IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

A blind policy for equalizing cumulative idleness

Rami Atar, Yair Y. Shaki, Adam Shwartz

CCIT Report #720
February 2009

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT REPORT #720 February 2009

A blind policy for equalizing cumulative idleness

Rami Atar* Yair Y. Shaki† Adam Shwartz‡

Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

February 17, 2009

Abstract

We consider a system with a single queue and multiple server pools of heterogenous exponential servers. The system operates under a policy that always routes a job to the pool with longest cumulative idleness among pools with available servers, in an attempt to achieve fairness toward servers. It is easy to find examples of a system with a fixed number of servers, for which fairness is not achieved by this policy in any reasonable sense. Our main result shows that in the many-server regime of Halfin and Whitt, the policy does attain equalization of cumulative idleness, and that the equalization time, defined within any given precision level, remains bounded in the limit. An important feature of this policy is that it acts 'blindly', in that it requires no information on the service or arrival rates.

Keywords: Blind control; Diffusion limits; Halfin-Whitt regime; Fairness; Many-server systems

1 Introduction

The performance and optimization of systems with a large number of servers has attracted much attention in recent years. This is due to their applicability—for example to call centers—as well as to their interesting structure. Since exact analysis proves impossible in most cases, large part of research has focused on asymptotics. Particularly, the many-server diffusion regime introduced by Halfin and Whitt [9] has been widely studied and is the subject of ongoing research. Under this regime, both the arrival rates and the number of servers are scaled up

*Research supported in part by the ISF (Grant 1349/08) and the fund for promotion of research at the Technion

†Research supported in part by the Viterbi Postdoctoral Fellowship and the ISF (Grant 1349/08)

‡The Julius M. and Bernice Neiman Chair in Engineering. Research supported in part by the fund for promotion of research at the Technion.

in a fixed proportion, while maintaining a critically loaded system. A unique property of this regime is that it allows for the total number of customers to sometimes rise above and sometimes fall behind the number of servers, with non-negligible probability. For this reason it is often considered more realistic for some applications than the conventional heavy traffic regime (where the service rates are scaled up rather than the number of servers) under which the number of customers exceeds that of servers except for a negligible set of times.

In this paper we consider a system with a single queue and a fixed number, l, of server pools of heterogenous exponential servers. For $i = 1, \ldots, l$ we denote by I_i the *idleness process* for pool i, representing the number of pool-i servers that are free, and by $J_i = \int_0^{\cdot} I_i(t)dt$ the *cumulative idleness process* for pool i, representing the overall duration of time when a server has been idle since time zero, summed over all pool-i servers. The system operates under a policy that always routes a job to the pool with longest cumulative idleness among pools with available servers (more generally, we analyze the longest *weighted* cumulative idleness policy). This is done in an attempt to drive the processes $J_i(t)$ to take nearly equal value for all i, as t becomes large, thus achieving a certain form of fairness toward servers. We refer to this as the *Longest Idle Pool First* (LIPF) policy. It is easy to exhibit examples with a fixed number of servers, in which fairness is not achieved by LIPF in any reasonable sense (see Example 2.1). Our main result (Theorems 2.2 and 2.3) is that in the Halfin-Whitt regime, the policy does attain equalization of cumulative idleness, and that the equalization time, defined within any given precision level, remains bounded in the limit.

A feature of LIPF that appears to be convenient, is that it requires no information on the service or arrival rates. In applications, it is a common practice to use a reminiscent of this policy, namely one that always routes to the *server* whose cumulative idleness is longest, as a job assignment paradigm [1]. In call center applications the motivation for fairness comes from operational considerations of human servers, but the same idea may be justified also in computer systems, where the notion of load balancing is standard. Following common nomenclature, we call a policy that does not use any information on the parameters a *blind* policy. This term may also refer to a situations where the control does not have access to the complete information about the system state, and in our model both apply.

Fairness towards *customers* in queueing systems is a rather broad and well-studied area: see e.g. [5, 13] and references therein. Very few papers have treated the subject of fairness toward *servers* (the reader is referred to [1] and references therein). Armony and Ward [1] consider the problem of optimal routing to minimize steady state delay costs subject to constraints that ensure idleness is distributed among the pools according to a given proportion. They find a policy that is asymptotically optimal for the problem in the Halfin-Whitt regime. Although [1] and the present paper are both motivated by fairness toward servers, the contributions of the two papers are quite different. First, the mathematical problem formulations are obviously different, as [1] finds a policy that optimizes a criterion whereas the current paper's contribution is on the performance analysis of a single policy, namely LIPF. A particular aspect of the problem formulation of [1] is that it allows the policy to access system parameters (such as service and arrival rates), and the parameters are in fact crucially used in the solution to the problem. In contract, LIPF is a blind policy. Another major difference is that [1] formulate the problem in steady state, whereas, as will be seen in the next section, the present paper analyzes

the model in the Halfin-Whitt asymptotics on a large, but fixed time horizon. Understanding the transient behavior, by analyzing a given model over a finite time horizon, may for some purposes be more accurate than steady state analysis.

Tseytlin [12] considers a policy where each arriving customer joins pool i with probability $I_i/\sum_j I_j$, and explicitly solves the model in steady state. The motivation of [12] as well is to be fair to servers by achieving what is referred to there as *idleness balancing*. Atar [3] analyzes a policy where jobs are routed according to the length of idleness period *of each server since last time of service*, identifying the diffusion limit, and showing that the length of last idle period is nearly equalized by the policy, among servers that are idle at a given time.

