Instability of natural load balancing in large-scale flexible-server systems

Alexander L. Stolyar and Elena Yudovina

***Abstract*— We consider large-scale service systems with several customer classes and several server pools. Mean service time of a customer depends both on the customer class and the server type. The routing is restricted to a fixed set of "activities," i.e. (customer-class, server-type) pairs. We assume that the bipartite graph with vertices being customer-classes and server-types, and edges being the activities, is a tree. The system behavior under a natural load balancing routing/scheduling rule, Longest-queue freest-server (LQFS-LB), is studied in both fluid-limit and Halfin-Whitt asymptotic regimes. We show that, quite surprizingly, LQFS-LB may render the system unstable in the vicinity of the equilibrium point. Such instability cannot occur in systems with "small" number of customer classes. We prove stability in one important special case.**

I. INTRODUCTION

Large-scale service systems (such as call centers) with heterogeneous customer and server (agent) populations bring up the need for efficient dynamic control policies that dynamically match arriving (or waiting) customers and available servers. In this setting, two goals are desirable. On the one hand, customers should not be kept waiting if this is possible. On the other hand, idle time should be distributed fairly among the servers. Furthermore, we would like the control policy to depend only on the current system state and not on knowledge of the arrival rates or mean service times.

The *Shadow Routing* policy proposed in [12], [13] achieves the load balancing objective without a priori knowledge of the arrival rates. However, it does need to "know" the service rates.

In this paper we consider the case when the possible routing choices are restricted to a fixed set of *activities* — (customer-class, server-type) pairs — which form a tree, defined precisely in Section II. The specific control rule we analyze in this paper can be seen as a special case of the Queue-and-Idleness-Ratio rule considered in [7]. Within the given activity tree, if an arriving customer sees multiple available servers (that can serve it), it will choose the server pool with the smallest load; while if a server sees several customers waiting in queues (that it can serve), it will take a customer from the longest queue. We call this rule *Longest-queue freest-server* (LQFS-LB).

A similar set-up has been considered by Gurvich and Whitt [7], Atar-Shaki-Shwartz [2], Armony and Ward [1], and others. Some of the results in these papers prove optimal behavior of simple load balancing schemes on a *finite time interval*. Our results show that the *stationary distribution* of such systems may not be "concentrated" near the equilibrium point, which is the optimal operating point.

More precisely, we consider the many-server asymptotic regime, such that the input rate of class i customers is $\lambda_i r + o(r)$, the number of servers in pool j is $\beta_j r$, where λ_i and β_j are some positive constants, $r \to \infty$ is a scaling parameter, and service rates μ_{ij} (of a class i customer by a type j server) remain constant. Our key results show that the *fluid limit* of the system process (obtained via space-scaling by $1/r$) can be unstable in the vicinity of the equilibrium point.

Using the fluid limit local instability (when such occurs), we prove that the sequence of stationary distributions of *diffusion-scaled* processes (measuring $O(\sqrt{r})$ deviations from the equilibrium point) may be non-tight, and in fact may escape to infinity.

In addition to the instability examples, we prove that in several cases the fluid limit will be (at least locally) stable. In the case when the service rate depends only on the server type (but not customer type, as long as it can be served), we show more — the global stability of the fluid limit.

General results on the asymptotics of stationary distributions (most importantly — their tightness), especially in the many-server systems' diffusion limit, are notoriously difficult to derive. (For recent results in this direction see [5], [6], [4].) In the special case when the service rate depends only on the server type, we prove that under the LQFS-LB policy the sequence of stationary distributions of diffusion-scaled processes is tight, and the limit of stationary distributions is the stationary distribution of the limiting diffusion process.

More details, including omitted proofs, may be found in [14]. Also, in this paper we mostly consider *underloaded* case, where the optimal average system utilization ρ is strictly less than 1; paper [14] contains results for the *critical load* case, $\rho = 1$, as well.

II. MODEL

A. The model; Static Planning (LP) Problem

Consider the model in which there are I customer classes, or types, labelled $1, 2, \ldots, I$, and J server (agent) pools, or classes, labelled $1, 2, \ldots, J$ (generally, we will use the subscripts i, i' for customer classes, and j, j' for server pools). The sets of customer classes and server classes will be denoted by $\mathcal{I}$ and $\mathcal{J}$ respectively.

We are interested in the scaling properties of the system as it grows large. Namely, we consider a sequence of

The second author was supported by the NSF Graduate Research Fellowship.

A. Stolyar is with Bell Labs, Alcatel-Lucent, Murray Hill, NJ, USA stolyar@research.bell-labs.com

E. Yudovina is with Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK e.yudovina@statslab.cam.ac.uk

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 $\lambda_i^r = r\lambda_i + o(r)$, 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 $\mu_{ij} = 0$.

We would like to balance the proportion of busy servers across the server pools, while keeping the system operating efficiently. Let λ_{ij}^r be the average rates at which type i customers are routed to server pools j. We would like the system state to be such that λ_{ij}^r are close to $\lambda_{ij}r$, where $\{\lambda_{ij}\}$ is an optimal solution to the following *static planning problem* (SPP), which is the following linear program:

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

subject to

$$\lambda_{ij} \geq 0, \quad \forall i, j \tag{2}$$

$$\sum\nolimits_j \lambda_{ij} = \lambda_i, \quad \forall i \tag{3}$$

$$\sum\nolimits_i \lambda_{ij}/(\beta_j\mu_{ij}) \leq \rho, \quad \forall j. \tag{4}$$

We assume that the SPP has a unique optimal solution $\{\lambda_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\}, \rho$; and it is such that the *basic activities*, i.e. those pairs, or edges, (ij) for which $\lambda_{ij} > 0$, form a (connected) tree in the graph with vertices set $\mathcal{I} \cup \mathcal{J}$. The set of basic activities is denoted $\mathcal{E}$. These assumptions constitute the *complete resource pooling* (CRP) condition, which holds "generically"; see [13, Theorem 2.2]. For a customer type i, let $\mathcal{S}(i) = \{j : (ij) \in \mathcal{E}\}$; for a server type j, let $\mathcal{C}(j) = \{i : (ij) \in \mathcal{E}\}$.

Note that under the CRP condition, all ("server pool capacity") constraints (4) are binding; in other words, the optimal solution to SPP minimizes and "perfectly balances" server pool loads. Optimal dual variables $\nu_i,\ i \in \mathcal{I}$, and $\alpha_j,\ j \in \mathcal{J}$, corresponding to constraints (3) and (4), respectively, are unique and all strictly positive; ν_i is interpreted as the "workload" associated with one type i customer, and α_j is interpreted as the (scaled by $1/r$) maximum rate at which server pool j can process workload. The following relations hold:

$$\alpha_j = \max_i \nu_i\beta_j\mu_{ij} \qquad \nu_i = \min_j \alpha_j/(\beta_j\mu_{ij})$$
$$\sum\nolimits_j \alpha_j = 1 \qquad \sum\nolimits_i \lambda_i\nu_i = \rho\sum\nolimits_j \alpha_j = \rho.$$

If $\rho < 1$, the system is called *underloaded*; if $\rho = 1$, the system is called *critically loaded*. Most of this paper is devoted to $\rho < 1$, with the exception of Section VII. More results for the critical load can be found in [14].

