi^r_{ij}(t) \to \sum_i \psi_{ij}(t) = \beta_j \ \text{ uniformly in } [T_1, T_2], \ \ \forall j,$$

$$d^r_{ij}(T_2) - d^r_{ij}(T_1) \to \int_{T_1}^{T_2} \mu_{ij}\psi_{ij}(t)dt, \ \ \forall (ij),$$

and therefore

$$\sum_{(ij)} \nu_i(d^r_{ij}(T_2) - d^r_{ij}(T_1)) \to \sum_{(ij)} \nu_i \int_{T_1}^{T_2} \mu_{ij}\psi_{ij}(t)dt =$$
$$\sum_{(ij)} \int_{T_1}^{T_2} \frac{\alpha_j}{\beta_j}\psi_{ij}(t)dt = \sum_j \frac{\alpha_j}{\beta_j} \int_{T_1}^{T_2} [\sum_i \psi_{ij}(t)]dt = T_2 - T_1.$$

□

Proof of Theorem 14. Recall that $\nu_i > 0$ is the workload associated with a single request of type i; i.e., the optimal dual variable associated with (1c) for type i. We consider the quantity

$$W^r(t) = \sum_i \nu_i x^r_i(t)$$

(where $x_i^r(t) = \sum_i q_i^r(t) + \sum_{ij} \psi_{ij}^r(t)$), the total workload of the system. We will argue that the quantity

$$\mathcal{L}^r(t) = [W^r(t)]^2$$

will serve as a Lyapunov function for the rth system. Namely, the following property holds: there exist positive constants K, T, C_1, C_2, C_3 such that, for all sufficiently large r (and all t),

$$\text{if } \mathcal{L}^r(t) > K \text{ then } \mathbb{E}[\mathcal{L}^r(t+T) - \mathcal{L}^r(t) | \mathcal{L}^r(t)] < -C_1 W^r(t) + C_2 \tag{5}$$

and

$$\text{if } \mathcal{L}^r(t) \le K \text{ then } \mathbb{E}[\mathcal{L}^r(t+T) - \mathcal{L}^r(t) | \mathcal{L}^r(t)] < C_3. \tag{6}$$

(The proof of (5)-(6) is given after we complete the theorem proof.) It is then a standard application of the Foster-Lyapunov criteria to conclude that for all sufficiently large r the system Markov process is positive recurrent, and moreover, the stationary distributions are such that $\mathbb{E}W^r = \sum_i \nu_i \mathbb{E}x_i^r$ remains uniformly (in r) bounded. Indeed, for any fixed initial state of the process, consider the embedded chain at times $0, T, 2T, \ldots$. It easily follows (using the fact that each input flow is Poisson, and fluid scaling is applied) that $0 \le \mathbb{E}\mathcal{L}^r(nT) = \mathbb{E}[W^r(nT)]^2 < \infty$ for all $n = 0, 1, 2, \ldots$. Also, WLOG, by rechoosing if necessary C_1 and C_2, we can assume that the "then" part of (5) holds for any $\mathcal{L}^r(t)$. We see that, for any n,

$$\mathbb{E}\mathcal{L}^r((n+1)T) - \mathbb{E}\mathcal{L}^r(nT) \le -C_1 \mathbb{E}W^r(nT) + C_2;$$

from here the positive recurrence and steady-state bound $\mathbb{E}W^r \le C_2/C_1$ easily follow, because the opposite would imply $\mathbb{E}\mathcal{L}^r(nT) \to -\infty$ as $n \to \infty$.

Uniform bound on $\mathbb{E}W^r$ implies tightness of invariant distributions. The tightness together with Proposition 13 imply that the sequence of invariant distributions must weakly converge to the point mass at equilibrium.

It remains to show property (5)-(6). First, it is easy to see (and is a standard observation) that

$$\forall T > 0, \;\; \mathbb{E}[W^r(t+T) - W^r(t)]^2 \text{ are uniformly bounded across all } r \text{ and } t. \tag{7}$$

This guarantees (6) for any fixed K. To prove (5), we fix $T_1 > 0$ as in Lemma 15, and then choose a large fixed $T > T_1$. Note that

$$(\min_{i \in \mathcal{I}} \nu_i)(\max_{i \in \mathcal{I}} q_i^r(t)) \le W^r(t) \le (\max_{i \in \mathcal{I}} \nu_i)(I \max_{i \in \mathcal{I}} q_i^r(t) + \sum_j \beta_j);$$

in particular, the condition $\max_{i\in\mathcal{I}} q_i^r(0) \to \infty$ in Lemma 15 is equivalent to $W^r(0) \to \infty$. If we fix a sufficiently small $\epsilon' > 0$ and apply Lemma 15, we obtain the following fact:
for a sufficiently large fixed $K > 0$ (as a function of T), uniformly on all $\mathcal{L}^r(0) > K$ and all large r,

$$\mathbb{P}\{W^r(T) - W^r(0) \le 2\rho T_1 - \frac{1}{2}(1-\rho)(T-T_1)\} \ge 1-\epsilon'. \tag{8}$$

