Performance Balancing Size-Interval Routing Policies

Josu Doncel
University of the Basque Country, UPV/EHU

josu.doncel@ehu.eus

Abstract

We study a parallel-queue system with Poisson arrivals, in which a dispatcher sends the incoming traffic to K queues using Size Interval Task Assignment (SITA) policy that aims to equalize the performance of all queues. We study existence and uniqueness of the allocation thresholds for a large set of performance functions of the queues. We also provide a family of performance functions of the queues such that the performance of the system is characterized. For a particular case of the latter family of functions, we show that the performance of the SITA policy we study coincides with the performance of the SITA policy in which the load are balanced. We investigate the optimality of the SITA policy under study and, according to our numerical experiments, for FCFS queues and Bounded Pareto distributed job sizes, the SITA policy we study is almost optimal, when we consider the mean queue length and $\alpha < 1$ as well as when we consider the mean slowdown.

1 Introduction

1.1 Motivation

We consider a data center with K servers and a single dispatcher. It is known that, in modern data centers, a short amount of jobs form almost the half of the system load [20], i.e., the size of incoming tasks follows a heavy-tailed distribution. Thus, if the servers give service in order of arrival, i.e., they are FCFS queues, the performance of these systems is clearly improved if shorts and long jobs are executed in different servers. This is indeed the idea behind the Size Interval Task Assignment (SITA) policy. More precisely, the SITA policy is an open-loop load-balancing technique where the job size distribution is divided into intervals and all the jobs whose size are in a given interval are dispatched to the same server. SITA policies are interesting from the practical point of view since they do not require signaling between the dispatcher and the servers.

The analytical study of the optimal SITA policy (in the sense that the intervals are chosen so as to minimize the mean response time of jobs in the system) seems to be an impossible task (we provide in Section 1.3 a detailed discussion of SITA policies). In this article, we study a SITA policy where the goal is to balance the performance of the queues, which is easier to deal with than the optimal SITA and, as we will see, it provides better performance than the SITA policy that equalizes the load of the servers.

Result	**Examples**
Existence and Uniqueness	Mean waiting time of FCFS queues Mean queue length of FCFS queues Mean number of customers of PS queues
Characterization	Mean waiting time of FCFS queues Mean number of customers of PS queues
Comparison with SITA-E	Mean number of customers of PS queues

Table 1: Examples of the performance functions that satisfy the main results of this article.

1.2 Contributions

We study the performance of a parallel-server system operating under the SITA policy such that the performance of the queues is the same. The main contributions of this work are summarized as follows:

(C1) We provide a family of performance functions such that the SITA policy that balances the performance of the queues exists and is unique (see Section 3).

(C2) We present a family of performance functions whose performance can be characterized (see Section 4).

(C3) We give a family of functions such that the performance of the SITA policy that balances this performance metric coincides with the performance of the SITA policy that equalizes the load of the serves (see Section 5).

(C4) We study the optimality of the SITA policy that balances performance and, for FCFS servers and Bounded Pareto distributed job sizes with parameter α, we present numerical experiments that show that, when the goal is to optimize the mean queue length, the SITA policy that balances the mean queue length is almost optimal (see Section 6).

We show in Table 1 examples of performance functions that are covered with our main results. We first observe that examples of the performance functions that satisfy the condition for the existence and uniqueness of the thresholds that balance the performance of the queues are the mean waiting time of First Come First Served (FCFS) queues, the mean queue length of FCFS queues and the mean number of customers of Processor Sharing (PS) queues. It can be also seen in Table 1 that the performance of the system can be characterized, when it is balanced the mean queue length of FCFS queues or the mean number of customers of PS queues. We also present that, when we consider the mean queue length of PS queues, the performance of the system coincides with the performance of the SITA policy that balances the load.

We present numerical experiments to compare the performance of the SITA policy under analysis in this paper with other SITA policies of the literature for Bounded Pareto distribution with parameter α. We first consider the mean queue length of FCFS queues and we show that, when $\alpha < 1$, the performance of the SITA policy that balances the performance of the queues is very close to that of the SITA policy with optimal thresholds. Besides, we consider the mean slowdown of FCFS servers and we show the performance of the SITA policy that balances the performance of the queues is always smaller than that of the SITA policy that equalizes the load of the system and it is very close to that of the SITA policy with optimal thresholds.

Finally, we study the SITA policy that balances the performance of the queues when the system is formed by servers of different speeds and we explain how the results obtained in this article extend to heterogeneous servers.

1.3 Related Work

A system formed by a set of parallel servers and a single dispatcher is often used to study the performance a large variety of systems, such as data-centers, see [12] for a recent book in this topic. It is known that when the number of customers in all the servers is known and the service time distribution is exponential or has a non-decreasing hazard rate the Join-the-Shortest-Queue policy minimizes the response of jobs [21, 19]. However, one practical issue in this type of is that server states may not be observable or it is too expensive to provide this information regularly. Therefore, in the last decades, the interest of researchers in studying non-observable dispatching policies has increased a lot, see the recent survey [17] and the references therein.

In this article we consider an non-observable dispatching policy in which the size of the incoming jobs is known. The idea of executing jobs of different sizes in different servers has been investigated first in [14, 13] under the name Size Interval Task Assignment with Equal load (SITA-E). In their model, the thresholds that determine the interval of job sizes that each server executes is chosen to equalize the load of the servers. The authors show that, when the servers are FCFS, this SITA-E improves the performance of the system comparing with other non-observable policies such as Round-Robin or Random Splitting. An important result for SITA policies is provided in [10], where the authors show that, when the service demand is known but the queues are non-observable and the servers are FCFS, the SITA policy with optimal thresholds optimizes the performance of the system. The analytical expression of the optimal thresholds is known only under the assumption of Bounded Pareto distributed job sizes [6, 18, 3]. Therefore, the results regarding the optimal SITA policy for an arbitrary distribution are scarce. We here present a couple of examples. The authors in [16] study a system formed by two server and explore how the load should be unbalanced in order to optimize the performance. Another recent result is [15], where the authors show that the performance of the optimal SITA can be much worse than the performance of the Least-Work-Left policy when the coefficient of variation of incoming jobs is high. The authors in [2] characterize the optimal SITA policy in the asymptotic regime where the arrival rate and the number of servers tend to infinity.