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.)

Let $\psi_{ij}^* = \lambda_{ij}/\mu_{ij}$. Continuing our interpretation of the optimal operating point of the system, let $\Psi_{ij}^r(t)$ be the number of servers of type j serving customers of type i at time t. It is desirable to have $\Psi_{ij}^r(t) = r\psi_{ij}^* + o(r)$. Later on we will be also interested in the question of whether or not the $o(r)$ term can in fact be $O(\sqrt{r})$.

B. Longest-queue, freest-server load balancing algorithm (LQFS-LB)

For the rest of the paper, we analyze the performance of the following intuitive load balancing algorithm.

We introduce the following notation (for the system with scaling parameter r):
$\Psi_{ij}^r(t)$, the number of servers of type j serving customers of type i at time t;
$\Psi_j^r(t) = \sum_i \Psi_{ij}^r(t)$, the total number of busy servers of type j at time t;
$\Psi_i^r(t) = \sum_j \Psi_{ij}^r(t)$, the total number of servers serving type i customers at time t;
$\Xi_j^r(t) = \Psi_j^r(t)/\beta_j$, the instantaneous load of server pool j at time t;
$Q_i^r(t)$, the number of customers of type i waiting for service at time t;
$X_i^r(t) = \Psi_i^r(t) + Q_i^r(t)$, the total number of customers of type i in the system at time t.

The algorithm 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: If an arriving customer of type i sees any unoccupied servers in server classes in $\mathcal{S}(i)$, it will pick a server in the least loaded server pool, i.e. $j \in \arg\min_{j' \in \mathcal{S}(i)} \Xi_{j'}^r(t)$. (Ties are broken in an arbitrary Markovian manner.)

Scheduling: If a server of type j, upon completing a service, sees a customer of a class in $\mathcal{C}(j)$ in queue, it will pick the customer from the longest queue, i.e. $i \in \arg\max_{i' \in \mathcal{C}(j)} Q_{i'}^r$. (Ties are broken in an arbitrary Markovian manner.)

By [7, Remark 2.3], the LQFS-LB algorithm described here is a special case of the algorithm proposed by Gurvich and Whitt, with constant probabilities $p_i = \frac{1}{I}$ (queues "should" be equal), $v_j = \frac{\beta_j}{\sum \beta_j}$ (the proportion of idle servers "should" be the same in all server pools).

We also note that *stochastic stability* (i.e. positive recurrence of the underlying Markov process, describing the system) under LQFS-LB is guaranteed, as long as system

is subcritically loaded, which is the case (for all large r) when $\rho < 1$ or when $\rho = 1$ and the asymptotic regime is as in Section VII. This is because the LQFS-LB scheduling rule can be equivalently written as: serve $i \in \arg\max_{i' \in \mathcal{C}(j)} \mu_{i'j}\nu_{i'}Q^r_{i'}$. (Here we use the fact that for a given j the value of $\mu_{ij}\nu_i$ is the same for all $i \in \mathcal{C}(j)$.) Therefore, LQFS-LB scheduling rule is a special case of "MaxWeight" scheduling rule (or, "Gcmu" with quadratic cost functions), which is known to ensure stochastic stability under subcritical load, regardless of the routing rule. Cf. [10] for details.

C. Basic notation

Vector $(\xi_i, i \in \mathcal{I})$, where ξ can be any symbol, is often written as (ξ_i) or $\xi_{\mathcal{I}}$; similarly, $(\xi_j, j \in \mathcal{J}) = (\xi_j) = \xi_{\mathcal{J}}$ and $(\xi_{ij},\ (ij) \in \mathcal{E}) = (\xi_{ij}) = \xi_{\mathcal{E}}$. In matrix expressions, $\xi_{\mathcal{I}}$, $\xi_{\mathcal{J}}$ and $\xi_{\mathcal{E}}$ are viewed as column-vectors; v' denotes transposition of v. 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, $(\xi_i(\cdot))$ and $\xi_{\mathcal{I}}(\cdot)$ both signify $((\xi_i(t), i \in \mathcal{I}),\ t \geq 0)$.

The symbol $\Rightarrow$ denotes convergence in distribution of either random variables in the Eucledian space $\mathbb{R}^d$ (with appropriate dimension d), or random processes in the Skorohod space $D^d[\eta, \infty)$ of RCLL (right-continuous with left limits) functions on $[\eta, \infty)$, for some constant $\eta \geq 0$. (Unless explicitly specified otherwise, $\eta = 0$.) We always consider the Borel σ-algebra on $\mathbb{R}^d$. The symbol $\rightarrow$ denotes ordinary convergence in $\mathbb{R}^d$. Standard Eucledian norm of a vector $x \in \mathbb{R}^d$ is denoted $|x|$.

III. FLUID MODEL

A. Definition

We now consider the behavior of fluid models (or, fluid paths) associated with this system. Informally speaking, fluid models form a set of trajectories that w.p.1 contains any limit of fluid-scaled trajectories of the original stochastic system, where fluid scaling in our context is, e.g., $(1/r)X^r_i(t)$ and similarly for other variables. (The proof of the fact that the set of fluid models we define just below, indeed satisfies the required property, is standard and is outlined in [14].)

For our system we define a *fluid model* as a set of Lipschitz continuous functions $\{(a_i(\cdot)), (x_i(\cdot)), (q_i(\cdot)), (\psi_{ij}(\cdot)), (\rho_j(\cdot))\}$, which satisfy the equations below. (Here $a_i(\cdot) = (a_i(t),\ t \geq 0)$, and similarly for other components.) The last two equations involving derivatives are to be satisfied at all regular points t, when the derivatives in question exist. The interpretation of the components is as follows: $a_i(t)$ is the total "amount" (i.e. the number, scaled by $1/r$) of arrivals of type i customers into the system by time t; $x_i(t)$ is the "amount" of customers of type i in the system at time t, i.e. the limit of $(1/r)X^r_i(t)$; and, analogously, $q_i(t)$, $\psi_{ij}(t)$, $\rho_j(t)$ are the limits of $(1/r)Q^r_i(t)$, $(1/r)\Psi^r_{ij}(t)$, $\Xi^r_j(t)/r$, respectively.

$$a_i(t) = \lambda_i t, \quad \forall i \in \mathcal{I} \tag{5a}$$

$$x_i(t) = q_i(t) + \sum\nolimits_j \psi_{ij}(t), \quad \forall i \in \mathcal{I} \tag{5b}$$

$$x_i(t) = x_i(0) + a_i(t) - \sum\nolimits_j \int_0^t \mu_{ij}\psi_{ij}(s)ds, \quad \forall i \in \mathcal{I} \tag{5c}$$

$$\rho_j(t) = \frac{1}{\beta_j} \sum\nolimits_i \psi_{ij}(t), \quad \forall j \in \mathcal{J} \tag{5d}$$

$$\rho_j(t) = 1 \text{ if } q_i(t) > 0 \text{ for any } i \in \mathcal{C}(j), \quad \forall j \in \mathcal{J} \tag{5e}$$

