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

Jonatha Anselmi

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

Jonatha Anselmi

**Abstract**—When dispatching jobs to parallel servers, or queues, the highly scalable round-robin (RR) scheme reduces the variance of interarrival times at all queues to a great extent but has no impact on the variances of service processes. Contrariwise, size-interval task assignment (SITA) routing has little impact on the variances of interarrival times but makes the service processes as deterministic as possible. In this paper, we unify both 'static' approaches to design a scalable load balancing framework able to control the variances of the arrival and service processes *jointly*. It turns out that the resulting combination significantly improves performance and is able to drive the mean job delay to zero in the large-system limit; it is known that this property is not achieved when both approaches are considered separately. Within realistic parameters, we show that the optimal number of size intervals that partition the support of the job size distribution is small with respect to the system size. This enhances the applicability of the proposed load balancing scheme at a large scale. In fact, we find that adding a little bit of information about job sizes to a dispatcher operating under RR improves performance a lot. Under the optimal scaling of size intervals and assuming highly variable job sizes, numerical simulations indicate that the proposed algorithm is competitive with the (less scalable) join-the-shortest-workload algorithm even when the system size grows large.

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

## 1 Introduction

A FUNDAMENTAL result from the queueing systems literature states that the mean (steady-state) waiting time experienced by customers, or *jobs*, joining a GI/GI/1 queue operating under the first-come first-served (FCFS) scheduling discipline, or any other discipline which does not affect the distribution of the number of jobs in the queue at any time [1], is upper bounded by

$$\frac{\lambda}{2}\frac{\sigma_A^2+\sigma_S^2}{1-\rho}, \tag{1}$$

where $\lambda > 0$ is the mean input traffic rate, $\rho \in [0,1)$ is the system load and $\sigma_A$ and $\sigma_S$ are the standard deviations of interarrival times and job sizes, respectively [2]. This bound also applies to the mean (steady-state) *workload* seen at the arrival times of a GI/GI/1 queue operating under any work-conserving scheduling discipline; the reader unfamiliar with queueing terminology may refer to Section 1.3.

When considering the problem of assigning incoming jobs to multiple parallel queues (each job needs to be routed to exactly one queue), then a natural objective would consist in designing a dispatching system able to minimize the numerator $\sigma_A^2+\sigma_S^2$ associated to each queue. Intuitively, this is equivalent to say that the arrival and service processes at each queue should be made as deterministic as possible. It is well known and not surprising that congestion phenomena are due to fluctuations in both processes [3].

On one extreme, the highly scalable round-robin (RR) scheme, which sends jobs to queues in a cyclic fashion, reduces the variance of interarrival times at all queues ($\sigma_A^2$) to a great extent but has no impact on the variances of service processes ($\sigma_S^2$) [4]. On another extreme, a form of size-based routing referred to in the literature as Size-Interval Task Assignment (SITA) routing, where each queue only accepts jobs of size belonging to a given interval, has little impact on the variances of interarrival times but can make the service processes as deterministic as possible [5]. Of all the dispatching strategies that have been proposed in the literature, both ideas are among the most effective ones for minimizing the variances of the arrival *or* service processes *separately*. However, each of both ideas does not take into account the benefits of the other. The main objective of this paper stands in combining these two static approaches to design a dispatching scheme able to minimize the variances of the arrival *and* service processes *jointly*, specifically $\sigma_A^2+\sigma_S^2$. Motivated by the huge sizes of real systems, our main goal is to understand whether it is possible to achieve *zero latency* in the large-system limit, that is the limiting regime where the network demand (average job arrival rate) grows to infinity proportionally with the capacity of processing resources. It is well known that this property is not obtained when RR and SITA are applied separately [6], [7].

### 1.1 SITA Routing and the "Zero-delay" Property

The development of load balancing schemes for parallel systems has a long history and several approaches continue to emerge in the literature, especially due to the constant introduction of new technologies. The celebrated join-the-shortest-queue (JSQ) and join-the-shortest-workload (JSW) algorithms, which respectively send each incoming job to the queue having the least number of jobs and workload (with ties broken randomly), are optimal, in a wide sense, under some assumptions [8], [9] but their applicability is often debated because of their little scalability, in part due to the high communication overhead between the queues and the dispatcher(s). We notice that both JSW and JSQ are *not* optimal within the assumptions that we will consider.

- *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. E-mail: jonatha.anselmi@inria.fr*

Given the size of real systems, much of the literature is currently investigating the problem of designing algorithms with a vanishing delay in the large-system limit described above but with a scalable communication overhead between the queues and the dispatcher(s) [7]. We refer to such algorithms as "asymptotically optimal". Load balancing schemes known to possess this property are the power-of-$d$-choices algorithm [10] when $d$, the number of queues to contact any time a new job arrives, grows to infinity with the system size [11], the power-of-$d$-choices algorithm *with memory* [12] provided that $d > \frac{1}{1-\lambda}$, the join-the-idle-queue (JIQ) algorithm [13], [14] and other similar pull-based approaches [7]. Essentially, these *dynamic* algorithms try to mimic the behavior of JSQ/JSW but with less information and for this reason it is to be expected that their performance can not be better than the one achieved by JSQ/JSW.

**Remark 1.** *The load balancing algorithm proposed in this paper is of a different nature: it does not mimic the behavior of JSQ/JSW, it will be scalable and asymptotically optimal, and numerical simulations show that under certain conditions it outperforms JSW even when the system size is large.*

SITA policies, defined in Section 2.3, are *static* dispatching rules and it is well known that they can outperform JSQ/JSW; see [5], [15], [16], [17], [18], [19], though the results in these references refer to "small" systems (e.g., with less than ten servers). The basic idea is that each server is assigned all jobs whose sizes belong to a distinct and continuous interval, and this can be achieved in several ways depending on the underlying architecture. For instance, each job may submit to the first level dispatcher an upper bound on its duration (as in, e.g., supercomputing systems) or the dispatcher may know a priori the identities of the servers hosting jobs of a given size (as in, e.g., web file transfers or content-aware load balancing [20], [21]), or the dispatcher may or may not be able to directly observe job sizes [16], [18], [22], [23]. The main reason for their performance benefits is commonly attributed to their ability to isolate small jobs from long ones, a phenomenon that does not necessarily occur with, e.g., JSW/JSQ. A comprehensive analytical comparison between SITA policies and JSW is shown in [19], where the authors exhibit several scenarios where one approach is better than the other. In such comparison, the system size is kept constant and the coefficient of variation of job sizes grows to infinity.

In contrast, in our approach we let the job size variability be fixed and let the system size grow to infinity. Within this framework, it has been shown that SITA policies are not asymptotically optimal in the sense above [6], see also [24], and thus eventually outperformed by the dynamic algorithms above in the large-system limit. The main objective of this work is to fill this gap. Towards this purpose, we enhance SITA policies with some RR routing to design a scalable dispatching system remaining competitive with JSW even for moderate to large system sizes.

## 1.2 Contribution

We view SITA and RR routings as two extreme points of a more general framework where each job is initially routed to one out of $d_K$ *virtual* second level dispatchers according to some SITA policy and each second level dispatcher applies RR to a set of (approximately) $K/d_K$ different queues, with $K$ being the overall number of queues (see Figure 1). Therefore, $d_K$ corresponds to the number of intervals that partition the support of the job size probability distribution and represents our control on the variances of both the service and arrival processes.

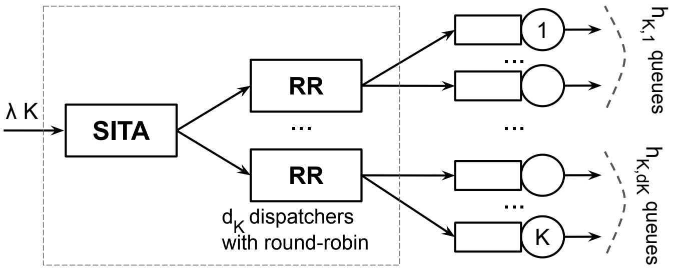

Fig. 1. Architecture of the proposed dispatching system.

If the SITA policy adopted by the first level dispatcher is SITA-E, which equalizes the loads of all queues, and $d_K$ grows to infinity sublinearly as $K \to \infty$, we rely on the upper bound (1) to show in a constructive manner that the mean waiting time or workload converges to zero in the large-system limit. This structural property is our main result (Theorem 1) and is proven under mild assumptions: essentially, interarrival times and job sizes are independent sequences of independent and identically distributed random variables, and queues operate under any work-conserving scheduling discipline. The choice of SITA-E is motivated by its optimality inside the set of SITA policies when $d_K = K$ and $K \to \infty$ [6]; see also [24].

Then, we use Theorem 1 to determine how to scale the control parameter $d_K$, the number of size intervals, to minimize the mean waiting time. If the support of the job size distribution is bounded, we show that the optimal scaling is asymptotic to $d_K^* = \gamma_1 K^{\frac{1}{3}}$, for some constant $\gamma_1$ that we explicit. Otherwise, if job sizes are Pareto distributed (with unbounded support), the optimal scaling is asymptotic to $d_K^* = \gamma_2 \sqrt{K}$, for some $\gamma_2$ that we explicit. This implies that the number of size intervals grows slowly with respect to the system size, which is convenient from a practical standpoint. In fact, we find that adding a little bit of information about job sizes to a dispatcher operating under RR improves performance a lot.

We also investigate the ratio between *i)* the minimum of the mean waiting times achieved by RR and the optimal SITA policy and *ii)* the minimum mean waiting time achievable with the proposed dispatching scheme, necessarily greater than or equal to one. When $K \to \infty$, Theorem 1 immediately implies that such ratio grows to infinity because RR and SITA policies are not asymptotically optimal. In Theorem 2, we show that such ratio can be arbitrarily large even when the system size $K$ is constant.

Finally, we have performed several simulations to assess the performance of the proposed algorithm with respect to the one achieved by others. When job sizes follow the bounded Pareto distribution with shape parameter $\alpha$ close to one, which is a case often found in empirical measurements of computing systems [25], [26], [27], our set of experiments indicates that it is always possible to find a set of $d_K$'s containing $d_K^*$ such that the proposed dispatching scheme outperforms RR, SITA and more surprisingly JSW.

### 1.3 Queueing Theoretic Terms

For quick reference and completeness, we provide the reader with a list of terms borrowed from queueing theory:

1) *service time* (of a job): the time it takes to process the job by a server operating at full speed with no interruptions; since all servers operate at constant speed 1, it is equivalent to its *job size*;
2) *waiting time* (of a job): the amount of time between the job arrival to the system and the beginning of its processing by some server;
3) *workload*: the total time the single server has to work to clear the system;
4) *work-conserving scheduling discipline*: a scheduling discipline that leaves the server idle only when there are no job to process;
5) *arrival/service process* (of queue $i$): the sequence of interarrival/service times of jobs at $i$.

### 1.4 Organization

This paper is organized as follows. Section 2 introduces the proposed load balancing algorithm together with modeling assumptions, SITA policies and performance metrics of interest. Section 3 is dedicated to the presentation of our main result (Theorem 1). The optimal scaling for the number of second level dispatchers $d_K$ is given in Section 4. Section 5 evaluates the synergies of the proposed combination of RR and SITA policies when the system size $K$ is fixed (Theorem 2). Section 6 is devoted to the presentation of numerical results. Finally, Section 7 draws the conclusions.