Taking into account the difficulty of the analytical study of the optimal SITA policy, some size-interval variants have been presented in the literature, see for example [11, 4] for the task assignment by guessing size and [7] for the application to web servers. Other authors have studied SITA-E instead of the optimal SITA policy, see for instance [8, 9]. In this article, we also consider a variant which consists of choosing the thresholds in a way to equalize general performance functions of the queues. The authors in [5] study the thresholds that balance the queue length of the servers and give an upper bound on the mean waiting time of jobs when these thresholds are selected. In this work, we provide examples which show that, when the goal is to optimize the mean queue length, there are instances where balancing the mean queue length is almost optimal. Another related work is [1], where the authors prove the existence and uniqueness of the thresholds that balance the mean waiting time and mean queue length in FCFS servers as well as the mean response time and the number of customers in PS servers. In the present work, we extend the results of [1] to general performance functions.

1.4 Organization

The rest of the article is organized as follows. In Section 2, we describe the model and we present the existence and uniqueness results in Section 3. In Section 4 we present on a family of functions such that the performance of the system can be characterized and in Section 5 we compare the performance of the SITA policy that balances the performance with the SITA policy that equals the load of the servers. We present numerical experiment to study the mean queue length and the mean slowdown of FCFS servers in Section 6 to study the optimality of the SITA policy under study. In Section 7 we discuss how our results extend to heterogeneous servers. Finally, we give the conclusions of our work in Section 8.

2 Model Description

We analyze a system of K parallel homogeneous servers and a single dispatcher. We assume that service times of incoming jobs form an i.i.d. sequence with a common distribution denoted by X and $\mathbb{E}(X)$ denotes its first moment. Let $F(x) = \mathbb{P}(X \leq x)$. We assume $F(x)$ to be differentiable and we write $f(x) = \frac{dF(x)}{dx}$. We denote by x_{min} and x_{max}, respectively, the minimum and maximum size of the incoming jobs to the system.

Servers are FCFS queues and the dispatcher handles all the incoming traffic, which arrives to the system according to a Poisson process of rate λ. The total load in the system is denoted by $\rho = \lambda \cdot \mathbb{E}(X)$. For stability reasons, we assume $\rho < K$.

We denote by λ_i the arrival rate to queue i and let X_i be the service time of jobs to be executed in queue i, where its first and second moments are denoted by $\mathbb{E}(X_i)$ and $\mathbb{E}(X_i^2)$, respectively.

We consider that routers implements a size-based policy called SITA. For this policy, there are $K+1$ thresholds that are denoted by $x_0, \dots, x_K$ and we have that $x_{min} = x_0 < x_1 < \dots < x_{K-1} < x_K = x_{max}$. Jobs ranging in size from x_{i-1} to x_i are executed in queue i. Therefore, the arrival rate to each queue is a Poisson process with rate $\lambda_i = \lambda(F(x_i) - F(x_{i-1}))$.

In this work, we consider that the performance of a queue that executes jobs ranging in size from x_i to x_{i+1} is a function denoted by $P(x_i, x_{i+1})$ and we study the SITA-BAL policy, which is a SITA policy that equalizes the performance of the queues. In other words, in the SITA-BAL policy, the goal is to find the thresholds $x_1, \dots, x_{K-1}$ such that, for a given performance function $P()$, the following condition is verified:

$$P(x_{min}, x_1) = P(x_1, x_2) = \dots = P(x_{K-1}, x_{max}).$$

Throughout this paper, we shall also be interested in other size-based policies, such as SITA with optimal thresholds, SITA-OPT, and SITA-E, that we briefly present next. For the optimal SITA, the thresholds are chosen to optimize the performance of the system. Unfortunately, to the best of our knowledge, there is no closed-form expression for these thresholds for an arbitrary distribution, even for a two-queues system. For the SITA-E policy, the thresholds are chosen in a way such that the following condition is satisfied:

$$\int_{x_{min}}^{x_1} xf(x)dx = \int_{x_1}^{x_2} xf(x)dx = \dots = \int_{x_{K-1}}^{x_{max}} xf(x)dx.$$

From this definition, it follows that, when SITA-E is implemented, the load in all the queues is the same.

In a system operating under the a SITA policy, once the thresholds are computed, the performance of the system can be easily obtained using the Pollaczek-Khinchine formula. As we shall see in Section 4, the performance of a system that implements the SITA policy we present in this paper can be characterized without computing the thresholds. Prior to that, we study the existence and uniqueness of the thresholds that balance the performance of the queues in the next section.

3 Existence and Uniqueness

In this section, we analyze the thresholds such that the performance of the queues is balanced. Let $w < y$ and positive. We consider two queues where one queue executes jobs whose size are in the interval $[w, x)$ and the other in the interval $[x, y]$. For a given w and y, we write $\nu(w, x, y) = P(w, x) - P(x, y)$. We aim to find the value of x such that the performance of both queues is the same, i.e., $P(w, x) = P(x, y)$ and this is equivalent to find the root of ν. For any w and y such that $w < y$, we assume that

(A1) $\nu(w, x, y)$ is continuous and increasing on x,

(A2) $\nu(w, w, y) < 0 < \nu(w, y, y)$.

It is immediate to show that, under these assumptions, the function ν has a unique root. In the following result, we focus on a two queues system and we show the existence of a threshold that balances their performance.

Proposition 1. *In a system with two queues, for any performance metric that satisfies (A1)-(A2), there exists a unique threshold value such that the performance of both queues is equal.*

Proof. In a system with two queues, there is a single threshold x that determines that the jobs ranging in size between x_{min} and x are executed in Queue 1 and the jobs ranging in size between x and x_{max} in Queue 2. Therefore, the performance of Queue 1 is $P(x_{min}, x)$ and of Queue 2 $P(x, x_{max})$. Moreover, since $x_{min} < x_{max}$, (A1)-(A2) imply that $\nu(x_{min}, x, x_{max})$ is continuous and increasing on x and $\nu(x_{min}, x_{min}, x_{max}) < 0 < \nu(x_{min}, x_{max}, x_{max})$. As a result, $\nu(x_{min}, x, x_{max})$ has a root in (x_{min}, x_{max}) and therefore, there exists a unique value of x such that the performance of both queues is equal. □

From this result, we know that there exists a unique value $x \in (x_{min}, x_{max})$ such that $\nu(x_{min}, x, x_{max}) = 0$. Thus, we have that $x = \phi(x_{min}, x_{max})$ is a bijective function such that, for any x_{min} and x_{max}, $\phi(x_{min}, x_{max})$ gives the value of x such that $\nu(x_{min}, x, x_{max}) = 0$.

