Combining Size-Based Load Balancing with Round-Robin for Scalable Low Latency

Jonatha Anselmi

Asymptotically Optimal Size-Interval Task Assignments

Jonatha Anselmi and Josu Doncel

Abstract—Size-based routing provides robust strategies to improve the performance of computer and communication systems with highly variable workloads because it is able to isolate small jobs from large ones in a static manner. The basic idea is that each server is assigned all jobs whose sizes belong to a distinct and continuous interval. In the literature, dispatching rules of this type are referred to as SITA (Size Interval Task Assignment) policies. Though their evident benefits, the problem of finding a SITA policy that minimizes the overall mean (steady-state) waiting time is known to be intractable. In particular it is not clear when it is preferable to balance or unbalance server loads and, in the latter case, how. In this paper, we provide an answer to these questions in the celebrated limiting regime where the system capacity grows linearly with the system demand to infinity. Within this framework, we prove that the minimum mean waiting time achievable by a SITA policy necessarily converges to the mean waiting time achieved by SITA-E, the SITA policy that equalizes server loads, provided that servers are homogeneous. However, within the set of SITA policies we also show that SITA-E can perform arbitrarily bad if servers are heterogeneous. In this case we prove that there exist exactly $C!$ asymptotically optimal policies, where C denotes the number of server *types*, and all of them are linked to the solution of a single strictly convex optimization problem. It turns out that the mean waiting time achieved by any of such asymptotically optimal policies does not depend on how job-size intervals are mapped to servers. Our theoretical results are validated by numerical simulations with respect to realistic parameters and suggest that the above insights are also accurate in small systems composed of a few servers, i.e., ten.

Index Terms—Dispatching policies, size-based routing, performance, asymptotic optimality

1 Introduction

The distributed architecture under investigation in this paper has the structure illustrated in Figure 1, where $M \geq 1$ denotes the number of *dispatchers* and $K \gg 1$ the number of *servers*. Work units, or *jobs* in the following,

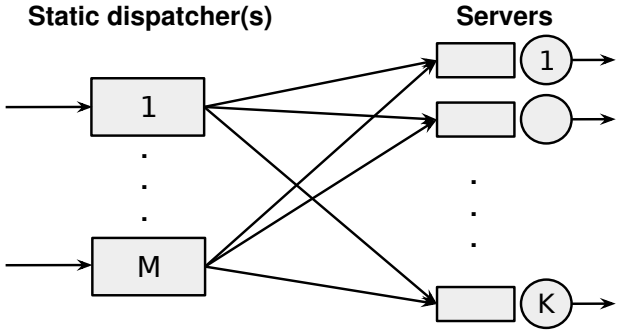

Fig. 1. Architecture of the parallel and distributed system under investigation for assigning jobs to servers.

join the dispatchers over time and are routed to servers for processing according to some dispatching policy. This is a typical scenario encountered in data centers, server farms, supercomputing systems and call centers. A fundamental performance-related question is how to allocate jobs to servers to achieve low latency when job sizes or processing times are highly variable. This is particularly interesting when servers locally schedule jobs in a first-come first-served (FCFS) manner, where a single long job may block many short jobs behind, thus deteriorating the overall average latency significantly.

A number of dispatching algorithms are based on feedback information *dynamically* flowing over time from the servers to the dispatchers, e.g., join-the-shortest-queue [1], power-of-d-choices schemes [2], pull-based techniques [3], [4], or across the servers themselves, e.g., job replication and/or redundancy [5], work-stealing and/or job migration [6]. Though dynamic information allows one to develop low latency dispatching schemes in large-scale clusters, the intrinsic price that these algorithms have to pay stands in the unavoidable communication overhead due to control messages together with the development of an ad-hoc communication protocol. Tradeoffs between communication overhead and latency have been recently investigated in [7].

Rather, we are interested in *static* assignment policies: routing decisions do not depend on past or current observations of system states, e.g., server idleness, queue lengths, workloads, and do not require the management of a local dynamic memory. Essentially, existing static policies try to exploit the fact that determinism minimizes the mean waiting time in the G/G/1 queue [8], [9]. This is achieved by reducing the variance of either the arrival or service process associated with each server: approaches such as Round-Robin (RR) [10] or generalized RR strategies [11], [12], [13] make the sequence of interarrival times at each server as *regular* as possible, while size-based routing [14], [15], [16] attempts to make the sequence of service times at each server as deterministic as possible. In principle, if the job sizes are more variable than the interarrival times, then size-based routing strategies are preferable, and vice versa.

- *J. Anselmi is with INRIA Bordeaux Sud Ouest, Team: CQFD, 33405 Talence, France and Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. E-mail: jonatha.anselmi@inria.fr*
- *J. Doncel is with the University of the Basque Country, UPV/EHU, Barrio sarriena s/n, 48940 Leioa, Spain. E-mail: josu.doncel@ehu.eus*

Manuscript received April 19, 2005; revised August 26, 2015.

With respect to the architecture in Figure 1, RR is widely applicable but it performs poorly in the case of multiple dispatchers because it becomes difficult to control the arrival process at each server. On the other hand, size-based routing is versatile enough to be implemented in a decentralized manner across multiple dispatchers without compromising performance. This can be proven analytically, e.g., under Poisson assumptions for the arrival processes.

1.1 Size-Interval Task Assignment

In this paper, we focus on a class of size-based routing policies referred to as SITA (Size Interval Task Assignment) policies; see, e.g., [14]. The basic idea is that each server is assigned all jobs whose sizes belong to a distinct and continuous interval. From a practical standpoint, a SITA policy can be implemented in several ways, depending on the underlying architecture [17], [18]: for instance, each job may submit to the dispatcher an upper bound on its duration (as in, e.g., supercomputing), or the dispatcher itself may either know the identities of the servers hosting jobs of a given size (as in, e.g., web files transfers) or just be able to directly observe job sizes.

The main benefits of SITA policies are attributed to their ability to isolate small jobs from long ones, which reduces the variance of the service processes of all servers. This is particularly noticeable in large-scale clusters because the length of each interval shrinks more and more as the number of servers increases. It is well known that they can outperform dynamic policies such as Join-the-Shortest-Workload, in which a job is assigned to the server with the least remaining work [14], [15], [19], [20], [21]. A comprehensive analytical comparison between the performance of SITA policies and Join-the-Shortest-Workload is shown in [21], where the authors present several scenarios where one approach can be better than the other.

The problem of finding a SITA policy that minimizes the overall mean steady-state waiting time is known to be intractable [16], [22], [23], [24], [25]. In particular, it is not clear when it is preferable to balance or unbalance server loads and, in the latter case, how. From an analytical standpoint, this problem is already difficult with only two homogeneous servers: in [26], some conditions are given to establish when the short or long job host should be underloaded.

To find the optimal SITA policy, one should solve a two-level optimization problem. First, one needs to understand whether 'short' or 'long' jobs should be mapped to the fastest or slowest servers. Then, fixed such mapping, one needs to find the associated optimal cutoffs, which is equivalent to solving a continuous nonlinear optimization problem. If K is the number of servers, possibly operating at different processing speeds, then $K-1$ is the number of cutoffs and $K!$ is the number of mappings, in fact *permutations*, between cutoff intervals and servers. For any choice of the cutoffs, a brute force approach to finding a corresponding optimal permutation is computationally expensive. It turns out that the optimal permutation, i.e., the permutation that achieves the lowest mean waiting time, strongly depends on the job size distribution [16]; this will be discussed in Section 4.2. To the best of our knowledge, the only analytical results available in the literature that determine whether one permutation is better, with respect to the overall mean waiting time, than another are [24, Theorem 6.6], which is only valid for the Bounded Pareto distribution, and [16, Proposition 6], which is only valid for two queues.

1.2 Our Contribution

Motivated by the size of large-scale clusters, in this paper we investigate the performance of SITA policies in the celebrated limiting regime where the system capacity grows linearly with the system demand to infinity. In our first result, we prove that the minimum mean waiting time achievable by a SITA policy *necessarily* converges to the mean waiting time achieved by SITA-E, the SITA policy that equalizes server loads, provided that servers are homogeneous. In other words, SITA-E is the unique asymptotically optimal SITA policy. Then, we consider heterogeneous systems and prove that in this case there exist exactly $C!$ asymptotically optimal policies, where C denotes the number of server *types*, and all of them are linked to the solution of a single strictly convex optimization problem. It turns out that the mean waiting time achieved by any of these asymptotically optimal policies does not depend on the particular permutation chosen between cutoff intervals and servers, though each of these permutations provides permutation-dependent cutoffs. Finally, we also evaluate the ratio between the mean waiting time achieved with SITA-E and the one of an asymptotically optimal policy. By means of a competitive analysis, we show that this efficiency ratio, when applied to a heterogeneous system, can be arbitrarily large, though in the average case is surprisingly close to one.

Our results provide efficient methods for computing near-optimal SITA policies in the prelimit and are validated by numerical simulations with respect to realistic parameters, suggesting that the above insights are also valid for small systems composed of a few servers, i.e., ten.

The closest reference to our work is [27], where the authors follow a similar approach to investigate the performance of SITA policies. The main difference with respect to our work is that we study the limit of the optimal policy instead of the optimal policy in the limit. Technically speaking, we are interested in a "$\liminf$-problem" rather than an "$\inf\lim$-problem", which is more difficult as the latter always provides an upper bound on the former. Our proof is also based on different arguments. Nonetheless, our results for homogeneous servers are consistent with those of that reference, which indicates that both the $\inf$ and $\lim$ operators can be exchanged.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we present the dispatching model, define SITA policies for homogeneous servers and state the technical problem. In Section 3, we show and prove our first result, Theorem 1, i.e., the asymptotic optimality of SITA-E in case of homogeneous servers. Section 4, devoted to heterogeneous servers, discusses the impact of the mapping between size intervals and servers, and presents our second result, Theorem 2, which gives the structure of asymptotically optimal SITA policies. Finally, Section 5 presents numerical results and Section 6 draws the conclusions of our work.

2 Dispatching Model

We consider a parallel system composed of K servers (or queues), namely $\{1,\dots,K\}$, and one central controller that is in charge of routing jobs to servers according to some policy, or decision rule; in fact, our model will be analytically equivalent to the setting in Figure 1 with multiple dispatchers. Jobs arrive to the central controller following a Poisson process of rate λK. Servers are modeled by FCFS queues, or any other discipline which does not affect the distribution of the number of jobs in the queue at any time, and operate at constant speed $\mu = 1$; in Section 4, we will deal with the case where speeds are different.