## 2 Dispatching Model

We consider the dispatching model indicated in Figure 1. Each job initially gets access to a first level dispatcher, which operates under a Size-Interval Task Assignment (SITA) policy (defined in Section 2.3), and then it is routed to one out of $d_K$ second level dispatchers. The second level dispatcher $i \in \{1, \dots, d_K\}$ applies round-robin (RR) on a set of $h_{K,i}$ queues, where each queue has an infinite buffer, processes jobs with unit rate and operates under any work-conserving scheduling discipline (different queues may have different disciplines). Therefore, the $n$-th job arriving to the second level dispatcher $i$ is sent to queue $(n \bmod h_{K,i}) + 1$ of its set. We assume that

$$\sum_{i=1}^{d_K} h_{K,i} = K \tag{2}$$

and that each queue can only receive jobs from a specific second level dispatcher. In particular, the second level dispatcher $i$ sends jobs to queues $H_{K,i-1}+1, \dots, H_{K,i}$ where $H_{K,i} \overset{\text{def}}{=} \sum_{j=1}^{i} h_{K,j}$ if $i > 0$ and $H_{K,0} \overset{\text{def}}{=} 0$. This implies that the total number of queues is $K$.

It will be clear from the definition of SITA policies that $d_K$ represents the number of intervals that partition the support of the job size distribution. In the extreme cases where $d_K = 1$ (no partitioning) and $d_K = K$, our dispatching model boils down to pure RR routing and pure SITA routing, respectively.

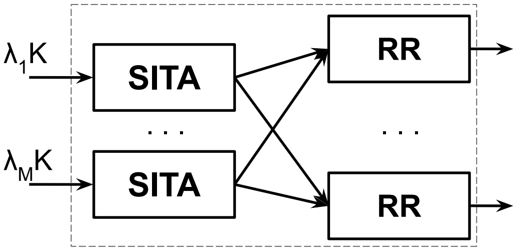

Fig. 2. A distributed implementation of the proposed dispatching system.

### 2.1 Multiple Dispatchers

Since real systems may be composed of hundreds of servers, it is often desirable that a load balancing algorithm is versatile enough to be implemented in a decentralized manner across multiple entry points; see, e.g., [13], [14], [28]. The dashed box in Figure 1 provides an abstraction for the structure of the proposed dispatching algorithm. From a practical standpoint, first and second level dispatchers may be all located on a centralized machine or distributed across multiple servers (one for each dispatcher). Specifically, in Figure 2 there are multiple first level dispatchers ($M$) and each of them adopts the same SITA policy and communicate with all the $d_K$ second level dispatchers (deployed on different machines). Under some assumptions, e.g., Poisson arrival processes at first level dispatchers, it can be shown analytically that this distributed implementation provides the same performance of its counterpart with only one first level dispatcher, provided that $\sum_{i=1}^{M} \lambda_i = \lambda$. This flexibility makes the proposed dispatching scheme more versatile and scalable than standard dynamic algorithms because a distributed implementation (with multiple dispatchers) of JSW or JSQ requires a non-negligible amount of control messages between the queues and the dispatchers *even when the dispatchers themselves can observe job sizes*. Within the same conditions, the proposed dispatching scheme requires no control message.

### 2.2 Stochastic Assumptions

We will consider a sequence of systems indexed by $K$, where the $K$-th system refers to the system with $K$ queues and $d_K$ second level dispatchers. All the random variables that follow belong to a fixed underlying probability space.

Let $(T_{K,n})_{n\in\mathbb{N}}$ and $(S_n)_{n\in\mathbb{N}}$ be independent sequences of independent and identically distributed random variables. These are the driving sequences of the $K$-th system in the sense that they represent the only source of randomness. Specifically, $T_{K,n} \in \mathbb{R}_+$ represents the interarrival time between the $(n-1)$-th and the $n$-th jobs joining the first level dispatcher and $S_n \in \mathbb{R}_+$ represents the size, or the service time, of the $n$-th job joining the first level dispatcher. The $T_{K,n}$'s have the same distribution of a random variable $T_K$, for which we assume that $\mathbb{E}[T_K] = \frac{1}{\lambda K}$. The $S_n$'s have the same distribution of a random variable $S$, which is assumed to have a Lipschitz continuous density function

$f(x)$ defined on $[x_m, x_M)$, $0 < x_m < x_M \leq \infty$, and such that $\frac{1}{xf(x)}$ is also Lipschitz. An important job size distribution that satisfies these assumptions is the *bounded Pareto distribution*, obtained when $x_M < \infty$ and

$$f(x) = \frac{C}{x^{\alpha+1}}, \quad C \stackrel{\text{def}}{=} \frac{\alpha x_m^\alpha}{1 - \left(\frac{x_m}{x_M}\right)^\alpha}. \tag{3}$$

It is well known that such distribution generates "highly variable" job sizes and that it is often found in empirical measurements of computing systems, especially when the "shape" parameter $\alpha$ is close to one [5], [25], [26], [27].

To apply the upper bound (1), we also require that $\rho \stackrel{\text{def}}{=} \lambda \mathbb{E}[S] < 1$, which is necessary to ensure stability.

## 2.3 SITA Policies

The first level dispatcher assigns jobs to second level dispatchers according to a SITA policy.

**Definition 1.** *A* SITA policy *when the number of second level dispatchers is $d_K$ is a cadlag, non-decreasing and surjective mapping $R : [x_m, x_M) \to \{\frac{1}{d_K}, \frac{2}{d_K}, \ldots, 1\}$ such that $R^{-1}(i/d_K)$ is an interval, for all $i \in \{1, \ldots, d_K\}$.*

Let $\mathcal{R}_{d_K}$ be the set of SITA policies for the $K$-th system when the number of second level dispatchers is $d_K$. The SITA policy $R \in \mathcal{R}_{d_K}$ is a piece-wise constant function with exactly $d_K - 1$ points of discontinuity, and the interpretation is that a controller adopting $R$ sends a job of size $x \in [x_m, x_M)$ to dispatcher $d_K R(x)$.

Given $R_{d_K} \in \mathcal{R}_{d_K}$, let $x_{K,i} \stackrel{\text{def}}{=} x_{K,i}(R_{d_K})$ denote its $i$-th discontinuity point, for all $i = 1, \ldots, d_K - 1$. Let also $x_{K,0} \stackrel{\text{def}}{=} x_m$, $x_{K,d_K} \stackrel{\text{def}}{=} x_M$. The points $(x_{K,i})_{i=0,\ldots,d_K}$ are said *thresholds*, or cutoffs, for assigning jobs to second level dispatchers. We notice that $(x_1, \ldots, x_{d_K-1}) \in \mathbb{R}^{d_K-1}$ such that $x_m < x_1 < \cdots < x_{d_K-1} < x_M$ uniquely constructs a SITA policy for the $K$-th system. Let also $S_{K,i} \stackrel{\text{def}}{=} S_{K,i}(R_{d_K})$ denote a random variable having the same distribution of the independent and identically distributed random variables representing the sizes of jobs joining the second level dispatcher $i$. Then, $S_{K,i} \in [x_{K,i-1}, x_{K,i})$ and after conditioning we obtain

$$\mathbb{E}[S_{K,i}^j] = \frac{\int_{x_{K,i-1}}^{x_{K,i}} x^j f(x)\,\mathrm{d}x}{\mathbb{P}(S \leq x_{K,i}) - \mathbb{P}(S \leq x_{K,i-1})}. \tag{4}$$

## 2.4 Performance Metrics

Given thresholds $(x_{K,i})_{i=0,\ldots,d_K}$, let $p_{K,i} \stackrel{\text{def}}{=} \mathbb{P}(S \leq x_{K,i}) - \mathbb{P}(S \leq x_{K,i-1})$, which corresponds to the probability of sending an incoming job to the second level dispatcher $i$.

Given that the arrival process at the first level dispatcher is renewal and that the thinning of a renewal process generates a renewal process (recall that the sequences $(T_{K,n})_{n\in\mathbb{N}}$ and $(S_n)_{n\in\mathbb{N}}$ are independent), the arrival process at each dispatcher $i$ is a renewal process with rate $\lambda K p_{K,i}$. Thus, let $(A_{K,i,n})_{n\in\mathbb{N}}$ be the sequence of i.i.d. random variables representing the interarrival times at the second level dispatcher $i$. By construction, we notice that

$$A_{K,i,n} =_{st} A_{K,i} \stackrel{\text{def}}{=} \sum_{m=1}^{Z_{K,i}} T_{K,m} \tag{5}$$

where $=_{st}$ denotes equality in distribution and $Z_{K,i} = \min\{n > 0 : S_n \in [x_{K,i-1}, x_{K,i})\}$ is a random variable independent of the $T_{K,m}$'s that follows a geometric distribution with parameter $p_{K,i}$. Therefore,

$$\begin{aligned}\mathrm{Var}(A_{K,i}) &= \mathrm{Var}(T_K)\mathbb{E}[Z_{K,1}] + \mathrm{Var}(Z_{K,1})(\mathbb{E}[T_K])^2 \\ &= \frac{\mathrm{Var}(T_K)}{p_{K,i}} + \frac{1 - p_{K,i}}{(\lambda K p_{K,i})^2}.\end{aligned}$$

Since RR is used by second level dispatchers, the arrival process of each queue controlled by dispatcher $i$ is a renewal process (possibly delayed) with rate $(\mathbb{E}[A_{K,i}]h_{K,i})^{-1} = \lambda K p_{K,i}/h_{K,i}$. In fact, the interarrival times at each queue controlled by $i$ have the same distribution of the random variable

$$A_{K,i}^{RR} \stackrel{\text{def}}{=} \sum_{n=1}^{h_{K,i}} A_{K,i,n} \tag{6}$$

where the $A_{K,i,j}$'s are independent and with the same distribution of $A_{K,i}$. By independence,

$$\mathrm{Var}(A_{K,i}^{RR}) = \mathrm{Var}(A_{K,i})h_{K,i}. \tag{7}$$

Let $W_K(d_K, R_{d_K}, h_K)$ denote the mean steady-state workload seen by jobs at their arrival times when the number of second level dispatchers is $d_K$, the first level dispatcher adopts the SITA policy $R_{d_K} \in \mathcal{R}_{d_K}$, and the number of queues controlled by each second level dispatcher is $h_K = (h_{K,1}, \ldots, h_{K,d_K})$. In the specific case where queues adopt the FCFS service discipline, it is known that $W_K(d_K, R_{d_K}, h_K)$ also corresponds to the mean steady-state waiting time experienced by jobs.

Since we have constructed a set of $K$ GI/GI/1 queues, we can use (1), specifically [2, Theorem 2], to bound from above $W_K(d_K, R_{d_K}, h_K)$. Provided that the necessary and sufficient stability condition

$$\frac{p_{K,i}\lambda K}{h_{K,i}}\mathbb{E}[S_{K,i}] < 1, \quad \forall i = 1, \ldots, d_K \tag{8}$$

is satisfied, we obtain

$$W_K(d_K, R_{d_K}, h_K) \tag{9a}$$

$$\leq \mathcal{W}_K(d_K, R_{d_K}, h_K) \tag{9b}$$

$$\stackrel{\text{def}}{=} \sum_{i=1}^{d_K} p_{K,i} \times \frac{p_{K,i}\lambda K}{2h_{K,i}} \frac{\mathrm{Var}(A_{K,i}^{RR}) + \mathrm{Var}(S_{K,i})}{1 - \frac{p_{K,i}\lambda K}{h_{K,i}}\mathbb{E}[S_{K,i}]}. \tag{9c}$$

## 2.5 Problem Statement

Given $d_K \in \mathbb{N}$, a SITA policy $R_{d_K} \in \mathcal{R}_{d_K}$ and $h_K \in \mathbb{R}_+^{d_K}$ such that (2) holds true, we refer to $(d_K, R_{d_K}, h_K)$ as a *dispatching scheme*.