Indeed, the $2\rho T_1$ is a crude upper bound on $W^r(T_1) - W^r(0)$, which holds with high probability (w.h.p.) for large r, since by (2) new workload arrives at average rate ρ (in the fluid-scaled system). The term $-(1/2)(1-\rho)(T-T_1)$ is an upper bound on $W^r(T) - W^r(T_1)$, also holding w.h.p., because by Lemma 15 the average rate at which workload leaves the system is w.h.p. close to 1 in $[T_1, T]$; and recall that $\rho < 1$. This proves (8). The RHS of the first inequality in (8) can be made negative, with arbitrarily large absolute value, by choosing T large enough. This implies (5), because we have the identity $\mathcal{L}^r(t+T)-\mathcal{L}^r(t) = 2W^r(t)(W^r(t+T)-W^r(t))+[W^r(t+T)-W^r(t)]^2$ and (7), which in particular implies that $|W^r(t+T) - W^r(t)|$ is uniformly integrable. □

5. Proof of Theorem 10(ii).

5.1. *Preliminaries.* In the previous section we have shown that the process $(\Psi_{ij}^r(\cdot), Q_i^r(\cdot))$ is positive recurrent, and then has unique stationary (or invariant) distribution for all large r (which proved Theorem 10(i)). Moreover,

$$r^{-1}(\Psi_{ij}^r - \psi_{ij}^* r, Q_i^r) \implies 0. \tag{9}$$

Here and in the rest of the paper (Ψ_{ij}^r, Q_i^r) means "$(\Psi_{ij}^r(t), Q_i^r(t))$ in stationary regime."

So, we know that Theorem 10(ii) is true for $\epsilon = 1/2$, and our goal is to prove it for any $\epsilon > 0$. In what follows $0 < \epsilon < 1/2$ is fixed.

From (9), for an arbitrarily small fixed $\delta > 0$, we can choose a positive function $g(r) = o(r)$, such that,

$$\mathbb{P}\{\left|(\Psi_{ij}^r - r\psi_{ij}^*, Q_i^r)\right| \le g(r)\} \ge 1-\delta. \tag{10}$$

Without loss of generality, assume $r^{-1/2-\epsilon} g(r) \to \infty$.

We will prove that there exist positive constants C and T, such that for any fixed $\delta_1 > 0$ the following holds for all sufficiently large r:

$$\text{(11)} \quad \left|\left(\Psi_{ij}^r(0) - r\psi_{ij}^*, Q_i^r(0)\right)\right| \le g(r) \text{ implies}$$
$$\mathbb{P}\{\left|\left(\Psi_{ij}^r(T\log r) - r\psi_{ij}^*, Q_i^r(T\log r)\right)\right| \le Cr^{1/2+\epsilon}\} \ge 1-\delta_1.$$

This fact, along with (10), implies that for all large r, in steady-state,

$$\mathbb{P}\{\left|\left(\Psi_{ij}^r - r\psi_{ij}^*, Q_i^r\right)\right| \le Cr^{1/2+\epsilon}\} \ge (1-\delta)(1-\delta_1).$$

This clearly proves Theorem 10(ii), because δ, δ_1 can be chosen, and ϵ re-chosen, to be arbitrarily small. So, the rest of Section 5 is the proof of (11), with the final part of the proof given in Section 5.4.

We will need FSLLN-type results, which can be obtained from a strong approximation of Poisson processes, available e.g. in [3, Chapters 1 and 2]:

Proposition 17. *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, such that $\forall T > 1$ and $\forall u \ge 0$*

$$\mathbb{P}\left(\sup_{0\le t\le T} |\Pi(t) - t - W(t)| \ge C_1 \log T + u\right) \le C_2 e^{-C_3 u}.$$

From here, for the unit rate Poisson processes $\Pi_i^{(a)}(\cdot)$ and $\Pi_{ij}^{(s)}(\cdot)$, driving exogenous arrivals and departures, we obtain the following fact. (For $\Pi_i^{(a)}(\cdot)$, for example, we replace t with $\lambda_i rt$; T with $\lambda_i rT\log r$; and u with $r^{1/4}$.)

Proposition 18. *For any fixed $T > 0$ and any subsequence of $r \to \infty$, we can find a further subsequence (with r increasing sufficiently fast), such that:*
for each i

$$\sup_{0\le t\le T\log r} r^{-1/2-\epsilon/2}\left|\Pi_i^{(a)}(\lambda_i rt) - \lambda_i rt\right| \to 0, \quad \textit{w.p.1},$$

and for each (ij)

$$\sup_{0\le t\le T\log r} r^{-1/2-\epsilon/2}\left|\Pi_{ij}^{(s)}(\mu_{ij}\beta_j rt) - \mu_{ij}\beta_j rt\right| \to 0, \quad \textit{w.p.1}.$$

Let $F^r(t)$ be the process of (unscaled) deviations from equilibrium; that is,

$$F^r(t) = (\Psi^r_{ij}(t) - r\psi^*_{ij}, Q^r_i(t)).$$

Suppose we have a function $h(r)$, such that $r^{1/2+\epsilon} \le h(r) \le g(r)$. (The quantity $h(r)$ will be the "scale" of $|F^r(0)|$; sometimes, we simply use $h(r) = |F^r(0)|$, but not necessarily.) We will establish properties of $F^r(\cdot)$ under two different scalings, called hydrodynamic and local-fluid.