Let $(X_i)_i$ be the sequence of i.i.d random variables representing the sizes of the incoming jobs, assumed independent of the arrival process at the controller. We also assume that the X_i's have the same distribution of a random variable X and a density function $f(x)$ defined over $[x_m, x_M]$, with $0 < x_m < x_M < \infty$, such that on this interval $\frac{1}{xf(x)}$ is Lipschitz. This implies that x_m and x_M are the minimum and maximum sizes of the incoming jobs, respectively. We also assume that

$$\rho \stackrel{\text{def}}{=} \frac{\lambda \mathbb{E}[X]}{\mu} < 1,$$

which will be necessary for stability of the Markov process underlying the K-th system. Let also $F(x) = P(X \le x)$.

2.1 SITA Policies

The dispatching rule adopted by the controller is assumed to only depend on the size of each incoming job. In particular, it is assumed that each server is assigned all jobs whose sizes belong to a distinct and continuous interval. Following queueing theory parlance, we refer to these deterministic dispatching rules as SITA policies, defined as follows.

Definition 1. *For a system with K servers, a* SITA policy *is a surjective mapping $R_K : [x_m, x_M] \to \{\frac{1}{K}, \frac{2}{K}, \dots, 1\}$ such that $R_K^{-1}(i/K)$ is an interval, for all $i \in \{1,\dots,K\}$.*

In other words, a SITA policy R_K is a piece-wise constant function with exactly $K-1$ points of discontinuity, and the interpretation is that a controller implementing R_K sends a job of size $x \in [x_m, x_M]$ to server $KR_K(x)$.

Let $x_{K,0} \stackrel{\text{def}}{=} x_m$ and $x_{K,K} \stackrel{\text{def}}{=} x_M$. For $i = 1,\dots,K-1$, let $x_{K,i} \stackrel{\text{def}}{=} x_{K,i}(R_K)$ denote the i-th discontinuity point of R_K. Given R_K, the points $(x_{K,i})_{i=1}^{K-1}$ are said *thresholds* (or cutoffs) of R_K. If $X_{K,i}$ denotes the random variable representing the size of jobs joining queue i, then after conditioning we obtain that the distribution of $X_{K,i}$ is

$$F_{K,i}(x) \stackrel{\text{def}}{=} \frac{\int_{x_{K,i-1}}^{x} f(x)\,\mathrm{d}x}{F(x_{K,i}) - F(x_{K,i-1})} \tag{1}$$

and

$$\mathbb{E}[X_{K,i}^n] = \frac{\int_{x_{K,i-1}}^{x_{K,i}} x^n f(x)\,\mathrm{d}x}{F(x_{K,i}) - F(x_{K,i-1})}. \tag{2}$$

Since the arrival process at the dispatcher is Poisson, a trivial consequence of the SITA policy R_K is that the arrival process at each server i is a Poisson process with rate $(F(x_{K,i}) - F(x_{K,i-1}))\lambda K$. Furthermore, using the Pollaczek–Khinchine formula (see, e.g., [28]) and provided that the (necessary and sufficient) stability condition

$$\lambda K(F(x_{K,i}) - F(x_{K,i-1}))\mathbb{E}[X_{K,i}] < \mu, \quad \forall i = 1,\dots,K \tag{3}$$

is satisfied, the mean steady-state waiting time experienced by the incoming jobs and achieved with R_K, say $W_K(R_K)$, is given by

$$W_K(R_K) = \frac{\lambda K}{2\mu^2} \sum_{i=1}^{K} \frac{(F(x_{K,i}) - F(x_{K,i-1}))^2\, \mathbb{E}[X_{K,i}^2]}{1 - \frac{\lambda}{\mu} K(F(x_{K,i}) - F(x_{K,i-1}))\mathbb{E}[X_{K,i}]}. \tag{4}$$

If (3) is not satisfied, we let $W_K(R_K) = +\infty$.

2.2 Problem Statement

We are interested in finding the SITA policies that minimize (4). As discussed in the Introduction, this is a known difficult problem and we aim at providing answers in the large-network limiting regime where $K \to \infty$. Towards this purpose, we will use the assumption below, which constructs the set of SITA policies of interest with respect to a sequence of systems indexed by K.

Let $\mathcal{S}$ denote the set of differentiable, increasing and Lipschitz functions $R : [x_m, x_M] \to [0,1]$ such that $R(x_m) = 0$, $R(x_M) = 1$ and

$$\sup_K \sup_{i=1,\dots,K} \lambda K \int_{R^{-1}\left(\frac{i-1}{K}\right)}^{R^{-1}\left(\frac{i}{K}\right)} x f(x)\mathrm{d}x < \mu. \tag{5}$$

Assumption 1. *The thresholds of the K-th system are given by $x_{K,i} = R^{-1}(i/K)$, for all $i = 1,\dots,K-1$, for some $R \in \mathcal{S}$.*

Given $R \in \mathcal{S}$, the previous assumption allows us to construct a sequence of SITA policies indexed by K: the SITA policy for the K-th system is a cadlag, piece-wise constant function $R_K : [x_m, x_M] \to \{\frac{1}{K}, \frac{2}{K}, \dots, 1\}$ with exactly $K-1$ points of discontinuity and such that $R_K^{-1}(i/K) = R^{-1}(i/K)$, for all $i \in \{1,\dots,K\}$.

The fact that R is assumed increasing is not a loss of generality because the processing speeds of the servers are identical for now. In fact, given a SITA policy R_K for the K-system, it is always possible to find a permutation of servers such that the resulting mapping between job sizes and servers is increasing without changing the mean waiting time $W_K(R_K)$. The requirement (5), obtained by using (2) and Assumption 1 in (3), is necessary and sufficient for stability, and allows us to use (4) for the mean waiting time.

We also notice that our construction of SITA policies is meant to exclude the set of the so called *nested* SITA policies [16], where size intervals may overlap. This is not a loss of generality for homogeneous systems because the optimal nested SITA policy provides the same mean waiting time of the optimal SITA policy with non-overlapping size intervals [16, Theorem 3].

Remark 1. *With a slight abuse of notation, given $R \in \mathcal{S}$ we will write $W_K(R)$ to refer to $W_K(R_K)$ where R_K is such that $R_K^{-1}(i/K) = R^{-1}(i/K)$.*

Our objective is to establish asymptotic optimality results within the set of SITA policies satisfying Assumption 1.

Specifically, we are interested in the problem of finding a function $R^* \in \mathcal{S}$ such that

$$\lim_{K\to\infty} W_K(R^*) = \lim_{K\to\infty} \inf_{R\in\mathcal{S}} W_K(R). \tag{6}$$

To the best of our knowledge, this mathematical problem has not been investigated in the literature; see the Introduction for related works.

3 Asymptotic Optimality of SITA-E

A particular SITA policy is obtained when server loads are equalized. When servers are homogeneous, such condition identifies a unique SITA policy, usually referred to as SITA-E in the literature, which for the K-th system we denote by R^*_K. The thresholds of R^*_K, namely $(x^*_{K,i})_{i=1}^{K-1}$, are uniquely determined by the following set of equations:

$$\int_{x^*_{K,i-1}}^{x^*_{K,i}} xf(x)\mathrm{d}x = \int_{x_m}^{x^*_{K,1}} xf(x)\mathrm{d}x, \quad \forall i = 2,\dots,K. \tag{7}$$

The next result shows that SITA-E is *the* asymptotically optimal SITA policy in the limiting regime where $K \to \infty$.

Theorem 1. *Let R^{*-1} be the unique solution of the initial value problem*

$$zf(z)z' = \mathbb{E}[X] \tag{8a}$$
$$z(0) = x_m. \tag{8b}$$

Then,

$$R^{*}_K{}^{-1}(\tfrac{i}{K}) = R^{*-1}(\tfrac{i}{K}), \quad \forall i = 1,\dots,K \tag{9}$$

and

$$\lim_{K\to\infty} \inf_{R\in\mathcal{S}} W_K(R) = \lim_{K\to\infty} W_K(R^*) = \frac{\lambda}{2}\frac{\mathbb{E}[X]^2}{1-\rho}. \tag{10}$$

Furthermore, no $R \in \mathcal{S}\setminus\{R^*\}$ exists such that

$$\lim_{K\to\infty} W_K(R) = \lim_{K\to\infty} W_K(R^*).$$

Proof. For $i = 0, 1, 2$, let

$$M_i(x) \stackrel{\text{def}}{=} \int x^i f(x)\mathrm{d}x \tag{11}$$

Given $R(\cdot)$ and substituting $x_{K,i} = R^{-1}(i/K)$ in (2), we obtain

$$\mathbb{E}[X^n_{K,i}] = \frac{\int_{R^{-1}\left(\frac{i-1}{K}\right)}^{R^{-1}\left(\frac{i}{K}\right)} x^n f(x)dx}{F(R^{-1}\left(\frac{i}{K}\right)) - F(R^{-1}\left(\frac{i-1}{K}\right))}.$$

and substituting in (4), we rewrite the mean waiting time achieved by the K-th system as

$$W_K(R) = \frac{\lambda}{2}\sum_{i=1}^{K} K\frac{\prod_{j\in\{0,2\}}\left(M_j(R^{-1}\left(\frac{i}{K}\right)) - M_j(R^{-1}\left(\frac{i-1}{K}\right))\right)}{1-\lambda K\left(M_1(R^{-1}\left(\frac{i}{K}\right)) - M_1(R^{-1}\left(\frac{i-1}{K}\right))\right)}. \tag{12}$$

Since F and R are differentiable, for $j = 0, 1, 2$ we have

$$M_j(R^{-1}\left(\tfrac{i-1}{K}\right)) = M_j(R^{-1}\left(\tfrac{i}{K}\right)) - \frac{1}{K}\left.\frac{\mathrm{d}M_j(R^{-1}(x))}{\mathrm{d}x}\right|_{x=\frac{i}{K}} + O(K^{-2}) \tag{13}$$

where

$$\frac{\mathrm{d}M_j(R^{-1}(x))}{\mathrm{d}x} = \left(R^{-1}(x)\right)^j f\left(R^{-1}(x)\right)\frac{\mathrm{d}R^{-1}(x)}{\mathrm{d}x}. \tag{14}$$