Gurvich and Whitt [8] study a parallel server system of a more general structure, that allows for multiple buffers, and consider a class of routing policies, referred to as *Queue- and Idleness-Ratio* (QIR) controls, that attempt to bring both the queuelength and idleness processes to a predetermined proportion. This is shown to be achieved in the Halfin-Whitt regime. Specializing to the case studied here, of a single buffer (and multiple pools), and letting u_i, $i = 1, \ldots, l$ be a given vector of proportions, their result asserts that under a suitable QIR policy, for each i, the process $I_i - u_i \sum_i I_i$, normalized in diffusion scale, converges to zero in probability, uniformly on compact time intervals. This clearly implies an analogous statement about the processes J_i, and thus fairness is in fact achieved by QIR in a sense very similar to what is achieved by LIPF. We mention upfront that achieving fairness has not been the goal of [8] in developing their result. Nevertheless, it is important to explain how our result relates to such an interpretation of [8], in view of the fact that it can be used for this purpose. We make several comments about this comparison. First, LIPF appears to be natural for the purpose of equalizing the *cumulative* idleness processes, since these processes are its observables. At the modeling level, our setting allows for each pool to contain heterogenous servers and for the arrivals to be according to a renewal process, assumptions not covered by [8] (nor by [1], [12]). The LIPF seems to enjoy an advantage over QIR in terms of robustness: the processes I_i change much faster than the cumulative idleness J_i, requiring any implementation of the latter a high rate of information to be transferred between the server stations and the system control unit. This is not the case with the LIPF policy, that is robust in that small delays between the time when the processes are observed and the time when the information is received by the control unit, lead to small degradation in terms of the fairness metric. We make a precise statement about such a robustness property in the last section (Theorem 4.2). Finally, we note that the proof in [8] is quite involved, and in turn relies on additional general state space collapse results from [7] as well as results from [2]. The proof offered here, regarding LIPF, is in contrast elementary and short, and illustrates that the analysis of this policy, at least in the more limited setting of a single buffer, does not require convergence or state space collapse, but is a direct consequence of properties of the policy and some basic technical results such as tightness.

We would like to further emphasize the aspect of blind control because we believe it is of interest in a much wider context. One of the key assumptions in many recent works on large queueing systems is that all model parameters are known exactly by the controller. Such an assumption about service rate, for example, is obviously invalid for call center applications, where servers are human. In fact, this assumption rarely holds even when servers are comput-

ers, since there are always external influences which affect service and arrival rates. Ideally, one would hope for an optimal policy that requires minimal information about the parameters, and will not be sensitive to their values.

In the context of the conventional heavy traffic theory, since the number of servers is fixed and time is accelerated, one can obtain accurate information on service time distribution by sampling service periods in a fixed, short time interval, so that the efficacy of policies that use such estimates seems natural. However, in the Halfin-Whitt limit, since time is not accelerated, one cannot hope to obtain good estimators for all system parameters over a fixed interval of time. This provides motivation for the present study, where the policy does not even attempt to estimate parameters. This was also the motivation for the approach taken in Atar and Shwartz [4], that relies on partial sampling from the service time distribution, demonstrating how nearly optimal blind policies can be constructed based on the size of the server population. Recent work of Stolyar and Tezcan [11] provides an alternative look, proposing a robust routing scheme for a multi-buffer multi-pool setting in the Halfin-Whitt regime, which optimally balances load on the server pools without the knowledge of the input rates.

As an application of our results we describe a model of a distributed collection of pools with a central control, but where the information needed for LIPF to operate is mostly local, and only minimal information about idleness is passed to the central controller.

The rest of this paper is organized as follows. The setting and main results appear in Section 2. Section 3 contains the proof. Finally, Section 4 provides extensions of the main result under relaxed assumptions on the arrival process as well as to the case of delayed information, and discusses implementation in a distributed set up.

2 Setting, notation and main result

Customers arrive at a system according to a renewal process denoted by $A(t)$. It is assumed that the inter-arrival times are positive and have finite second moment. Each arrival has a single noninterruptible service requirement. Arriving customers are routed to one of l pools, and within the pool to a particular server, according to a routing policy, provided a free server is available: if not, they are queued in a buffer with infinite room. Customers from the buffer are routed to servers according to a first-come-first-served rule. We consider work conserving routing policy, so that no server may be idle when at least one customer is in the buffer. Each customer leaves the system when its service requirement is fully processed.

There are N servers, arranged in l pools, so that the number of servers in pool i is N_i, for $i \in L := \{1, \dots, l\}$. The servers are labeled $1, \dots, N$, and the set of k's for which server k is in pool i is denotes by K_i. We write K for $\{1, \dots, N\}$, so $\cup_i K_i = K$, and $|K_i| = N_i$, $i \in L$. Server k serves according to an exponential service time distribution with rate μ_k.

All processes defined below are assumed to have right-continuous sample paths.

For $k \in K$ and $t \geq 0$, let $I_{(k)}(t)$ take the value 1 if server k is idle at time t, and let it be 0

otherwise. Set $Z_{(k)} = 1 - I_{(k)}$. Let

$$J_{(k)}(t) = \int_0^t I_{(k)}(s)ds, \qquad k \in K, t \geq 0.$$

Denote by $I_i(t)$ the number of idle servers from pool i at time t, for $i \in L$ (see (1)) and define $Z_i(t)$ in a similar manner. Then both I_i and Z_i are stochastic processes taking values in $[0, N_i]$, and

$$I_i = \sum_{k \in K_i} I_{(k)} = N_i - Z_i, \qquad i \in L. \tag{1}$$

The modeling of service completions will require usage of standard Poisson processes S_i, $i \in L$. The number of service completions by pool-i servers until time t is denoted by $D_i(t)$, and is represented as

$$D_i(t) = S_i(T_i(t)),$$

where T_i is defined as

$$T_i(t) = \sum_{k \in K_i} \mu_k \int_0^t Z_{(k)}(s)ds, \qquad i \in L, t \geq 0. \tag{2}$$

The number-in-system and the number-in-buffer processes are denoted by X and Q, respectively. The initial configuration, namely,

$$(\{I_{(k)}(0), k \in K\}, Q(0)),$$

and the processes A and S_i, $i \in L$, are assumed to be mutually independent $l + 2$ entities.

We will say that a routing policy is *work conserving* if for all $t \geq 0$,

$$Q(t) = (X(t) - N)^+, \quad \text{or equivalently} \quad I(t) = (N - X(t))^+.$$

Note that this imposes an assumption on the initial configuration as well as on the policy.