We remark that the use of multiple scalings (in addition to the "standard" fluid scaling) is typical in the analysis of systems in many-server asymptotic regime, cf. [4] and references therein. However, our hydrodynamic and local-fluid scalings are somewhat unusual in that the scaling factor $h(r)$ is strictly "between" r and $r^{1/2}$. (When $h(r) = r$, both local-fluid and hydrodynamic scalings become the standard fluid scaling; if $h(r) = r^{1/2}$, the local-fluid scaling becomes the standard diffusion scaling.) Also, the system behavior, of course, depends on the control discipline, LAP in our case; and so our analysis of LAP under various scalings is new. Most importantly, the way we use these multiple scalings for the purposes of proving tightness of *stationary* distributions is novel, to the best of our knowledge.

5.2. *Hydrodynamic scaling.* Consider the process under the following scaling and centering:

$$(12)\quad (\overline{\psi}^r_{ij}(t), \overline{q}^r_i(t), \overline{x}^r_i(t), \overline{a}^r_i(t), \overline{d}^r_{ij}(t), \overline{\xi}^r_{ij}(t)) = \\ h(r)^{-1}\Big(\Psi^r_{ij}((h(r)r^{-1}t) - r\psi^*_{ij}, Q^r_i(h(r)r^{-1}t), X^r_i(h(r)r^{-1}t) - r\sum_j \psi^*_{ij}, \\ A^r_i(h(r)r^{-1}t), D^r_{ij}(h(r)r^{-1}t), \Xi^r_{ij}(h(r)r^{-1}t)\Big).$$

THEOREM 19. *Consider a sequence of deterministic realizations, such that the driving realizations satisfy FSLLN conditions, namely:*

$$(13)\qquad (\overline{a}^r_i(t),\ t \ge 0) \to (\lambda_i t,\ t \ge 0), \quad u.o.c.,\ \forall i$$

$$(14)\quad \Big(h(r)^{-1}\big(D^r_{ij}(h(r)r^{-1}t) - \mu_{ij}\int_0^{h(r)r^{-1}t} \Psi^r_{ij}(s)ds\big),\ t \ge 0\Big) \to 0, \quad u.o.c.,\ \forall (ij).$$

Suppose $(\overline{\psi}^r_{ij}(0), \overline{q}^r_i(0)) \to (\overline{\psi}_{ij}(0), \overline{q}_i(0))$.
Then, for any subsequence of r *there exists a further subsequence along which* $(\overline{\psi}^r_{ij}(\cdot), \overline{q}^r_i(\cdot), \overline{x}^r_i(\cdot), \overline{a}^r_i(\cdot), \overline{d}^r_{ij}(\cdot), \overline{\xi}^r_{ij}(\cdot))$ *converges uniformly on compact sets to a set* $(\overline{\psi}_{ij}(\cdot), \overline{q}_i(\cdot), \overline{x}_i(\cdot), \overline{a}_i(\cdot), \overline{d}_{ij}(\cdot), \overline{\xi}_{ij}(\cdot))$ *of Lipschitz continuous functions*

satisfying conditions (15)*. (The conditions involving derivatives are to be satisfied at regular time points $t \geq 0$ of the limiting set of functions.)*

The hydrodynamic model conditions are:

$$\bar{q}_i(t) \geq 0, \quad \forall i \in \mathcal{I}; \qquad \sum_i \bar{\psi}_{ij}(t) \leq 0, \quad \forall j \in \mathcal{J} \tag{15a}$$

$$\bar{a}_i(t) = \lambda_i t, \quad \forall i \in \mathcal{I}; \qquad \bar{d}_{ij}(t) = \mu_{ij}\psi^*_{ij} t, \quad \forall (ij) \in \mathcal{E} \tag{15b}$$

$$\begin{aligned} \bar{q}_i(t) &= \bar{q}_i(0) + \bar{a}_i(t) - \sum_j \bar{\xi}_{ij}(t), \quad \forall i \in \mathcal{I}; \\ \bar{\psi}_{ij}(t) &= \bar{\psi}_{ij}(0) + \bar{\xi}_{ij}(t) - \bar{d}_{ij}(t), \quad \forall i \in \mathcal{I} \end{aligned} \tag{15c}$$

$$\bar{x}_i(t) = \bar{q}_i(t) + \sum_j \bar{\psi}_{ij}(t) \equiv \bar{x}_i(0), \quad \forall i \in \mathcal{I} \tag{15d}$$

$$\sum_i \bar{\psi}_{ij}(t) = 0, \quad \text{whenever } \bar{q}_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j) \tag{15e}$$

$$\frac{d}{dt}\bar{\xi}_{ij}(t) = 0, \quad \text{whenever } \bar{q}_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j),\ i' < i \tag{15f}$$

$$\frac{d}{dt}\bar{\xi}_{ij}(t) = 0, \quad \text{whenever } \sum_k \bar{\psi}_{kj'}(t) < 0 \text{ for at least one } (ij') < (ij) \tag{15g}$$

$$\begin{aligned} \frac{d}{dt}\bar{\xi}_{ij}(t) &= \sum_{i'} \mu_{i'j}\psi^*_{i'j} - \sum_{(i'j)<(ij)} \frac{d}{dt}\bar{\xi}_{i'j}(t), \\ &\text{whenever } \bar{q}_i(t) > 0 \text{ (and then necessarily } \sum_k \bar{\psi}_{kj}(t) = 0) \end{aligned} \tag{15h}$$

$$\begin{aligned} \frac{d}{dt}\bar{\xi}_{ij}(t) &= \lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\bar{\xi}_{ij'}(t), \\ &\text{whenever } \sum_k \bar{\psi}_{kj}(t) < 0 \text{ (and then necessarily } \bar{q}_i(t) = 0) \end{aligned} \tag{15i}$$