**Definition 2.** *We say that a sequence of dispatching schemes $(d_K, R_{d_K}, h_K)_K$ is* asymptotically optimal *if*

$$\lim_{K\to\infty} W_K(d_K, R_{d_K}, h_K) = 0. \tag{10}$$

This notion of asymptotic optimality is thus related to the ideal situation where jobs can be always dispatched to empty queues (in the limit). Our main objective consists in constructing asymptotically optimal dispatching schemes.

### 2.6 Summary of Notation

For quick reference, we provide a summary of the most common symbols used to define our model and that will be used in the following:

- $K$: number of queues;
- $d_K$: number of intervals that partition the support of the job size distribution function;
- $f(x)$ and $F(x)$: job size density and distribution function, respectively;
- $S$: a random variable having distribution $F$;
- $h_K = (h_{K,1}, \dots, h_{K,d_K})$: a partition of the set of queues;
- $x_m$ and $x_M$: minimum and maximum job sizes;
- $\delta = x_M/x_m$: job variability ratio;
- $\lambda$: overall arrival rate of jobs;
- $\rho \overset{\text{def}}{=} \lambda \mathbb{E}[S]$;
- $\mathcal{R}_{d_K}$: set of SITA policies with $d_K$ intervals;
- $R$: generic SITA policy;
- $p_{K,i}(R)$: probability of dispatching a job to queue $i$ when policy $R$ is adopted;
- $(d_K, R_{d_K}, h_K)$: generic dispatching scheme;
- $W_K(d_K, R_{d_K}, h_K)$: mean steady-state waiting time achieved by dispatching scheme $(d_K, R_{d_K}, h_K)$;
- $\mathcal{W}_K(d_K, R_{d_K}, h_K)$: the upper bound on $W_K(d_K, R_{d_K}, h_K)$ given in (9);
- $\alpha$: shape parameter of the Pareto distribution.

## 3 Asymptotic Optimality

In this section we present our main result (Theorem 1). Towards this purpose, we first construct the set of dispatching schemes that we will show to be asymptotically optimal.

### 3.1 Balanced Subdivisions

For the distribution of queues among second level dispatchers, we introduce the concept of "balanced subdivision".

**Definition 3.** *Given a sequence $(d_K)_K$, a* balanced subdivision *is a triangular array $(h_{K,1}, \dots, h_{K,d_K})_{K\in\mathbb{N}}$ such that* (2) *holds true for all $K$, and*

$$h_{K,i} = \frac{K}{d_K} + \bar{h}_{K,i} \in \mathbb{N}, \quad \forall K, i = 1, \dots, d_K, \tag{11}$$

*where $\bar{h}_{K,i} \in \mathbb{R}$ and $\sup_K \sup_{i=1,\dots,d_K} |\bar{h}_{K,i}| < \infty$.*

A balanced subdivision ensures that the number of queues controlled by each second level dispatcher is sufficiently close to $K/d_K$. Balanced subdivisions exist; see Section 6.1.

### 3.2 SITA-E

Let $R^*_{d_K} \in \mathcal{R}_{d_K}$ denote the SITA policy for the $K$-th system that equalizes server loads, i.e., ensuring that

$$\frac{\lambda K p_{K,i} \mathbb{E}[S_{K,i}]}{h_{K,i}} = \rho, \quad \forall i = 1, \dots, d_K. \tag{12}$$

Following common queueing theory parlance [5], we refer to $R^*_{d_K}$ as SITA-E. By definition and using (4), $R^*_{d_K}$ is uniquely determined by the thresholds $(x^*_{K,i})_{i=0}^{d_K}$, given by the unique solution of the following system of equations

$$\int_{x^*_{K,i-1}}^{x^*_{K,i}} x f(x) \mathrm{d}x = \frac{h_{K,i}}{K} \mathbb{E}[S], \quad \forall i = 1, \dots, d_K. \tag{13}$$

These can be easily precomputed by iteration, starting for instance from $x^*_{K,1}$ and using that $x^*_{K,0} = x_m$. We also notice that their computation does not require the knowledge of $\lambda$.

The following lemma connects $R^*_{d_K}$ with some function $g$ independent of $K$.

**Lemma 1.** *Let $g : [0,1) \to [x_m, x_M)$ be the unique solution of the initial value problem*

$$z f(z) z' = \mathbb{E}[S] \tag{14a}$$

$$z(0) = x_m. \tag{14b}$$

*Then, $x^*_{K,i} = g\left(\frac{H_{K,i}}{K}\right)$, for all $i = 1, \dots, d_K - 1$.*

*Proof.* The uniqueness of solutions of (14) follows by the Picard–Lindelöf theorem because $\frac{1}{xf(x)}$ is Lipschitz continuous by assumption. Integrating both sides of (14a) and using a change of variable, we obtain

$$\frac{h_{K,i}}{K}\mathbb{E}[S] = \int_{\frac{H_{K,i-1}}{K}}^{\frac{H_{K,i}}{K}} \mathbb{E}[S] \mathrm{d}x = \int_{\frac{H_{K,i-1}}{K}}^{\frac{H_{K,i}}{K}} g(x) f(g(x)) \mathrm{d}g(x)$$

$$= \int_{g\left(\frac{H_{K,i-1}}{K}\right)}^{g\left(\frac{H_{K,i}}{K}\right)} x f(x) \mathrm{d}x \tag{15}$$

for all $i = 1, \dots, d_K$. Then, we notice that the choice $x^*_{K,i} = g\left(H_{K,i}/K\right)$ satisfies (13). □

### 3.3 Main Result

The following theorem is our main result and shows in a constructive manner that is indeed possible to obtain, within the proposed load balancing scheme, the zero-delay property discussed in the Introduction. Essentially, this structural property is achieved when the first level dispatcher applies SITA-E and when $(h_K)_K$ is a balanced subdivision.

**Definition 4.** *We write $f(K) \approx g(K)$ if $\lim_{K\to\infty} \frac{f(K)}{g(K)} = 1$.*

**Theorem 1.** *Let $(h_K)_K$ be a balanced subdivision, let $g(\cdot)$ be as in Lemma 1 and assume that*

$$d_K = o(K), \quad \lim_{K\to\infty} d_K = +\infty. \tag{16}$$

*If $x_M < \infty$, then*

$$\frac{2}{\lambda}(1-\rho)\,\mathcal{W}_K(d_K, R^*_{d_K}, h_K) \tag{17a}$$

$$\approx \frac{d_K - 1}{\lambda^2 K} + K\mathrm{Var}(T_K) + \frac{\mathbb{E}[S]^2}{12 d_K^2} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x. \tag{17b}$$

*If $x_M = \infty$ and $S$ is Pareto distributed with $\mathbb{E}[S^2] < \infty$, then*

$$\frac{2}{\lambda}(1-\rho)\,\mathcal{W}_K(d_K, R^*_{d_K}, h_K) \tag{18a}$$

$$\approx \frac{d_K - 1}{\lambda^2 K} + K\mathrm{Var}(T_K) + \frac{1}{d_K}\frac{\mathbb{E}[S]^2}{12(\alpha-1)^2}\frac{\pi^2}{6}. \tag{18b}$$

*Therefore, if in the scenarios above*

$$\lim_{K\to\infty} K\,\mathrm{Var}(T_K) = 0, \tag{19}$$

*then the sequence of dispatching schemes $(d_K, R^*_{d_K}, h_K)_K$ is asymptotically optimal.*

*Proof.* See Appendix A. □

Some comments are in order.

The assumption (16) rules out the cases where $d_K = K$ (pure SITA-E routing) and $d_K = 1$ (pure RR routing), and is necessary to ensure that $\mathcal{W}_K(d_K, R^*_{d_K}, h_K) \to 0$ as $K \to \infty$.

The RHS terms (17b) and (18b) are related to the variance of the arrival (first two terms) and service (third term) processes. Depending on the asymptotic behavior of $\mathrm{Var}(T_K)$ and $d_K$, they indicate whether it is the variance of the arrival or service process that eventually has a major influence on performance.

The adoption of SITA-E is justified by its optimality *inside the set of SITA policies* when $d_K = K$ and $K \to \infty$ [6], [24]. However, it is not the optimal policy in $\mathcal{R}_{d_K}$ and thus other SITA policies that improve the asymptotic estimate in (17) may exist. We do not investigate this in this paper. One advantage of SITA-E with respect to other SITA policies is that the identification of the cutoffs $x^*_{K,i}$ does not depend on the arrival process (see (13)).

In the case of a Poisson arrival process at the first level dispatcher, i.e., $T_K$ is exponentially distributed with rate $\lambda K$, (19) is clearly satisfied. This is also the case if for instance $T_K$ has a phase-type distribution where the size of the underlying transition matrix does not depend on $K$.

In the proof of Theorem 1, we show that our asymptotic approximations are related to the analysis of the sum in (48) as $K \to \infty$. When $x_M < \infty$, the key observation is to recognize that such sum is connected to a Riemann sum, regardless of the job size distribution. The case $x_M = \infty$ is different and does not seem possible to construct a 'general' asymptotic estimate unless a particular structure on the job size distribution is assumed.

We also observe that the system designer can actually control the parameter $d_K$. While letting $(h_K)_K$ remain any balanced subdivision, it is then natural to search for a $d_K$ that minimizes $\mathcal{W}_K(d_K, R^*_{d_K}, h_K)$, as it would impact the convergence speed of $W_K(d_K, R^*_{d_K}, h_K)$ to zero as well as the applicability of the proposed method itself at a large scale (it is clear that the smaller $d_K$ is, the better). Finding the optimal scaling for $d_K$ is the subject of Section 4.

In the RHS of (17), we notice that the integral

$$\mathcal{G} \overset{\mathrm{def}}{=} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x \tag{20}$$

does not seem to admit an explicit formula unless $g(x)$ takes a particular form. If job sizes follow the bounded Pareto distribution with parameter $\alpha$ (see (3)), then $g(x)$ must satisfy the ODE $z' = z^\alpha \mathbb{E}[S]/C$ with $z(0) = x_m$. Integrating both sides, assuming $\alpha \neq 1$ and recalling that $\mathbb{E}[S] = C\,(x_m^{1-\alpha} - x_M^{1-\alpha})/(\alpha - 1)$, one can first verify that

$$g(x) = \left(x_m^{1-\alpha} + x\left(x_M^{1-\alpha} - x_m^{1-\alpha}\right)\right)^{\frac{1}{1-\alpha}}. \tag{21}$$

Then,

$$\mathcal{G} = \frac{(x_M^{1-\alpha} - x_m^{1-\alpha})^2}{(1-\alpha)^2} \int_0^1 \left(x_m^{1-\alpha} + x\left(x_M^{1-\alpha} - x_m^{1-\alpha}\right)\right)^{-2} \mathrm{d}x \tag{22a}$$

$$= \frac{\left(x_M^{1-\alpha} - x_m^{1-\alpha}\right)\left(x_m^{\alpha-1} - x_M^{\alpha-1}\right)}{(1-\alpha)^2}. \tag{22b}$$

For the case where $\alpha = 1$, we notice that $\mathcal{G}$, $\mathbb{E}[S]$ and $g$ are all continuous in $\alpha$.