We now analyze a system with $K > 2$ queue. For this instance, there are $K + 1$ values satisfying $x_{min} = x_0 < x_1 < \cdots < x_{K-1} < x_K = x_{max}$. In the next result, we show the existence and uniqueness of these thresholds that balances the performance of all the queues.

Proposition 2. *In any system with K queues whose performance satisfies (A1)-(A2) for all $w < y$, there exist a unique set of thresholds $x_1, x_2, \ldots, x_{K-1}$ such that the performance of the queues is equalized.*

Proof. Assuming that (A1)-(A2) for any $w < y$, it follows for all $i = 1, \dots, K-1$, that

$$\begin{aligned} x_i &= \phi(x_{i-1}, x_{i+1}) \\ &= \phi(\phi(x_{i-2}, x_i), \phi(x_i, x_{i+2})) \\ &= \phi(\phi(\phi(x_{i-3}, x_{i-1}), x_i), \phi(x_i, \phi(x_{i+1}, x_{i+3}))), \end{aligned}$$

and, by applying this recursion until we write x_i as a function of only x_0, x_i and x_K, the previous expression can be written as $x_i = h(x_0, x_i, x_K)$, where $h(\cdot)$ is the required composition of function $\phi()$ to write x_i as a function of only x_0, x_i and x_K. Since x_0 and x_K are fixed and ϕ is a continuous mapping, $h(\cdot)$ is a function with one variable that satisfies Brouwer's fixed point theorem. Therefore, $h(\cdot)$ has a unique fixed point and the threshold x_i exists and is unique.

□

Before going further, we note the conditions (A1)-(A2) are very general and include, not only the performance metrics considered in [1], but also a wide range of performance functions such as the variance of the waiting time of FCFS queues as well as the performance functions presented in Table 1.

In the following sections, we consider particular performance functions that satisfy (A1)-(A2). First, we present a family of performance functions (that includes the mean waiting time of FCFS queues and the mean number of customers of PS queues) such that the performance of the system can be easily characterized. Then, we consider another family of performance functions (that includes the mean number of customers of PS queues) and we show that its performance coincides with that of SITA-E.

In the next remark, we provide an example where one of the conditions required for the existence is not satisfied.

Remark 1. *We consider that the job size distribution is U(0,1) and we aim to balance the mean response time of jobs in a system with two Processor Sharing queues. For this case, we write $\nu(0, x, 1) = \mu(x)$ and we obtain*

$$\mu(x) = \frac{\frac{1}{2}x}{1 - \frac{\lambda}{2}x^2} - \frac{\frac{1}{2}(1+x)}{1 - \frac{\lambda}{2}(1-x^2)},$$

which is continuous and increasing on x and $\mu(0) < 0$. And we observe that (A2) is not satisfied if $\lambda < 1$ since $\mu(1)$ is negative in this case.

We note that the above result is correct since the mean response time of a Processor Sharing queue is the mean job size of jobs executed in that queue divided by one minus the load of that queue.

4 Characterization

We consider in this section that the performance of a queue executing jobs ranging in size from x_{i-1} to x_i is given by:

$$P(x_{i-1}, x_i) = \frac{\sum_{j=1}^{n} (q_j(x_i, \lambda) - q_j(x_{i-1}, \lambda))}{r_0 + \sum_{j=1}^{m} (r_j(x_i, \lambda) - r_j(x_{i-1}, \lambda))}, \quad (1)$$

where

- q_j and r_j are continuous functions from $[x_{min}, x_{max}] \times \mathbb{R}^+$ to $\mathbb{R}^+$ such that $q_i(x_{min}, \lambda) = 0$ for all $i = 1, \dots, n$ and $r_i(x_{min}, \lambda) = 0$ for all $i = 1, \dots, m$,
- r_0 is a function from $\mathbb{R}$ to $\mathbb{R}$,
- (1) satisfies assumptions (A1)-(A2).

We remark that the family of functions we consider in this section covers a wide range of performance metrics. For instance, the mean waiting time of a FCFS queue that executes jobs ranging in size from x_{j-1} to x_j, is given by

$$\frac{\lambda \int_{x_{j-1}}^{x_j} x^2 f(x)dx}{2(1 - \lambda \int_{x_{j-1}}^{x_j} xf(x)dx)}.$$

This expression can be clearly written in the form of (1) with $m = 1,\ n = 1,\ r_0 = 2,$ $r_1(x) = -2\lambda \int_{x_{min}}^{x} uf(u)du$ and $q_1(x) = \lambda \int_{x_{min}}^{x} u^2 f(u)du.$

We study the performance of a system where the dispatcher implements the SITA routing that balances the performance of the servers, which is given by (1). Thus, we have that

$$\frac{\sum_{i=1}^n q_i(x_1, \lambda)}{r_0 + \sum_{i=1}^m r_i(x_1, \lambda)} = \frac{\sum_{i=1}^n (q_i(x_2, \lambda) - q_i(x_1, \lambda))}{r_0 + \sum_{i=1}^m (r_i(x_2, \lambda) - r_i(x_1, \lambda))} = \frac{\sum_{i=1}^n (q_i(x_{max}, \lambda) - q_i(x_{K-1}, \lambda))}{r_0 + \sum_{i=1}^m (r_i(x_{max}, \lambda) - r_i(x_{K-1}, \lambda))}. \tag{2}$$