(15j)

$$\frac{d}{dt}\bar{\xi}_{ij}(t) = \min\left(\lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\bar{\xi}_{ij'}(t), \sum_{i'} \mu_{i'j}\psi^*_{i'j} - \sum_{(i'j)<(ij)} \frac{d}{dt}\bar{\xi}_{i'j}(t)\right)$$
$$\text{whenever } \bar{q}_i(t) = 0 \text{ and } \sum_k \bar{\psi}_{kj}(t) = 0.$$

There is a clear correspondence between the hydrodynamic model and fluid model conditions. This is not surprising, of course, – the hydrodynamic limit is also an FSLLN-type limit, but on a different, finer time and space scale. We omit the proof of Theorem 19 – it is analogous to that of Proposition 11.

We call any Lipschitz solution $(\bar{\psi}_{ij}(\cdot), \bar{q}_i(\cdot), \bar{x}_i(\cdot), \bar{a}_i(\cdot), \bar{d}_{ij}(\cdot), \bar{\xi}_{ij}(\cdot))$ of (15) a *hydrodynamic model* (HM) of the system with initial state $(\bar{\psi}_{ij}(0), \bar{q}_i(0))$; a set $(\bar{\psi}_{ij}(\cdot), \bar{q}_i(\cdot))$, which is a projection of an HM we often call a hydrodynamic model as well. Also, we sometimes use shorter notations $\bar{f}^r(\cdot) = (\bar{\psi}^r_{ij}(\cdot), \bar{q}^r_i(\cdot))$, $\bar{f}(\cdot) = (\bar{\psi}_{ij}(\cdot), \bar{q}_i(\cdot))$.

We have the following corollary of Theorem 19 which we record for future reference.

Corollary 20. *For any fixed $T > 0$, $K > 0$ and $\delta_2 > 0$, there exists a sufficiently small $\delta_3 > 0$, such that the following holds. Uniformly on all $|\bar{f}^r(0)| \le K$ and all sufficiently large r, conditions*

$$\max_i \sup_{[0,T]} |\bar{a}^r_i(t) - \lambda_i t| \le \delta_3, \tag{16}$$

$$\max_{(ij)} \sup_{[0,T]} |h(r)^{-1}\big(D^r_{ij}(h(r)r^{-1}t) - \mu_{ij}\int_0^{h(r)r^{-1}t} \Psi^r_{ij}(s)ds\big)| \le \delta_3, \tag{17}$$

imply

$$\sup_{[0,T]} |\bar{f}^r(t) - \bar{f}(t)| \le \delta_2, \tag{18}$$

where $\bar{f}(\cdot)$ is a hydrodynamic model.

Proof. Suppose not. Fix T, K, δ_2. There must exist a sequence $\delta_3 \downarrow 0$, and a corresponding sequence $r = r(\delta_3) \uparrow \infty$, such that (16), (17) and the convergence $\bar{f}^r(0) \to \bar{f}(0)$ of initial states hold, but (18) fails for *any* hydrodynamic model. This, however, is impossible, because according to

Theorem 19 (or rather its version, specialized to a finite time interval, to be precise) we can choose a further subsequence of r along which $\overline{f}^r(t) \to \overline{f}(t)$, uniformly in $[0,T]$, where $\overline{f}(\cdot)$ is a hydrodynamic model starting from $\overline{f}(0)$. □

THEOREM 21. *For any $K > 0$ there exists a finite time $T = T(K)$ and constant $C = C(K) > 0$ such that all hydrodynamic models with $\left|\overline{f}(0)\right| \leq K$ satisfy the following conditions: $\sum_i \overline{\psi}_{ij}(T) = 0, \forall j < J$, $\overline{q}_i(T) = 0, \forall i \in \mathcal{I}$, and $\overline{f}(t) \equiv \overline{f}(T)$ for all $t \geq T$; $\max_{t\geq 0} |\overline{f}(t)| \leq CK$. Moreover, $(\overline{\psi}_{ij}(T), \overline{q}_i(T)) = L(\overline{\psi}_{ij}(0), \overline{q}_i(0))$, where L is a fixed linear mapping.*

PROOF. Consider a fixed HM $\overline{f}(\cdot)$. Consider the highest priority activity $(1j)$. There are two possible cases: j is a leaf or 1 is a leaf.
Case a: If j is a leaf, then $\overline{\psi}_{1j}(t) \leq 0$ at all times, and $\overline{\psi}_{1j}(t)$ must increase at positive rate, bounded away from 0, until it reaches 0 within a finite time. Thereafter, $\overline{\psi}_{1j}(t)$ will stay at 0. (The argument is very similar to Case b in the proof of Proposition 13.)
Case b: If type 1 is a leaf, then $\overline{q}_1(t)$ must decrease and $\overline{\psi}_{1j}(t)$ increase at the same rate (positive, bounded away from 0), until the entire queue (if any) "relocates into" $\overline{\psi}_{1j}$; after that time, $\overline{\psi}_{1j}(t)$ and $\overline{q}_1(t) = 0$ will not change.
We see that in either case a or b, after a finite time, the highest priority activity $(1j)$ can be in a sense "ignored". This allows us to proceed by induction on the activities, from the highest priority to the lowest, to check that by some finite time T (depending on K) the hydrodynamic model gets into a state $\overline{f}(T)$, satisfying conditions of the theorem, and will stay in this state for all $t \geq T$. Since all HMs are uniformly Lipschitz, we obviously have a uniform bound $\max_{t\geq 0} |\overline{f}(t)| \leq CK$ for some C.