# 4 Optimal Tradeoff between SITA-E and RR

In this section, we use Theorem 1 to determine the optimal scaling for the number of second level dispatchers $d_K$, or equivalently the number of size intervals. Specifically, we are interested in studying the optimization problem

$$\min_{d_K \in \{1,\dots,K\}, h_K} \mathcal{W}_K(d_K, R^*_{d_K}, h_K) \tag{23}$$

with respect to a sequence of systems indexed by $K$. In principle, this is a difficult non-linear combinatorial optimization problem and for this reason we look for efficient and practical approximations.

For $K$ large enough, Theorem 1 ensures that all balanced subdivisions are asymptotically equivalent, in the sense that $\mathcal{W}(d_K, R^*_{d_K}, h_K) \approx \mathcal{W}(d_K, R^*_{d_K}, \hat{h}_K)$ as $K \to \infty$ for any two balanced subdivisions $(h_K)_K$ and $(\hat{h}_K)_K$, and optimal. Thus, with respect to a sequence of systems indexed by $K$, we approximate the optimization in (23) by

$$\min_{d_K \in \{1,\dots,K\}} \mathcal{W}(d_K, R^*_{d_K}, h_K) \tag{24}$$

where $(h_K)_K$ is any balanced subdivision.

## 4.1 Bounded Support

Let us assume that $x_M < \infty$. Since the second term in the RHS of (17) does not depend on $d_K$, we approximate an optimizer of (24) by some $d_K \in \{1,\dots,K\}$ that minimizes

$$\frac{d_K}{K} + \frac{\rho^2}{12 d_K^2} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x. \tag{25}$$

Since (25) is a strictly convex function in $d_K$, there exists a unique minimizer, say $d^*_K$. Assuming $d_K$ a continuous variable and imposing the derivative to zero, we obtain the condition

$$\frac{1}{K} = \frac{\rho^2}{6(d^*_K)^3} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x, \tag{26}$$

which gives

$$d^*_K = K^{\frac{1}{3}} \left(\frac{\rho^2}{6} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x\right)^{\frac{1}{3}}. \tag{27}$$

**Remark 2.** *It turns out that $d^*_K$ as given in* (27) *provides a very accurate approximation for the $d_K$ that solves the optimization in* (24) *even when $K$ is relatively small. This will be shown numerically in Section 6.*

Figure 3 illustrates the behavior of $d^*_K$ when $\lambda = 0.9$ and $S$ follows the bounded Pareto distribution with shape parameter $\alpha \in [\frac{1}{2}, \frac{3}{2}]$. We also assume $x_M = 10^5\, x_m$ and $\mathbb{E}[S] = 1$. We notice that $d^*_K$ is actually very small and is minimized when $\alpha = 1$, a value that is often found in empirical measurements of computing systems [5], [26], [27]. When $K = 100$ (respectively, $K = 10^4$) the optimal tradeoff between SITA and RR is obtained when job sizes are partitioned in only 12 (56) intervals. This choice of $d_K$

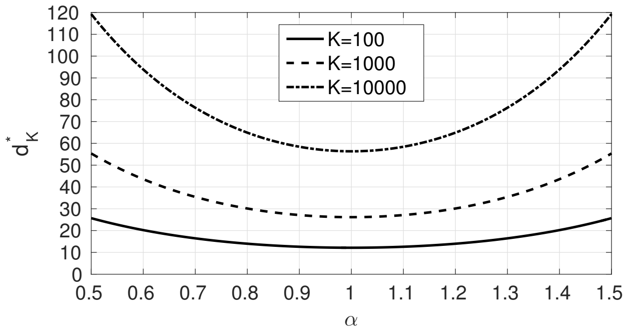

Fig. 3. Behavior of $d^*_K$ when $S$ follows the bounded Pareto distribution.

provides much better results than the cases where SITA-E and RR are applied separately, i.e., $d_K = K$ and $d_K = 1$ respectively (see Section 6). Interestingly, we also observe a symmetry around $\alpha = 1$, justified by (22) and consistent with the duality theory developed in [29].

### 4.2 Pareto Job Sizes

Let us assume that $x_M = \infty$ and that $S$ is Pareto distributed. Using (18) and proceeding as above, we approximate an optimizer of (24) by some $d_K \in \{1, \dots, K\}$ that minimizes the strictly convex function

$$\frac{d_K}{K} + \frac{1}{d_K} \frac{\rho^2}{12(\alpha-1)^2} \frac{\pi^2}{6}. \tag{28}$$

Imposing the derivative to zero, for the unique optimizer $d^*_K$ we obtain

$$d^*_K = \sqrt{K} \, \frac{\rho}{\sqrt{2}(\alpha-1)} \frac{\pi}{6}. \tag{29}$$

### 4.3 Optimal Performance and Convergence Speed

Substituting $d_K = d^*_K + \bar{d}_K$ in (17), where $(\bar{d}_K)_K$ is any uniformly bounded sequence such that $d^*_K + \bar{d}_K \in \mathbb{N}$ for all $K$, we obtain

$$\begin{aligned} \frac{2}{\lambda}(1-\rho)\, \mathcal{W}_K(d_K, R^*_{d_K}, h_K) \approx K \mathrm{Var}(T_K) - \frac{1}{\lambda^2 K} \\ + K^{-\frac{2}{3}} \frac{3}{2\lambda^2} \left( \frac{\rho^2}{6} \int_0^1 \left( \tfrac{g'(x)}{g(x)} \right)^2 \mathrm{d}x \right)^{\frac{1}{3}}, \end{aligned} \tag{30}$$

if $x_M < \infty$, and

$$\begin{aligned} \frac{2}{\lambda}(1-\rho)\, \mathcal{W}_K(d_K, R^*_{d_K}, h_K) \approx K \mathrm{Var}(T_K) - \frac{1}{\lambda^2 K} \\ + \frac{1}{\sqrt{K}} \frac{\sqrt{2}}{6} \frac{\pi \rho}{\lambda^2(\alpha-1)} \end{aligned} \tag{31}$$

if $x_M = \infty$ and $S$ is Pareto distributed. These formulas provide simple approximations for the minimum steady-state workload achievable with dispatching schemes of the form $(d_K, R^*_{d_K}, h_K)$.

In the case where $T_K$ follows a phase-type distribution having the rate of each phase proportional to $K$ and the size of the underlying transition matrix independent of $K$, then $\mathrm{Var}(T_K) = O(K^{-2})$ and $\mathcal{W}(R^*_{d_K}, d^*_K + \bar{d}_K, h_K)$ converges to zero with speeds $K^{2/3}$ ($x_M < \infty$) and $\sqrt{K}$ ($x_M = \infty$).

## 5 Synergies of the Combination

In this section, we analytically show that the performance gain of the proposed dispatching scheme $(d_K, R_{d_K}, h_K)$ with respect to both pure RR and SITA routings can be made arbitrarily large regardless of the system size $K$. Towards this purpose, we assume FCFS queues and, letting $R^{RR}$ be the only element of $\mathcal{R}_1$ (the degenerate map $R^{RR}(x) = 1$) and $\mathbf{1} = (1, \dots, 1)$, we fix $K$ and define the ratio

$$\mathcal{E}(K) \overset{\text{def}}{=} \frac{\min\left\{ W_K(1, R^{RR}, K), \inf\limits_{R \in \mathcal{R}_K} W_K(K, R, \mathbf{1}) \right\}}{\inf\limits_{d_K \in \{1,\dots,K\}, R \in \mathcal{R}_{d_K}, h_K} W_K(d_K, R, h_K)} \geq 1,$$

that is the ratio between the minimum of the mean steady-state waiting times achieved by RR and the optimal SITA policy for the $K$-th system when $d_K = K$ and the minimum mean steady-state waiting time achievable by the proposed dispatching scheme, which is necessarily no less than one.

Within the setting of Theorem 1, it is not difficult to show that the efficiency ratio $\mathcal{E}(K) \to \infty$ when $K \to \infty$. This holds true because both RR and SITA policies are not asymptotically optimal. The following result shows that $\mathcal{E}$ can grow unboundedly even in the case where the system size $K$ is kept constant. To achieve this, it will be sufficient to consider a scenario where the interarrival times $(T_{K,n})$ are constant and job sizes $(S_n)$ are highly variable.

**Theorem 2.** *Let $K \geq 3$ be fixed. Then, $\sup \mathcal{E} = +\infty$ where the* sup *is taken over the set of probability distributions of $S$ and $T_K$.*

*Proof.* See Appendix B. □

In Theorem 2, the case $K = 2$ is excluded because $(d_2, R_2, h_2)$ clearly boils down to either RR ($d_2 = 1$) or SITA ($d_2 = 2$), in which case $\mathcal{E} = 1$.

## 6 Performance and Accuracy Assessment

Assuming that queues operate under the FCFS scheduling discipline, in this section we present the results of several numerical simulations aimed at showing:

- how the average long-run waiting time achieved with $(d_K, R^*_{d_K}, h_K)_K$ compares with the ones achieved by join-the-shortest-workload (JSW), join-the-idle-queue (JIQ), RR and SITA-E;
- how the performance of $(d_K, R^*_{d_K}, h_K)_K$ varies with $d_K$;
- the accuracy of $d^*_K$ as in (27), our approximation for the optimal choice of $d_K$ developed in Section 4.

### 6.1 Regular Subdivisions

Our simulations have been performed under the assumption that $(h_K)$ is a *regular subdivision*.

**Definition 5.** *Given a sequence $(d_K)_K$, we say that the triangular array $((h_{K,1}, \dots, h_{K,d_K}))_{K \in \mathbb{N}}$ is a* regular subdivision *if* (2) *and* (11) *hold true with $|\bar{h}_{K,i}| \in [0, 1]$.*

Regular subdivisions can be constructed as follows. First, we choose a target second level dispatcher, say $i^* \in \{1, \dots, d_K\}$, and let $h_{K,i^*} = K - (d_K - 1)\lfloor \frac{K}{d_K} \rfloor$ and $h_{K,i} = \lfloor \frac{K}{d_K} \rfloor$ for all $i \in \{1, \dots, d_K\}$ such that $i \neq i^*$. At this point, if $h_{K,i^*} > \lceil \frac{K}{d_K} \rceil$, then necessarily $h_{K,i^*} - \lceil \frac{K}{d_K} \rceil \le d_K - 1$ and we distribute the $h_{K,i^*} - \lceil \frac{K}{d_K} \rceil$ queues in excess at $i^*$ among $h_{K,i^*} - \lceil \frac{K}{d_K} \rceil$ different second level dispatchers. This increases the number of queues controlled by each dispatcher $i \neq i^*$ at most by one.

### 6.2 Simulation Framework

We assume FCFS queues and that $(h_K)_K$ is a regular subdivision (see above) with $h_{K,i}$ nonincreasing in $i$ (this choice is for uniqueness and has a negligible impact). The arrival process at the first level dispatcher is Poisson and job sizes follow the bounded Pareto distribution with shape parameter $\alpha \in [\frac{1}{2}, \frac{3}{2}]$. Such values of $\alpha$, especially those in [1,1.3], are realistic [5, Section 2.2]; see also [25], [26], [27]. We assume $\rho = 0.9$ and that $(x_m, x_M)$ is chosen such that $\mathbb{E}[S] = 1$ and $x_M = 10^5 x_m$. With these assumptions, the thresholds of SITA-E are $x^*_{K,i} = g\left(\frac{H_{K,i}}{K}\right)$ with $g$ given by (21).