Let

$$J_i := \sum_{k \in K_i} J_{(k)}, \quad i \in L, \quad J = \sum_{i \in L} J_i.$$

Then $J_i(t)$ represents the overall idleness time accumulated by servers from pool i until time t. A vector $u = (u_1, \ldots, u_l)$, where $u_i \in (0, 1)$ and $\sum_i u_i = 1$, will be called a *target vector*. Given a target vector u, we will consider a family of policies that keep track of the J_i processes and attempt to drive the relative idleness $J_i(t)/J(t)$ toward u_i, for each i. More precisely, let a target vector u be given, and let $v_i = u_i^{-1}$, $i \in L$. We say that a policy is *u-greedy* if

- The policy is *work conserving*; i.e., when a server becomes available and there is a customer in the queue, a customer is routed to the server. When a customer arrives to find some available servers, it is routed to one of them.
- If a customer is to be routed at time t to an available server, and $AV(t) \subset L$ denotes the set of pools containing available servers at this time, it is routed to any one of the available servers from a pool $i \in AV(t)$ for which

$$v_i J_i(t) \geq v_j J_j(t), \qquad \text{for all} \quad j \in AV(t). \tag{3}$$

Example 2.1 **(a)** Consider a system with Poisson arrivals at rate $\lambda \in (0,1)$, and two servers with deterministic service times, 1 and r. We argue that for large values of r a policy that is work conserving and always routes to the server with longest cumulative idleness first, is unfair toward the slow server. It will be clear that the argument can be generalized to any finite number of servers, and to quite general service time distributions, but we do not provide these details. A quasi-stationary analysis of the system within a service period of the slow server, say $[t, t+r]$, shows that the cumulative idleness time of the fast server during the interval is proportional to r. If the slow server completes service at some time t_1, and t_2 denotes the first time after t_1 when two arrivals occur within less than a unit of time, then the distribution of $t_2 - t_1$ does not depend on r. Moreover, by time t_2 the slow server will necessarily be assigned a new job. This shows that, as r becomes large, the fast and, respectively, slow server enjoys idleness at a proportion of $O(1)$ and $O(1/r)$ on average.

(b) Next, consider two pools of servers, where pool 1 [resp., 2] contains a fixed number, n, of servers with deterministic service time 1 [resp., r, where r is some large number]. Consider again Poisson arrivals, where now $\lambda \in (0,n)$ is fixed. The system is assumed to work under LIPF. Consider a specific server from pool 2. If it completes service at some time t_1, then it must enter a new service cycle no later than when $2n$ new arrivals occur within a window of one unit of time; thus similarly to case (a), the duration of vacation following t_1 is stochastically bounded by a distribution that is independent of r. As a result, the idleness proportion for each of the slow servers is $O(1/r)$ for large values of r. On the other hand, since the system is sub-critically loaded, the average idleness proportion for pool-1 servers must be bounded below as r becomes large. We conclude that LIPF does not achieve fairness in the situation described here. $\diamond$

This example provides motivation to study under what conditions a u-greedy policy achieves equalization with respect to a given target vector u. Our asymptotic results regard this question in a many-server heavy traffic regime.

To formulate the notion of a large number of servers, we consider a sequence of systems, parameterized by n, where the number of servers in the nth system is proportional to n; particularly, $N_i^n = \lfloor \nu_i n \rfloor$, where ν_i are some positive constants. In the sequel, the notation of all processes and system parameters introduced thus far will be used with a superscript n, denoting dependence on the parameter; there is no need however to parameterize the standard Poisson processes, and they will still be denoted by S_i.

The rate of arrival λ^n is assumed to satisfy $\lambda^n/n \to \lambda \in (0,\infty)$ and moreover,

$$\widehat{\lambda}^n =: \frac{\lambda^n - \lambda n}{\sqrt{n}} \to \widehat{\lambda} \in \mathbb{R}. \tag{4}$$

The parameters μ_k^n are assumed to satisfy

$$\underline{\mu} \le \mu_k^n \le \overline{\mu}, \qquad k \in K^n,\, n \in \mathbb{N}, \tag{5}$$

where $0 < \underline{\mu} < \overline{\mu} < \infty$ are constants. In addition, it is assumed that the limits

$$\bar{\mu}^n := \frac{1}{n} \sum_{k \in K^n} \mu_k^n \to \mu \in [\underline{\mu}, \overline{\mu}], \tag{6}$$

and

$$\widehat{\mu}^n := \frac{1}{\sqrt{n}} \sum_{k \in K^n} (\mu_k^n - \mu) \to \widehat{\mu} \in \mathbb{R}, \tag{7}$$

exist. The 'heavy traffic' assumption makes the system critically loaded by relating the arrival and service rates as

$$\lambda = \mu. \tag{8}$$

We will also denote $\beta^n = \widehat{\lambda^n} - \widehat{\mu}^n$ and assume its limit satisfies

$$\widehat{\beta} := \widehat{\lambda} - \widehat{\mu} < 0. \tag{9}$$

Note that the random variable $X^n(0)$ is given by $Q^n(0) + Z^n(0)$. The 'second order asymptotics' of $X^n(0)$ is assumed to satisfy

$$\widehat{X}^n(0) := n^{-\frac{1}{2}}(X^n(0) - N^n) \text{ is a tight sequence of random variables.} \tag{10}$$

Define $\widehat{J}^n := n^{-\frac{1}{2}} J^n$. Given $\varepsilon > 0$ let $\gamma^n(\varepsilon) := \inf\{t : \widehat{J}^n(t) \geq \varepsilon\}$. Our main result states that under a u-greedy policy, given any level of precision, equalization of the cumulative idleness processes is achieved soon after $\gamma^n(\varepsilon)$, in the large n limit, with large probability.

Theorem 2.2 *Let u be a given target vector, and let π be any u-greedy policy. Then under π, for every $\varepsilon > 0$ and $T > 0$,*

$$\lim_{n \to \infty} P\Big\{ \max_{i,j \in L, i \neq j} \sup_{s \in [0,T]} \Big| v_i \widehat{J}_i^n(s) - v_j \widehat{J}_j^n(s) \Big| \geq \varepsilon \Big\} = 0.$$

Moreover, for any $t \geq 0$, the random variables $\gamma^n = \gamma^n(\varepsilon)$ are tight, and one has

$$\liminf_{n \to \infty} P\Big\{ \gamma^n < \infty \quad \textit{and} \quad \max_{i \in L} \Big| \frac{J_i^n(\gamma^n + t)}{J^n(\gamma^n + t)} - u_i \Big| \leq \varepsilon \Big\} \geq 1 - \varepsilon.$$

Note that measuring fairness in terms of ratios is meaningful only when $J^n > 0$. This is why the formulation of the last assertion above involves γ^n. As will be clear from the proof of the result, in case that the random variables $\widehat{X}^n(0)$ (10) are further assumed to be bounded above by some $-\delta < 0$, the random times $\gamma^n(\varepsilon)$ will be small with probability tending to 1 (as $n \to \infty$), provided that ε is sufficiently small. In this case, the above result asserts that equalization is attained soon after time zero.