For any set of server types $\mathcal{J}^* \subseteq \mathcal{J}$ and any set of customer types $\mathcal{I}^* \subseteq \mathcal{I}$ such that $q_i(t) > 0$ for all $i \in \mathcal{I}^*$, and $q_i(t) > q_{i'}(t)$ whenever $i \in \mathcal{I}^*$, $i' \notin \mathcal{I}^*$ and $\mathcal{S}(i) \cap \mathcal{S}(i') \cap \mathcal{J}^* \neq \emptyset$,

$$\sum_{i \in \mathcal{I}^*} \sum_{j \in \mathcal{S}(i) \cap \mathcal{J}^*} \dot{\psi_{ij}}(t) = \sum_{j \in \cup_{i \in \mathcal{I}^*} \mathcal{S}(i) \cap \mathcal{J}^*} \sum_{i' \in \mathcal{C}(j)} \mu_{ij}\psi_{ij}(t) - \sum_{i \in \mathcal{I}^*} \sum_{j \in \mathcal{S}(i) \cap \mathcal{J}^*} \mu_{ij}\psi_{ij}(t) \tag{5fa}$$

For any sets of customer types $\mathcal{I}_* \subseteq \mathcal{I}$ and server types $\mathcal{J}_* \subseteq \mathcal{J}$ such that $\rho_j(t) < 1$ for all $j \in \mathcal{J}_*$, and $\rho_j(t) < \rho_{j'}(t)$ whenever $j \in \mathcal{J}_*$, $j' \notin \mathcal{J}_*$, and $\mathcal{C}(j) \cap \mathcal{C}(j') \cap \mathcal{I}_* \neq \emptyset$,

$$\sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \dot{\psi_{ij}}(t) = \sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \lambda_i - \sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \mu_{ij}\psi_{ij}(t) \tag{5fb}$$

We comment on (5f) as the least intuitive. Consider the meaning of (5fa) for $\mathcal{I}^* = \{i^*\}, \mathcal{J}^* = \{j^*\}$. Then the customer type $i^* \in \mathcal{C}(j^*)$ has the longest queue among all of the customer types that can be served by j^* (i.e. $q_{i^*}(t) > q_i(t)$ for all other $i \in \mathcal{C}(j^*)$). Equation (5fa) then asserts that all of the departures from server pool j^* are replaced by customers from queue i^*. In general, (5fa) generalizes this fact for multiple queues of maximal length. The second equation, (5fb), generalizes the fact that, if for some customer type i_* ($\mathcal{I}_* = \{i_*\}$), a server type $j_* \in \mathcal{S}(i_*)$ ($\mathcal{J}_* = \{j_*\}$) has minimal load (i.e. $\rho_{j_*}(t) < \rho_j(t)$ for all other $j \in \mathcal{S}(i_*)$), then all of the arrivals to type i_* are directed to server pool j_*.

B. Behavior in the vicinity of equilibrium point

We define the *equilibrium (invariant)* point of the underloaded ($\rho < 1$) fluid model to be the state $\psi_{ij} = \psi^*_{ij}$ and $q_i = 0$ for all $i \in \mathcal{I}$, $j \in \mathcal{J}$. Clearly, fluid model with all $\psi_{ij}(t) \equiv \psi^*_{ij}$ and all $q_i(t) \equiv 0$ (and other components defined accordingly) is indeed a stationary fluid model. Desirable system behavior would be to have $(\psi_{ij}(t)) \rightarrow (\psi^*_{ij})$ as $t \rightarrow \infty$.

Note that if the initial system state is in the vicinity of the equilibrium point (with $\rho < 1$), then there is no queueing in the system, and we can describe the system with just the

variables $(\psi_{ij}(t))$. This will be true for at least some time (depending on ρ and the initial distance to the equilibrium point), because the fluid model is Lipschitz.

The following is a "state space collapse" result for the underloaded fluid model in the neighborhood of the equilibrium point.

Theorem 3.1: Let $\rho < 1$. There exists a sufficiently small $\epsilon > 0$, depending only on the system parameters, such that for all sufficiently small δ the following holds. There exist $T_1 = T_1(\delta)$ and $T_2 = T_2(\delta)$, $0 < T_1 < T_2$, such that if the initial system state $(\psi_{ij}(0))$ satisfies

$$\left|(\psi_{ij}(0)) - (\psi^*_{ij})\right| < \delta,$$

then for all $t \in [T_1, T_2]$ the system state satisfies

$$\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon, \rho_j(t) = \rho_{j'}(t) \text{ for all } j, j' \in \mathcal{J}.$$

Moreover, $T_1 \downarrow 0$ and $T_2 \uparrow \infty$ as $\delta \downarrow 0$. The evolution of the system on $[T_1, T_2]$ is described by a linear ODE, specified below by (10).

In the rest of this section and the paper, the values associated with a stationary fluid model, "sitting" at an equilibrium point, are referred to as *nominal*. For example, ψ^*_{ij} is the nominal occupancy (of pool j by type i), λ_i is the nominal arrival rate, λ_{ij} is the nominal routing rate (along activity (ij)), $\psi^*_{ij}\mu_{ij} = \lambda_{ij}$ is the nominal service rate (of type i in pool j), $\sum_j \psi^*_{ij}\mu_{ij} = \lambda_i$ is the nominal total service rate (of type i), ρ is the nominal total occupancy (of each pool j), etc.

Proof: [Proof of Theorem 3.1] Let us choose a suitably small $\epsilon > 0$ (we will specify how small later). Then, we can always choose some $T_2 > 0$ such that for all sufficiently small $\delta > 0$ we can guarantee that $\left|(\psi_{ij}(0)) - (\psi^*_{ij})\right| < \delta$ implies $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$ for all $t \le T_2$. We will show that $\rho_j(t) = \rho_{j'}(t)$ for all $j, j' \in \mathcal{J}$, in $[T_1, T_2]$ for some T_1 depending on δ.

Consider $\rho_*(t) = \min_j \rho_j(t)$, $\rho^*(t) = \max_j \rho_j(t)$, and assume $\rho_*(t) < \rho^*(t)$. Let $\mathcal{J}_*(t) = \{j : \rho_j(t) = \rho_*(t)\}$. Then the total arrival rate to servers of type $j \in \mathcal{J}_*$ are $\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*} \lambda_i$, which is strictly greater (by a constant) than the nominal arrivals $\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*} \lambda_{ij}$. Since their total service rate is $\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*} \mu_{ij}\psi_{ij}(t)$ and $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$, we see that arrivals exceed services by at least a constant. (This determines what "suitably small" means for ϵ in terms of the system parameters.) Consequently, the minimal load $\rho_*(t)$ is increasing at a rate bounded below by a constant. Similarly, $\rho^*(t)$ is decreasing at a rate bounded below by a constant. Thus, in finite time $T_1 = T_1(\delta)$ we will arrive at a state $\rho_*(t) = \rho^*(t)$. (Clearly, $T_1(\delta) \to 0$ as $\delta \to 0$.) This equality will continue to hold while $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$, in particular for $T_1 \le t \le T_2$.