We have independently generated 400 sequences of the form $(T_{1,n}(\omega), S_n(\omega))$ for $n = 1, \dots, 10^8$, representing the interarrival times and job sizes of $10^8$ jobs for the base system where $K = 1$; this has been done using the C language function `srand`(seed), where seed $= 1, \dots, 400$. The 400 sequences associated to the $K$-th system, $K > 1$, have the form $(T_{K,n}(\omega), S_n(\omega))_{n=1,\dots,10^8}$ where $T_{K,n}(\omega) = KT_{1,n}(\omega)$. Within this coupling, both our algorithm and JSW are compared "$\omega$-per-$\omega$", i.e., within the same events.

With respect to each of the 400 sequences above and using Lindley's equation, we have computed the average waiting time of jobs starting from an empty system and without taking into account the first $4 \times 10^5$ jobs to eliminate some transitory effects that may bias the results (as in [5]). For the $K$-th system, we refer to such average as $W_K^{JSW}$ (respectively, $W_K^{SR(d_k)}$) if the load balancing algorithm used is JSW (the dispatching scheme $(d_K, R^*_{d_K}, h_K)$).

### 6.3 Comparison with Join-the-shortest-workload

Within the simulation setup described above, we assess the performance of $(d_K, R^*_{d_K}, h_K)$ with respect to the one achieved with JSW by measuring the ratio

$$\mathcal{R}_K\left(\frac{d_K}{K}\right) \overset{\text{def}}{=} \frac{W_K^{JSW}}{W_K^{SR(d_k)}}. \tag{32}$$

**Remark 3.** *The JSW algorithm is the ideal benchmark to test the performance of our dispatching scheme. However, this comparison is not completely fair because as discussed in Section 2.1 JSW is less scalable.*

Since $W_K^{JSW}$ does not vary with $d_K$, $\mathcal{R}_K$ also provides information about the performance gain of $(d_K, R^*_{d_K}, h_K)$ with respect to both SITA-E ($d_K = K$) and RR ($d_K = 1$).

Figure 4 illustrates the average and the standard deviation of $\mathcal{R}_K$ by increasing the number of second level dispatchers $d_K$ from 1 to $K$, for $K = 20, 50, 100$ and when $\alpha = 1$. The $x$-axis represents $d_K/K$ and indicates our approximation for the optimal scaling, $d^*_K/K$: $A = d^*_{20}/20 = 0.3549$, $B = d^*_{50}/50 = 0.1927$ and $C = d^*_{100}/100 = 0.1214$. Each point marked in both plots refers to 400 samples.

First, we notice that if $d_K/K > 0.05$, the proposed dispatching scheme always outperforms the best between pure RR and pure SITA routings. In addition, our approximation for the optimal $d_K$, i.e. $d^*_K$, is very close to the exact $d_K$ that maximizes $\mathcal{R}_K$ (equivalently, that minimizes $W_K^{SR(d_k)}$) and we observe that just a few size intervals help reducing $W_K^{SR(d_k)}$ a lot: when moving from $d_K = 1$ (RR) to $d^*_K$, the magnitude of $W_K^{SR(d_k)}$ always reduces remarkably. We also notice that the optimal tradeoff is achieved with a small number of size intervals as, e.g., $[d^*_{100}] = 12$.

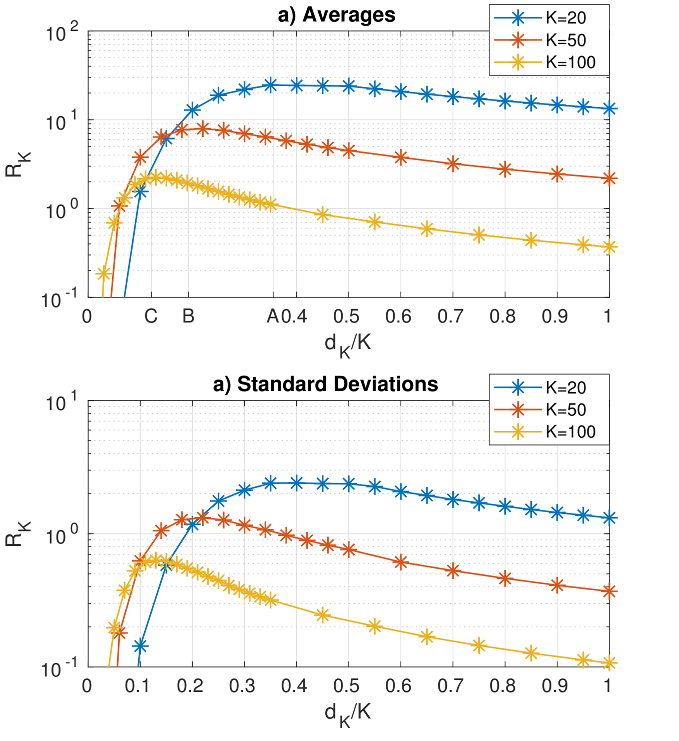

Fig. 4. Averages and standard deviations of $\mathcal{R}_K$ $(d_K/K)$ by increasing $d_K$, where $A = \frac{d^*_{20}}{20}$, $B = \frac{d^*_{50}}{50}$ and $C = \frac{d^*_{100}}{100}$. The maximum performance gain achievable by $(d_K, R^*_{d_K}, h_K)$ is indeed when $d_K$ is close to $d^*_K$, where $(d_K, R^*_{d_K}, h_K)$ outperforms JSW.

**Remark 4.** *The scenarios where RR and SITA routings are considered separately are eventually outperformed by JSW as $K$ increases; note that $\mathbb{E}[\mathcal{R}_{100}(1)] = 0.39$ and $\mathbb{E}[\mathcal{R}_{100}(\frac{1}{100})] \le 5 \times 10^{-5}$. This is to be expected because the mean steady-state waiting times achieved with both approaches remain bounded away from zero in the limit where $K \to \infty$ [6], [7].*

**Remark 5.** *It is always possible to find a set of $d_K$'s containing $d^*_K$ such that $\mathcal{R}_K > 1$, i.e., the proposed dispatching scheme $([d^*_K], R^*_{[d^*_K]}, h_K)$ outperforms JSW.*

We now illustrate the behavior of $\mathcal{R}_K\left(\frac{[d^*_K]}{K}\right)$ in two orthogonal scenarios as a function of the job size variability while keeping the system size large but constant. In the first, we increase the shape parameter $\alpha$ from 0.5 to 1.4 with step 0.1 while keeping fixed the ratio $\delta \overset{\text{def}}{=} \frac{x_M}{x_m} = 10^5$, and in the second we fix $\alpha = 1$ and let $\delta = 10^i$ for $i \in \{2, \dots, 7\}$. In both scenarios, we fix $K = 100$ and adjust parameters to ensure that $\mathbb{E}[S] = 1$; given the structure of the bounded Pareto distribution, $\mathbb{E}[S^2]$ clearly varies in $\delta$

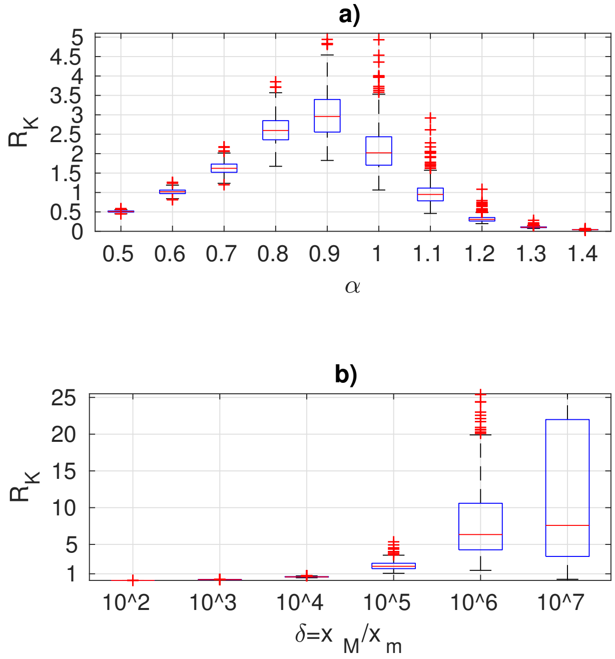

Fig. 5. Behavior of $\mathcal{R}_K\left(\frac{[d_K^*]}{K}\right)$ by varying *a)* the shape parameter $\alpha$ and *b)* the variability ratio $\delta$.

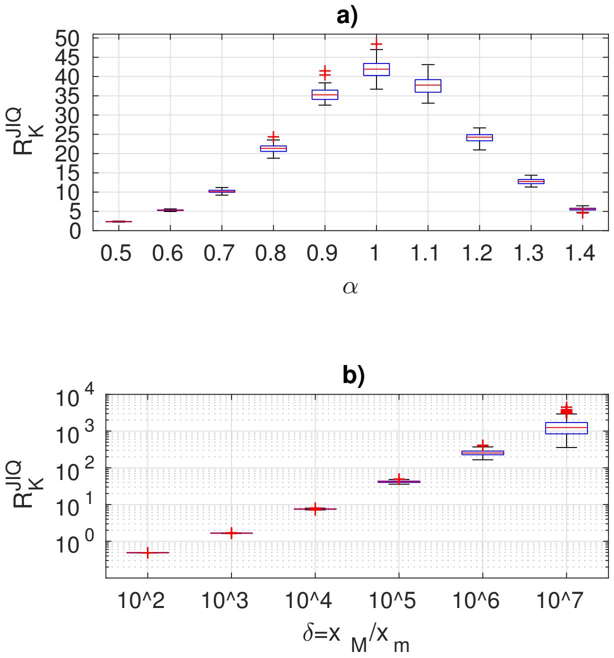

Fig. 6. Behavior of $\mathcal{R}_K^{JIQ}\left(\frac{[d_K^*]}{K}\right)$ by varying *a)* the shape parameter $\alpha$ and *b)* the variability ratio $\delta$.

and $\alpha$, and we notice that $\mathbb{E}[S^2] \to \infty$ as $\delta \to \infty$. The results of both scenarios are shown in Figure 5 by means of the Matlab's `boxplot` command, which indicates the median, the 25th and 75th percentiles (the edges of the box), the most extreme datapoints considered to be not outliers and the outliers (red '+' signs). Each box refers to 400 samples. Figure 5.a) shows that $\mathcal{R}_K$ is very sensitive to $\alpha$ and that the benefits of $(d_K^*, R_{[d_K^*]}^*, h_K)$ increase as $\alpha$ is around one, which is the case of practical interest. When $\alpha \in [0.6, 1.1]$, $(d_K^*, R_{[d_K^*]}^*, h_K)$ outperforms JSW but outside that interval JSW performs better. Figure 5.b) shows that in average $\mathcal{R}_K$ increases in $\delta$ significantly, with $\mathcal{R}_K > 1$ for all $\delta \geq 10^5$. This suggests that JSW is more sensitive to $\mathbb{E}[S^2]$ than $(d_K^*, R_{[d_K^*]}^*, h_K)$. We could not test for higher values of $\delta$ due to the cost of simulation: the evidently high variance appearing when $\delta = 10^7$ comes from the difficulty of simulating JSW, which is not able to isolate small jobs from long ones. Contrariwise, the simulation of $(d_K^*, R_{[d_K^*]}^*, h_K)$ is robust as when $\delta = 10^7$ we found that $\mathbb{E}[W_K^{SR([d_k^*])}] = 0.849$ with a small standard deviation equal to $0.0543$.