Furthermore, since all $\overline{x}_i(t)$ do not change with time, the linear mapping L is as follows: $L(u_{ij}, w_i) = (c_{ij}, 0)$ where (c_{ij}) is the unique solution to

$$\sum_j u_{ij} + w_i = \sum_j c_{ij}, \quad \forall i \in \mathcal{I} \tag{19a}$$

$$\sum_i c_{ij} = 0, \quad \forall j < J \tag{19b}$$

□

REMARK 22. Examination of the proof of Theorem 21 reveals that the HM for any initial state is in fact unique. Moreover, with a little further argument, it is easy to show that an HM depends on the initial state continuously. Furthermore, the HMs are scalable: if $(\overline{f}(t),\ t \geq 0)$ is an HM, then so is $(\overline{f}(ct)/c,\ t \geq 0)$ for any $c > 0$. From here, it is easy to find that the theorem statement holds for a constant C independent of K and for $T = CK$. We will not need these stronger properties in this paper.

For future reference, note that $L(u_{ij}, w_i) = (c_{ij}, 0)$ is a function only of the vector (z_i), where $z_i = w_i + \sum_j u_{ij}$. The corresponding linear mapping from (z_i) to (c_{ij}), we denote L'.

5.3. *Local-fluid scaling.* The process under *local fluid scaling* is as follows. For each r consider

$$(\tilde{\psi}^r_{ij}(t), \tilde{q}^r_i(t)) \equiv \tilde{f}^r(t) = h(r)^{-1} F^r(t).$$

We will also denote $\tilde{x}^r_i(t) = h(r)^{-1}[X^r_i(t) - \sum_j \psi^*_{ij} r] \equiv \tilde{q}^r_i(t) + \sum_j \tilde{\psi}^r_{ij}(t)$.

Since $\tilde{\psi}^r_{ij}(\cdot)$ (as well as $\overline{\psi}^r_{ij}(\cdot)$) is centered before it is scaled in space, we in particular have (by Assumption 7) that $\sum_i \tilde{\psi}^r_{ij}(t) \leq 0$ for all $j < J$ at all times t.

Theorem 23. *Consider a sequence of deterministic realizations, such that the driving realizations satisfy FSLLN conditions, namely:*

$$(h(r)^{-1}(A^r_i(t) - \lambda_i r t),\ t \geq 0) \to 0, \quad u.o.c.,\ \forall i \tag{20}$$

$$\Big(h(r)^{-1}\big(D^r_{ij}(t) - \mu_{ij} \int_0^t \Psi^r_{ij}(s) ds\big),\ t \geq 0\Big) \to 0, \quad u.o.c.,\ \forall (ij). \tag{21}$$

Assume that the initial states converge to a fixed vector $(\tilde{\psi}^r_{ij}(0), \tilde{q}^r_i(0)) \to (\tilde{\psi}_{ij}(0), \tilde{q}_i(0))$. Further assume that $\tilde{q}_i(0) = 0, \forall i$, and $\sum_i \tilde{\psi}_{ij}(0) = 0$ for all $j < J$. (In other words, $(\tilde{\psi}_{ij}(0), \tilde{q}_i(0)) = L(\tilde{\psi}_{ij}(0), \tilde{q}_i(0))$.) Then, for any subsequence of r there exists a further subsequence along which

$$(\tilde{\psi}^r_{ij}(\cdot), \tilde{q}^r_i(\cdot)) \to (\tilde{\psi}_{ij}(\cdot), \tilde{q}_i(\cdot)), \quad u.o.c., \tag{22}$$

where $(\tilde{\psi}_{ij}(\cdot), \tilde{q}_i(\cdot)$ is a set of Lipschitz functions, with initial condition $(\tilde{\psi}_{ij}(0), \tilde{q}_i(0))$, satisfying (local fluid model) conditions (24)*. Moreover, these limit trajectories depend continuously on the initial state and are such that, uniformly on all of them,*

$$|(\tilde{\psi}_{ij}(t), \tilde{q}_i(t))| \leq |(\tilde{\psi}_{ij}(0), \tilde{q}_i(0))| c_1 e^{-c_2 t}, \quad \forall t \geq 0, \tag{23}$$

where $c_1, c_2 > 0$ are fixed constants.

The local fluid model conditions are as follows:

$$\tilde{q}_i(t) = 0, \quad \forall i \in \mathcal{I} \tag{24a}$$

$$\sum_j \tilde{\psi}_{ij}(t) = \sum_j \tilde{\psi}_{ij}(0) - \sum_j \int_0^t \mu_{ij}\tilde{\psi}_{ij}(s)ds, \quad \forall i \in \mathcal{I} \tag{24b}$$

$$\sum_i \tilde{\psi}_{ij}(t) = 0, \quad \forall j < J \tag{24c}$$

The $I+J-1$ equations for the $I+J-1$ functions $(\tilde{\psi}_{ij}(\cdot))$ can be solved sequentially, in order of decreasing activity priority, since the highest unsolved-for priority will always correspond to either a customer-type or a server-type leaf of the remaining activity tree. Any Lipschitz trajectory satisfying (24) we will call a *local fluid model* (LFM). Conditions (24) reduce to a system of linear ODEs for $(\tilde{\psi}_{ij}(t))$, which of course implies the continuous dependence on initial state; the fact that each LFM converges to 0 is easily established, again by induction on activities; therefore, we obtain the uniform exponential bound (23).