Substituting (13) in (12) and using that the $O(K^{-2})$ terms in (13) are uniformly bounded in i for fixed K (recall that R is Lipschitz continuous and defined on a compact set), we obtain

$$\lim_{K\to\infty}\inf_{R\in\mathcal{S}} W_K(R) \tag{15}$$
$$= \frac{\lambda}{2}\lim_{K\to\infty}\inf_{R\in\mathcal{S}}\sum_{i=1}^{K}\frac{1}{K}\frac{\prod_{j\in\{0,2\}}\frac{\mathrm{d}}{\mathrm{d}x}M_j(R^{-1}(x))\big|_{x=\frac{i}{K}}}{1-\lambda\frac{\mathrm{d}}{\mathrm{d}x}M_1(R^{-1}(x))\big|_{x=\frac{i}{K}}} \tag{16}$$
$$= \frac{\lambda}{2}\lim_{K\to\infty}\inf_{R\in\mathcal{S}}\sum_{i=1}^{K}\frac{1}{K}\frac{g^2(\frac{i}{K})}{1-\lambda g(\frac{i}{K})} \tag{17}$$

where

$$g(x) \stackrel{\text{def}}{=} \frac{\mathrm{d}M_1(R^{-1}(x))}{\mathrm{d}x} = R^{-1}(x)f\left(R^{-1}(x)\right)\frac{\mathrm{d}R^{-1}(x)}{\mathrm{d}x}. \tag{18}$$

Letting $u(x) \stackrel{\text{def}}{=} \frac{x^2}{1-\lambda x}$, we have

$$\inf_{R\in\mathcal{S}}\sum_{i=1}^{K}\frac{1}{K}\frac{g^2(\frac{i}{K})}{1-\lambda g(\frac{i}{K})} = \inf_{\alpha\in\mathbb{R}_+}\inf_{\substack{R\in\mathcal{S}:\\ \sum_{i=1}^{K}\frac{1}{K}g(\frac{i}{K})=\alpha}}\sum_{i=1}^{K}\frac{u\left(g\left(\frac{i}{K}\right)\right)}{K} \tag{19}$$
$$\geq \inf_{\alpha\in\mathbb{R}_+}\inf_{\substack{R\in\mathcal{S}:\\ g(\frac{i}{K})=\alpha,\forall i}}\sum_{i=1}^{K}\frac{u\left(g\left(\frac{i}{K}\right)\right)}{K} \tag{20}$$
$$= \inf_{\alpha\in\mathbb{R}_+}\inf_{\substack{R\in\mathcal{S}:\\ g(\frac{i}{K})=\alpha,\forall i}} u(g(1)) \tag{21}$$
$$= \inf_{\substack{R\in\mathcal{S}:\\ g(\frac{1}{K})=g(\frac{2}{K})=\cdots=g(1)}} u(g(1)). \tag{22}$$

In (19), we have used that $g(x) \geq 0$ because R, and thus R^{-1}, is increasing. In (20), we have used the convexity of u and applied Karamata's inequality as the vector $\left(g\left(\frac{1}{K}\right), g\left(\frac{2}{K}\right), \dots, g(1)\right)$ majorizes the vector $(\alpha,\dots,\alpha)$, provided that $\sum_{i=1}^{K}\frac{1}{K}g\left(\frac{i}{K}\right) = \alpha$; this is the key observation of our proof.

Now, let $\mathcal{S}^* \stackrel{\text{def}}{=} \{R \in \mathcal{S} : g(x) \text{ is constant}\}$. We notice that $R \in \mathcal{S}^*$ if and only if $R \in \mathcal{S}$ and for some $c \in \mathbb{R}_+$

$$g(x) = R^{-1}(x)f\left(R^{-1}(x)\right)\frac{\mathrm{d}R^{-1}(x)}{\mathrm{d}x} = c. \tag{23}$$

Given c and since $\frac{1}{xf(x)}$ is Lipschitz continuous by assumption, the Picard–Lindelöf theorem ensures that the ODE problem $zf(z)z' = c$, with $z(0) = x_m$, admits a unique solution. This implies that there exists a unique R^{-1}, and thus a unique R, that satisfies (23), say $R = R_c$ to stress the dependency on c. We notice that R_c is a member of $\mathcal{S}^*$ if and only if $R_c(x_M) = 1$. Since the derivative of $R_c^{-1}(x)$ is positive and proportional to c, there exists a unique c, say $c = h$, such that $R_c(x_M) = 1$. This proves that $\mathcal{S}^*$ is composed of one element only, say R^*. Now, given K, fix $R \in \mathcal{S}$ such that $g(\frac{1}{K}) = g(\frac{2}{K}) = \cdots = g(1)$ and $R \notin \mathcal{S}^*$. For all

$K' > K$ large enough, it is clear that such R will not be an element of $\{R \in \mathcal{S} : g(\frac{1}{K'}) = g(\frac{2}{K'}) = \cdots = g(1)\}$. This proves that we can exclude such functions in the infimum of the RHS term of (22). Therefore, given that $\mathcal{S}^*$ is a singleton, we obtain

$$\lim_{K\to\infty} \inf_{R\in\mathcal{S}} W_K(R) \geq \frac{\lambda}{2} \lim_{K\to\infty} \inf_{\substack{R\in\mathcal{S}: \\ g(\frac{1}{K})=g(\frac{2}{K})=\cdots=g(1)}} \frac{g^2(1)}{1-\lambda g(1)} \tag{24a}$$

$$= \frac{\lambda}{2} \lim_{K\to\infty} \inf_{R\in\mathcal{S}^*} \frac{g^2(1)}{1-\lambda g(1)} \tag{24b}$$

$$= \frac{\lambda}{2} \frac{h^2}{1-\lambda h}. \tag{24c}$$

Integrating (23) when $c = h$ and using a change of variable, we observe that

$$h = \int_0^1 R^{*-1} f(R^{*-1}) dR^{*-1}(x) \tag{25a}$$

$$= \int_{R^{*-1}(0)}^{R^{*-1}(1)} x f(x) \mathrm{d}x = \int_{x_m}^{x_M} x f(x) \mathrm{d}x = \mathbb{E}[X]. \tag{25b}$$

Thus, we obtain the lower bound

$$\lim_{K\to\infty} \inf_{R\in\mathcal{S}} W_K(R) \geq \frac{\lambda}{2} \frac{\mathbb{E}[X]^2}{1-\rho}. \tag{26}$$

To conclude the proof, we exhibit a matching upper bound. Towards this purpose, it is enough to show that $(R^{*-1}(i/K))_i$, for all $i = 1, \ldots, K-1$, are the thresholds of R^*_K, the SITA-E policy applied to the K-th system. Given (7), this amounts to prove that R^* satisfies

$$\int_{R^{*-1}(\frac{i-1}{K})}^{R^{*-1}(\frac{i}{K})} x f(x) \mathrm{d}x = \int_{x_m}^{R^{*-1}(\frac{1}{K})} x f(x) \mathrm{d}x, \quad \forall i = 2, \ldots, K. \tag{27}$$

i.e., for all $i = 2, \ldots, K$

$$M_1(R^{*-1}(\tfrac{i-1}{K})) - M_1(R^{*-1}(\tfrac{i}{K})) = M_1(R^{*-1}(\tfrac{1}{K})) - M_1(x_m) \tag{28}$$

Given (18) and (23), the derivative w.r.t. x of $M_1(R^{*-1}(x))$ is constant and therefore (28) must hold true. □

Theorem 1 provides a constructive method to find the asymptotically optimal SITA policy, SITA-E, by solving the ODE system (8). A general explicit expression for its solution depends on the probability distribution of the job sizes $F(x)$. For instance, if $F(x)$ is the Bounded Pareto distribution with shape parameter $t \in \mathbb{R}$, meaning that

$$f(x) = \frac{b}{x^{t+1}}, \quad b \stackrel{\text{def}}{=} \frac{t x_m^t}{1 - \left(\frac{x_m}{x_M}\right)^t} \tag{29}$$

then R^* takes a simple form. Integrating both sides of (8), assuming for simplicity $t \neq 1$, we obtain that $R^{*-1}(x)$ must satisfy the equation

$$x\mathbb{E}[X] = \frac{b}{1-t}\left(R^{*-1}(x)^{1-t} - x_m^{1-t}\right), \tag{30}$$

which holds true if and only if

$$R^{*-1}(x) = \left(x_m^{1-t} + (x_M^{1-t} - x_m^{1-t})x\right)^{\frac{1}{1-t}}. \tag{31}$$

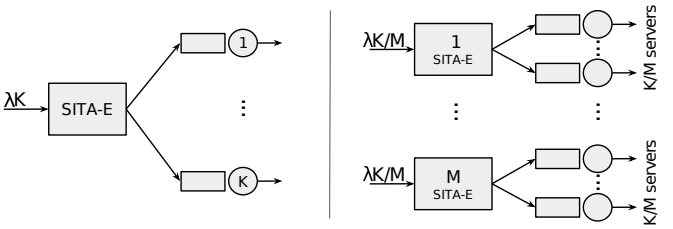

Fig. 2. Queueing systems for economies of scale.

We notice that the thresholds (31) are identical to the ones identified in [14, Theorem 3], though in that reference they have been obtained with a different method. Finally, inverting the previous function, we obtain

$$R^*(x) = \frac{x^{1-t} - x_m^{1-t}}{x_M^{1-t} - x_m^{1-t}}. \tag{32}$$

The case $t = 1$ is treated similarly to obtain

$$R^*(x) = \frac{x_m}{\mathbb{E}[X]} \left(1 - \frac{x_m}{x_M}\right)^{-1} \ln \frac{x}{x_m}.$$

3.1 Comparison with Round-Robin

We stress that SITA-E reduces the variance of the service process at each server: as K grows, the length of each size interval converges to zero so that in the limit each server sees deterministic job sizes coming in. The 'complementary' approach is given by the Round-Robin (RR) policy, another static policy that routes the n-th job to server $1 + (n \bmod K)$. In fact, RR does exactly the opposite: it reduces the variance of the arrival process at each server. In the limit where $K \to \infty$, each server sees jobs coming in at deterministic interarrival times. It is well-known that the resulting limiting mean waiting time, W^{RR}, corresponds to the mean waiting time of D/G/1 queue. Using a classic heavy-traffic approximation, see [29, Formula 2.10], this gives $W^{RR} \approx \frac{\lambda}{2}\frac{\mathrm{Var}(X)}{1-\rho}$. Comparing this formula with the asymptotic mean waiting time of SITA-E (the RHS of (10)), $W^{SITA} \stackrel{\text{def}}{=} \lim_{K\to\infty} W_K(R^*)$, we obtain

$$\frac{W^{RR}}{W^{SITA}} \approx \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2} - 1. \tag{33}$$

Thus, in the limit, SITA-E outperforms RR if and only if the square coefficient of variation of X, i.e. $\mathbb{E}[X^2]/\mathbb{E}[X]^2$, is greater than or equal to 2, which is indeed the case if X is highly variable.