### 6.4 Comparison with Join-the-idle-queue

Within the simulation setup described above, we also assess the performance of $([d_K^*], R_{[d_K^*]}^*, h_K)$ with respect to the one achieved with the join-the-idle-queue (JIQ) algorithm [13]. Within JIQ, an incoming job is sent to an idle queue if an idle queue exists, otherwise to a random server. Since JIQ try to mimic the dynamics of JSW but with less information, one expects that our approach gives better performance gains than the ones presented in previous section.

Within the same settings used to obtain Figure 5, Figure 6 illustrates the behavior of $\mathcal{R}_K^{JIQ}\left(\frac{[d_K^*]}{K}\right) \overset{\text{def}}{=} \frac{W_K^{JIQ}}{W_K^{SR(d_k)}}$. It turns out that the resulting performance gains have qualitatively the same shape but they are significantly amplified. For instance, as a function of the shape parameter $\alpha$, the results in Figure 6.a) reveal a 10-fold improvement with respect to the results in Figure 5.a). Furthermore, Figure 6.b) shows that JIQ is much more sensitive to the job variability ratio $\delta$ than the proposed dispatching scheme $([d_K^*], R_{[d_K^*]}^*, h_K)$, which performs orders of magnitude better.

### 6.5 An Upper Bound on the Optimal Performance

The purpose of this section is to show that our asymptotic estimate for $\mathcal{W}_K([d_K^*], R_{[d_K^*]}^*, h_K)$, i.e., (30), actually provides an upper bound on the system performance. Towards this purpose, we evaluate by simulation the ratio $\mathcal{E}_K \overset{\text{def}}{=} \mathcal{W}_K^*/W_K^{SR([d_k^*])}$. Table 1 reports the behavior of $\mathcal{E}_K$ by changing the job size variability parameters $\alpha$ and $\delta = x_M/x_m$ (as done above) and shows that the simulated long-run average waiting time achieved with our dispatching scheme is smaller than our asymptotic estimate, as $\mathcal{E}_K > 1$. We claim that this insight is robust because Table 1 shows that standard deviations are significantly smaller than the corresponding averages. This suggests that the analytical formula (30) may be further used in the context of capacity dimensioning or admission control of computer systems where Quality-of-Service (QoS) guarantees need to be taken in to account; see, e.g., [30].

| $\delta = 10^5$ | | $\alpha = 1.0$ | |
|---|---|---|---|
| $\alpha$ | $\mathcal{E}_{100}$ | $\delta$ | $\mathcal{E}_{100}$ |
| 0.6 | 1.1482 ± 0.015 | $10^3$ | 1.4290 ± 0.010 |
| 0.8 | 1.1672 ± 0.026 | $10^4$ | 1.3623 ± 0.022 |
| 1.0 | 1.3462 ± 0.049 | $10^5$ | 1.3462 ± 0.049 |
| 1.2 | 1.3496 ± 0.090 | $10^6$ | 1.3919 ± 0.079 |
| 1.4 | 1.2534 ± 0.155 | $10^7$ | 1.4971 ± 0.092 |

TABLE 1
Averages ± standard deviations for $\mathcal{E}_K$ by increasing the job size variability parameters $\alpha$ and $\delta$.

## 7 Concluding Remarks

We have unified two 'dichotomic' load balancing schemes, namely Round-Robin (RR) and Size Interval Task Assignment (SITA), in a single dispatching algorithm. The synergies that come out from our combination allow one to *jointly* control the variances of both the arrival and service processes overcoming the limitations of both approaches when applied separately. We have proven that such scheme achieves zero latency in the large-system limit, shown that the performance gain with respect to pure RR and SITA routings can be arbitrarily large, and numerically shown that its performance is competitive with the join-the-shortest-workload algorithm, which as discussed in Section 2.1 does not possess the same scalability properties.

We also notice that the generic dispatching scheme $(d_K, R_{d_K}, h_K)$ makes job assignments with minimal computational requirements: $O(d_K)$ memory cells are needed to store the size thresholds, and for each size-$x$ job assignment one needs to search for the corresponding size interval to identify the right second level dispatcher, which can be done in $O(\log d_K)$ steps with a binary search.

With respect to realistic choices for the job size distribution, we have shown in Sections 4 and 6.3 that the optimal number of size intervals that partition the support of the job size distribution is 'small'. This enhances the applicability of the proposed load balancing scheme at a large scale because RR is known to be highly scalable and only a small number of cutoffs need to be estimated in practice.

The proposed dispatching scheme admits a bilevel interpretation where a first level dispatcher applies SITA to a set of second level dispatchers that in turn apply RR on non-overlapping sets of queues. If the roles of first and second level dispatchers were inverted, the zero-delay property in the large-system limit would not hold. This is intuitive because even if the arrival process at the first level dispatcher is deterministic, the second level dispatcher still randomizes over job sizes, making the arrival process at each queue renewal and non-deterministic.

## Appendix A
## Proof of Theorem 1

Unless otherwise specified, the hidden constants in the big-$O$ terms that follow will not depend on $i \in \{1, \dots, d_K\}$.

Let $g(x)$ be as in Lemma 1,

$$M_j(x) \overset{\text{def}}{=} \int x^j f(x) \mathrm{d}x, \tag{33}$$

and $\bar{H}_{K,i} \overset{\text{def}}{=} H_{K,i}/K$.

We treat the cases $x_M < \infty$ and $x_M = \infty$ separately. First, let us assume that $x_M < \infty$.

We notice that

$$M_j(x^*_{K,i}) - M_j(x^*_{K,i-1}) \tag{34a}$$
$$= M_j\left(g\left(\bar{H}_{K,i}\right)\right) - M_j\left(g\left(\bar{H}_{K,i-1}\right)\right) \tag{34b}$$
$$\begin{aligned} &= \tfrac{h_{K,i}}{K}\mathbb{E}[S]\, g^{j-1}\left(\bar{H}_{K,i}\right) \\ &\quad - \frac{1}{2}\left(\tfrac{h_{K,i}}{K}\right)^2 \mathbb{E}[S](j-1)g^{j-2}\left(\bar{H}_{K,i}\right) g'\left(\bar{H}_{K,i}\right) \\ &\quad + \frac{1}{3!}\left(\tfrac{h_{K,i}}{K}\right)^3 \mathbb{E}[S](j-1)g^{j-3}\left(\bar{H}_{K,i}\right) \\ &\qquad \times \Big((j-2)\left(g'\left(\bar{H}_{K,i}\right)\right)^2 + g\left(\bar{H}_{K,i}\right) g''\left(\bar{H}_{K,i}\right)\Big) \\ &\quad + O\left(\tfrac{h^4_{K,i}}{K^4}\right). \end{aligned} \tag{34c}$$

In (34b) we have used Lemma 1. In (34c) we have used a Taylor expansion of $M_j(g(\cdot))$ in $\bar{H}_{K,i}$ together with

$$\tfrac{\mathrm{d}}{\mathrm{d}x} M_j(g(x)) = (g(x))^j f(g(x))\, g'(x) = \mathbb{E}[S] g^{j-1}(x) \tag{35a}$$
$$\tfrac{\mathrm{d}^2}{\mathrm{d}x^2} M_j(g(x)) = \mathbb{E}[S](j-1)g^{j-2}(x)g'(x) \tag{35b}$$
$$\begin{aligned} \tfrac{\mathrm{d}^3}{\mathrm{d}x^3} M_j(g(x)) &= \mathbb{E}[S](j-1)g^{j-3}(x) \\ &\quad \times \left((j-2)(g'(x))^2 + g(x)g''(x)\right), \end{aligned} \tag{35c}$$

where the second equality in (35a) follows by the definition of $g(x)$ given in Lemma 1, and that $g$ is twice differentiable.

Now, substituting (13) in (9), we obtain

$$\begin{aligned} &\mathcal{W}_K(d_K, R^*_{d_K}, h_K) \\ &= \frac{\lambda}{2}\frac{1}{1-\rho}\sum_{i=1}^{d_K} \frac{K}{h_{K,i}} p^2_{K,i}\left(\mathrm{Var}(A^{RR}_{K,i}) + \mathrm{Var}(S_{K,i})\right) \end{aligned} \tag{36}$$

where, using (4) and (7),

$$p_{K,i} = M_0(x^*_{K,i}) - M_0(x^*_{K,i-1}) \tag{37}$$
$$\mathrm{Var}(A^{RR}_{K,i}) = h_{K,i}\left(\frac{\mathrm{Var}(T_K) - \frac{1}{\lambda^2 K^2}}{p_{K,i}} + \frac{1}{(\lambda K p_{K,i})^2}\right) \tag{38}$$
$$\mathrm{Var}(S_{K,i}) = \frac{M_2(x^*_{K,i}) - M_2(x^*_{K,i-1})}{p_{K,i}} \tag{39}$$
$$- \frac{(M_1(x^*_{K,i}) - M_1(x^*_{K,i-1}))^2}{p^2_{K,i}}. \tag{40}$$

For the variances of the interarrival times, we obtain

$$\begin{aligned} \sum_{i=1}^{d_K} \frac{K p^2_{K,i}}{h_{K,i}} \mathrm{Var}(A^{RR}_{K,i}) &= \sum_{i=1}^{d_K} K p_{K,i}\mathrm{Var}(T_K) - \frac{p_{K,i}}{\lambda^2 K} + \frac{1}{\lambda^2 K} \\ &= \frac{d_K}{\lambda^2 K} + K\mathrm{Var}(T_K) - \frac{1}{\lambda^2 K} \end{aligned} \tag{41a}$$

and for the variances of the service processes, we obtain

$$\begin{aligned} &p^2_{K,i}\mathrm{Var}(S_{K,i}) = -\left(M_1(x^*_{K,i}) - M_1(x^*_{K,i-1})\right)^2 + \\ &\left(M_0(x^*_{K,i}) - M_0(x^*_{K,i-1})\right)\left(M_2(x^*_{K,i}) - M_2(x^*_{K,i-1})\right) \end{aligned} \tag{42}$$

where, using (34) and that $(h_K)$ is a balanced subdivision,

$$\begin{aligned} (M_1(x^*_{K,i}) - M_1(x^*_{K,i-1}))^2 &= \left(\tfrac{\mathbb{E}[S]h_{K,i}}{K} + O\left(\tfrac{1}{d^4_K}\right)\right)^2 \\ &= \tfrac{\mathbb{E}[S]^2 h^2_{K,i}}{K^2} + O\left(\tfrac{1}{d^5_K}\right) \end{aligned} \tag{43}$$

and

$$
\begin{aligned}
&\frac{1}{\mathbb{E}[S]^2} \prod_{j\in\{0,2\}} \left(M_j(x^*_{K,i}) - M_j(x^*_{K,i-1})\right)\\
&= \Bigg( \tfrac{h_{K,i}}{K} \tfrac{1}{g(\bar{H}_{K,i})} + \tfrac{1}{2}\left(\tfrac{h_{K,i}}{K}\right)^2 \tfrac{g'(\bar{H}_{K,i})}{g^2(\bar{H}_{K,i})}\\
&\quad - \tfrac{h^3_{K,i}}{3!K^3} \tfrac{g(\bar{H}_{K,i})g''(\bar{H}_{K,i}) - 2(g'(\bar{H}_{K,i}))^2}{g^3(\bar{H}_{K,i})} + O\left(\tfrac{1}{d_K^4}\right)\Bigg) \times\\
&\quad \Bigg( \tfrac{h_{K,i}}{K} g(\bar{H}_{K,i}) - \tfrac{1}{2}\left(\tfrac{h_{K,i}}{K}\right)^2 g'(\bar{H}_{K,i})\\
&\qquad\qquad + \tfrac{h^3_{K,i}}{3!K^3} g''(\bar{H}_{K,i}) + O\left(\tfrac{1}{d_K^4}\right)\Bigg)\\
&= \tfrac{h^2_{K,i}}{K^2} + \tfrac{1}{12}\left(\tfrac{g'(\bar{H}_{K,i})}{g(\bar{H}_{K,i})}\right)^2 \left(\tfrac{h_{K,i}}{K}\right)^4 + O\left(\tfrac{1}{d_K^5}\right). \qquad (44)
\end{aligned}
$$