It remains to derive the differential equation, and to show that T_2 can be chosen depending on δ so that $T_2 \uparrow \infty$ as $\delta \downarrow 0$.

Once we are confined to the manifold $\rho_j(t) = \rho_{j'}(t) = \rho(t)$ for all t, the system evolution is determined in terms of only I independent variables. Decreasing ϵ if necessary to ensure that there is no queueing while $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$, we can take the I variables to be $\psi_i(t) := \sum_j \psi_{ij}(t)$. Given $(\psi_i(t))$ we know $\rho(t)$ as $(\sum_i \psi_i(t))/(\sum_j \beta_j)$. Consequently, we know $\sum_i \psi_{ij}(t) = \rho(t)\beta_j$ and $\sum_j \psi_{ij}(t) = \psi_i(t)$. On a tree, this allows us to solve for $\psi_{ij}(t)$; the relationship will clearly be linear, i.e.

$$(\psi_{ij}(t)) = M(\psi_i(t)) \tag{7}$$

for some matrix M. For future reference, we define the ("load balancing") linear mapping M from $y \in \mathbb{R}^I$ to $z = (z_{ij}, (ij) \in \mathcal{E}) \in \mathbb{R}^{I+J-1}$ as follows: $z = My$ is the unique solution of

$$\eta = \frac{\sum_i y_i}{\sum_j \beta_j}; \quad \sum\nolimits_i z_{ij} = \eta\beta_j, \forall j; \quad \sum\nolimits_j z_{ij} = y_i, \forall i. \tag{8}$$

The evolution of $\psi_i(t)$ is given by

$$\dot{\psi}_i(t) = \lambda_i - \sum\nolimits_j \mu_{ij}\psi_{ij}(t), \ \forall i. \tag{9}$$

Then, by the above arguments we see that this entails (in matrix form)

$$(\dot{\psi}_i(t)) = (\lambda_i) + A_u(\psi_i(t)), \tag{10}$$

where A_u is an $I \times I$ matrix, $A_u = GM$. Here, G is a $I \times (I + J - 1)$ matrix with entries $G_{i,(kj)} = -\mu_{ij}$ if $i = k$, and $G_{i,(kj)} = 0$ otherwise.

It remains to justify the claim that $T_2(\delta) \uparrow \infty$ as $\delta \downarrow 0$. This follows from the fact that, as long as $t \ge T_1$ and $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$, the evolution of the system is described by the linear ODE above. At worst, the solutions of the ODE are escaping away from equilibrium exponentially fast. Therefore, the time it takes the solution of the linear ODE to escape the set $\left|(\psi_{ij}(t)) - (\psi^*_{ij})\right| < \epsilon$ increases to ∞ as $\left|(\psi_{ij}(T_1)) - (\psi^*_{ij})\right|$ decreases to 0. Since as $\delta \downarrow 0$ we have $T_1 \downarrow 0$ and the system is Lipschitz, we also get $\left|(\psi_{ij}(T_1)) - (\psi^*_{ij})\right| \downarrow 0$, and hence we can choose $T_2(\delta) \uparrow \infty$. ∎

Since activities form a tree, if we know $(\psi_i(t))$, we can easily find the common value of $\rho(t)$ and then $(\psi_{ij}(t))$ by working from the leaves inwards. Since in underload we have $\dot{\psi}_i(t) = \lambda_i - \sum_j \mu_{ij}\psi_{ij}(t)$, we obtain an expression for A_u, given in Lemma 3.2(i) just below.

Lemma 3.2: (i) The entries $(A_u)_{ii'}$ of the matrix A_u (for the underload case, $\rho < 1$) are as follows. The coefficient of ψ_i in $\dot{\psi}_i$ is

$$(A_u)_{ii} = -\frac{1}{\sum_j \beta_j} \sum_{j \in \mathcal{S}(i)} \mu_{ij} \sum_{j' \preceq (j,i)} \beta_{j'}.$$

The coefficient of $\psi_{i'}$ in $\dot{\psi}_i$ is

$$(A_u)_{ii'} = \frac{1}{\sum_j \beta_j} \left[- \sum_{j \in \mathcal{S}(i), j \neq j_{ii'}} \mu_{ij} \sum_{j' \preceq (j,i)} \beta_{j'} + \mu_{ij_{ii'}} \sum_{j' \preceq (i, j_{ii'})} \beta_{j'} \right] = (A_u)_{ii} + \mu_{ij_{ii'}}.$$

The relation $\preceq$ is defined as follows. Suppose we disconnect the basic activity tree by removing the edge (i_0, j_0). Then for any node k (either customer type or server type) we say $k \preceq (i_0, j_0)$ if it falls in the same component as i_0; otherwise, $k \preceq (j_0, i_0)$. The server $j_{ii'} \in \mathcal{S}(i)$ is the neighbor of i such that, after removing the edge $(i, j_{ii'})$ from the basic activity tree, nodes i and i' will be in different connected components. (Such a node is unique, since there is a unique path along the tree from i to i'.)

(ii) The matrix A_u is non-singular.

(iii) The matrix A_u depends only on (β_j), (μ_{ij}) and the basic activity tree structure $\mathcal{E}$, and does *not* depend on (λ_i) and (ψ^*_{ij}).

C. Definition of local stability

We say that the (fluid) system is *locally stable*, if for any $C > 0$, all fluid models starting in a sufficiently small neighborhood of the equilibrium point (which is unique for $\rho < 1$) are such that

$$|(\psi_{ij}(t)) - (\psi^*_{ij})| \le C \quad \text{and} \quad |(\psi_{ij}(t)) - (\psi^*_{ij})| \to 0.$$

By Theorem 3.1 we see that the local stability is determined by the stability of a linear ODE, which in turn is governed by the eigenvalues of the matrix A_u. We will call matrix A_u stable if all its eigenvalues have negative real part.[1] In this terminology, *the local stability of the system is equivalent to the stability of the matrix A_u*. On the other hand, if A_u has an eigenvalue with positive real part, the ODE has solutions diverging from equilibrium (ψ^*_i) exponentially fast; if A_u has (a pair of conjugate) pure imaginary eigenvalues, the ODE has oscillating, never converging solutions.