The following is almost an immediate consequence of the above result. Its purpose is to emphasizes that equalization is in fact achieved (with high probability) after sufficiently large time.

Theorem 2.3 *Under the hypotheses of Theorem 2.2, for every $\varepsilon > 0$ there exists T such that for every $T_1 \in [T, \infty)$,*

$$\liminf_{n \to \infty} P\Big\{ J^n(T) > 0 \quad \textit{and} \quad \max_{i \in L} \sup_{s \in [T, T_1]} \Big| \frac{J_i^n(s)}{J^n(s)} - u_i \Big| \leq \varepsilon \Big\} \geq 1 - \varepsilon.$$

3 Proof

The main result will be proved by diffusion scale analysis. To this end, we define processes at diffusion scale, as follows. We denote centered, normalized versions of the processes, for $i \in L$ and $t \geq 0$, by

$$\widehat{I}_i^n(t) = \frac{I_i^n(t)}{\sqrt{n}}, \qquad \widehat{Q}^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \tag{11}$$

$$\widehat{A}^n(t) = \frac{A^n(t) - \lambda^n t}{\sqrt{n}}, \qquad \widehat{S}_i^n(t) = \frac{S_i(nt) - nt}{\sqrt{n}}, \qquad \widehat{X}^n(t) = \frac{X^n(t) - N^n}{\sqrt{n}}, \tag{12}$$

and

$$\widehat{I}^n(t) = \frac{I^n(t)}{\sqrt{n}}, \qquad \widehat{J}^n(t) = \frac{J^n(t)}{\sqrt{n}}. \tag{13}$$

The fluid-scale process

$$\bar{T}_i(t) = \frac{1}{n} T_i^n(t), \quad i \in L \tag{14}$$

will also be used.

Lemma 3.1 *Define*

$$F^n(t) = \frac{1}{\sqrt{n}} \sum_{k \in K} \mu_k \int_0^t I_{(k)}^n(s) ds, \tag{15}$$

$$W^n(t) = \widehat{A}^n(t) - \sum_{i=1}^{l} \widehat{S}_i^n(\bar{T}_i^n(t)) \tag{16}$$

$$\widehat{\beta}^n = \frac{1}{\sqrt{n}} \Big(\lambda^n - \sum_{k \in K} \mu_k \Big). \tag{17}$$

Then

$$\widehat{X}^n(t) - \widehat{X}^n(0) = W^n(t) + \widehat{\beta}^n t + F^n(t), \qquad t \geq 0. \tag{18}$$

Proof: We have for every t that $X^n(t) = X^n(0) + A^n(t) - \sum_{i \in L} D_i^n(t)$, by definition of these processes. Thus

$$X^n(t) = X^n(0) + A^n(t) - \sum_{i=1}^{l} S_i(T_i^n(t)). \tag{19}$$

Hence

$$X^n(t) - N^n - (X^n(0) - N^n) = [A^n(t) - \lambda^n t] + \lambda^n t - \sum_{i=1}^{l} [S_i(T_i^n(t)) - T_i^n(t)] - \sum_{i=1}^{l} T_i^n(t).$$

Since $Z_{(k)}^n + I_{(k)}^n = 1$,

$$\sum_{i=1}^{l} T_i^n(t) = \sum_{k \in K} \mu_k \int_0^t Z_{(k)}^n(s)\, ds = \sum_{k \in K} \mu_k - \sum_{k \in K} \mu_k \int_0^t I_{(k)}^n(s)\, ds,$$

and dividing by $\sqrt{n}$,

$$\widehat{X}^n(t) - \widehat{X}^n(0) = \widehat{A}^n(t) - \sum_{i=1}^{l} \widehat{S}_i^n(\bar{T}_i^n(t)) + \left(\frac{\lambda^n t}{\sqrt{n}} - \sum_{k \in K} \mu_k \right) + \frac{1}{\sqrt{n}} \sum_{k \in K} \mu_k \int_0^t I_{(k)}^n(s)\, ds$$

which yields (18). ■

Throughout, we let $|f|_t^* = \sup_{0 \le s \le t} |f(s)|$. Denote the modulus of continuity of a function f by

$$w_\theta(f, \delta) := \sup_{0 \le s \le t \le (s+\delta) \wedge \theta} |f(t) - f(s)|, \qquad f : [0, \theta] \to \mathbb{R},\ \delta > 0.$$

A sequence of processes defined on $[0, \theta]$, with sample paths in the Skorohod space, is said to be *C-tight* if it is tight, and every subsequential limit has continuous sample paths with probability one. C-tightness of, say $\{X^n\}$, implies tightness of $|X^n|_\theta^*$ and the convergence in probability $w_\theta(X^n, \delta) \to 0$, for every δ (see [6, Section 18]). These facts will be used in the sequel in conjunction with the application of the following lemma.

Lemma 3.2 *Given any $\theta \in (0, \infty)$, the sequence of random variables*

$$|\widehat{A}^n|_\theta^* \vee \max_i |\widehat{S}_i^n \circ \bar{T}_i^n|_\theta^* \vee \max_i |\widehat{I}_i^n|_\theta^*, \quad n \in \mathbb{N},$$

is tight. In fact, $\{\widehat{A}^n\}_{n \in \mathbb{N}}$ and, for every $i \in L$, $\{\widehat{S}_i^n \circ \bar{T}_i^n\}_{n \in \mathbb{N}}$, are C-tight. Furthermore, given any $\varepsilon_1, \varepsilon_2 > 0$ there exists t_1 such that

$$\limsup_{n \to \infty} P(|R^n|_t^* \ge \varepsilon_1 t) \le \varepsilon_2, \qquad t \ge t_1, \tag{20}$$

where R^n is any one of the processes $\widehat{A}^n$ or $\widehat{S}_i^n \circ \bar{T}_i^n$.