3.2 Economies of Scale

In [30], the economies of scale in a parallel system composed of a single dispatcher operating under SITA-E are investigated when $K \to \infty$ by means of the *degradation factor*, $D_{K,M}$. This is meant to compare the ratio between the delays achieved by the queueing systems depicted on the right and on the left of Figure 2. This is equivalent to the ratio between the mean waiting time of a system with K/M servers (assuming that M divides K) and arrival rate $\lambda K/M$ and the mean waiting time of a system with K servers and arrival rate λK. It is shown that $D_{K,M} \leq D_{K,K}$

and the main conclusion is that the worst case degradation factor can increase without bound with the variability of the job size distribution. However, an explicit form for the degradation factor is not given and the previous conclusion is only shown with respect to some distribution functions. The following corollary of Theorem 1 is straightforward and gives the explicit structure of the worst-case degradation factor for any distribution.

Corollary 1.

$$\lim_{K\to\infty} D_{K,K} = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2}. \tag{34}$$

Proof. By definition, $D_{K,K}$ corresponds to the ratio between the mean waiting time of a M/G/1 queue with arrival rate λ and service time X and $W(R_K^*)$. Using the Pollaczek-Khinchine formula for the numerator, we obtain

$$D_{K,K} = \frac{\lambda}{2}\frac{\mathbb{E}[X^2]}{1-\rho} \times \frac{1}{W(R_K^*)} \to \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2}. \tag{35}$$

□

Since SITA-E is asymptotically optimal (Theorem 1), $\mathbb{E}[X^2]/\mathbb{E}[X]^2$ is also an upper bound on the degradation factor obtained when dispatchers operate under any SITA policy.

4 Heterogeneous Servers

In this section, we treat the case where servers may have different processing speeds. In this case, we will show that SITA-E is no longer asymptotically optimal. However, we will use Theorem 1 to identify asymptotically optimal SITA policies by means of convex programming.

Let $\alpha_0 = 0$ and $\alpha_1, \ldots, \alpha_C \in \mathbb{Q}_+$ such that $\sum_{c=1}^C \alpha_c = 1$. Let also $\beta_c \stackrel{\text{def}}{=} \sum_{c'=0}^{c} \alpha_{c'}$, for all $c = 0, \ldots, C$. We assume that there are K servers in total and that they belong to C types, in the sense that we let servers[1] $\beta_{c-1}K+1, \ldots, \beta_c K$ operate at constant speed μ_c, for all $c = 1, \ldots, C$. We also assume that $\lambda\mathbb{E}[X] < \sum_c \alpha_c\mu_c$, which is necessary for stability.

4.1 Construction of SITA Policies and Assumptions

In the case of heterogeneous servers, the construction of the SITA policies of interest is not trivial. Given $(a,b] \subseteq [x_m, x_M]$, let $\mathcal{R}^c_{[a,b]}$ denote the set of differentiable, increasing and Lipschitz functions $R : (a,b] \to (\beta_{c-1}, \beta_c]$ such that $R(a) \stackrel{\text{def}}{=} \lim_{x\downarrow a} R(x) = \beta_{c-1}$, $R(b) = \beta_c$ and

$$\sup_K \sup_{i=\beta_{c-1}K+1,\ldots,\beta_c K} \lambda K \int_{R^{-1}\left(\frac{i-1}{K}\right)}^{R^{-1}\left(\frac{i}{K}\right)} x f(x)\mathrm{d}x < \mu_c. \tag{36}$$

Let

$$\mathcal{Y} \stackrel{\text{def}}{=} \Big\{(y_0, \ldots, y_C) \in \mathbb{R}_+^{C+1} : \lambda \int_{y_{c-1}}^{y_c} x f(x)\mathrm{d}x \le \mu_c\alpha_c, \forall c \text{ and } x_m = y_0 \le y_1 \le \cdots \le y_C = x_M\Big\},$$

which will be interpreted as the set of threshold vectors associated to queues of different types: jobs of size in $(y_{c,1}, y_c]$ will be only sent to queues of type c', for some c' not necessarily equal to c.

Given $y \in \mathcal{Y}$ and a permutation π over $\{1, \ldots, C\}$, let

$$\mathcal{R}_{y,\pi} \stackrel{\text{def}}{=} \Big\{R : [x_m, x_M] \to [0,1] : \ R \text{ is bijective and } R_{|(y_{c-1},y_c]} \in \mathcal{R}^{\pi(c)}_{(y_{c-1},y_c]}, \forall c = 1, \ldots, C\Big\}$$

where $R_{|A}$ denotes the restriction of R to $A \subseteq \mathbb{R}$.

Finally, let $\mathcal{R}_y \stackrel{\text{def}}{=} \bigcup_\pi \mathcal{R}_{y,\pi}$ and $\mathcal{R} \stackrel{\text{def}}{=} \bigcup_{y\in\mathcal{Y}} \mathcal{R}_y$.

Assumption 2. *The thresholds of the K-th system are given by $x_{K,i} = R^{-1}(i/K)$, for all $i = 1, \ldots, K-1$, for some $R \in \mathcal{R}$.*

As done in Section 2.2, the previous assumption defines the set of SITA policies of interest for the K-th system and we will establish asymptotic optimality results within the set of SITA policies satisfying Assumption 2. Again, the fact that $R \in \mathcal{R}_{y,\pi}$ is assumed increasing when restricted over $[y_{c-1}, y_c]$ is not a loss of generality because all the thresholds included in $[y_{c-1}, y_c]$ belong to type-$\pi(c)$ servers only.

Using the Pollaczek–Khinchine formula and the notation (11), the mean steady-state waiting time experienced by the incoming jobs and achieved with $R \in \mathcal{R}_{y,\pi}$, say $W_K(R)$, is given by

$$W_K(R) = \frac{\lambda}{2}\sum_{c=1}^{C} \frac{1}{\mu^2_{\pi(c)}} \sum_{i=\beta_{\pi(c)-1}K+1}^{\beta_{\pi(c)}K} \frac{K\prod_{j\in\{0,2\}}\left(M_j(R^{-1}\left(\frac{i}{K}\right)) - M_j(R^{-1}\left(\frac{i-1}{K}\right))\right)}{1 - \frac{\lambda}{\mu_{\pi(c)}}K\left(M_1(R^{-1}\left(\frac{i}{K}\right)) - M_1(R^{-1}\left(\frac{i-1}{K}\right))\right)} \tag{37}$$

because the stability condition

$$\lambda K\left(M_1\left(R^{-1}\left(\tfrac{i}{K}\right)\right) - M_1\left(R^{-1}\left(\tfrac{i-1}{K}\right)\right)\right) < \mu_{\pi(c)} \tag{38}$$

is satisfied for all $i = \beta_{\pi(c)-1}K+1, \ldots, \beta_{\pi(c)}K$ by construction.

4.2 On the Impact of Permutations

Figure 3 illustrates the role of permutations when $C = 2$, where the two possible permutations are $\pi(c) = c$ and $\varpi(c) = 3 - c$, and $\alpha_1 = \alpha_2 = \frac{1}{2}$. Fixed $y \in \mathcal{Y}$, the function R^π, represented by a continuous line, is an element of $\mathcal{R}_{y,\pi}$ that forces jobs of size $x \in [x_m, y_1]$ to be mapped to servers $1, \ldots, \beta_1 K$ and jobs of size $x \in [y_1, x_M]$ to be mapped to servers $\beta_1 K + 1, \ldots, K$. However, for the same y, we can also map jobs of size $x \in [x_m, y_1]$ to servers $\beta_1 K + 1, \ldots, K$ with profile $\alpha_1 + R^\pi_{|(x_m,y_1]}$ and jobs of size $x \in [y_1, x_M]$ to servers $1, \ldots, \beta_1 K$ with profile $R^\pi_{|(y_1,x_M]} - \alpha_1$. This constructs a SITA policy in $\mathcal{R}_{y,\varpi}$, represented in Figure 3 by a dashed line. To take into account all these different allocations given y, it is clear that one should handle all the possible mappings between the server types and the intervals $[y_{c-1}, y_c]$ for all c.

The construction above ends up with two different allocations with, in principle, different mean waiting times. This is also true if y depends on the particular chosen permutation, as the following remark states.

1. We assume that $\beta_c K \in \mathbb{Z}_+$ for all c. When taking limits as $K \to \infty$, for simplicity we will implicitly consider subsequences such that $\beta_c K \in \mathbb{Z}_+$ for all c, which exist because the β_c's are rational numbers.

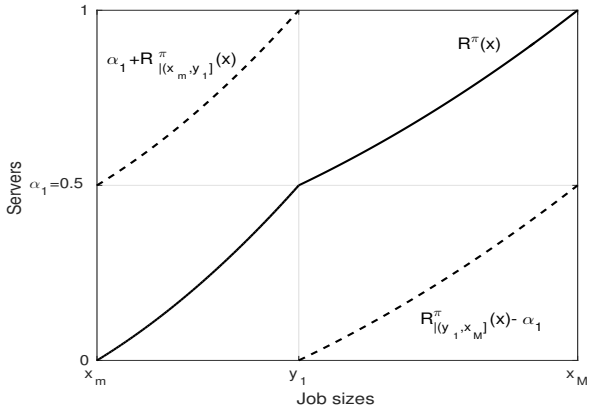

Fig. 3. Profiles of two SITA policies: $R_\pi \in \mathcal{R}_{y,\pi}$ (continuous line) and $R_\varpi \in \mathcal{R}_{y,\varpi}$ (dashed line), where $\pi(c) = c$ and $\varpi(c) = 3 - c$, respectively.