This expression can be written as the following system of equations:

$$\left\{ \begin{array}{lcl}
\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} & = & \frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} \\
\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} & = & \frac{\sum_{i=1}^n (q_i(x_2,\lambda)-q_i(x_1,\lambda))}{r_0+\sum_{i=1}^m (r_i(x_2,\lambda)-r_i(x_1,\lambda))} \\
\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} & = & \frac{\sum_{i=1}^n (q_i(x_3,\lambda)-q_i(x_2,\lambda))}{r_0+\sum_{i=1}^m (r_i(x_3,\lambda)-r_i(x_2,\lambda))} \\
\vdots & & \\
\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} & = & \frac{\sum_{i=1}^n (q_i(x_{K-1},\lambda)-q_i(x_{K-2},\lambda))}{r_0+\sum_{i=1}^m (r_i(x_{K-1},\lambda)-r_i(x_{K-2},\lambda))} \\
\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)} & = & \frac{\sum_{i=1}^n (q_i(x_{max},\lambda)-q_i(x_{K-1},\lambda))}{r_0+\sum_{i=1}^m (r_i(x_{max},\lambda)-r_i(x_{K-1},\lambda))},
\end{array} \right.$$

We rearrange both sides of the equations and we obtain:

$$\left\{ \begin{array}{l}
\left(\sum_{i=1}^n q_i(x_1, \lambda)\right) \left(r_0 + \sum_{i=1}^m r_i(x_1, \lambda)\right) = \left(r_0 + \sum_{i=1}^m r_i(x_1, \lambda)\right) \left(\sum_{i=1}^n q_i(x_1, \lambda)\right) \\
\left(\sum_{i=1}^n q_i(x_1, \lambda)\right) \left(r_0 + \sum_{i=1}^m (r_i(x_2, \lambda) - r_i(x_1, \lambda))\right) = \left(r_0 + \sum_{i=1}^m r_i(x_1, \lambda)\right) \left(\sum_{i=1}^n (q_i(x_2, \lambda) - q_i(x_1, \lambda))\right) \\
\\
\left(\sum_{i=1}^n q_i(x_1, \lambda)\right) \left(r_0 + \sum_{i=1}^m (r_i(x_3, \lambda) - r_i(x_2, \lambda))\right) = \left(r_0 + \sum_{i=1}^m r_i(x_1, \lambda)\right) \left(\sum_{i=1}^n (q_i(x_3, \lambda) - q_i(x_2, \lambda))\right) \\
\\
\quad \vdots \\
\\
\left(\sum_{i=1}^n q_i(x_1, \lambda)\right) \left(r_0 + \lambda \sum_{i=1}^m (r_i(x_{max}, \lambda) - r_i(x_{K-1}, \lambda))\right) \\
\qquad\qquad\qquad = \left(r_0 + \sum_{i=1}^m r_i(x_1, \lambda)\right) \left(\sum_{i=1}^n (q_i(x_{max}, \lambda) - q_i(x_{K-1}, \lambda))\right).
\end{array} \right.$$

As it can be observed, the LHS of all the equations has the common factor $\sum_{i=1}^n q_i(x_1,\lambda)$ and the RHS the factor $r_0+\sum_{i=1}^m r_i(x_1,\lambda)$. Therefore, summing the K equations and simplifying, it results that

$$\left(\sum_{i=1}^n q_i(x_1,\lambda)\right)\left(Kr_0+\sum_{i=1}^m r_i(x_{max},\lambda)\right)=\left(r_0+\sum_{i=1}^m r_i(x_1,\lambda)\right)\left(\sum_{i=1}^n q_i(x_{max},\lambda)\right),$$

which is equivalent to

$$\frac{\sum_{i=1}^n q_i(x_1,\lambda)}{r_0+\sum_{i=1}^m r_i(x_1,\lambda)}=\frac{\frac{1}{K}\sum_{i=1}^n q_i(x_{max},\lambda)}{r_0+\frac{1}{K}\sum_{i=1}^m r_i(x_{max},\lambda)}.$$

We observe that the LHS of this expression coincides with the first term of (2). As a result, the RHS of this expression characterizes the performance of the system if the performance of the servers is of the form of (1).

Proposition 3. *Consider a parallel-server system whose performance is in the form of* (1)*. The performance of a system with arrival rate* λ *and* K *servers, where the dispatcher implements the SITA routing that balances the performance is*

$$\frac{\frac{1}{K}\sum_{i=1}^n q_i(x_{max},\lambda)}{r_0+\frac{1}{K}\sum_{i=1}^m r_i(x_{max},\lambda)}.$$

Since the mean waiting time of FCFS queues is a particular case of the function $P(\cdot)$ under consideration, the result of Proposition 3 can be applied for this case.

Corollary 1. *The mean waiting time of a system that operates under the SITA policy that balances the mean waiting time of FCFS queues is*

$$\frac{\frac{\lambda}{K}\int_{x_{min}}^{x_{max}} x^2 f(x)dx}{2(1-\frac{\lambda}{K}\int_{x_{min}}^{x_{max}} xf(x)dx)},$$

which coincides with the mean waiting time of a M/G/1 queue with arrival rate λ/K*.*

5 Performance Comparison

We now consider the performance function (1) where $q_i(x_i,\lambda)$ (resp. $r_i(x_i,\lambda)$) is a function $s_i(\lambda\int_{x_{min}}^{x_i} xf(x)dx)$ (resp. $w_i(\lambda\int_{x_{min}}^{x_i} xf(x)dx)$) that is affine, i.e., $s_i(y)=a_1y+b_1$, for some a_1 and b_1 (resp. $w_i(y)=a_2y+b_2$ for some a_2 and b_2). Hence, the following family of performance functions are considered here:

$$\frac{\sum_{i=1}^n s_i(\lambda\int_{x_{min}}^{x_i} xf(x)dx)-s_i(\lambda\int_{x_{min}}^{x_{i-1}} xf(x)dx)}{r_0+\sum_{i=1}^m w_i(\lambda\int_{x_{min}}^{x_i} xf(x)dx)-w_i(\lambda\int_{x_{min}}^{x_{i-1}} xf(x)dx)}. \tag{3}$$

According to the result of Proposition 3, it follows that the performance of the system when it is implemented the SITA policy that balances the performance of the queues is

$$\frac{\frac{1}{K}\sum_{i=1}^n s_i(\lambda\int_{x_{min}}^{x_{max}} xf(x)dx)}{r_0+\frac{1}{K}\sum_{i=1}^m w_i(\lambda\int_{x_{min}}^{x_{max}} xf(x)dx)}.$$