Proof: First, we note by (2) that for any $t \le \theta$,

$$\bar{T}_i^n(t) \le \frac{\overline{\mu}}{n} \theta N_i^n \le \overline{\mu} \theta \nu_i,$$

since Z_i^n is bounded by N_i^n. Hence

$$|\widehat{S}_i^n(\bar{T}_i^n)|_\theta^* = \sup_{0 \le t \le \theta} |\widehat{S}_i^n(\bar{T}_i^n(t))| \le |\widehat{S}_i^n|_{\overline{\mu}\theta\nu_i}^*.$$

It is well known that the scaled renewal processes $\widehat{A}^n$ and $\widehat{S}_i^n$ converge in distribution, uniformly on compacts, to independent zero mean Brownian motions with diffusion coefficients $\lambda^{1/2}$ and, respectively, 1 [6, Section 17]. Thus $\{|\widehat{A}^n|_\theta^*\}_n, \{|\widehat{S}_i^n \circ \bar{T}_i^n|_\theta^*\}_n$ are tight.

We next prove that $\{|\widehat{I}_i^n|_\theta^*\}_n$ is tight. By (5) and (15),

$$0 \le F^n(t) \le \overline{\mu} \int_0^t \widehat{I}^n(s) ds = \overline{\mu} \int_0^t \widehat{X}^n(s)^- ds.$$

Thus by (18), given θ, we have for every $t \in [0, \theta]$

$$|\widehat{X}^n(t)| \leq |\widehat{X}^n(0)| + |W^n|^*_\theta + |\widehat{\beta}^n|\theta + \overline{\mu} \int_0^t |\widehat{X}^n(s)|ds.$$

By Gronwall's inequality ([10] page 36)

$$|\widehat{X}^n|^*_\theta \leq (|\widehat{X}^n(0)| + |W^n|^*_\theta + |\widehat{\beta}^n|\theta)\, e^{\overline{\mu}\theta}.$$

Since we already established tightness of the sup norm of terms in the sum defining W^n, since $\widehat{\beta}^n$ are bounded (cf. (4), (7), (8) and (17)), and $\widehat{X}^n(0)$ are tight by assumption, it follows that $|\widehat{X}^n|^*_\theta$ are tight, and thus so are $|\widehat{I}^n_i|^*_\theta$.

We argue that the processes $\widehat{S}^n_i \circ \bar{T}^n_i$ are C-tight. Denote $M^n_i = \sum_{k \in K_i} \mu^n_k/n$ and note that N^n_i are given as $\lfloor \nu_i n \rfloor$ and μ_k are bounded, whence M^n_i are bounded. By (2),

$$|\bar{T}^n_i(t) - M^n_i t| \leq \frac{\bar{\mu}\theta}{\sqrt{n}} |\widehat{I}^n_i|^*_\theta, \qquad t \in [0, \theta].$$

Consider a subsequence on which M^n_i converges to some $M \in [0, \infty)$. In view of the tightness of $|\widehat{I}^n_i|^*_\theta$, it follows that $\bar{T}_i$ converges in distribution to Mt. Combined with the convergence in distribution of $\widehat{S}^n_i$ to a Brownian motion and an application of the random change of time lemma [6, p. 151], this shows that $\widehat{S}^n_i \circ \bar{T}^n_i$ are C-tight.

Finally, (20) is an immediate consequence of the fact that uniformly-on-compacts subsequential limits of any of the processes $\widehat{A}^n$ and $\widehat{S}^n_i \circ \bar{T}^n_i$ are all Brownian motions with zero drift and bounded diffusion coefficient. ■

Lemma 3.3 *Consider a target vector u and any u-greedy policy. Fix $p, q \in L$, $p \neq q$ and $\theta \in (0, \infty)$. Let $\Delta^n(t) = v_p \widehat{J}^n_p(t) - v_q \widehat{J}^n_q(t)$. Then, for any θ, $|\Delta^n|^*_\theta \to 0$ in probability as $n \to \infty$.*

Proof: To simplify the notation, we remove the superscript n from most of the notation (there will be no confusion). We start by analyzing a scenario where no jobs are routed to a certain pool within a given interval. More precisely, fix $n \in \mathbb{N}$ and let η and ζ be $[0, \theta]$-valued random variables such that $\eta \leq \zeta$. Fix $i \in L$ and let H be any event under which

- $Q = 0$ within the interval $[\eta, \zeta]$; and
- no jobs are routed to pool i within the same interval.

Write L_i for $L \setminus \{i\}$. Then, with the notation $Y[a, b] = Y(b) - Y(a)$, we will show that, on the event H,

$$\sum_{j \in L_i} \widehat{I}_j[\eta, \zeta] + \widehat{A}[\eta, \zeta] - \sum_{j \in L_i} \widehat{S}_j \circ \bar{T}_j[\eta, \zeta] - \frac{1}{\sqrt{n}} \sum_{j \in L_i} \sum_{k \in K_j} \mu_k \int_\eta^\zeta Z_{(k)}(s)ds + \frac{\lambda^n}{\sqrt{n}}(\zeta - \eta) = 0. \quad (21)$$

By equation (19),

$$X[\eta,\zeta] = A[\eta,\zeta] - \sum_{j\in L} D_j[\eta,\zeta] = A[\eta,\zeta] - \sum_{j\in L} S_j \circ T_j[\eta,\zeta].$$

By definition of the processes X, Q and Z_j, we have $X = Q + \sum_{j\in L} Z_j$. Since Q vanishes on the interval $[\eta,\zeta]$, we have

$$X[\eta,\zeta] = \sum_{j\in L} Z_j[\eta,\zeta].$$

Also, since no jobs are routed to pool i, we have

$$Z_i[\eta,\zeta] = -D_i[\eta,\zeta] = -S_i \circ T_i[\eta,\zeta].$$

Combining the above three equations,

$$\sum_{j\in L_i} Z_j[\eta,\zeta] = A[\eta,\zeta] - \sum_{j\in L_i} S_j \circ T_j[\eta,\zeta].$$

Using $Z_j + I_j = N_j$, $j \in L$, we have

$$\begin{aligned} -\sum_{j\in L_i} I_j[\eta,\zeta] \;=\; & [A[\eta,\zeta] - \lambda^n(\zeta-\eta)] + \lambda^n(\zeta-\eta) \\ & - \sum_{j\in L_i} [S_j(n\bar{T}_j(\zeta)) - n\bar{T}_j(\zeta)] + \sum_{j\in L_i} [S_j(n\bar{T}_j(\eta)) - n\bar{T}_j(\eta)] - \sum_{j\in L_i} T_j[\eta,\zeta], \end{aligned}$$

and dividing by $\sqrt{n}$ and using the definitions of the processes involved (2), (11)–(14), yields (21).