IV. SPECIAL CASES IN WHICH FLUID MODELS ARE STABLE

In this section we analyze two special cases of the system parameters, for which we demonstrate convergence results. In Section IV-A we consider the case when there exists a set of positive μ_j, $j \in \mathcal{J}$, such that $\mu_{ij} = \mu_j$ for $(ij) \in \mathcal{E}$ (i.e. the service rate μ_{ij} is constant across all $i \in \mathcal{C}(j)$); we show global convergence of fluid models to equilibrium. In Section IV-B we consider the case when there exists a set of positive μ_i, $i \in \mathcal{I}$, such that $\mu_{ij} = \mu_i$ for $(ij) \in \mathcal{E}$ (i.e. the service rate μ_{ij} is constant across all $j \in \mathcal{S}(i)$); we show local stability of the fluid model (i.e. stability of A_u).

A. Global stability in the case $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$.

We call the system *globally stable* if any fluid model, with arbitrary initial state, converges to an equilibrium point as $t \to \infty$. (This of course implies $\rho_j(t) \to \rho$ for all $j \in \mathcal{J}$ and $\psi_{ij}(t) \to \psi^*_{ij}$ for all $i \in \mathcal{I}$, $j \in \mathcal{J}$. Note that, in underload, the definition necessarily implies $q_i(t) \to 0$ for all $i \in \mathcal{I}$.)

Theorem 4.1: The system with $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$, is globally stable for $\rho < 1$. In addition, the system is locally stable as well (i.e. the matrix A_u is stable).

[1] A symmetric matrix A is stable if and only if it is negative definite, but A_u is not, in general, symmetric.

The proof is somewhat similar in style to the proof of Theorem 3.1.

B. Local stability in the case $\mu_{ij} = \mu_i$, $(ij) \in \mathcal{E}$.

Theorem 4.2: Assume $\rho < 1$ and $\mu_{ij} = \mu_i$ for $(ij) \in \mathcal{E}$. Then the system is locally stable (i.e. A_u is stable).

Proof: We have

$$\dot{\psi}_i(t) = \lambda_i - \mu_i \psi_i(t)$$

and A_u is simply a diagonal matrix with entries $-\mu_i$. ■

V. FLUID MODELS FOR GENERAL μ_{ij}: LOCAL INSTABILITY EXAMPLES

In Sections IV-A, IV-B we have shown that the matrix A_u is stable in the cases $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$ and $\mu_{ij} = \mu_i$, $(ij) \in \mathcal{E}$. Since the entries of A_u depend continuously on μ_{ij} via Lemma 3.2, and the eigenvalues of a matrix depend continuously on its entries, we know that the matrix will be stable for all parameter settings sufficiently close to those special cases. Therefore, there exists a non-trivial parameter domain of local stability. One might consider it to be a reasonable conjecture that local stability holds for any parameters. It turns out, however, that this conjecture is false. We will now construct examples to demonstrate that, in general, the system can be locally unstable.

Local instability example. Consider a system with 3 customer types A, B, C and 4 server types 1 through 4, connected $1 - A - 2 - B - 3 - C - 4$. Set $\beta_1 = 0.97$ and $\beta_2 = \beta_3 = \beta_4 = 0.01$. Set $\mu_{A1} = \mu_{B2} = \mu_{C3} = 1$, and $\mu_{A2} = \mu_{B3} = \mu_{C4} = 100$. (See Figure 1.) We compute by Lemma 3.2

$$A_u = \begin{pmatrix} -1.99 & -0.99 & -.99 \\ 97.02 & -2.98 & -1.98 \\ 96.03 & 96.03 & -3.97 \end{pmatrix}$$

with eigenvalues $\{-17.8, 4.45 \pm 23.4i\}$. Therefore by Theorem 3.1, the system with these parameters is described by an unstable ODE in the neighborhood of its equilibrium point.

We can show that this is a minimal instability example, in the sense made precise by the following

Lemma 5.1: Consider an underloaded system, $\rho < 1$.

(i) Let $I \ge 2$. Any customer type i that is a leaf in the basic activity tree, does not affect the local stability of the system. Namely, let us modify the system by removing type i, and then modifying (if necessary) input rates λ_k of the remaining types $k \in \mathcal{I} \setminus i$ so that the basic activity tree of the modified system is $\mathcal{E} \setminus (ij)$, where (ij) is the (only) edge in $\mathcal{E}$ adjacent to i. (Note that such modification is easy to construct, and recall that matrix A_u depends only on pool sizes $(\beta_{j'})$, service rates $(\mu_{i'j'})$ and the basic activity tree structure, and *not* on input rates $(\lambda_{i'})$ or nominal occupancies $(\psi^*_{i'j'})$.) Then, the original system is locally stable if and only if the modified one is.

(ii) A system with two (or one) non-leaf customer types is locally stable.

It is possible to construct an instability example with more "reasonable" values of β_j, μ_{ij}, although it will be bigger. The

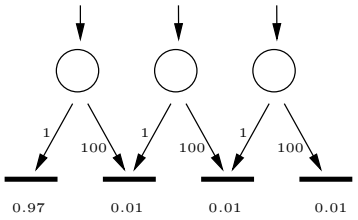

Fig. 1. System with three customer types whose underload equilibrium is unstable

associated matrix A_u and its eigenvalues can be found online [15]. We do not have an explicit characterization of the local instability domain, beyond the necessity of $I \geq 3$.

While sufficiently small systems are at least locally stable, we will show now that, in the underload case, any sufficiently large system is locally unstable for some parameter settings.

Lemma 5.2: In underload ($\rho < 1$), any shape of basic activity tree that includes a locally unstable system (i.e., with A_u having an eigenvalue with positive real part) as a subset will, with some set of parameters (β_j), (μ_{ij}), become locally unstable. In particular, any shape of basic activity tree that includes the above instability examples (for $\rho < 1$) will be locally unstable for some set of parameters β_j, μ_{ij}.

The idea of proof is to show that, if U is unstable and the graph of U is a subset of the graph of S, then by taking β_j to be small on $S \setminus U$ we will obtain a small perturbation of U, which will still be unstable.

VI. DIFFUSION LIMIT OF AN UNDERLOADED SYSTEM. POSSIBLE EVANESCENCE OF INVARIANT DISTRIBUTIONS

Above we have shown that on a fluid scale, around the equilibrium point, the system converges to a subset of its possible states, on which it evolves according to a differential equation, possibly divergent. This strongly suggests that the stochastic system is in fact "never" at equilibrium. Our goal in this section is to demonstrate that it is the case at least on the diffusion scale. More precisely, we consider the system in underload, i.e. $\rho < 1$, and show that the sequence of stationary distributions of the system on the $O(\sqrt{r})$ scale is such that the measure of any compact set vanishes, if the associated fluid model is divergent.