Analogously to $\tilde{f}^r(\cdot) = (\tilde{\psi}^r_{ij}(\cdot), \tilde{q}^r_i(\cdot))$, we will use shorter notation $\tilde{f}(\cdot) = (\tilde{\psi}_{ij}(\cdot), \tilde{q}_i(\cdot))$.

Proof of Theorem 23. The non-trivial part of the proof is showing the Lipschitz property of the limit $\tilde{f}(\cdot)$, because it is no longer a simple consequence of the FSLLN for the driving processes (as it was for the fluid and hydrodynamic limits). This is because the arrival and service rates in the system (with index r) are $O(r)$, while the space is scaled down by $h(r) = o(r)$. For the same reason, it is also not "automatic" that the limit queues $\tilde{q}_i(\cdot)$ stay at 0. This difficulty is resolved as follows. Consider arbitrary number $C_4 > \|(\tilde{\psi}_{ij}(0))\|$, and the random time $\tau(r) = \min\{t \mid \|(\tilde{\psi}^r_{ij}(t))\| \geq C_4\}$. Then, for each i, the sequence of trajectories $\tilde{x}^r_i(\cdot)$ is "asymptotically Lipschitz" in the interval $[0, \tau(r)]$ with the Lipschitz constant $\eta = C_4\|(\mu_{ij})\|$; namely, if we consider each trajectory stopped at the corresponding time $\tau(r)$, i.e. $\tilde{x}^r_i(\min\{t, \tau(r)\})$, then any subsequence of trajectories contains a further subsequence, converging u.o.c. to a Lipschitz trajectory with constant η. This is because in $[0, \tau(r))$ the scaled difference of arrival and departure rates, $h(r)^{-1}|\lambda_i r - \sum_j \mu_{ij}\Psi^r_{ij}(t)| = |\sum_j \mu_{ij}\tilde{\psi}^r_{ij}(t)|$, is upper bounded by η, and we have (20)-(21). (Recall that $\lambda_i = \sum_j \mu_{ij}\psi^*_{ij}$ and $\tilde{\psi}^r_{ij}(t) = h(r)^{-1}[\Psi^r_{ij}(t) - \psi^*_{ij}r]$.) Similarly, each queue length trajectory $\tilde{q}^r_i(\cdot)$ is asymptotically Lipschitz in $[0, \tau(r)]$.

From the asymptotic Lipschitz properties described above, we obtain the following. If $\tau(r) \to 0$ along some subsequence, then (denoting $\tilde{x}_i(0) = \sum_j \tilde{\psi}_{ij}(0)$)

$$\sup_{[0,\tau(r)]} \|(\tilde{x}_i^r(t)) - (\tilde{x}_i(0))\| \to 0, \qquad \sup_{[0,\tau(r)]} \|(\tilde{q}_i^r(t)) - (\tilde{q}_i(0))\| \to 0. \tag{25}$$

If $\liminf \tau(r) > \epsilon_4 > 0$ along some subsequence, then there exists a further subsequence along which

$$(\tilde{x}_i^r(\cdot)) \to (\tilde{x}_i(\cdot)), \qquad (\tilde{q}_i^r(\cdot)) \to (\tilde{q}_i(\cdot)), \tag{26}$$

where the convergences are uniform in $[0, \epsilon_4]$, and each function $\tilde{x}_i(\cdot)$ and $\tilde{q}_i(\cdot)$ is Lipschitz with constant η in $[0, \epsilon_4]$.

In the case $\tau(r) \to 0$, as a consequence of (25), we also must have

$$\sup_{[0,\tau(r)]} \|(\tilde{\psi}_{ij}^r(t)) - (\tilde{\psi}_{ij}(0))\| \to 0. \tag{27}$$

Indeed, if (27) does not hold, then for some fixed $\delta > 0$, we can choose a subsequence of r and a corresponding sequence of times $\tau_1(r) \in [0, \tau(r)]$ such that $\|(\tilde{\psi}_{ij}^r(t)) - (\tilde{\psi}_{ij}(0))\| \geq \delta$ for the first time. Fix $T > 0$ (we will specify the choice later), and consider the sequence of times $\tau_1(r) - Th(r)/r$. Suppose first that $\tau_1(r) - Th(r)/r \geq 0$ for infinitely many r; then, we consider a further subsequence along which this holds, and the trajectory on the time interval $[\tau_1(r) - Th(r)/r, \tau_1(r)]$. Now, if we reset the time origin to $\tau_1(r) - Th(r)/r$ and "stretch" the interval $[\tau_1(r) - Th(r)/r, \tau_1(r)]$ by the factor $r/h(r)$, we will obtain hydrodynamic-scaled trajectories in the interval $[0, T]$. We then choose a further subsequence of r along which these trajectories converge (u.o.c.) to an HM. This HM $\overline{f}(\cdot)$ will be such that $(\overline{x}_i(0)) = (\tilde{x}_i(0))$, $\|(\overline{\psi}_{ij}(0)) - (\tilde{\psi}_{ij}(0))\| \leq \delta$, all $\overline{q}_i(0) = 0$, and $\|(\overline{\psi}_{ij}(T)) - (\tilde{\psi}_{ij}(0))\| \geq \delta$. We now specify the choice of T: it is large enough so that $(\overline{\psi}_{ij}(T)) = L'(\overline{x}_i(0))$. But, $(\overline{x}_i(0)) = (\tilde{x}_i(0))$, which means $(\overline{\psi}_{ij}(T)) = L'(\tilde{x}_i(0)) = (\tilde{\psi}_{ij}(0))$ – a contradiction. The contradiction in the case when $\tau_1(r) - Th(r)/r < 0$ for all large r is obtained similarly, except for all r we use time interval $[0, Th(r)/r]$ to construct a contradicting HM. Thus, we proved (27). But, this leads to a contradiction with the definition of $\tau(r)$. We conclude that the case $\tau(r) \to 0$ is in fact impossible, and we always have the case $\liminf \tau(r) > \epsilon_4 > 0$ and (26).