Remark 2. *The optimal mean waiting time achieved with one permutation is generally different from the optimal mean waiting time achieved with another permutation.*

Assuming two servers such that $\mu_1 \leq \mu_2$, Section 4 of [16] presents numerical results suggesting that the optimal mean waiting time achieved with the descending permutation, i.e., $\pi(c) = 3 - c$, is always less than the optimal mean waiting time achieved with the ascending permutation, i.e., $\pi(c) = c$, provided that X follows a Bounded-Pareto distribution with shape parameter $t > 1$; this property has been proven in [24, Theorem 6.6]. Contrariwise, if X is Weibull distributed, then the ascending permutation works better than the descending one, and if X follows a lognormal distribution, then both permutations do not produce significantly different results; these additional properties are conjectured in [16].

This highlights that *the optimal permutation strongly depends on the job size distribution*, at least when K is finite. Since the cardinality of the space of permutations is $C!$, an exhaustive search is clearly intractable, unless C is small.

Remark 3. *We anticipate here a property shown in the following: when $K \to \infty$, the optimal mean waiting time is* insensitive *to server-type permutations, for any job size distribution.*

4.3 Asymptotic Optimality Results

Let

$$\varrho_{c,a,b} \stackrel{\text{def}}{=} \frac{\lambda}{\alpha_c \mu_c} \int_a^b x f(x) \mathrm{d}x, \tag{39}$$

interpreted as the aggregate load at all type-c servers provided that they only accept jobs of size in $[a, b]$.

The next proposition provides a combinatorial optimization framework for computing an asymptotically optimal SITA policy; we recall that π is a permutation over $\{1, \ldots, C\}$.

Proposition 1.

$$\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W_K(R) = \min_{y\in\mathcal{Y},\pi} \sum_{c=1}^{C} \frac{\alpha_{\pi(c)}}{2\lambda} \frac{\varrho^2_{\pi(c),y_{c-1},y_c}}{1 - \varrho_{\pi(c),y_{c-1},y_c}}. \tag{40}$$

Proof. By construction of the functions in $\mathcal{R}_{y,\pi}$, we first notice that

$$\begin{aligned}
&\inf_{R\in\mathcal{R}} W_K(R) \\
&= \inf_{y\in\mathcal{Y},\pi} \inf_{R\in\mathcal{R}_{y,\pi}} W_K(R) \\
&= \inf_{y\in\mathcal{Y},\pi} \sum_{c=1}^{C} \inf_{R\in\mathcal{R}^{\pi(c)}_{[y_{c-1},y_c]}} \sum_{i=\beta_{\pi(c)-1}K+1}^{\beta_{\pi(c)}K} \frac{\lambda}{2} \frac{1}{\mu^2_{\pi(c)}} \times \\
&\qquad \frac{K \prod_{j\in\{0,2\}} \left(M_j(R^{-1}\left(\frac{i}{K}\right)) - M_j(R^{-1}\left(\frac{i-1}{K}\right))\right)}{1 - \frac{\lambda}{\mu_{\pi(c)}} K \left(M_1(R^{-1}\left(\frac{i}{K}\right)) - M_1(R^{-1}\left(\frac{i-1}{K}\right))\right)}.
\end{aligned}$$

The second sum in the RHS of previous equation has the same structure of (12), and therefore the remainder of the proof adapts the same arguments used in the proof of Theorem 1.

Applying the same arguments used to prove (24), we obtain

$$\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W_K(R) \tag{41}$$

$$\geq \frac{\lambda}{2} \lim_{K\to\infty} \inf_{y\in\mathcal{Y},\pi} \sum_{c=1}^{C} \frac{\alpha_{\pi(c)}}{\mu^2_{\pi(c)}} \inf \frac{g^2(\beta_{\pi(c)})}{1 - \frac{\lambda}{\mu_{\pi(c)}} g(\beta_{\pi(c)})} \tag{42}$$

$$= \frac{\lambda}{2} \lim_{K\to\infty} \inf_{y\in\mathcal{Y},\pi} \sum_{c=1}^{C} \frac{\alpha_{\pi(c)}}{\mu^2_{\pi(c)}} \inf_{\substack{R\in\mathcal{R}^{\pi(c)}_{[y_{c-1},y_c]}: \\ g(x) \text{ is constant}}} \frac{g^2(\beta_{\pi(c)})}{1 - \frac{\lambda}{\mu_{\pi(c)}} g(\beta_{\pi(c)})} \tag{43}$$

$$= \frac{\lambda}{2} \inf_{y\in\mathcal{Y},\pi} \sum_{c=1}^{C} \frac{\alpha_{\pi(c)}}{\mu^2_{\pi(c)}} \frac{h^2_{\pi(c),y_{c-1},y_c}}{1 - \frac{\lambda}{\mu_{\pi(c)}} h_{\pi(c),y_{c-1},y_c}} \tag{44}$$

where

$$g(x) \stackrel{\text{def}}{=} \frac{\mathrm{d}}{\mathrm{d}x} M_1(R^{-1}(x)) = R^{-1}(x) f\left(R^{-1}(x)\right) \frac{\mathrm{d}R^{-1}(x)}{\mathrm{d}x}, \tag{45}$$

the first inf is taken over all $R \in \mathcal{R}^{\pi(c)}_{[y_{c-1},y_c]}$ such that

$$g\left(\frac{\beta_{\pi(c)-1}K + 1}{K}\right) = g\left(\frac{\beta_{\pi(c)-1}K + 2}{K}\right) = \cdots = g\left(\beta_{\pi(c)}\right),$$

and $h_{\pi(c),y_{c-1},y_c}$ is the unique solution of the boundary value problem

$$z f(z) z' = h_{\pi(c),y_{c-1},y_c} \tag{46a}$$

$$z(\beta_{\pi(c)-1}) = y_{c-1},\ z(\beta_{\pi(c)}) = y_c, \tag{46b}$$

and thus independent of K. Now, let R^{*-1} be the unique solution of the initial value problem

$$z f(z) z' = h_{\pi(c),y_{c-1},y_c} \tag{47a}$$

$$z(\beta_{\pi(c)-1}) = y_{c-1}. \tag{47b}$$

As similarly done in (25a), making a change of variable we notice that

$$\alpha_c h_{\pi(c),y_{c-1},y_c} = \int_{\beta_{\pi(c)-1}}^{\beta_{\pi(c)}} R^{*-1} f(R^{*-1}) dR^{*-1}(x) \tag{48a}$$

$$= \int_{R^{*-1}(\beta_{\pi(c)-1})}^{R^{*-1}(\beta_{\pi(c)})} x f(x) \mathrm{d}x = \int_{y_{c-1}}^{y_c} x f(x) \mathrm{d}x, \tag{48b}$$

and thus

$$\lambda h_{\pi(c),y_{c-1},y_c} = \mu_{\pi(c)} \varrho_{\pi(c),y_{c-1},y_c}. \quad (49)$$

We remark that the inequality sign in (41) can be replaced by an equality because $(R^{*-1}(i/K))_{i=\beta_{\pi(c)-1}K+1,\ldots,\beta_{\pi(c)}K}$ are the thresholds of R^*_K (when restricted over $[y_{c-1}, y_c]$). This can be shown as done in (27) and (28). Finally, the inf in (44) can be replaced by a minimum because of the continuity of $h_{\pi(c),y_{c-1},y_c}$ with respect to the initial condition $y \in \mathcal{Y}$ (see (48)), with $\mathcal{Y}$ compact. This and (49) give (40). □

By Theorem 1, we notice that

$$\frac{1}{2\lambda} \frac{\varrho^2_{\pi(c),y_{c-1},y_c}}{1-\varrho_{\pi(c),y_{c-1},y_c}} \quad (50)$$

is interpreted as the asymptotic mean waiting time achieved by SITA-E provided that job sizes belong to the interval $[y_{c-1}, y_c]$ and are mapped to servers $\beta_{\pi(c)-1}K + 1, \ldots, \beta_{\pi(c)}K$. Therefore, Proposition 1 shows that an asymptotically optimal SITA policy applies SITA-E to all servers of a given type; this property is actually shown in the proof of Theorem 2. However, such policy does not necessarily equalize the loads of all servers: servers of different types may have different loads. Below, we will show that SITA-E is no more asymptotically optimal in the case of heterogeneous servers.

In principle, the combinatorial optimization framework given in Proposition 1 remains intractable because the number of possible permutations is $C!$. What remains to understand is which permutation π and which vector y minimize the asymptotic mean waiting time.

The following result implies that all permutations become "equivalent" in the limit where $K \to \infty$. We recall that this does not hold true when K is finite; see Section 4.2.

Theorem 2. *For any pair of permutations π and ϖ over $\{1, \ldots, C\}$,*

$$\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W_K(R) = \min_{y\in\mathcal{Y}} \sum_{c=1}^{C} \frac{\alpha_{\pi(c)}}{2\lambda} \frac{\varrho^2_{\pi(c),y_{c-1},y_c}}{1-\varrho_{\pi(c),y_{c-1},y_c}} \quad (51)$$

$$= \min_{y\in\mathcal{Y}} \sum_{c=1}^{C} \frac{\alpha_{\varpi(c)}}{2\lambda} \frac{\varrho^2_{\varpi(c),y_{c-1},y_c}}{1-\varrho_{\varpi(c),y_{c-1},y_c}} \quad (52)$$

$$= \min_{v\in\mathbb{R}^C_+} \sum_{c=1}^{C} \frac{\lambda}{2\alpha_c\mu_c^2} \frac{v_c^2}{1-\frac{\lambda v_c}{\alpha_c\mu_c}} \quad (53)$$

$$\text{s.t.:} \sum_{c=1}^{C} v_c = H(x_M) - H(x_m) \quad (54)$$

$$\lambda v_c \le \alpha_c\mu_c, \quad \forall c \quad (55)$$

where $\varrho_{\pi(c),y_{c-1},y_c}$ and $\varrho_{\varpi(c),y_{c-1},y_c}$ are defined via (39)*, and $H(x) \stackrel{\text{def}}{=} \int x f(x)\mathrm{d}x$. Furthermore, if y^π is an optimizer of* (51)*, then*

$$v^*_{\pi(c)} = H(y^\pi_c) - H(y^\pi_{c-1}), \quad \forall c \quad (56)$$

where and v^* is the unique optimizer of the strictly convex optimization problem (53)-(55)*.*

Proof. First, we notice that

$$\varrho_{\pi(c),y_{c-1},y_c} = \frac{\lambda}{\alpha_{\pi(c)}\mu_{\pi(c)}} \int_{y_{c-1}}^{y_c} x f(x)\mathrm{d}x \quad (57)$$

$$= \frac{\lambda}{\alpha_{\pi(c)}\mu_{\pi(c)}} \left(H(y_c) - H(y_{c-1})\right). \quad (58)$$

Let W_π and W_ϖ denote the optimal objective function value of the minimizations in (51) and (52), respectively. Making a change of variable and using that $H(\cdot)$ is strictly increasing, we obtain

$$W_\pi = \min_{v\in\mathbb{R}^C_+} \sum_{c=1}^{C} \frac{\lambda}{2\alpha_{\pi(c)}\mu^2_{\pi(c)}} \frac{v_c^2}{1-\frac{\lambda v_c}{\alpha_{\pi(c)}\mu_{\pi(c)}}} \quad (59)$$

$$\text{s.t.:} \sum_{c=1}^{C} v_c = H(x_M) - H(x_m)$$

$$\lambda v_c \le \alpha_{\pi(c)}\mu_{\pi(c)}, \quad \forall c$$

i.e., a strictly convex optimization problem.