In what follows, fix an arbitrary $\varepsilon > 0$. Note that $\Delta(0) = 0$, and let

$$\tau = \tau^n = \inf\{t \mid \Delta(t) \geq \varepsilon\},$$

and $E = E^n = \{\tau \leq \theta\}$. To prove the lemma, it suffices to show that $P(E) \to 0$ as $n \to \infty$. Let us define on the event E

$$\sigma = \sigma^n = \sup\{t \mid t < \tau, \Delta(t) \leq \varepsilon/2\}, \qquad \kappa = \kappa^n = \inf\{t \in [\sigma,\tau] \mid I_p(t) = 0\}.$$

Note that on E we always have $\sigma \in [0,\tau]$. The random variable κ represents the first time between σ and τ when all servers from pool p are occupied. If this never happens within $[\sigma,\tau]$, we have, by definition, $\kappa = \infty$.

For $B \in \mathcal{F}$, we write $P_E(B) = P(E \cap B)$. The proof proceeds in three steps, where Steps 1 and 2 are based on (21).

Step 1. We will show that for every $\delta > 0$,

$$P_E(\kappa \wedge \tau - \sigma > \delta) \to 0 \quad \text{as} \quad n \to \infty. \tag{22}$$

Fixing δ, we will use the foregoing analysis concerning the event H, with

$$H = E \cap \{\kappa \wedge \tau - \sigma > \delta\}.$$

We take $\eta = \sigma$, and $\zeta = \kappa \wedge \tau$. Under H, by the definition of κ, at any time within $[\sigma, \kappa \wedge \tau)$ at least one server from pool p is idle. Consequently, on this time interval, no customer is in the queue (by work conservation), and pool q receives no jobs (by (3) and the fact $\Delta > 0$). This discussion shows that H satisfies both bullet conditions from the first part of the proof. Consequently (21) is valid with $i = q$.

Let R denote the sum of the first three terms on the l.h.s. of (21). From Lemma 3.2, the sequence of random variables $\{R\mathbf{1}_H\}_{n\in\mathbb{N}}$ is tight. Using (21) and the inequality $Z_{(k)} \leq 1$, we have on H,

$$R + \Big(\frac{\lambda^n}{\sqrt{n}} - \frac{1}{\sqrt{n}} \sum_{j\in L_q} \sum_{k\in K_j} \mu_k\Big)(\zeta - \eta) \leq 0.$$

Using the notation $\widehat{\beta}^n$ (17), and the inequality $\sum_{k\in K_q} \mu_k \geq \underline{\mu} N_q \geq \underline{\mu}\underline{\nu}_q n/2$,

$$R + (\widehat{\beta}^n + \sqrt{n}\underline{\mu}\underline{\nu}_q/2)(\zeta - \eta) \leq 0.$$

Since $R\mathbf{1}_H$ are tight, $\widehat{\beta}^n$ converge (cf. (4), (7)), and $\underline{\mu}\underline{\nu}_q > 0$, it follows that $P_E(\zeta - \eta > \delta) \to 0$, establishing (22).

Step2. We will show that for every $\gamma > 0$,

$$P_E(\tau > \kappa, \sup_{t\in[\kappa,\tau]} I_p(t) > n^{1/2}\gamma) \to 0. \tag{23}$$

We use again the analysis from the first part of the proof. This time let $H = E \cap \{\tau > \kappa, \sup_{t\in[\kappa,\tau]} I_p(t) > n^{1/2}\gamma\}$. On H define

$$\tau_1 = \inf\{t \in [\kappa, \tau] \,|\, I_p(t) > n^{1/2}\gamma\}, \qquad \sigma_1 = \sup\{t \in [\kappa, \tau_1] \,|\, I_p(t) < n^{1/2}\gamma/2\},$$

and note that on H one has $\kappa \leq \sigma_1 \leq \tau_1 \leq \tau$. On the event H, within the time interval $[\sigma_1, \tau_1]$, $I_p > 0$, and so $Q = 0$, and no jobs are routed to pool q. Consequently, the bullet conditions about H hold true with $\eta = \sigma_1$ and $\zeta = \tau_1$. We can thus again use (21) with $i = q$. By definition of the times σ_1 and τ_1, we have that the first term in (21) is bounded below by $\gamma/2$, on H. We again use the inequality $Z_{(k)} \leq 1$ as well as the lower bound $\underline{\mu}$ on μ_k, $k \in K_q$, in (21), to obtain the inequality (on H):

$$\gamma/2 + \widetilde{W}[\eta, \zeta] + \underline{\mu}\underline{\nu}_q n^{1/2}(\zeta - \eta) \leq 0, \tag{24}$$

where

$$\widetilde{W}(t) := \widehat{A}(t) - \sum_{j\in L_q} \widehat{S}_j \circ \bar{T}_j(t) + \widehat{\beta}^n t.$$

Recall that $\widehat{\beta}^n \to \widehat{\beta}$. By Lemma 3.2, the processes $\{\widetilde{W}\}_{n\in\mathbb{N}}$ are C-tight. Let us denote $H_1 = H \cap \{\zeta - \eta \leq n^{-1/4}\}$ and $H_2 = H \cap \{\zeta - \eta > n^{-1/4}\}$. Then by (24), we have

$$P(H_1) \leq P(\gamma/2 - w_\theta(\widetilde{W}, n^{-1/4}) \leq 0)$$

and

$$P(H_2) \leq P(-2|\widetilde{W}|^*_\theta + \underline{\mu}\underline{\nu}_i n^{1/4} \leq 0).$$

In view of the C-tightness, it follows that both $P(H_1)$ and $P(H_2)$ converge to zero as $n \to \infty$. This shows that $P(H) \to 0$, and (23) follows.