Next, in addition to (26), we show that

$$|\tilde{f}^r(t) - L\tilde{f}^r(t)| \to 0, \quad \text{in particular } |(\tilde{\psi}_{ij}^r(t)) - L'(\tilde{x}_i^r(t))| \to 0, \tag{28}$$

uniformly in $[0,\epsilon_4]$. (This is, again, proved by contradiction. If (28) would not hold, we would be able to construct an HM violating the claim of Theorem 21 that $\overline{f}(t) = L\overline{f}(t)$ must hold after a finite time. We omit details which are analogous to those in the proof of (27) above.)

In $[0,\epsilon_4]$ we also have

$$\tilde{x}_i(t) = \tilde{x}_i(0) - \tilde{d}_i(t), \quad \forall i,$$

where the Lipschitz function $\tilde{d}_i(\cdot)$ is a limit (along a subsequence) of $\sum_j \int_0^t \mu_{ij}\tilde{\psi}_{ij}^r(s)ds$.

The above properties lead to conditions (24) on the interval $[0,\epsilon_4]$. Namely, we formally define $(\tilde{\psi}_{ij}(\cdot)) = L'(\tilde{x}_i(\cdot))$, obtain the convergence $(\tilde{\psi}_{ij}^r(\cdot)) \to (\tilde{\psi}_{ij}(\cdot))$ from (28), and then (24) follows.

Finally, as already observed earlier, the linear ODE (24) solutions satisfy condition (23). In particular, each local fluid model remains bounded in $[0,\infty)$. This in turn allows us to conclude that by choosing a sufficiently large C_4, the corresponding ϵ_4 can be arbitrarily large. This completes the proof. □

We will actually need a generalized version of Theorem 23.

THEOREM 24. *Consider a sequence of deterministic realizations, such that the driving realizations satisfy (20)-(21). Assume that the initial states converge to a fixed vector $\tilde{f}^r(0) \to \tilde{f}^\circ(0)$. (We do* not *assume $\tilde{f}^\circ(0) = L\tilde{f}^\circ(0)$.) Then, for any subsequence of r there exists a further subsequence along which*

$$\tilde{f}^r(\cdot) \to \tilde{f}(\cdot), \tag{29}$$

uniformly on compact subsets of $[0,\infty)$ not containing 0, where $\tilde{f}(\cdot)$ is a local fluid model with initial state $\tilde{f}(0) = L\tilde{f}^\circ(0)$. (Recall that (23) holds for any LFM.) In addition, for any $K > 0$ there exists $C = C(K) > 0$ such that $|\tilde{f}^\circ(0)| \leq K$ implies

$$\limsup_{r\to\infty} \sup_{0\leq t\leq 1} |\tilde{f}^r(t)| \leq CK. \tag{30}$$

PROOF. The proof is a slight generalization of that of Theorem 23. For a fixed $T_5 > 0$ consider the interval $[0, T_5 h(r)/r]$, and the corresponding hydrodynamic-scaled trajectories in the interval $[0, T_5]$; T_5 is chosen large enough so that the hydrodynamic model reaches state $\tilde{f}(0) = L\tilde{f}^\circ(0)$ by time

T_5. Then, we must have $\tilde{f}^r(T_5 h(r)/r) \to \tilde{f}(0)$; moreover, by Theorem 21 and Corollary 20,

$$\limsup_{r\to\infty} \sup_{0\le t\le T_5 h(r)/r} |\tilde{f}^r(t)| \le CK,$$

for some C, when $|\tilde{f}^\circ(0)| \le K$.

Then, we consider local fluid scaled trajectories starting time point $T_5 h(r)/r$ (as opposed to 0), and the rest of the proof is essentially same as that of Theorem 23. □

The following corollary is derived from Theorem 24 analogously to the way Corollary 20 was derived from Theorem 19.