Let v^π be the (unique) minimizer of the last optimization problem.

For simplicity, in the remainder of the proof we assume that $\pi(i) = i$ and $\varpi(i) = C + 1 - i$, for all $i = 1, \ldots, C$. This is not a loss of generality because we do not require the μ_i's to follow a particular ordering *a priori*. We have

$$W_\pi = \sum_{c=1}^{C} \frac{\lambda}{2\alpha_{\varpi(c)}\mu^2_{\varpi(c)}} \frac{(v^\pi_{C-c+1})^2}{1-\frac{\lambda}{\mu_{\varpi(c)}} v^\pi_{C-c+1}}$$

$$\ge \sum_{c=1}^{C} \frac{\lambda}{2\alpha_{\varpi(c)}\mu^2_{\varpi(c)}} \frac{(v^\varpi_c)^2}{1-\frac{\lambda}{\mu_{\varpi(c)}} v^\varpi_c} = W_\varpi$$

Repeating this argument exchanging π and ϖ, we obtain $W_\pi = W_\varpi$. □

Theorem 2 implies that there exist exactly $C!$ asymptotically optimal SITA policies, one for each permutation. The policy associated to the generic permutation π can be iteratively constructed using (56) to find y^π; the uniqueness of v^* and the monotonicity of $H(\cdot)$ imply that (56) uniquely identifies y^π (once v^* is given). Then, the thresholds of type-$\pi(c)$ servers are the ones of SITA-E when the minimum and maximum job sizes are y^π_{c-1} and y^π_c, respectively.

Being v^* the unique solution of a strictly convex optimization problem, it can be efficiently computed in polynomial time by applying standard algorithms [31].

Remark 4. *The $C!$ asymptotically optimal SITA policies induce the same loads on servers (and the same mean waiting time) but this is achieved with* different *thresholds.*

Figure 4 illustrates the structure of the two asymptotically optimal SITA policies achieved with permutations $\pi(c) = c$ and $\varpi(c) = 3 - c$, $R^*_\pi(x)$ (continuous line) and $R^*_\varpi(x)$ (dashed line), respectively. It is assumed that $C = 2$, X uniformly distributed over [0.5,1.5] and $\lambda = \mu_1 = \mu_2/2 = 1$. The threshold points $y^\pi_1 = 0.919822$ and $y^\varpi_1 = 1.28605$ have been obtained by numerically solving the optimizations in (51). Though different, these two SITA policies induce the same vector of loads: with π and ϖ, the loads of type-1 servers are 0.149018 and the loads of type-2 servers are 0.175491. Both permutations induce an asymptotic mean waiting time equal to 0.63324.

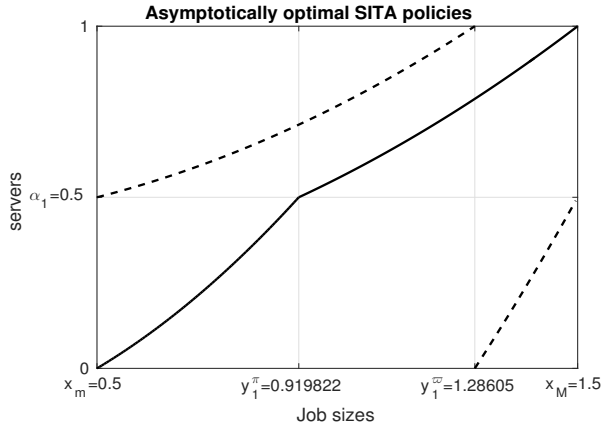

Fig. 4. Profiles of the SITA policies $R^*_\pi(x)$ (continuous line) and $R^*_\varpi(x)$ (dashed line).

4.4 Suboptimality of SITA-E

The KKT conditions of (53) imply that the following equations, obtained by differentiating the associated Lagrangian function, need to be satisfied by $v \in \mathbb{R}^C_+$ to be optimal (assuming that v is in the interior of the feasibility region):

$$\frac{\lambda^2 v_c}{\alpha_c \mu_c^2} \frac{2 - \frac{\lambda v_c}{\alpha_c \mu_c}}{\left(1 - \frac{\lambda v_c}{\alpha_c \mu_c}\right)^2} = \ell, \quad \forall c \tag{60}$$

where $\ell \in \mathbb{R}$ is the Lagrangian multiplier associated to (54). By contradiction, if $\frac{\lambda v_c}{\alpha_c \mu_c} = \gamma$ for all c, i.e., the loads are all equal, this means that necessarily $\gamma = \lambda \mathbb{E}[X] / \sum_c \alpha_c \mu_c$, and previous conditions boil down to

$$\frac{\lambda \gamma}{\mu_c} \frac{2 - \gamma}{(1-\gamma)^2} = \ell, \quad \forall c \tag{61}$$

which holds true if and only if μ_c is constant. This proves that *SITA-E is no more asymptotically optimal when servers are heterogeneous*.

By means of a competitive analysis, we now investigate how bad the performance of SITA-E can be. The following result says that the ratio between the mean waiting time achieved by SITA-E and the mean waiting time achieved by an asymptotically optimal SITA policy can be made arbitrarily large; in the case of heterogeneous servers, there are multiple SITA policies able to equalize server loads, as this depends on which mapping between servers and intervals is used, and we let R^*_K denote any of such policies.

Theorem 3. *Let the number of server types, C, be fixed. Then,*

$$\sup_{\lambda,\alpha,\mu,F} \frac{\lim_{K\to\infty} W(R^*_K)}{\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W(R)} \geq C. \tag{62}$$

Proof. Let $\rho \stackrel{\text{def}}{=} \lambda \mathbb{E}[X] / \sum_c \alpha_c \mu_c$, $W_E \stackrel{\text{def}}{=} \lim_{K\to\infty} W(R^*_K)$ and $W^* \stackrel{\text{def}}{=} \lim_{K\to\infty} \inf_{R\in\mathcal{R}} W(R)$. Using (53)-(55),

$$W^* = \frac{1}{2\lambda} \min_{v\in\mathbb{R}^C_+} \sum_{c=1}^{C} \frac{\lambda^2}{\alpha_c \mu_c^2} \frac{v_c^2}{1 - \frac{\lambda v_c}{\alpha_c \mu_c}} \tag{63a}$$

$$\text{s.t.:} \sum_{c=1}^{C} v_c = H(x_M) - H(x_m) \tag{63b}$$

$$\lambda v_c \leq \alpha_c \mu_c, \quad \forall c. \tag{63c}$$

Assume $\lambda = \alpha_c = \frac{1}{C}$ for all c, $\mu_1 = 1$, $\mu_2 = \cdots = \mu_C = \epsilon$, and X uniformly distributed over $[\frac{1}{2}, \frac{3}{2}]$, so that $\mathbb{E}[X] = 1$, $H(x_M) - H(x_m) = x_M^2 - x_m^2 = 1$ and $\rho = \frac{1}{(C-1)\epsilon+1}$. Within these conditions, we obtain

$$W_E = \frac{1}{2\lambda} \frac{\rho^2}{1-\rho} = \frac{C}{2} \frac{1}{((C-1)\epsilon+1)} \frac{1}{(C-1)\epsilon}, \tag{64}$$

where to obtain the first equality we have added in (53)-(55) the constraint that server loads are equal. Adding $v_2 = \cdots = v_C$ to the constraints of the optimization problem in (63), we also obtain

$$W^* \leq \frac{1}{2} \min_{\substack{v_1\in[0,1] \\ v_2\in[0,\epsilon]}} \frac{v_1^2}{1-v_1} + \frac{v_2^2}{\epsilon^2} \frac{C-1}{1-\frac{v_2}{\epsilon}} \tag{65}$$

$$\text{s.t.: } v_1 + (C-1)v_2 = 1$$

$$= \frac{1}{2} \min_{v_1} \frac{v_1^2}{1-v_1} + \frac{1}{\epsilon^2} \frac{C-1}{1-\frac{1}{\epsilon}\frac{1-v_1}{(C-1)}} \frac{(1-v_1)^2}{(C-1)^2} \tag{66}$$

$$\text{s.t.: } 1 - \epsilon(C-1) \leq v_1 \leq 1.$$

For $\phi \in (0,1)$, $v_1 = 1 - \phi\epsilon(C-1)$ is a feasible point for the last optimization, and therefore for such choice we obtain

$$W^* \leq \frac{1}{2} \frac{(1-\phi\epsilon(C-1))^2}{\phi\epsilon(C-1)} + \frac{1}{2} \frac{\phi^2(C-1)}{1-\phi} \tag{67}$$

$$= \frac{1}{2} \frac{(1-\phi\epsilon(C-1))^2 + \frac{\phi^3}{(1-\phi)}(C-1)^2\epsilon}{\phi\epsilon(C-1)}. \tag{68}$$

Combining this with (64), we obtain

$$\frac{W_E}{W^*} \geq \frac{C}{(C-1)\epsilon+1} \frac{1}{(C-1)\epsilon} \times \frac{\phi\epsilon(C-1)}{(1-\phi\epsilon(C-1))^2 + \frac{\phi^3}{(1-\phi)}(C-1)^2\epsilon}$$

and therefore, fixed C

$$\sup_{\lambda,\alpha,\mu,F} \frac{\lim_{K\to\infty} W(R^*_K)}{\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W(R)} \geq \sup_{\lambda,\alpha,\mu,F} \frac{W_E}{W^*}$$

$$\geq \sup_{\phi\in(0,1)} \lim_{\epsilon\to 0} \frac{C}{(C-1)\epsilon+1} \frac{\phi}{(1-\phi\epsilon(C-1))^2 + \frac{\phi^3(C-1)^2\epsilon}{(1-\phi)}}$$

$$= \sup_{\phi\in(0,1)} C\phi = C.$$

This concludes the proof. □

5 Numerical Results

We present numerical results to validate our theoretical findings and understand whether they can be applied in the prelimit where K is finite and relatively small. Unless otherwise specified, in the following we let X follow the Bounded-Pareto distribution with shape parameter $t = 1.5$ (see (29)) and such that $\mathbb{E}[X] = 1$ and $x_M/x_m = 10^4$ as in [16]. The Bounded-Pareto distribution with parameter $t \in (1,2)$ generates 'highly variable' job sizes and is common in empirical measurements of computing systems; see [14, Section 2.2].