A. Transient behavior of diffusion scaled process. State space collapse

In this section we cite the diffusion limit result (for the process transient behavior) that we will need from [7]. Again, we consider a sequence of systems indexed by r, with the input rates being $\lambda_i^r = r\lambda_i$, server pool sizes being $\beta_j r$, and the service rates μ_{ij} unchanged with r. The notation for the unscaled processes is the same as in the previous section; however, we are now interested in a different — diffusion — scaling. We define

$$\hat{\Psi}_{ij}^r(t) = \frac{\Psi_{ij}^r(t) - r\psi_{ij}}{\sqrt{r}}, \quad \hat{\Psi}_i^r(t) = \sum_j \hat{\Psi}_{ij}^r(t),$$
$$\hat{\Psi}_j^r(t) = \sum_i \hat{\Psi}_{ij}^r(t) = \frac{\Psi_j^r(t) - \rho r \beta_j}{\sqrt{r}} \quad (11)$$

We will denote by M' the linear mapping from $z = (z_{ij}, (ij) \in \mathcal{E}) \in \mathbb{R}^{I+J-1}$ to $y = (y_i) \in \mathbb{R}^I$, given by $\sum_j z_{ij} = y_i$. (So, $(\hat{\Psi}_i^r(t)) \equiv M'(\hat{\Psi}_{ij}^r(t))$.) There is the obvious relation between M' and the operator M defined by (8): $M'My = y$ for any $y \in \mathbb{R}^I$. Let us define $\mathcal{M} := \{My \mid y \in \mathbb{R}^I\}$, an I-dimensional linear subspace of $\mathbb{R}^{I+J-1}$; equivalently, $\mathcal{M} = \{z \in \mathbb{R}^{I+J-1} \mid z = MM'z\}$.

Theorem 6.1: [This is essentially a corollary of Theorem 3.1 and Theorem 4.4 in [7].] Let $\rho < 1$. Assume that as $r \to \infty$, $\hat{\Psi}_{\mathcal{E}}^r(0) \to \hat{\Psi}_{\mathcal{E}}(0)$ where $\hat{\Psi}_{\mathcal{E}}(0)$ is deterministic and finite. (Consequently, $\hat{\Psi}_{\mathcal{I}}^r(0) \to \hat{\Psi}_{\mathcal{I}}(0) = M'\hat{\Psi}_{\mathcal{E}}(0)$.) Then,

$$\hat{\Psi}_{\mathcal{I}}^r(\cdot) \implies \hat{\Psi}_{\mathcal{I}}(\cdot) \text{ in } D^I[0,\infty), \quad (12)$$

and for any fixed $\eta > 0$,

$$\hat{\Psi}_{\mathcal{E}}^r(\cdot) \implies M\hat{\Psi}_{\mathcal{I}}(\cdot) \text{ in } D^{I+J-1}[\eta,\infty), \quad (13)$$

where $\hat{\Psi}_{\mathcal{I}}(\cdot)$ is the unique solution of the SDE

$$\hat{\Psi}_i(t) = \hat{\Psi}_i(0) - \sum_{j \in \mathcal{S}(i)} \mu_{ij} \int_0^t (M\hat{\Psi}_{\mathcal{I}}(s))_{ij} ds + \sqrt{2\lambda_i} B_i(t), \quad (14)$$

for $i \in \mathcal{I}$, and the processes $B_i(\cdot)$ are independent standard Brownian motions.

Recalling the definition of matrix A_u (see (10)), (14) can be written as

$$\hat{\Psi}_{\mathcal{I}}(t) = \hat{\Psi}_{\mathcal{I}}(0) + \int_0^t A_u \hat{\Psi}_{\mathcal{I}}(s) ds + (\sqrt{2\lambda_i} B_i(t)). \quad (15)$$

The meaning of Theorem 6.1 is simple: the diffusion limit of the process $\hat{\Psi}_{\mathcal{I}}^r(\cdot)$ is such that, at initial time 0, it "instantly jumps" to the state $MM'\hat{\Psi}_{\mathcal{E}}(0)$ on the manifold $\mathcal{M}$ (where $MM'\hat{\Psi}_{\mathcal{E}}(0) = \hat{\Psi}_{\mathcal{E}}(0)$ only if $\hat{\Psi}_{\mathcal{E}}(0) \in \mathcal{M}$); after this initial jump, the process stays on $\mathcal{M}$ and evolves according to SDE (15). Theorem 6.1 is "essentially a corollary" of results in [7], because the setting in [7] is such that $\rho = 1$, while we assumed $\rho < 1$. However, our Theorem 6.1 can be proved the same way, and in a sense is easier, because when $\rho < 1$, the queues vanish in the limit (which is why the queue length process is not even present in the statement of Theorem 6.1).

B. Evanescence of invariant measures

In this section we show that if the matrix A_u has eigenvalues with positive real part, the stationary distribution of the (diffusion scaled) process $\hat{\Psi}_{\mathcal{E}}^r(\cdot)$ escapes to infinity as $r \to \infty$. Namely, we prove the following

Theorem 6.2: Suppose $\rho < 1$. Consider a sequence of systems as defined in Section VI-A, and denote by μ^r the stationary distribution of the process $\hat{\Psi}_{\mathcal{E}}^r(\cdot)$, a probability measure on $\mathbb{R}^{I+J-1}$. Let $b_K = \{|z| \leq K\} \subset \mathbb{R}^{I+J-1}$. Suppose the matrix A_u has eigenvalues with positive real

parts and no pure imaginary eigenvalues. Then for any K, $\mu^r(b_K) \to 0$ as $r \to \infty$.

A detailed proof is given in [14]. Here we only outline its key points. Let $\mathcal{C} \subset \mathcal{M}$ denote the submaniford of convergence (stability) of the linear ODE $(d/dt)z = (MA_uM')z$ on $z \in \mathcal{M}$. (In other words, $\mathcal{C} = M\mathcal{C}_{\mathcal{I}}$, where $\mathcal{C}_{\mathcal{I}}$ is the submanifold of convergence of ODE $(d/dt)y = A_u y$ on $\mathbb{R}^I$.) Given assumptions of the theorem on A_u, the solutions to $(d/dt)z = (MA_uM')z$ converge to 0 exponentially fast if $z(0) \in \mathcal{C}$, and go to infinity exponentially fast if $z(0) \in \mathcal{M} \setminus \mathcal{C}$. Since SDE (15) is linear with deterministic initial condition, its solution is a Gaussian process, whose mean is given by the solution to ODE $(d/dt)y = A_u y$, and the covariance matrix is non-singular for any $t > 0$. Now, according to Theorem 6.1, when r is large, the diffusion-scaled process $\hat{\Psi}^r_{\mathcal{E}}(\cdot)$ is "close" to $M\hat{\Psi}_{\mathcal{I}}(\cdot)$, where $\hat{\Psi}_{\mathcal{I}}(\cdot)$ is the solution to (15). From these facts, we can draw two "conclusions". (a) Uniformly on bounded initial states $\hat{\Psi}^r_{\mathcal{E}}(0)$ and large r, after some finite time, $\hat{\Psi}^r_{\mathcal{E}}(t)$ will be close to manifold $\mathcal{M}$, but "away" from submanifold $\mathcal{C}$; (b) If $\hat{\Psi}^r_{\mathcal{E}}(0)$ is close to manifold $\mathcal{M}$, but "away" from submanifold $\mathcal{C}$, then after some finite time, $|\mathbb{E}\hat{\Psi}^r_{\mathcal{E}}(t)|$ is very large, and therefore the probability of $\hat{\Psi}^r_{\mathcal{E}}(t)$ being within an a priori fixed bounded set is small. In turn, (a) and (b) allow us to show that, for large r, the measure μ^r of any a priori fixed bounded set must be arbitrarily small.