We now focus on the performance of a system operating under SITA-E routing. Let $x_1^E,\dots,x_{K-1}^E$ be the cutoffs of the SITA-E policy in a system with K servers. Besides, the performance of a server that executes jobs ranging in size between x_{i-1}^E and x_i^E is given by

$$\frac{\sum_{i=1}^n s_i(\lambda\int_{x_{min}}^{x_i^E} xf(x)dx) - s_i(\lambda\int_{x_{min}}^{x_{i-1}^E} xf(x)dx)}{r_0+\sum_{i=1}^m w_i(\lambda\int_{x_{min}}^{x_i^E} xf(x)dx) - w_i(\lambda\int_{x_{min}}^{x_{i-1}^E} xf(x)dx)}.$$

Since s_i and w_i are linear, the previous expression is equal to the following one:

$$\frac{\sum_{i=1}^n s_i(\lambda\int_{x_{i-1}^E}^{x_i^E} xf(x)dx)}{r_0+\sum_{i=1}^m w_i(\lambda\int_{x_{i-1}^E}^{x_i^E} xf(x)dx)}. \tag{4}$$

In a system that implements SITA-E policy the load of each server is the same, that is,

$$\int_{x_{i-1}^E}^{x_i^E} uf(u)du = \frac{1}{K}\int_{x_{min}}^{x_{max}} uf(u)du,\ \forall i=1,\dots,K.$$

Using this in (4), we obtain that the following result:

Proposition 4. *When the performance of the queues is given by* (3)*, the performance of the SITA-E routing and of the SITA routing that balances the performance is the same.*

The mean number of customers of server j, which is a PS queue with load ρ_j, is given by $\frac{\rho_j}{1-\rho_j}$. We observe that it is of the form of (3) with $m=1$, $n=1$, $r_0=1$, $s_1(x)=x$ and $w_1(x)=-x$. Note that, if server j executes jobs ranging in size between x_{j-1} and x_j, then $\rho_j=\lambda\int_{x_{j-1}}^{x_j} xf(x)dx$. Therefore, the results of Proposition 4 can be applied for this instance.

Corollary 2. *The mean number of customers of a system that implements the SITA policy that balances the mean number of customers in PS queues is*

$$\frac{\frac{\rho}{K}}{1-\frac{\rho}{K}},$$

which coincides with the mean number of customers of a M/G/1 queue with arrival rate λ/K and also of a system that with arrival rate λ and K servers that implements SITA-E policy.

6 Numerical Experiments

We investigate the optimality of the size-based routing policy that balances the performance of the queues. We assume that the job size distribution is Bounded Pareto with parameter α, that is, if $x_{min}\le x\le x_{max}$, then

$$f(x)=\frac{\alpha\ x_{min}^{\alpha}}{1-(x_{min}/x_{max})^{\alpha}}\ x^{-\alpha-1},$$

and $f(x)=0$ otherwise, where α is the tail of the distribution. We consider that the smallest job size is 1, i.e., $x_{min}=1$. We have performed the optimization using Newtonian search with a tolerance of 10^{-8}.

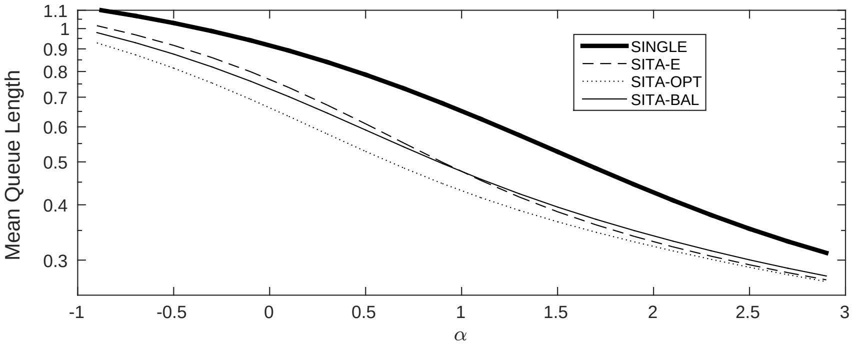

Figure 1: Mean queue length comparison in a two server system with Bounded Pareto distributed job sizes varying α from -1 to 3. $x_{min} = 1$, $x_{max} = 10$ and $\rho = 0.5$.

In the plots we present in this section, we analyze the performance of four policies: the solid and narrow line represents the performance of the SITA routing that balances the performance of the servers (SITA-BAL), the dotted line represents the SITA routing that optimizes the performance (SITA-OPT), the dashed line represents the SITA routing that equalizes the load of the servers (SITA-E) and the solid and wide line represents the single server with arrival rate λ/K (SINGLE), where K is the number of servers.

We have carried out extensive numerical experiments considering different parameters of the system. The figures we present in this section are representative of the general pattern.

6.1 Mean Queue Length

We study the performance of the SITA policy that balances the mean queue length of FCFS servers. This performance function is not of the form of (1). Hence, the result of Proposition 3 and Proposition 4 do not apply for this case and, therefore, it seems that one can only study the performance of this routing policy numerically.

We first study a system with two servers. In Figure 1, we illustrate the mean queue length of these four policies when α varies from -1 to 3 and the system is at low load ($\rho = 0.5$) and $x_{min} = 1$ and $x_{max} = 10$. We observe that, when α is close to 3, the performance of SITA-E and of the SITA policy that balances the mean queue length are close to the optimal performance. Besides, when $\alpha < 1$, the SITA policy that balances the mean queue length outperforms SITA-E. This figure confirms the result of [16] that show that for the optimal SITA policy the loads are very unbalanced (and therefore the SITA-E is far from being optimal) and shows that for the optimal SITA policy the performance of the servers can be very similar. This plot also shows that, for all α, the performance of a single server with arrival rate λ/K is much worse that the rest of the policies.

We note that, when $\alpha = -1$, the Bounded Pareto distribution coincides with the uniform distribution. Therefore, as we observe in Figure 1, the SITA policy that balances the mean queue length outperforms SITA-E for the uniform distribution.