Corollary 25. *For any fixed $K > 0$, there exists $C = C(K) > 0$ such that the following holds. For any fixed $T > 0$, $\delta_2 > 0$ and $\epsilon_2 > 0$, there exists a sufficiently small $\delta_3 > 0$ such that: uniformly on all $|\tilde{f}^r(0)| \le K$ and all sufficiently large r, conditions*

$$\max_i \sup_{[0,T]} |h(r)^{-1}(A_i^r(t) - \lambda_i r t)| \le \delta_3, \tag{31}$$

$$\max_{(ij)} \sup_{[0,T]} |h(r)^{-1}\big(D_i^r(t) - \mu_{ij}\int_o^t \Psi_{ij}^r(s)ds\big)| \le \delta_3, \tag{32}$$

imply

$$\sup_{[0,T]} |\tilde{f}^r(t)| \le (K+1)C, \tag{33}$$

$$\sup_{[\epsilon_2,T]} |\tilde{f}^r(t) - \tilde{f}(t)| \le \delta_2, \tag{34}$$

where $\tilde{f}(\cdot)$ is a local fluid model with $|\tilde{f}(0) - L\tilde{f}^r(0)| \le \delta_2$.

5.4. *Proof of Theorem 10(ii).* We are now in position to prove (11), and then Theorem 10(ii). The basic idea is to consider the process in the interval $[0, T\log r]$, subdivided into $\log r$ intervals, each being T-long. (To be precise, we need to consider an integer number, say $\lfloor \log r \rfloor$, of subintervals. This does not cause any difficulties besides making notation cumbersome.) Then, using the local fluid limit results, we show that, with high probability, in each of the T-long subintervals, the norm $|F^r(t)|$ decreases by a factor $\delta_6 \in (0,1)$, unless the norm $|F^r(t)|$ at the beginning of the subinterval is below $r^{1/2+\epsilon}$ – in this case $|F^r(t)|$ will be bounded above by $3Cr^{1/2+\epsilon}$ in the entire subinterval (where C is as in Corollary 25 with $K = 2$). Now, if δ_6 is small enough, so that

$$\delta_6^{\log r} < r^{1/2+\epsilon}/r, \quad \text{that is} \quad \delta_6 < e^{-1/2+\epsilon}, \tag{35}$$

this means that $|F^r(t)|$ must "dip" below $r^{1/2+\epsilon}$ at least once, and therefore $|F^r(T \log r)| \leq 3Cr^{1/2+\epsilon}$ (with high probability). We proceed with the details.

Let us choose $\delta_6 > 0$ satisfying (35), and then $\delta_2 > 0$ such that $2\delta_2 < \delta_6$. Denote by $|L|$ the norm of the linear operator L (defined in Theorem 21), i.e. the maximum of absolute values of its eigenvalues. Let us choose $T > 0$ large enough so that (see Theorem 24) $|L|c_1 e^{-c_2 T} < \delta_2$.

Suppose, for each r the initial state is as in (11). To prove (11) it suffices to show that from any subsequence of r we can find a further subsequence, along which (11) holds. So, consider any fixed subsequence, and a fixed $\delta_1 > 0$.

In each of the subintervals $[(i-1)T, iT]$, $i = 1, 2, \dots, \log r$, we consider the process with the time origin reset to $(i-1)T$ and the corresponding initial state $F^r((i-1)T)$; and if $|F^r((i-1)T)| \leq g(r)$, then we set $h(r) = \max(|F^r((i-1)T)|, r^{1/2+\epsilon})$. (If $|F^r((i-1)T)| > g(r)$ we set $h(r) = g(r)$ for completeness.) By Proposition 18, we can choose a further subsequence, with r increasing sufficiently fast, so that, w.p.1, conditions (31) and (32) hold for all large r, *simultaneously* on each of the subintervals $[0, T]$, $[T, 2T]$, ..., $[T(\log r - 1), T \log r]$. We consider the corresponding local fluid scaled processes $\tilde{f}^r(\cdot)$, with their corresponding $h(r)$, on each of the subintervals; and apply Corollary 25. We see that, with probability 1, for all large r, the following holds for each interval $[(i-1)T, iT]$, $i = 1, 2, \dots, \log r$:
if $|F^r((i-1)T)| \in [r^{1/2+\epsilon}, g(r)]$ then $|F^r(iT)| \leq 2\delta_2 |F^r((i-1)T)|$;
if $|F^r((i-1)T)| < r^{1/2+\epsilon}$ then $|F^r(iT)| \leq 3Cr^{1/2+\epsilon}$.
Since $2\delta_2 < \delta_6$ we must have $|F^r(iT)| < r^{1/2+\epsilon}$ for at least one i. Finally, we conclude that condition $|F^r(T \log r)| \leq 3Cr^{1/2+\epsilon}$ must hold (w.p.1 for all large r). This obviously implies (11).

6. Acknowledgments. The authors would like to thank the referees for their helpful comments.

REFERENCES

[1] Armony, M. and Ward, A. (October 2011). Blind fair routing in large-scale service systems. Preprint. `http://www-bcf.usc.edu/~amyward/ArWa_10_6_11`.

[2] Atar, R., Shaki, 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.* Wiley.

[4] Gurvich, I. and Whitt, W. (May 2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of OR* **34**, 2, 363–396.

[5] Mandelbaum, A. and Stolyar, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research 52*, 836–855.

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

[7] Stolyar, A. L. and Tezcan, T. (2011). Shadow routing based control of flexible multi-server pools in overload. *Operations Research 59*, 1427–1444.

[8] Stolyar, A. L. and Yudovina, E. (December 2010). Systems with large flexible server pools: Instability of "natural" load balancing. Submitted. `arXiv:1012.4140`.

Murray Hill, NJ
E-mail: stolyar@research.bell-labs.com

Ann Arbor, MI
E-mail: yudovina@umich.edu