In Figure 2 are simulation results for the locally unstable system considered in Section V. We plot the instantaneous load $\rho^r_j(t) = \Xi^r_j(t)/r$ for all server pools; the nominal load is $\rho = 0.5$. We see that the system started from the equilibrium point does not stay "close" to it: the server pools' instantaneous loads fluctuate widely, and moreover, even *long-time* average loads of 3 pools out of 4 are much higher than nominal — the load balancing objective is clearly not achieved. We want to emphasize that such system behavior does *not* mean stochastic instability of the process (stochastic stability is guaranteed under LQFS-LB, see Section II-B), but rather is a consequence of local instability. For comparison, in Figure 3 we also present simulation results for the locally (and globally) stable system with the same activity tree and all $\mu_{ij} = 1$.

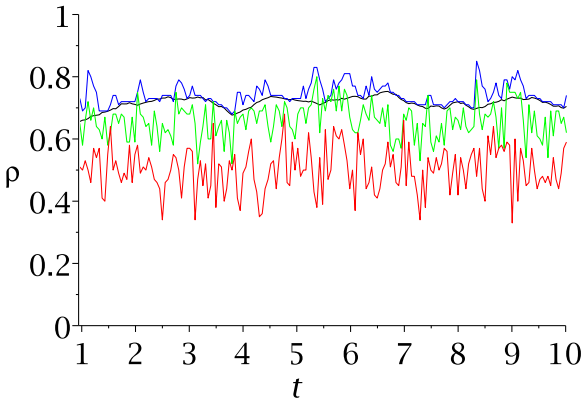

Fig. 2. Simulation of a locally unstable system.

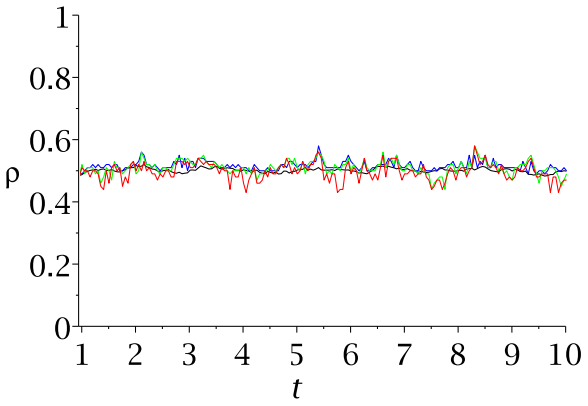

Fig. 3. Simulation of a locally (and globally) stable system.

VII. DIFFUSION SCALE TIGHTNESS OF STATIONARY DISTRIBUTIONS FOR THE CASE WHEN SERVICE RATE DEPENDS ON THE SERVER TYPE ONLY

In this section we consider a special case when there exists a set of positive rates $\{\mu_j\}$, such that $\mu_{ij} = \mu_j$ as long as $(ij) \in \mathcal{E}$. We demonstrate tightness of invariant distributions of the diffusion-scaled process, assuming the system is critically loaded on the fluid scale, i.e. $\rho = 1$. (An analogous result holds for the underload system.) This, in combination with the transient diffusion limit results, allows us to claim that the limit of invariant distributions is the invariant distribution of the limiting diffusion process.

We consider the following asymptotic regime. The optimal solution to SPP (1) is such that $\rho = 1$. As scaling parameter $r \to \infty$, assume that the server pool sizes are $r\beta_j$ (same as throughout the paper), and the input rates are $\lambda^r_i = r\lambda_i + \sqrt{r}l_i$, where the parameters (finite real numbers) $\{l_i\}$ are such that $\sum l_i\nu_i = -C < 0$. Denote by $\rho^r, \{\lambda^r_{ij}\}$ the optimal solution of SPP (1), with β_j's and λ_i's replaced by $r\beta_j$ and λ^r_i, respectively. (This solution is unique, as can be easily seen from the CRP condition.) Then, it is easy to check that $\rho^r = 1 + (\sum l_i\nu_i)/\sqrt{r} = 1 - C/\sqrt{r}$, which in turn easily implies that, for any r, the process is stable with the unique stationary distribution.

We use the definitions of (11), and add to them $\hat{Q}^r_i(t) = Q^r_i(t)/\sqrt{r}$, $Z^r_j(t) = \Psi^r_j(t) - r\beta_j$, $\hat{Z}^r_j(t) = Z^r_j(t)/\sqrt{r}$. Note that, although the optimal average occupancy of pool j is at $\rho^r r\beta_j$, the quantity $\hat{Z}^r_j(t)$ measures the deviation from full occupancy $r\beta_j$. Our choice of signs is such that $\hat{Q}^r_i \geq 0$ while $\hat{Z}^r_j \leq 0$.

Theorem 7.1: Suppose $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$ and $\rho = 1$. Consider a system under the LQFS-LB rule in the asymptotic regime defined above in this section. Then, for any real

$$\theta < \theta_0 := \frac{2\min_i \lambda_i}{\sum_i \lambda_i + (\max_j \mu_j)\sum_j \beta_j},$$

the stationary distributions are such that

$$\limsup_r \mathbb{E}[\sum_i \exp(\theta\hat{Q}^r_i) + \sum_j \beta_j \exp(\theta\hat{Z}^r_j/\beta_j)] < \infty.$$

Proof: [Outline of proof, details in [14]] Our method is partially based on that in [4]. We consider the embedded Markov chain taken at instants right after the transitions, and add virtual transitions to keep the total rate of all transitions from any state constant. The stationary distribution of the embedded Markov chain, in discrete time $\tau = 0, 1, 2, \ldots$, is the same as that of the original, continuous-time chain.

We work with the Lyapunov function

$$\mathcal{L}(\tau) := \sum_i \exp(\theta \hat{Q}_i^r(\tau)) + \sum_j \beta_j \exp(\theta \hat{Z}_j^r(\tau)/\beta_j). \quad (16)$$

A key step in the proof involves the following artificial scheduling/routing rule. We do not actually use this rule in the system; instead, we use it solely to construct an upper bound on the drift of $\mathcal{L}$ within one time-step.

Artificial scheduling/routing rule. We will use the following notation: $\mathcal{I}_+ = \mathcal{I}_+(\tau) := \{i : \hat{Q}_i^r(\tau) > 0\}$, $\mathcal{I}_0 = \mathcal{I}_0(\tau) := \{i : \hat{Q}_i^r(\tau) = 0\}$, $\mathcal{J}_- = \mathcal{J}_-(\tau) := \{j : \hat{Z}_j^r(\tau) < 0\}$, $\mathcal{J}_0 = \mathcal{J}_0(\tau) := \{j : \hat{Z}_j^r(\tau) = 0\}$.