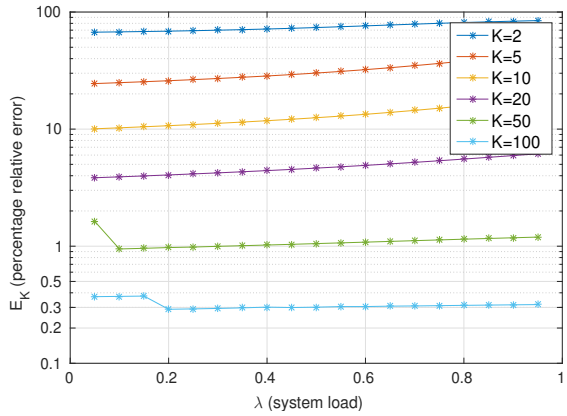

Fig. 5. Percentage relative error of the asymptotic approximation in Theorem 1 w.r.t. the optimal mean waiting time of the K-th system.

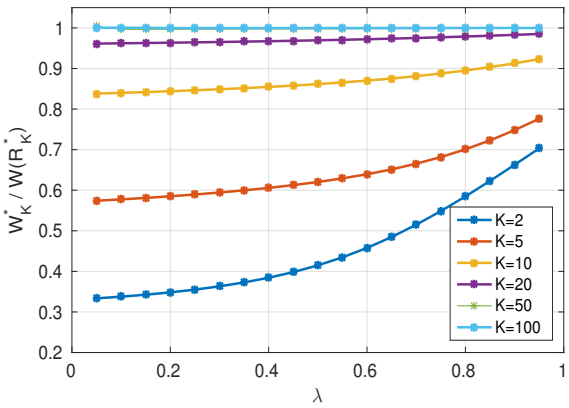

Fig. 6. Comparison between the mean waiting times of SITA-E and of the optimal SITA policy, respectively $W(R_K^*)$ and W_K^*.

5.1 Convergence Speed of Optimal Performance

We evaluate the asymptotic formula for the optimal mean waiting time, the RHS of (10), say W^*, as approximation of the optimal mean waiting time for the (finite) K-th system, say W_K^*. Though we are mainly interested in large systems, $K \geq 10^3$, we limit the presentation to moderate values of K because the exact computation of W_K^* is expensive and exhibits numerical instabilities. We assume homogeneous servers operating at speed $\mu = 1$.

On a logarithmic scale, Figure 5 plots the percentage relative error of W^* with respect to W_K^*, defined as

$$E_K \stackrel{\text{def}}{=} \frac{|W^* - W_K^*|}{W_K^*} \times 100\%, \tag{69}$$

by varying K and λ. It is quite surprising that the asymptotic approximation provides accurate results even when K is very small. When $K = 20$, the percentage relative error ranges between 3% and 6%. When $K = 100$, the asymptotic formula almost matches the optimal performance.

Similar results are obtained when comparing the mean waiting time achieved by SITA-E, $W(R_K^*)$, versus the one of the optimal SITA policy, W_K^*. Figure 6 plots the ratio $W_K^*/W(R_K^*)$ for increasing values of λ and K. It converges to 1 uniformly over λ fast.

5.2 Insensitivity to Server-type Permutations

We now focus on heterogeneous servers and investigate the insights given in Section 4.2 and Theorem 2 numerically. We fix $C = 2$, $\alpha_1 = \alpha_2 = 0.5$, $\mu_1 \leq 0.5$ and μ_2 such that $\mu_1 + \mu_2 = 1$. Furthermore, we let $\lambda = 0.4$ so that the system is 80% loaded for all $K \geq 2$. When $K = 2$, this setup is equivalent to the one used in Section 4 of [16].

Since $C = 2$, there are only two possible permutations: namely, the ascending mapping $A(i) = i$ and the descending mapping $D(i) = 3 - i$. Let $W_{K,\pi}^*$ be the 'optimal' mean waiting time achieved with permutation π; when $K > 2$, the computation of the optimal mean waiting time is numerically expensive and therefore we assume that SITA-E is applied among servers of the same type. Figure 7 illustrates the behavior of $W_{K,D}^*/W_{K,A}^*$ for increasing values of μ_1 and K. For all K and μ_1, we observe that the descending

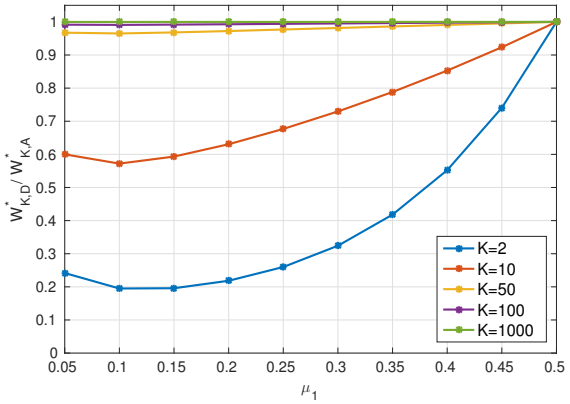

Fig. 7. Comparison between the ascending and descending server-type permutations: when K is large, both of them provide the optimal performance.

permutation always performs better than the ascending one, meaning that it is preferable to assign short (long) jobs to the fastest (slowest) servers. This is consistent with the results presented in [16], [24]. However, both permutations provide more and more similar results when K increases, almost identical already for $K = 50$. This is in agreement with Theorem 2 and also suggests that the convergence speed of the optimal SITA policy to $R^*(x)$ is "fast" even when servers are heterogeneous. Similar results can be obtained for other values of the shape parameter t and other distribution functions, which we omit.

5.3 Balancing vs Unbalancing

In Theorem 3, we have shown that SITA-E performs poorly in the worst-case scenario and when servers are heterogeneous, as the efficiency ratio

$$\mathcal{E} \stackrel{\text{def}}{=} \frac{\lim_{K\to\infty} W(R_K^*)}{\lim_{K\to\infty} \inf_{R\in\mathcal{R}} W(R)} \tag{70}$$

is not uniformly bounded over the set of model parameters. Since the argument used to prove this result is based on the analysis of a very unbalanced system in heavy-traffic, which may be pathological, we now investigate whether this holds

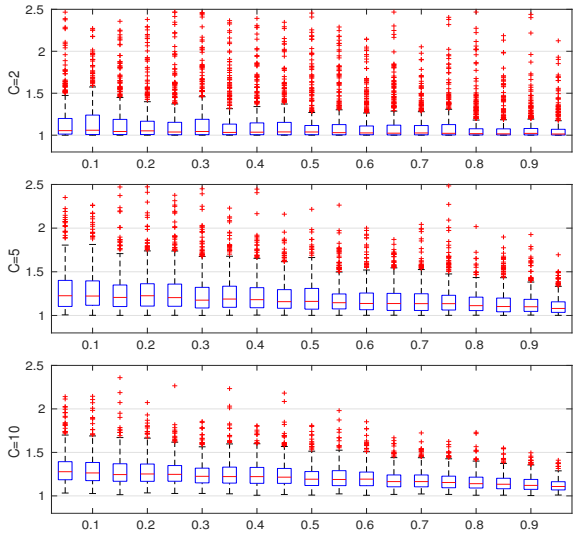

Fig. 8. `Boxplot`s for the efficiency ratio $\mathcal{E}$ by increasing the system load (x-axis) and for different values of C (number of server types).

true in the average-case scenario. Specifically, our objective is to understand whether the average efficiency ratio $\mathcal{E}$ scales as well (linearly) with C or other model parameters.

Towards this purpose, we have conducted an extensive numerical analysis based on thousands of randomly generated models. This analysis is enhanced by the simple form of the optimization in (53)-(55). The shape parameter t has been chosen in (1,2) and the μ_c's over $[0.01, 10]$, uniformly at random, together with the α_c's.

Using the Matlab's `boxplot` command, which indicates the median, the outliers ('+' signs), the 25th and 75th percentiles, Figure 8 illustrates the statistical behavior of $\mathcal{E}$ by varying the system load ρ and when $C \in \{2, 5, 10\}$. The data contained in each box refer to 1,000 random models, so that a total of $3 \times 19 \times 1000$ models have been generated.

In average, it turns out that the efficiency ratio is almost constant in C and surprisingly close to one. This makes the question 'balancing vs unbalancing' more intriguing because in practice a balanced allocation may be more robust to errors in the estimation of model parameters. We also notice that $\mathcal{E}$ is slightly decreasing in ρ. This is not surprising because our metric is the overall mean waiting time, i.e., a function that approaches zero when $\rho \to 0$. Under light load conditions, equalizing server loads is probably not the best choice because all traffic should be sent to the fastest server.