Therefore, using (43) and (44) in (42) we obtain

$$
\begin{aligned}
&\sum_{i=1}^{d_K} \frac{K}{h_{K,i}} p^2_{K,i} \mathrm{Var}(S_{K,i}) && (45a)\\
&= \sum_{i=1}^{d_K} \tfrac{\mathbb{E}[S]^2}{12} \left(\tfrac{g'(\bar{H}_{K,i})}{g(\bar{H}_{K,i})}\right)^2 \left(\tfrac{h_{K,i}}{K}\right)^3 + O\left(\tfrac{K}{h_{K,i}} \tfrac{1}{d_K^5}\right) && (45b)\\
&= O\left(\tfrac{1}{d_K^3}\right) + \tfrac{\mathbb{E}[S]^2}{12} \left(\tfrac{1}{d_K} + O(\tfrac{1}{K})\right)^2 \sum_{i=1}^{d_K} \tfrac{h_{K,i}}{K} \left(\tfrac{g'(\bar{H}_{K,i})}{g(\bar{H}_{K,i})}\right)^2 && (45c)\\
&= O\left(\tfrac{1}{d_K^3}\right) + \left(\tfrac{\mathbb{E}[S]^2}{12 d_K^2} + O\left(\tfrac{1}{d_K K}\right)\right) \Bigg(\int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x\\
&\qquad + O\left(\tfrac{1}{d_K}\right)\Bigg) && (45d)\\
&= O\left(\tfrac{1}{d_K^3}\right) + \tfrac{\mathbb{E}[S]^2}{12 d_K^2} \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x + O\left(\tfrac{1}{K d_K}\right). && (45e)
\end{aligned}
$$

In (45d), we have used that $\left(\frac{g'(x)}{g(x)}\right)^2$ is differentiable and Lipschitz (and thus Riemann integrable), and a crude error bound for Riemann sums, i.e.,

$$
\begin{aligned}
\sum_{i=1}^{d_K} \frac{h_{K,i}}{K} \left(\tfrac{g'(\bar{H}_{K,i})}{g(\bar{H}_{K,i})}\right)^2 &\le \int_0^1 \left(\tfrac{g'(x)}{g(x)}\right)^2 \mathrm{d}x\\
&+ \frac{1}{2} \max_{x\in[0,1]} \left| \frac{\mathrm{d}}{\mathrm{d}x} \left(\tfrac{g'(x)}{g(x)}\right)^2 \right| \times \max_{i=1,\dots,d_K} \tfrac{h_{K,i}}{K}.
\end{aligned}
$$

The Lipschitz property holds true because using that $g(x)$ is increasing with $g(0) = x_m > 0$ (by definition) we obtain

$$
\begin{aligned}
&\frac{1}{2} \left| \frac{\mathrm{d}}{\mathrm{d}x} \left(\tfrac{g'(x)}{g(x)}\right)^2 \right| = \tfrac{g'(x)}{g^3(x)} |g''(x) g(x) - (g'(x))^2|\\
&\le \tfrac{L_g}{x_m^3} |g''(x)| x_M + \tfrac{L_g^3}{x_m^3}\\
&= \tfrac{L_g x_M}{x_m^3} \mathbb{E}[S] \frac{|g' f(g) - g f'(g) g'|}{g^2 f^2(g)} + \tfrac{L_g^3}{x_m^3}\\
&\le \tfrac{L_g^2 x_M}{x_m^5} \mathbb{E}[S] \frac{\max_{x\in[x_m,x_M]} f(x) + x_m L_f}{\min_{x\in[x_m,x_M]} f^2(x)} + \tfrac{L_g^3}{x_m^3} < \infty
\end{aligned}
$$

where $L_f$ and $L_g$ are the Lipschitz constants of $g$ and $f$, respectively. In the penultimate inequality, we have used that $\min_{x\in[x_m,x_M]} f^2(x) > 0$, which holds true because otherwise $\frac{1}{x f(x)}$ would not be Lipschitz.

Finally, combining (41a) and (45) in (36), we obtain (17) and the asymptotic optimality of $(d_K, R^*_{d_K}, h_K)$ follows by the scaling assumptions on Var$(T_K)$ and $d_K$, i.e. (19) and (16).

We now assume that $x_M = \infty$ and that $S$ is Pareto distributed with $\mathbb{E}[S^2] < \infty$. We recall that $\mathbb{E}[S^2] < \infty$ if and only if $\alpha > 2$. In this case,

$$
M_j(x) = \int x^j \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \mathrm{d}x = \alpha x_m^\alpha \frac{x^{j-\alpha}}{j-\alpha} \qquad (46)
$$

and using that $\mathbb{E}[S] = \frac{\alpha x_m}{\alpha - 1}$, $g$ satisfies $g^{-\alpha} g' = x_m^{1-\alpha}/(\alpha - 1)$ with $g(0) = x_m$. Integrating both sides, we obtain

$$
g(x) = x_m (1-x)^{\frac{1}{1-\alpha}}. \qquad (47)
$$

What remains to show is the limit behavior of

$$
\sum_{i=1}^{d_K} \frac{K}{h_{K,i}} p^2_{K,i} \mathrm{Var}(S_{K,i}). \qquad (48)
$$

Here, we notice that the Taylor's expansion (34) fails at $i = d_K$ because in this case $\lim_{x\uparrow 1} g(x) = +\infty$, and therefore that the Riemann sum (45c) may diverge; indeed, by letting $x_M \to \infty$ in (22), we notice that such sum diverges. In the following, we show that such sum converges if scaled by $d_K/K$, though this observation depends on the particular structure of the Pareto distribution.

Now,

$$
\begin{aligned}
&\sum_{i=1}^{d_K-1} \frac{K}{h_{K,i}} p^2_{K,i} \mathrm{Var}(S_{K,i}) && (49a)\\
&= \sum_{i=1}^{d_K-1} \tfrac{\mathbb{E}[S]^2}{12} \left(\tfrac{g'(\bar{H}_{K,i})}{g(\bar{H}_{K,i})}\right)^2 \left(\tfrac{h_{K,i}}{K}\right)^3 + O\left(\tfrac{K}{h_{K,i}} \tfrac{1}{d_K^5}\right) && (49b)\\
&= O\left(\tfrac{1}{d_K^3}\right) + \sum_{i=1}^{d_K-1} \tfrac{\mathbb{E}[S]^2}{12(\alpha-1)^2} \tfrac{1}{(1-\bar{H}_{K,i})^2} \left(\tfrac{h_{K,i}}{K}\right)^3, && (49c)
\end{aligned}
$$

In (49b), we have used (43) and (44) in (42); in (49c), we have used (46) and (47).

Letting $\bar{h} \overset{\text{def}}{=} \sup_K \sup_i |\bar{h}_{K,i}|$, for all $K$ sufficiently large

$$
\begin{aligned}
&\sum_{i=1}^{d_K-1} \tfrac{1}{(1-\bar{H}_{K,i})^2} \left(\tfrac{h_{K,i}}{K}\right)^2 = \sum_{i=1}^{d_K-1} \left(\frac{h_{K,i}}{K - H_{K,i}}\right)^2\\
&= \sum_{i=1}^{d_K-1} \left(\frac{h_{K,i}}{\sum_{j=i+1}^{d_K} h_{K,j}}\right)^2 \le \sum_{i=1}^{d_K-1} \left(\frac{\frac{K}{d_K} + \bar{h}}{(d_K - i)\left(\frac{K}{d_K} - \bar{h}\right)}\right)^2\\
&= \left(\frac{\frac{K}{d_K} + \bar{h}}{\frac{K}{d_K} - \bar{h}}\right)^2 \sum_{i=1}^{d_K-1} \frac{1}{(d_K - i)^2} \approx \sum_{i=1}^{d_K-1} \frac{1}{i^2} \approx \frac{\pi^2}{6}.
\end{aligned}
$$

Replacing $\bar{h}$ by $-\bar{h}$ in the last inequality, we obtain

$$
\sum_{i=1}^{d_K-1} \frac{K}{h_{K,i}} p^2_{K,i} \mathrm{Var}(S_{K,i}) \approx \frac{1}{d_K} \frac{\mathbb{E}[S]^2}{12(\alpha-1)^2} \frac{\pi^2}{6}.
$$

To prove (18), it remains to show that the last term in (48) converges to zero (here, we use that $\mathbb{E}[S^2] < \infty$). Since

$$
p^2_{K,d_K} \mathrm{Var}(S_{K,d_K}) \le p^2_{K,d_K} \mathbb{E}[S^2_{K,d_K}]
$$

$$= \prod_{j\in\{0,2\}} \left(M_j(x^*_{K,d_K}) - M_j(x^*_{K,d_K-1})\right)$$

$$= (1-\bar{H}_{K,d_K-1})^{\frac{\alpha}{\alpha-1}} \times \alpha \frac{x_m^2(1-\bar{H}_{K,d_K-1})^{\frac{\alpha-2}{\alpha-1}}}{\alpha-2}$$

$$= \left(\tfrac{h_{K,d_K}}{K}\right)^2 \alpha \tfrac{x_m^2}{\alpha-2} \approx \tfrac{1}{d_K^2} \alpha \tfrac{x_m^2}{\alpha-2}$$

we obtain $\frac{K}{h_{K,d_K}} p^2_{K,d_K} \mathrm{Var}(S_{K,d_K}) = O(\frac{1}{d_K}) \to 0$ as desired.

# Appendix B Proof of Theorem 2

Let $W(A,B)$ denote the mean steady-state waiting time experienced by jobs in a GI/GI/1 queue where interarrival and service times are equal in distribution to random variables $A$ and $B$, respectively.