Scheduling: Departures from servers $j \in \mathcal{J}_-$ are processed normally, i.e. reduce the corresponding $Z_j^r(\tau)$ by 1. Whenever there is a departure from a server pool $j \in \mathcal{J}_0$, the server takes up a customer of type i with probability $\lambda_{ij}^r / \sum_i \lambda_{ij}^r$, keeping $Z_j^r(\tau+1) = 0$ and reducing $Q_i^r(\tau+1) = Q_i^r(\tau) - 1$. However, if it happens that the chosen i is such that $Q_i^r(\tau) = 0$, i.e. $i \in \mathcal{I}_0$, then we keep $Q_i^r(\tau+1) = Q_i^r(\tau) = 0$ and instead allow $Z_j^r(\tau+1) = -1$.

Routing: Arrivals to customer types $i \in \mathcal{I}_+$ are processed normally, i.e. increase the corresponding $Q_i^r(\tau)$ by 1. Whenever there is an arrival to a customer type $i \in \mathcal{I}_0$, it is routed to server pool j with probability $\lambda_{ij}^r / \lambda_i^r$, keeping $Q_i^r(\tau+1) = Q_i^r(\tau) = 0$ and increasing $Z_j^r(\tau+1) = Z_j^r(\tau) + 1$. However, if it happens that the chosen j is such that $Z_j^r(\tau) = 0$, i.e. $j \in \mathcal{J}_0$, then we keep $Z_j^r(\tau+1) = Z_j^r(\tau) = 0$ and instead allow $Q_i^r(\tau+1) = 1$.

For the artificial rule, we can write a convenient upper bound on the drift of $\mathcal{L}$ that it "produces". This upper bound, in turn, is also an upper bound on the drift produced by LQFS-LB. (In the special case when all β_j are equal, it is easy to observe that the drift under LQFS-LB cannot be larger than under the artificial rule. In the case of general β_j, this is "almost" true, which allows us to get a common upper bound under both rules.) Using this bound on the drift under LQFS-LB, along with the fact that (roughly speaking) the average drift of $\mathcal{L}$ is zero when the chain is in stationary regime, we obtain an upper bound on $\mathbb{E}\mathcal{L}$ in steady-state, which holds uniformly on all sufficiently large r. ∎

Corollary 7.2: The sequence of stationary distributions of the processes $\left((\hat{Q}_i^r(\cdot)), (\hat{Z}_j^r(\cdot))\right)$ has a weak limit, which is the unique stationary distribution of the limiting process $\left((\hat{Q}_i(\cdot)), (\hat{Z}_j(\cdot))\right)$, described as follows: $\hat{Q}_i(t) = \max\{\hat{Y}(t)/I, 0\}$, $\forall i$, $\hat{Z}_j(t) = \min\{\frac{\beta_j}{\sum_k \beta_k}\hat{Y}(t), 0\}$, $\forall j$, where $\hat{Y}(\cdot)$ is a one-dimensional diffusion process with constant variance parameter $2\sum_i \lambda_i$ and piece-wise linear drift, equal at point x to $-[\sum_j \mu_j][C + \min\{x, 0\}]$. The invariant distribution density is, then, a continuous function, which is a "concatenation" at point 0 of exponential (for $x \geq 0$) and Gaussian (for $x \leq 0$) distribution densities.

Finally, we remark that a tightness result analogous to Theorem 7.1 holds for the underloaded system, $\rho < 1$, and can be proved essentially the same way.

The asymptotic regime in this case is such that $\lambda_i^r = r\lambda_i$ (there is no point in considering $O(\sqrt{r})$ terms in λ_i^r when $\rho < 1$). We denote $Z_j^r(t) = \Psi_j^r(t) - r\beta_j\rho$ (which is consistent with the definition given earlier in this section for $\rho = 1$), and keep notation $Q_i^r(t)$ for the queue length. We work with the following Lyapunov function:

$$\mathcal{L} := \sum\nolimits_i \left[\exp(\theta(1-\rho)\sqrt{r} + \theta\hat{Q}_i^r) - \exp(\theta(1-\rho)\sqrt{r})\right] + \sum\nolimits_j \beta_j \exp(\theta\hat{Z}_j^r/\beta_j). \quad (17)$$

The same approach as in the proof of Theorem 7.1 leads to the following result: for any real θ,

$$\limsup_r \mathbb{E}[\sum\nolimits_j \exp(\theta\hat{Z}_j^r)] < \infty.$$

The limiting process for $(\hat{Z}_j^r(\cdot))$ is $(\hat{Z}_j(\cdot)) = (\frac{\beta_j}{\sum_k \beta_k}\hat{Y}(\cdot))$, with $\hat{Y}(\cdot)$ being a one-dimensional Ornstein-Uhlenbeck process, with Gaussian stationary distribution. The limit of stationary distributions of $(\hat{Z}_j^r(\cdot))$ is the stationary distribution of $(\hat{Z}_j(\cdot))$.

REFERENCES

[1] M. Armony, A. Ward. Blind Fair Routing in Large-Scale Service Systems. February 2010, preprint. http://www.stern.nyu.edu/om/faculty/armony/research/blind_fair_routing.pdf

[2] R. Atar, Y. Shaki, A. Shwartz. A blind policy for equalizing cumulative idleness. February 2010, preprint. http://webee.technion.ac.il/people/atar/equalization.pdf

[3] M. Farkas. *Dynamical Models in Biology*. Academic Press 2001.

[4] D. Gamarnik, A. L. Stolyar. Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime. Asymptotics of the stationary distribution. arXiv:1105.0635

[5] D. Gamarnik, A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability* vol. 16, 2006, pp.56-90.

[6] D. Gamarnik, P. Momcilovic. Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Advances in Applied Probability* vol. 40, 2008, pp.548-577.

[7] I. Gurvich, W. Whitt. Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of OR* vol. 34 no. 2, May 2009, pp. 363-396.

[8] I. Karatzas, S. Shreve. *Brownian Motion and Stochastic Calculus (2nd ed.)*. Springer 1996.

[9] R. Sh. Liptser, A. N. Shiryaev. *Theory of Martingales*. Kluwer Academic Publishers 1989 (translated from Russian by K. Dzjaparidze; in Russian Nauka 1986)

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

[11] C. D. Meyer. *Matrix analysis and applied linear algebra, Volume 1*. SIAM 2000.

[12] A. L. Stolyar, T. Tezcan. Control of systems with flexible multi-server pools: A shadow routing approach. *Queueing Systems*, vol. 66, 2010, pp. 1-51.

[13] A. L. Stolyar, T. Tezcan. Shadow-routing based control of flexible multi-server pools in overload. *Operations Research*, to appear.

[14] A. L. Stolyar, E. Yudovina. Systems with large flexible server pools: Instability of "natural" load balancing. Submitted. arXiv:1012.4140.

[15] Supporting computations. http://www.statslab.cam.ac.uk/~ey221/LQFS-LB/