Step 3. We conclude by combining Steps 1 and 2. On E, we clearly have

$$\widehat{J}_p(\tau) - \widehat{J}_p(\sigma) \le (\kappa \wedge \tau - \sigma)|\widehat{I}_p|^*_\theta + \mathbf{1}_{\{\tau > \kappa\}}\, \theta \sup_{[\kappa,\tau]} \widehat{I}_p. \tag{25}$$

Now, using the fact that $\widehat{J}_q$ is nondecreasing and then (25), denoting $\varepsilon' = \varepsilon/(4v_p)$,

$$\begin{aligned} P(E) &= P(\tau^n \le \theta) = P_E(\Delta(\tau) - \Delta(\sigma) \ge \varepsilon/2) \\ &\le P_E(\widehat{J}_p(\tau) - \widehat{J}_p(\sigma) \ge 2\varepsilon') \\ &\le P_E((\kappa \wedge \tau - \sigma)|\widehat{I}_p|^*_\theta \ge \varepsilon') + P_E(\tau > \kappa,\, \theta \sup_{[\kappa,\tau]} \widehat{I}_p \ge \varepsilon'). \end{aligned}$$

By Step 1 and the tightness of $|\widehat{I}_p|^*_\theta$ the first term converges to 0, and by Step 2 so does the second term. ■

In fact, the proof of Lemma 3.3 establishes the following.

Corollary 3.4 *Fix $p, q \in L$, $p \ne q$, $\theta \in (0,\infty)$ and $\varepsilon > 0$. Define $\Delta^n(t) = v_p \widehat{J}^n_p(t) - v_q \widehat{J}^n_q(t)$. Then under any (work conserving) policy that gives priority to pool p over pool q whenever $\Delta^n(t) > \varepsilon$, for any $\delta > 0$,*

$$P\left(|(\Delta^n)^+|^*_\theta > \varepsilon + \delta\right) \to 0$$

as $n \to \infty$.

Lemma 3.5 *For every $\eta, \varepsilon > 0$, there is $t_1 > 0$ such that for all $t > t_1$,*

$$\liminf_n P(\widehat{J}^n(t) \ge \eta) \ge 1 - \varepsilon.$$

Proof: Recall equations (15)–(18). On the event $E_{n,t} = \{\widehat{J}(t) < \eta\}$ one has

$$F^n(t) = \sum_{k \in K} \mu_k \widehat{J}_{(k)}(t) \le \bar{\mu}\eta$$

and

$$\int_0^t \widehat{X}^-(s)ds = \int_0^t \widehat{I}(s)ds = \widehat{J}(t) \le \eta.$$

Hence

$$-\eta \le \int_0^t \widehat{X}(s)ds \le \widehat{X}(0)t + \int_0^t W^n(s)ds + \frac{1}{2}\hat{\beta}^n t^2 + \bar{\mu}\eta t,$$

and dividing by t, we have on $E_{n,t}$

$$-\frac{\eta}{t} - \widehat{X}(0) - \bar{\mu}\eta - \frac{1}{2}\hat{\beta}^n t \le \frac{1}{t}[\,|\widehat{A}|^*_t + \sum_i |\widehat{S}_i \circ \bar{T}_i|^*_t\,]. \tag{26}$$

Recalling that $\widehat{\beta}^n \to \widehat{\beta} < 0$ by (9), and that $\widehat{X}(0)$ are tight, the result follows on applying the last assertion of Lemma 3.2. ■

Proof of Theorems 2.2 and 2.3: Let $\varepsilon > 0$ and $t \geq 0$ be given. The first statement is an immediate consequence of Lemma 3.3. Tightness of $\gamma^n(\varepsilon)$ follows from Lemma 3.5. Let θ be so large that $\gamma^n = \gamma^n(\varepsilon) < \theta - t$ with probability at least $1 - \varepsilon/2$, for all sufficiently large n. Fix $\theta_1 \geq \theta$. Let $\delta > 0$ be given, and consider the event $\max_{i,j} |v_i \widehat{J}_i^n - v_j \widehat{J}_j^n|^*_{\theta_1} \leq \delta$. By Lemma 3.3, this event has probability at least $1 - \varepsilon/2$, for all sufficiently large n. A simple calculation shows that if $a_i \geq 0$, $i \in L$ are given numbers that sum up to $a > 0$, and $\max_{i,j} |v_i a_i - v_j a_j| \leq \delta$ then $\max_i |(a_i/a) - u_i| \leq \delta/a$. Hence for all sufficiently large n, with probability of at least $1 - \varepsilon$, one has

$$\gamma^n + t < \theta \quad \text{and} \quad \max_{i \in L} \Big| \frac{\widehat{J}_i^n(\gamma^n + t)}{\widehat{J}^n(\gamma^n + t)} - u_i \Big| \leq \frac{\delta}{\widehat{J}^n(\gamma^n + t)} \leq \frac{\delta}{\varepsilon},$$

where we used the fact $\widehat{J}^n(\gamma) = \varepsilon$ and that the process $\widehat{J}^n$ is nondecreasing. Theorem 2.2 follows upon setting $\delta = \varepsilon^2$.

To prove Theorem 2.3, we take $t = 0$ in the above argument, and set $T = \theta$. Then we fix some $T_1 \geq T$ and set $\theta_1 = T_1$. On the event analyzed in the previous paragraph, for all sufficiently large n, with probability at least $1 - \varepsilon$, we obtain

$$T \geq \gamma^n \text{ (hence } \widehat{J}^n(T) \geq \varepsilon) \quad \text{and} \quad \max_{i \in L} \sup_{s \in [T, T_1]} \Big| \frac{\widehat{J}_i^n(s)}{\widehat{J}^n(s)} - u_i \Big| \leq \frac{\delta}{\widehat{J}^n(T)} \leq \frac{\delta}{\varepsilon} = \varepsilon.$$

■

4 Extensions and discussion

We begin by noting that the proofs hold under assumptions on the arrival process that are weaker than the renewal structure. We then discuss two aspects of implementation, namely delayed information transmission, and a distributed set up.