Let us assume that $T_K = \frac{1}{\lambda K}$, $\frac{1}{2} < \lambda < \frac{2}{3}$, and that

$$S = x_1 \mathbb{I}_{\{U\le p\}} + x_2 \mathbb{I}_{\{U>p\}} \tag{50}$$

where $U$ is uniformly distributed over $[0,1]$ and $p \in [0,1]$. We also assume that $x_1 = \frac{1}{p}$ and $x_2 = \frac{1}{\sqrt{1-p}}$, which implies

$$\mathbb{E}[S] = x_1 p + x_2(1-p) = 1 + \sqrt{1-p}$$

$$\begin{aligned}\mathrm{Var}(S) &= x_1^2 p + x_2^2(1-p) - \mathbb{E}[S]^2 \\ &= \frac{1}{p} - 1 + p - 2\sqrt{1-p}.\end{aligned} \tag{51}$$

Within these conditions and also if $K \ge 3$, for the numerator of $\mathcal{E}$ we can show that (see below for a proof)

$$\lim_{p\uparrow 1} \min\left\{ W_K(1, R^{RR}, K), \inf_{R_K\in\mathcal{R}_K} W_K(K, R_K, \mathbf{1}) \right\} > 0. \tag{52}$$

Within the same conditions as above and with respect to some choice of $d_K$ and $R_{d_K}$, in the remainder of the proof we show that the denominator of $\mathcal{E}$ converges to zero when $p \uparrow 1$. This will conclude the proof.

Assume that $d_K = 2$, $h_K = (K-1, 1)$ and that $R_{d_K} \in \mathcal{R}_{d_K}$ sends jobs of size $x_i$ to second level dispatcher $i$. Then,

$$W_K(d_K, R_{d_K}, h_K) = pW(A_1^{RR}, x_1) + (1-p) W\left(\tfrac{Z_2}{\lambda K}, x_2\right) \tag{53}$$

where $Z_2 \sim$ Geometric$(1-p)$, $A_1^{RR}(p) \overset{\text{def}}{=} \sum_{n=1}^{K-1} \frac{Z_{1,n}}{\lambda K}$ with $Z_{1,n} \sim$ Geometric$(p)$ and the $Z_{1,n}$'s are independent.

Applying the upper bound (1) (recall that $\lambda < 2/3$) and using that $x_1 = \frac{1}{p}$, we obtain

$$W(A_1^{RR}, x_1) \le \frac{\lambda K p}{2(K-1)} \frac{1}{1 - \frac{\lambda K}{K-1}} \underbrace{\frac{K-1}{(\lambda K)^2} \frac{1-p}{p^2}}_{=\mathrm{Var}(A_1^{RR})} \xrightarrow[p\uparrow 1]{} 0. \tag{54}$$

Now, we also develop an upper bound for $W\left(\frac{Z_2}{\lambda K}, x_2\right)$; here, we cannot use again (1) as in (54) because this does not yield a sufficiently tight bound. Let $(E_n)_n$ be an i.i.d. sequence of exponentially distributed random variables with rate $\lambda K$ independent of everything else. Since $Z_2/\lambda K \le_{cx} \sum_{n=1}^{Z_2} E_n$, where $\le_{cx}$ denotes the convex order (see [31, Theorem 3.A.15]), we can apply [32, Corollary 5.2] to obtain

$$W\left(\tfrac{Z_2}{\lambda K}, x_2\right) \le W\left(\sum_{n=1}^{Z_2} E_n, x_2\right). \tag{55}$$

Furthermore, since $\sum_{n=1}^{Z_2} E_n$ is an exponentially distributed random variable with rate $\lambda K(1-p)$, we have bounded $W\left(\frac{Z_2}{\lambda K}, x_2\right)$ in terms of an M/G/1 queue. Applying the Pollaczek–Khinchine formula [32] to the RHS of (55) and using that $x_2 = \frac{1}{\sqrt{1-p}}$, for $p$ sufficiently large $\lambda K(1-p)x_2 < 1$ and

$$W\left(\tfrac{Z_2}{\lambda K}, x_2\right) \le \frac{\lambda K(1-p)}{2} \frac{x_2^2}{1 - \lambda K(1-p)x_2} \xrightarrow[p\uparrow 1]{} \frac{\lambda K}{2}. \tag{56}$$

Finally, using (54) and (56) in (53), we obtain that $W_K(d_K, R_{d_K}, h_K) \to 0$ as $p \uparrow 1$, as desired.

## B.1 Proof of (52)

First, we notice that $W_K(1, R^{RR}, K) = W(\frac{1}{\lambda}, S)$, and applying the lower bound in [33, Formula 2.51] to $W(\frac{1}{\lambda}, S)$, we obtain

$$W_K(1, R^{RR}, K) \ge \frac{\lambda}{2} \frac{\mathrm{Var}(S)}{1 - \lambda \mathbb{E}[S]} - \frac{\mathbb{E}[S]}{2}. \tag{57}$$

Given (51), $\lim_{p\uparrow 1} W_K(1, R^{RR}, K) > 0$ if $\lambda > \frac{1}{2}$.

We now show that $\lim_{p\uparrow 1} \inf_{R_K} W_K(K, R_K, \mathbf{1}) > 0$. Since $S$ has not a density function, we need to adapt the definition of SITA policy given in Definition 1 (here, we could consider a perturbed version of the probability distribution of our choice for $S$, say $F_\epsilon$, such that $F = F_0$ and $F_\epsilon$ is differentiable for all $\epsilon > 0$ and apply what follows, but we omit this for simplicity). When $S$ is given by (50), SITA policies have the following structure: there exists some $d < K$ such that the dispatcher sends jobs of size $x_1$ (respectively, $x_2$) randomly to all queues $i \le d$ ($i > d$). Therefore,

$$\inf_{R_K\in\mathcal{R}_K} W_K(K, R_K, \mathbf{1}) =$$

$$\min_{d\in\{1,\dots,K-1\}} \sum_{i=1}^{K} q_i W\left(\tfrac{Z_i}{\lambda K}, x_1 \mathbb{I}_{\{i\le d\}} + x_2 \mathbb{I}_{\{i>d\}}\right) \overset{\text{def}}{=} W^S$$

where $q_i = \frac{p}{d}\mathbb{I}_{\{i\le d\}} + \frac{1-p}{K-d}\mathbb{I}_{\{i>d\}}$ and $Z_i \sim$ Geometric$(q_i)$. Furthermore,

$$\begin{aligned} W^S &= \min_{d\in\{1,\dots,K-1\}} p\,W\left(\tfrac{Z_1}{\lambda K}, x_1\right) + (1-p) W\left(\tfrac{Z_K}{\lambda K}, x_2\right) \\ &\ge \min_{d\in\{1,\dots,K-1\}} p\,W\left(\tfrac{Z_1}{\lambda K}, x_1\right). \end{aligned} \tag{58a}$$

In order to have $W\left(\frac{Z_1}{\lambda K}, x_1\right)$ finite, the stability condition $\lambda K x_1 < \mathbb{E} Z_1$, i.e., $\lambda < d/K$ needs to be satisfied. Thus, an optimizer of (58a) necessarily satisfies $d > \lambda K$. For any $d > \lambda K$, applying again the lower bound in [33, Formula 2.51], we obtain

$$\begin{aligned} & p\,W\left(\tfrac{Z_1}{\lambda K}, x_1\right) \\ &\ge p\left( \frac{\lambda K}{2\mathbb{E} Z_K} \frac{\mathrm{Var}(\frac{Z_1}{\lambda K})}{1 - \lambda K/d} - \frac{1}{2}\left( x_1 + \frac{\mathrm{Var}(\frac{Z_1}{\lambda K})}{\mathbb{E}[Z_1]/\lambda K} \right) \right) \end{aligned}$$

$$= \frac{1}{2\lambda K}\frac{d-p}{1-\lambda K/d} - \frac{1}{2}\left(1+\frac{d-p}{\lambda K}\right)$$
$$= \frac{d-p}{2\lambda K}\frac{\lambda K/d}{1-\lambda K/d} - \frac{1}{2} \geq \frac{1}{2}\frac{\lambda K-1}{d-\lambda K} > 0$$

where the last inequality follows by using that $K \geq 3$ and $\lambda > \frac{1}{2}$, and thus $\lambda K > 1$.

## References

[1] J. F. C. Kingman, "The effect of queue discipline on waiting time variance," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, no. 1, p. 163164, 1962.

[2] ——, "Some inequalities for the queue gi/g/1," *Biometrika*, vol. 49, no. 3-4, pp. 315–324, 1962.

[3] P. Humblet, *Determinism Minimizes Waiting Time in Queues*, ser. LIDS-P-1207. Laboratory for Information and Decision Systems, M.I.T., 1982.

[4] Z. Liu and R. Righter, "Optimal load balancing on distributed homogeneous unreliable processors," *Operations Research*, vol. 46, no. 4, pp. 563–573, 1998.

[5] 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.

[6] J. Anselmi and J. Doncel, "Asymptotically optimal size-interval task assignments," *IEEE Transactions on Parallel and Distributed Systems*, to appear.

[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] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, no. 1, pp. 181–189, 1977.

[9] R. R. Weber, "On the optimal assignment of customers to parallel servers," *J. of App. Prob.*, vol. 15, no. 2, pp. 406–413, 1978.

[10] 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.

[11] D. Mukherjee, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting, "Asymptotic Optimality of Power-of-$d$ Load Balancing in Large-Scale Systems," *ArXiv e-prints*, Dec. 2016.

[12] J. Anselmi and F. Dufour, "Power-of-d-choices with memory: Fluid limit and optimality," *Mathematics of Operations Research*, to appear.

[13] 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.

[14] A. L. Stolyar, "Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers," *Queueing Syst. Theory Appl.*, vol. 85, no. 1-2, pp. 31–65, Feb. 2017.

[15] M. El-Taha and B. Maddah, "Allocation of service time in a multiserver system," *Management Science*, vol. 52, no. 4, pp. 623–637, 2006.

[16] Q. Zhang, A. Riska, W. Sun, E. Smirni, and G. Ciardo, "Workload-aware load balancing for clustered web servers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 3, pp. 219–233, March 2005.

[17] 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.

[18] 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.

[19] 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.

[20] 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, Apr. 2004.

[21] L. Cherkasova and M. Karlsson, "Scalable web server cluster design with workload-aware request distribution strategy ward," in *Proc. 3rd Int. Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems. WECWIS 2001*, June 2001, pp. 212–221.

[22] M. Harchol-Balter, "Task assignment with unknown duration," *J. ACM*, vol. 49, no. 2, pp. 260–288, 2002.

[23] A. Riska, W. Sun, E. Smirni, and G. Ciardo, "Adaptload: effective balancing in clustered web servers under transient load conditions," in *Proceedings 22nd International Conference on Distributed Computing Systems*, July 2002, pp. 104–111.

[24] 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.

[25] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, Feb 1997.

[26] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec 1997.

[27] M. E. Crovella, M. S. Taqqu, and A. Bestavros, "A practical guide to heavy tails," R. J. Adler, R. E. Feldman, and M. S. Taqqu, Eds. Cambridge, MA, USA: Birkhauser Boston Inc., 1998, ch. Heavy-tailed Probability Distributions in the World Wide Web, pp. 3–25.

[28] M. Mitzenmacher, "Analyzing distributed join-idle-queue: A fluid limit approach," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 312–318.

[29] E. Bachmat and H. Sarfati, "Analysis of sita policies," *Perform. Eval.*, vol. 67, no. 2, pp. 102–120, Feb. 2010.

[30] J. Almeida, V. Almeida, D. Ardagna, talo Cunha, C. Francalanci, and M. Trubian, "Joint admission control and resource allocation in virtualized servers," *Journal of Parallel and Distributed Computing*, vol. 70, no. 4, pp. 344 – 362, 2010.

[31] M. Shaked and J. G. Shanthikumar, *Stochastic orders and their applications*. Academic Pr, 1994.

[32] S. Asmussen, *Applied Probability and Queues*. Wiley, 1987.

[33] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*. Wiley, 1976.

**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 – BCAM, 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.