We now focus on the influence on the performance when the value of x_{max} varies from 2 to 100, the value of α is fixed and the system is at low load ($\rho = 0.5$). In Figure 2, we consider $\alpha = 0.2$ and $K = 2$ and we depict the performance of the four policies under consideration. First, we observe that, when x_{max} is small, the performance of all the policies turns out

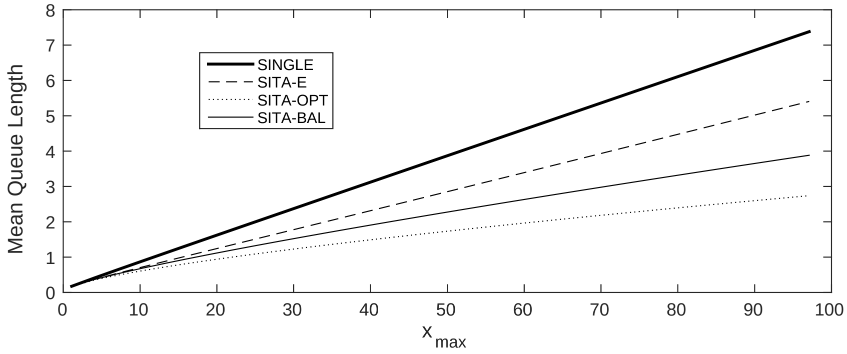

Figure 2: Mean queue length comparison in a two-server system with Bounded Pareto distributed job sizes with $\alpha = 0.2$, $x_{min} = 1$, $\rho = 0.5$ and x_{max} varying from 2 to 100.

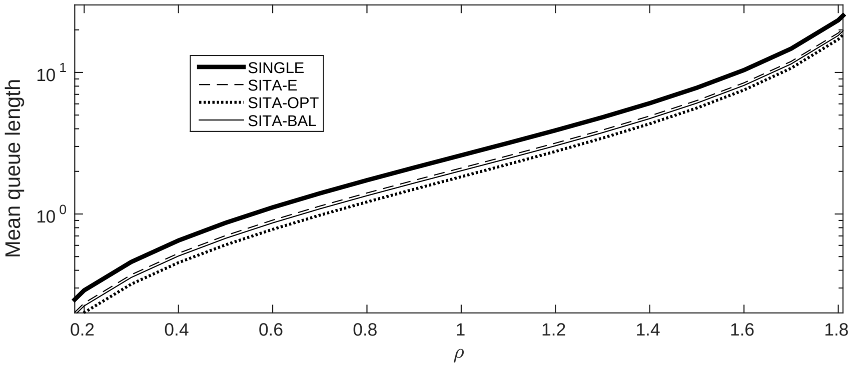

Figure 3: Mean queue length comparison in a two-server system with Bounded Pareto distributed job sizes with $\alpha = 0.2$, $x_{min} = 1$, $x_{max} = 10$ and ρ varying from 0.2 to 1.8 (y-axis in logarithmic scale).

to be very similar. However, when x_{max} grows, the difference on the performance between the SITA policy that balances performance and the optimal SITA increases. Besides, the difference on the performance is also high when we compare the SITA-E policy with the SITA policy that balances the performance of the servers. This simulation suggests that, when the difference between the largest and the smallest job size increases, the optimal SITA policy performs better than the SITA policies under consideration in this section.

We also study the influence of the load of the system on the performance of the policies under consideration in this section when the rest of the parameters are fixed and $K = 2$. Hence, we set $x_{min} = 1$, $x_{max} = 10$, $\alpha = 0.2$ and we vary ρ from 0.2 and 1.8 and, as it can be seen in Figure 3, the performance of the SITA policy that balances the performance is better than the performance of SITA-E and of a single server. Besides, the difference between SITA-E and the SITA policy that balances performance do not vary with the load of the system.

We now study compare these policies when the number of servers is higher than two. In Figure 4 and Figure 5, we consider the mean queue length of a system with four and eight servers, respectively, where the parameters are the same as in Figure 1. We observe that,

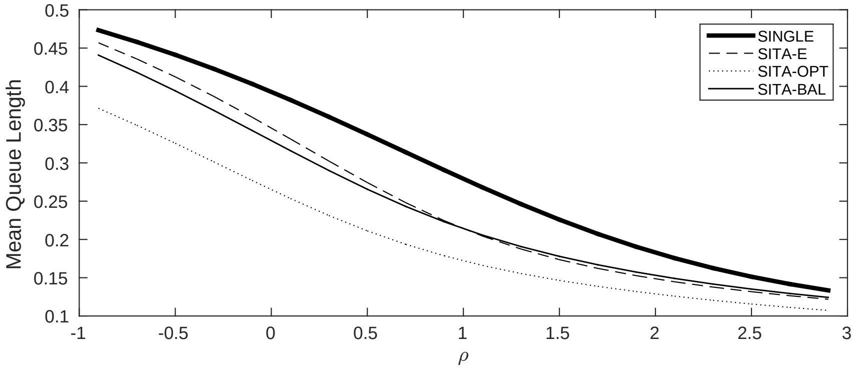

Figure 4: Mean queue length comparison in system with four servers and Bounded Pareto distributed job sizes varying α from -1 to 3. $x_{min} = 1$, $x_{max} = 10$ and $\rho = 0.5$.

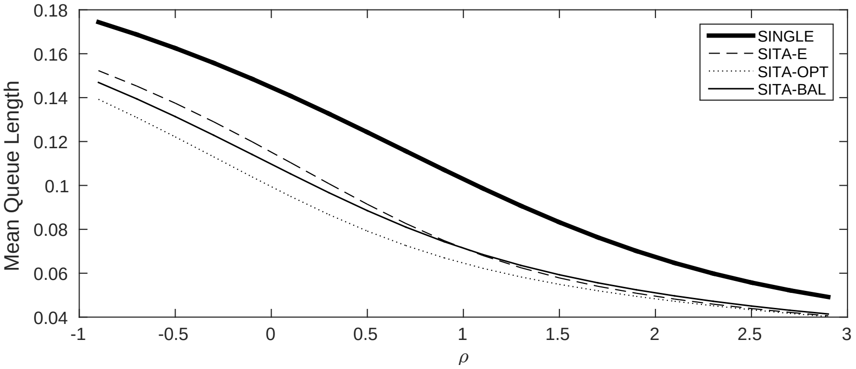

Figure 5: Mean queue length comparison in system with eight servers and Bounded Pareto distributed job sizes varying α from -1 to 3. $x_{min} = 1$, $x_{max} = 10$ and $\rho = 0.5$.

in both cases, SITA-BAL outperforms SITA-E when α is less than one and vice versa when α is higher than one. Regarding the optimality of SITA-BAL, we see that when there are eight servers, the performance of SITA-BAL is very close to the optimal one.