Similar results for the efficiency ratio are obtained even in the case of realistic parameters. The online repository [32], which contains workload logs collected from large scale parallel systems in production use in various places around the world, allows us to parameterize our model with respect to C, K, α, μ and X; to handle the data contained in the repository. We used data from the RICC and CEA data centers, which appear to be the most unbalanced in terms of processor speeds. This led to the following parameterization. For the RICC data center: $K = 9441$, $C = 3$, $\alpha = (0.9728, 0.0271, 0.0001)$, $\mu = (1.66, 2.93, 3)$,

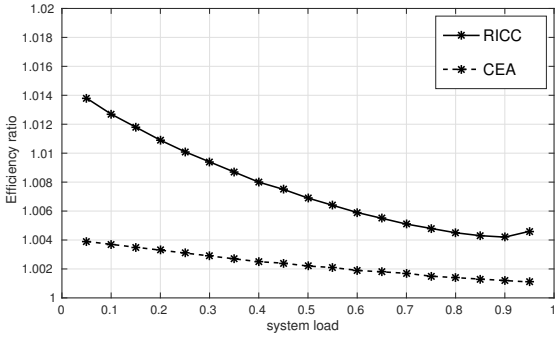

Fig. 9. Efficiency ratio with respect to realistic parameters.

$x_m = 1$, $x_M = 259200$, $\mathbb{E}[X] = 6.0907\text{e+}10$ and $\mathbb{E}[X^2] = 1.4546\text{e+}16$. For the CEA data center: $K = 18864$, $C = 3$, $\alpha = (0.6107, 0.1221, 0.2672)$, $\mu = (2.3, 2.66, 2.7)$, $x_m = 1$, $x_M = 79920$, $\mathbb{E}[X] = 96299030$ and $\mathbb{E}[X^2] = 2.1329\text{e+}12$. Figure 9 shows the behavior of $\mathcal{E}$ w.r.t. such parameters.

6 Conclusions

We have studied a class of size-based routing strategies in large-scale multi-server distributed queueing systems with highly variable workloads. When servers have identical processing speeds, we have shown that the minimum mean waiting time achievable by a Size-Interval Task Assignment (SITA) policy converges, in the limit where the system size grows to infinity, to the mean waiting time achieved by SITA-E, the SITA policy that equalizes server loads (Theorem 1). On the other hand, we have also shown that SITA-E can perform arbitrarily bad when servers are heterogeneous (Theorem 3). In this case, we have proven that the mean waiting time achieved by an asymptotically optimal policy does not depend on how job-size intervals are mapped to servers (Theorem 2), though optimal cutoffs do. We observe that there might be practical reasons to prefer one permutation rather than another: for instance, in some architectures it may be convenient to run long (short) jobs on more (less) reliable servers. Theorem 2 also allows for the *efficient* computation of the $C!$ asymptotically optimal policies as all of them are linked to the solution of a unique strictly convex optimization problem composed of C variables only. This problem may be further analyzed analytically and an explicit solution may be found in heavy traffic.

An extensive numerical analysis supports the claim that the insights found in the limit $K \to \infty$ hold true even for the original system where K is finite and that SITA-E performs almost optimally in the average-case scenario even when servers are heterogeneous.

In our analysis, we have assumed that the job size random variable X admits a density function. If X takes a finite number of values, one can always build an auxiliary random variable $\tilde{X}$, admitting a density function on a bounded support, arbitrarily "close" to the original X to obtain the framework considered in this work.

An other direction for further research consists in extending our asymptotic optimality results to *nested* SITA policies, which allow for size intervals that overlap. In homogeneous systems, it is known that the optimal nested SITA policy

provides the same mean waiting time of the optimal SITA policy with non-overlapping size intervals [16, Theorem 3] (and thus Theorem 1 also holds true with respect to this larger set of policies). However, this may not be the case in heterogeneous systems for finite K, though we may conjecture that an asymptotically optimal SITA policy with non-overlapping intervals is also asymptotically optimal within this larger set. This is left as future work.

ACKNOWLEDGMENTS

Research partially supported by the Marie Sklodowska-Curie grant agreement No 777778, the Basque Government, Spain, Consolidated Research Group Grant IT649-13, and the Spanish Ministry of Economy and Competitiveness project MTM2016-76329-R.

REFERENCES

[1] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, no. 1, pp. 181–189, 003 1977.
[2] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 12, no. 10, pp. 1094–1104, Oct. 2001.
[3] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Perform. Eval.*, vol. 68, no. 11, pp. 1056–1071, Nov. 2011.
[4] M. van der Boor, S. Borst, and J. van Leeuwaarden, "Load balancing in large-scale systems with multiple dispatchers," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.
[5] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky, "Redundancy-d: The power of d choices for redundancy," *Operations Research*, vol. 65, no. 4, pp. 1078–1094, 2017.
[6] W. Minnebo and B. V. Houdt, "A fair comparison of pull and push strategies in large distributed networks," *IEEE/ACM Transactions on Networking*, vol. 22, no. 3, pp. 996–1006, June 2014.
[7] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia, "Delay, memory, and messaging tradeoffs in distributed service systems," ser. SIGMETRICS '16. New York, NY, USA: ACM, 2016, pp. 1–12.
[8] P. Humblet, M. I. of Technology. Laboratory for Information, and D. Systems, *Determinism Minimizes Waiting Time in Queues*, ser. LIDS-P-. Laboratory for Information and Decision Systems, 1982.
[9] B. Hajek, "The proof of a folk theorem on queuing delay with applications to routing in networks," *J. ACM*, vol. 30, no. 4, pp. 834–851, Oct. 1983.
[10] Z. Liu and R. Righter, "Optimal load balancing on distributed homogeneous unreliable processors," *Journal of Operations Research*, vol. 46, no. 4, pp. 563–573, 1998.
[11] E. Altman, B. Gaujal, and A. Hordijk, "Multimodularity, convexity, and optimization properties." *Math. Oper. Res.*, vol. 25, no. 2, pp. 324–347, 2000.
[12] D. van der Laan, "Routing jobs to servers with deterministic service times." *Math. Oper. Res.*, vol. 30, no. 1, pp. 195–224, 2005.
[13] E. Altman, B. Gaujal, and A. Hordijk, "Balanced sequences and optimal routing," *J. ACM*, vol. 47, no. 4, pp. 752–775, Jul. 2000.
[14] M. Harchol-Balter, M. E. Crovella, and C. D. Murta, "On choosing a task assignment policy for a distributed server system," *Journal of Parallel and Distributed Computing*, vol. 59, no. 2, pp. 204 – 228, 1999.
[15] G. Ciardo, A. Riska, and E. Smirni, "Equiload: A load balancing policy for clustered web servers," *Perform. Eval.*, vol. 46, no. 2-3, pp. 101–124, Oct. 2001.
[16] H. Feng, V. Misra, and D. Rubenstein, "Optimal state-free, size-aware dispatching for heterogeneous m/g/-type systems," *Perform. Eval.*, vol. 62, no. 1-4, pp. 475–492, Oct. 2005.
[17] B. Schroeder and M. Harchol-Balter, "Evaluation of task assignment policies for supercomputing servers: The case for load unbalancing and fairness," *Cluster Computing*, vol. 7, no. 2, pp. 151–161, 2004.
[18] L. Cherkasova and M. Karlsson, "Scalable web server cluster design with workload-aware request distribution strategy ward," in *Proc. Third Int. W. on Advanced Issues of E-Commerce and Web-Based Information Systems. WECWIS 2001*, June 2001, pp. 212–221.
[19] K. Oida and K. Shinjo, "Characteristics of deterministic optimal routing for a simple traffic control problem," in *Proceedings of the IEEE International Performance Computing and Communications Conference, IPCCC 1999, Phoenix, Arizona, USA*, 1999, pp. 386–392.
[20] M. El-Taha and B. Maddah, "Allocation of service time in a multiserver system," *Management Science*, vol. 52, no. 4, pp. 623–637, 2006.
[21] M. Harchol-Balter, A. Scheller-Wolf, and A. R. Young, "Surprising results on task assignment in server farms with high-variability workloads," ser. SIGMETRICS '09. New York, NY, USA: ACM, 2009, pp. 287–298.
[22] M. E. Crovella, M. Harchol-Balter, and C. D. Murta, "Task assignment in a distributed system (extended abstract): Improving performance by unbalancing load," ser. SIGMETRICS '98/PERFORMANCE '98. New York, NY, USA: ACM, 1998, pp. 268–269.
[23] E. Bachmat and H. Sarfati, "Analysis of size interval task assignment policies," *SIGMETRICS Perform. Eval. Rev.*, vol. 36, no. 2, pp. 107–109, Aug. 2008.
[24] ——, "Analysis of sita policies," *Perform. Eval.*, vol. 67, no. 2, pp. 102–120, Feb. 2010.
[25] R. Vesilo, "Asymptotic analysis of load distribution for size-interval task allocation with bounded pareto job sizes," ser. ICPADS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 129–137.
[26] M. Harchol-balter and R. Vesilo, "To balance or unbalance load in size-interval task allocation," *Probab. Eng. Inf. Sci.*, vol. 24, no. 2, pp. 219–244, Apr. 2010.
[27] E. Bachmat and A. Natanzon, "Analysis of sita queues with many servers and spacetime geometry," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 3, pp. 92–94, Jan. 2012.
[28] S. Asmussen, *Applied Probability and Queues*. Wiley, 1987.
[29] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*. Wiley, 1976.
[30] J. Doncel, S. Aalto, and U. Ayesta, "Economies of scale in parallel-server systems," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.
[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.
[32] "Logs of real parallel workloads from production systems," http://www.cs.huji.ac.il/labs/parallel/workload/logs.html.

Jonatha Anselmi is a researcher at INRIA, France, since 2014. Prior to this, he was a full-time researcher at the Basque Center for Applied Mathematics, a postdoctoral research associate at INRIA and held visiting positions at IBM T.J. Watson and Caltech. He obtained a PhD in computer engineering from Politecnico di Milano, Italy, in 2009. His main research interests focus on the performance evaluation and optimization of distributed systems.

Josu Doncel is an assistant professor in the University of the Basque Country. He obtained from the same university the Industrial Engineering degree in 2007, the Mathematics degree in 2010 and, in 2011, the Master degree in Applied Maths and Statistics. He received in 2015 the PhD degree from Université de Toulouse (France). He has previously held research positions at LAAS-CNRS (France), INRIA Grenoble (France) and BCAM-Basque Center for Applied Mathematics (Spain) and teaching positions at ENSIMAG (France), INSA-Toulouse (France) and IUT-Blagnac (France).