Relaxed assumptions on arrivals. The probabilistic assumption on the arrival process can be much relaxed. Rather than assuming A^n are renewal processes, let us assume that the normalized processes $\widehat{A}^n$ (12) are C-tight (recall the definition from Section 3). In addition, assume that given $\varepsilon_1, \varepsilon_2 > 0$ there exists t_1 such that

$$\limsup_{n \to \infty} P(|\widehat{A}^n|^*_t \geq \varepsilon_1 t) \leq \varepsilon_2, \qquad t \geq t_1 \tag{27}$$

(compare with (20)). The assumptions on the parameters λ^n, namely (4), (8) and (9) are, of course, kept.

Corollary 4.1 *Under the relaxed assumptions on the arrival processes just described, the results of Theorems 2.2 and 2.3 are valid.*

Proof: A review of the proofs of Lemmas 3.1–3.3 shows that the C-tightness property suffices. Lemma 3.5 relies, in addition, on the last assertion of Lemma 3.2, which has been substituted by the assumption (27). Finally, the proof of the theorems holds verbatim. ■

Delayed information. Note that by Corollary 3.4, we can extend the results to policies which are only approximately u-greedy. This implies that we may perform the routing on the basis of (slightly) delayed information.

Theorem 4.2 *Let π be any work conserving policy that gives priority at every time $t > d_0$ to the pool with the largest value for $v_p J_p^n(t - d_0)$, where $d_0 > 0$ is a fixed delay. Then under π, for every $\varepsilon > 0$ and $T > 0$,*

$$\lim_{d_0 \to 0} \lim_{n \to \infty} P\Big\{ \max_{i,j \in L, i \neq j} \sup_{s \in [0,T]} \big| v_i \widehat{J}_i^n(s) - v_j \widehat{J}_j^n(s) \big| > \varepsilon \Big\} = 0.$$

The meaning of the result is that when the information on the processes J_i is obtained with a small delay, the effect with regard to the fairness performance is small.

Proof: The proof relies on the fact that we have a-priory bounds on the *slope* of Δ^n, and thus a small delay may cause an error with only small probability.

Note that Lemmas 3.1 and 3.2 are in force, since the only property of the policy that they use is that it is work conserving. Fix p, q, T, and $\varepsilon > 0$, and let Δ^n be as in Corollary 3.4. Consider the events

$$\Omega_1^n = \{\text{for every } t \in [d_0, T] \text{ one has } \Delta^n(t) > \varepsilon \text{ provided that } \Delta^n(t - d_0) > 2\varepsilon\},$$

$$\Omega_2^n = \{|\Delta^n|_{d_0}^* < \varepsilon\}.$$

Applying the corollary with $\delta = \varepsilon$ shows that

$$\lim_{n \to \infty} P(\Omega_1^n \cap \Omega_2^n \cap \{|(\Delta^n)^+|_T^* > 2\varepsilon\}) = 0.$$

Recalling that Δ^n is differentiable and null at zero, the probability of the complement of $\Omega_1^n \cap \Omega_2^n$ is bounded by

$$\begin{aligned} &P(\text{there exists } t \in [d_0, T] \text{ such that } \Delta^n(t) \leq \varepsilon \text{ and } \Delta^n(t - d_0) \geq 2\varepsilon; \text{ or } |\Delta^n|_{d_0}^* \geq \varepsilon) \\ &\quad \leq P\Big(\Big| \frac{d}{dt} \Delta^n \Big|_T^* \geq \frac{\varepsilon}{d_0} \Big). \end{aligned}$$

However, the derivative of Δ^n is bounded by $\max_i v_i \, |\widehat{I}^n|_T^*$, which, by Lemma 3.2 is tight. As a result, the bound in the above display tends to 0 as $d_0 \to 0$. Since p, q and ε are arbitrary, this completes the proof. ■

We remark that random, pool-dependent, time varying delays may be handled in the same way. If the maximal delay over the interval tends to 0 in probability, an analogue of the above result is valid.

Tree structure implementation. Consider now a system with a large number of pools. From the point of view of managing such systems, it is desirable to reduce the amount of information that is needed in order to decide on the routing, and possibly also reduce the amount of computation required by the central controller. In practice, this may be of particular importance if the facility is distributed geographically over many locations. As a model, consider the following tree-like system of pools. Each leaf is a pool, and each node represents a local processing center, the root being the central controller.

Since the only information the central controller requires is the index of the pool with largest value $v_i J_i$ that has free servers, a u-greedy policy can be implemented as follows. Each node sends to its parent (the closest node connected to it which is closer to the root node) a single number - the value of $v_i J_i$ for that pool among its offspring (including pools or nodes under it), with largest value $v_i J_i$ that has free servers. The decision by the root node is then between a small number of values—one for each sub-node or pool directly under it. This strategy has the advantage that most of the information is transmitted only locally—from each leaf to the node above it. The analysis of the case of delayed information applies also to the tree structure implementation.

References

[1] Armony, M., and Ward, A.R. (2008) Fair dynamic routing in large-scale heterogeneous-server systems. Preprint.

[2] Atar, R. (2005) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15 no. 4, 26062650

[3] Atar, R. (2008) Central limit theorem for a many-server queue with random service rates. *Ann. Appl. Probab.*, 18, no. 4, 1548-1568

[4] Atar, R., and Shwartz, A. (2008) Efficient routing in heavy traffic under partial sampling of service times. *Math. Op. Res.* 33: 899 - 909

[5] Avi-Itzhak, B. Levy, H. and Raz, D. (2004) Quantifying fairness in queueing systems: Principles and applications. Preprint

[6] Billingsley, P., *Convergence of Probability Measures*, John Wiley and Sons, Inc.

[7] Dai J. G. and Tezcan, T. State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.*, to appear.

[8] Gurvich I. and Whitt W. (2007) Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Math. Op. Res.*, to appear.

[9] Halfin, S., and Whitt, W. (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29, no. 3, 567-588.

[10] Hale, J.K. (1980) *Ordinary differential equations*, Robert E. Krieger publishing Company, Huntington, New York.

[11] Stolyar, A.L. Tezcan, T. (2008), Control of systems with flexible multi-server pools: A shadow routing approach. Preprint

[12] Tseytlin, Y. (2007), *Queueing systems with heterogeneous servers: Improving patients' flow in hospitals.* Research Proposal, The Faculty of Industrial Engineering and Management, Technion.

[13] Wierman, A. (2007) Fairness and classification. *Performance Evaluation Review*, 34(4), pp. 412.