6.2 Mean Slowdown

The slowdown is defined as the waiting time divided by the job size. We refer to [6] for an analysis of the mean slowdown with the optimal SITA policy and Bounded Pareto distribution. Here, we consider the mean slowdown of FCFS servers for the SITA policy that balances the mean response time. We first study the mean slowdown of the system as a function of the parameter α of the Bounded Pareto distribution and then as a function of the maximum job size.

In Figure 6, we plot the evolution of the mean slowdown over the parameter α of the Bounded Pareto distribution, when the size of the smaller and the largest job is respectively 1 and 10 and $\rho = 0.5$ and the system is formed by two servers. We first observe that the three SITA policies under consideration in this plot are symmetric with respect to $\alpha = 1$. Besides, when we consider the mean slowdown, the SITA policy that balances the mean

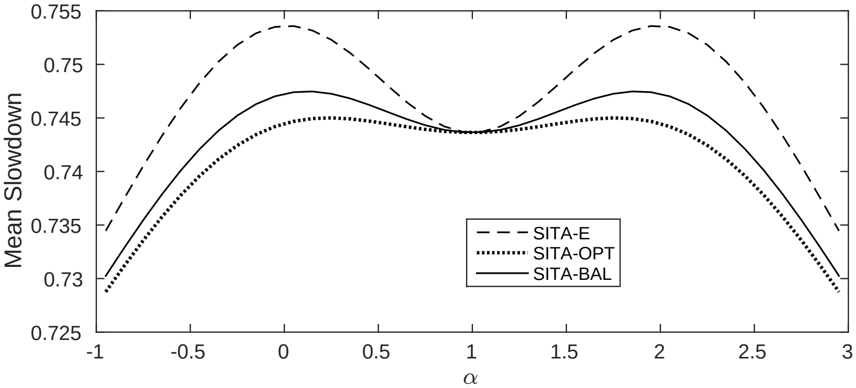

Figure 6: Mean slowdown comparison in a two-server system with Bounded Pareto distributed job sizes with $x_{min} = 1$, $x_{max} = 10$, $\rho = 0.5$ and α varying from -1 to 3.

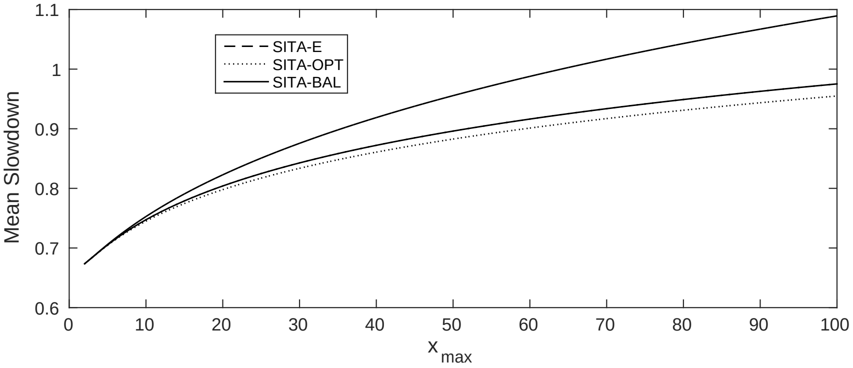

Figure 7: Mean slowdown comparison in a two-server system with Bounded Pareto distributed job sizes with $x_{min} = 1$, $\alpha = 0.2$, $\rho = 0.5$ and x_{max} varying from 2 to 100.

delay outperforms the SITA-E policy, is optimal when α is one and is almost optimal when α is close to -1 and 3. We also investigate the influence of x_{max} on the performance difference of the three SITA policies we consider in this section. In Figure 7, we consider $\alpha = 0.2$, $x_{min} = 1$, $\rho = 0.5$ and x_{max} varying from 2 to 100 and we observe that the SITA policy that balances the delay of the servers is very close the optimal performance when the difference between x_{min} and x_{max} increases, whereas it is not the case for the SITA-E policy.

7 Heterogeneous servers

We study the SITA policy that balances the performance of the servers when the system is formed by K heterogeneous servers and let μ_i be the service rate server i. For this instance, the performance of server i that executes jobs ranging in size from x_{i-1} and x_i is given by

$P(x_{i-1}, x_i, \mu_i)$ and, thus, the goal is to find a set of thresholds such that

$$c_1 P(x_0, x_1, \mu_1) = c_2 P(x_1, x_2, \mu_2) = \cdots = c_K P(x_{K-1}, x_K, \mu_K),$$

where c_j is the holding cost of server j.

First, we focus on the existence and uniqueness result presented in Section 3 and we explain how it can be adapted for heterogeneous systems. For any $w < y$ and any μ_i and μ_{i+1}, we define $\nu(w, x, \mu_i, y, \mu_{i+1}) = P(w, x, \mu_i) - P(x, y, \mu_{i+1})$. We give the following conditions:

(A3) $\nu(w, x, \mu_i, y, \mu_{i+1})$ is continuous and increasing on x,

(A4) $\nu(w, w, \mu_i, y, \mu_{i+1}) < 0 < \nu(w, y, \mu_i, y, \mu_{i+1})$.

Now, we note that the same arguments of the proof of Proposition 2 prove the existence and uniqueness of the thresholds that balance the performance in heterogeneous systems.

Regarding the performance analysis of the SITA policy that balances the performance of the servers, we consider that the performance of server with service rate μ_i that executes jobs in the interval $(x_{i-1}, x_i]$ is

$$P(x_{i-1}, x_i) = \frac{\sum_{i=1}^{n} (q_i(x_i, \lambda, \mu_i) - q_i(x_{i-1}, \lambda, \mu_i))}{r_0 + \sum_{i=1}^{m} (r_i(x_i, \lambda, \mu_i) - r_i(x_{i-1}, \lambda, \mu_i))},$$

where q_i and r_i are continuous functions from $[x_{min}, x_{max}] \times \mathbb{R}^+ \times \mathbb{R}^+$ to $\mathbb{R}^+$ such that $q_i(x_{min}) = 0$ for all $i = 1, \ldots, n$ and $r_i(x_{min}) = 0$ for all $i = 1, \ldots, m$. Using the same techniques as in Section 4, we can show that the performance of a system under the SITA policy that balances the performance of the servers is given by

$$\frac{\sum_{i=1}^{n} q_i(x_{max}, \lambda)}{r_0 \left(\sum_{i=1}^{K} \mu_i \right) + \sum_{i=1}^{m} r_i(x_{max}, \lambda)}.$$

We also extend the result of Section 5 to a system with heterogeneous servers. Indeed, one can easily show that the performance of SITA-E and of the SITA policy that balances the performance coincide when the performance of the queue that executes jobs ranging in size from x_{i-1} to x_i is

$$\frac{\sum_{i=1}^{n} (s_i(\frac{\lambda}{\mu_i} \int_{x_{min}}^{x_i} x f(x) dx) - s_i(\frac{\lambda}{\mu_i} \int_{x_{min}}^{x_{i-1}} x f(x) dx)}{r_0 + \sum_{i=1}^{m} (s_i(\frac{\lambda}{\mu_i} \int_{x_{min}}^{x_i} x f(x) dx) - s_i(\frac{\lambda}{\mu_i} \int_{x_{min}}^{x_{i-1}} x f(x) dx))},$$

8 Conclusions

In this work, we have investigated a queueing system formed by a dispatcher that sends the incoming traffic to K parallel queues. The dispatcher knows the size of each job, and carries out a SITA policy, which is a size-based load balancing, where the service time distribution of jobs is divided in intervals and all the jobs with size ranging in the same interval are sent to the same server. This routing policy has several advantages with respect to other routing policies. In particular, there is no communication requirement between the queues and the dispatcher. We have focused a SITA policy that, instead of searching the thresholds that optimize the performance of the system, seeks to find the thresholds which equalize the performance (e.g., mean waiting time of jobs or mean queue lengths) of all queues

We first study the existence and uniqueness of the thresholds of the SITA policy that balances the performance of the queues. We consider a large family of performance functions and we provide conditions on the performance function for the existence and uniqueness of the allocation thresholds. We then consider instances where the performance is of the form of (1) and, for this case, we characterize the performance of the system that implements the SITA policy that balances performance. Besides, for the mean waiting time of FCFS servers and mean number of customers of PS servers, we compare the performance of the SITA policy that balances performance with the performance of SITA-E. For the former case, we show that SITA-E outperforms the SITA policy that balances performance, whereas for the latter case, we show that the performance of both policies coincide. Finally, we present numerical experiments that show that SITA that balances performance outperforms SITA-E and it is almost optimal.

These results suggest that, when we aim to optimize a given performance metric (say P_1), a system that balances this performance metric might achieve an almost optimal performance and, therefore, we do not need to consider another performance metric (say P_2). As a future research, we think that an interesting extension of this work would be an asymptotic work for a system with a large number of servers. For this case, the range of the interval of job sizes that each server handles tends to zero, which might lead to a simplification of the problem we consider in this article. Another possible direction for future research is to extend the family of functions where the performance of a system implementing the SITA policy that balances performance can be characterized.

9 Acknowledgments

The authors are grateful to the anonymous referees for valuables suggestions that helped to improve the presentation of the paper. This work has been partially funded by the Project of the Spanish Ministry of Economy and Competitiveness with reference MTM2016-76329-R, the Marie Sklodowska-Curie grant agreement No 777778 and the Consolidated Research Group MATHMODE (IT1294-19).

References

[1] M. Abidini, O. Boxma, and J. Doncel. Size-based routing to balance performance of the queues. In *Proceedings of Valuetools*, 2017.

[2] J. Anselmi and J. Doncel. Asymptotically optimal size-interval task assignments. *IEEE Transactions on Parallel and Distributed Systems*, 30(11):2422–2433, 2019.

[3] E. Bachmat. *Mathematical adventures in performance analysis.* Springer-Birkhauser, 2014.

[4] E. Bachmat, J. Doncel, and H. Sarfati. Performance and stability analysis of the task assignment based on guessing size routing policy. In *2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 1–13. IEEE, 2019.

[5] E. Bachmat and A. Natanzon. Analysis of sita queues with many servers and spacetime geometry. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):92–94, 2012.

[6] E. Bachmat and H. Sarfati. Analysis of SITA policies. *Performance Evaluation*, 67(2):102–120, 2010.

[7] G. Ciardo, A. Riska, and E. Smirni. Equiload: a load balancing policy for clustered web servers. *Performance Evaluation*, 46(2), 2001.

[8] J. Doncel, S. Aalto, and U. Ayesta. Economies of scale in parallel-server systems. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017.

[9] J. Doncel, S. Aalto, and U. Ayesta. Performance degradation in parallel-server systems. *IEEE/ACM Transactions on Networking*, 27(2):875–888, 2019.

[10] H. Feng, V. Misra, and D. Rubenstein. Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems. *Performance Evaluation*, 62(1-4):475–492, Oct. 2005.

[11] M. Harchol-Balter. Task assignment with unknown duration. In *International Conference on Distributed Computing Systems*, 2000.

[12] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge Univ. Press, 2013.

[13] M. Harchol-Balter, M. Crovella, and C. Murta. Task assignment in a distributed system: Improving performance by load unbalancing. In *Proceedings of SIGMETRICS*, 1998.

[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*, 59(2):204 – 228, 1999.

[15] M. Harchol-Balter, A. Scheller-Wolf, and A. R. Young. Surprising results on task assignment in server farms with high-variability workloads. In *Proceedings of SIGMETRICS*, 2009.

[16] M. Harchol-Balter and R. Vesilo. To balance or unbalance load in size-interval task allocation. *Probability in the Engineering and Informational Sciences*, 24(2):219–244, Apr. 2010.

[17] F. Semchedine, L. Bouallouche-Medjkoune, and D. Aissani. Task assignment policies in distributed server systems: A survey. *Journal of Network and Computer Applications*, 34(4):1123 – 1130, 2011.

[18] R. Vesilo. Asymptotic analysis of load distribution for size-interval task allocation with bounded pareto job sizes. In *IEEE International Conference on Parallel and Distributed Systems.*, 2008.

[19] R. R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15(2):406–413, 1978.

[20] A. Williams, M. Arlitt, C. Williamson, and K. Barker. *Web Workload Characterization: Ten Years Later*, pages 3–21. Springer US, Boston, MA, 2005.

[21] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14(1):181–189, 1977.