Size-based Routing to Balance Performance of the Queues

M. Abidini
Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

O. Boxma
Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

J. Doncel
Department of Applied Mathematics and Statistics and Operational Research, University of the Basque Country, Spain

ABSTRACT

We study a queueing system with a Poisson arrival process, in which a dispatcher sends the jobs to K homogeneous queues. The dispatcher knows the size of each job, and can implement a size-aware policy. Instead of trying to optimize system performance, we propose a Size Interval Task Assignment (SITA) policy that aims to equalize the performance (mean waiting times, or mean queue lengths) of all queues by allocating the jobs to the queues according to size. Such SITA routing requires no communication between the servers and the dispatcher, and is hence easily implemented.

We study existence and uniqueness of the allocation thresholds. For FCFS and PS queues in heavy traffic, those thresholds coincide with those of a dispatching rule, SITA-E, in which loads are balanced. Preliminary numerical studies suggest that a SITA dispatching policy that equalizes performance is close to optimal when the difference between the size of the largest and the smallest job is small.

CCS CONCEPTS

• **Networks** → **Network performance modeling**; **Network performance analysis**; *Data center networks*;

ACM Reference format:
M. Abidini, O. Boxma, and J. Doncel. 2017. Size-based Routing to Balance Performance of the Queues. In *Proceedings of 11th EAI International Conference on Performance Evaluation Methodologies and Tools, Venice, Italy, December 5–7, 2017 (VALUETOOLS 2017),* 8 pages.

VALUETOOLS 2017, December 5–7, 2017, Venice, Italy

ACM ISBN 978-1-4503-6346-4/17/12…$15.00
https://doi.org/10.1145/3150928.3150948

https://doi.org/10.1145/3150928.3150948

1 INTRODUCTION

We study a queueing system with Poisson arrivals of jobs and a single dispatcher that handles all the incoming traffic and sends the jobs to K homogeneous queues. In some situations the dispatcher is not able to observe the queue lengths, or it is too expensive to provide this information regularly. In such cases the dispatcher might assign jobs to queues in a round-robin fashion (thus reducing interarrival time variance), or – if the sizes of jobs are known – the dispatcher might use a size-based policy to assign jobs to queues, thus reducing service time variance.

In this study we focus on the latter case and we assume that the dispatcher implements a Size Interval Task Assignment (SITA) routing, that is, a size-aware policy where the service times are divided into intervals and all the jobs with size in a given interval are dispatched to the same queue. Roughly speaking, we assume that long jobs and short jobs are executed by different servers, which, when we consider First-Come-First-Served (FCFS) queues, leads to a performance improvement in the system in comparison with other popular routing policies such as Bernoulli or Round Robin.

Motivated by the application of this model in data center analysis, most of the literature in this area studies load balancing schemes to minimize the expected waiting time of incoming jobs. However, there are instances where, rather than optimizing, it might be more convienient to *equalize* the performance of all the queues. Indeed, parallel queues often represent humans providing a certain service, such as in a supermarket or in a bank, and the system planner might be interested in equalizing the performance of the workers so as to implement a fair policy among them.

In this work, we study a SITA policy that balances the performance in the queues. In other words, we seek to analyze how the sizes of incoming jobs must be divided in intervals so as to equalize the performance in all the queues.

The main contributions of this work are summarized in Table 1. In a system with an arbitrary number of FCFS queues,

	SITA policy	Result	Number of Queues
FCFS	Balancing Mean Waiting Time	Existence and Uniqueness	Arbitrary
	Balancing Mean Queue Length	Existence and Uniqueness	Arbitrary
PS	Balancing Mean Response Time	Existence and Uniqueness	K=2
	Balancing Mean Queue Length	Existence, Uniqueness and Characterization	Arbitrary

Table 1: Summary of the main contributions of this paper.

we show the existence and uniqueness of the thresholds of the SITA routing that balance the mean waiting times of jobs and that balance the mean queue lengths. In a system with two processor sharing (PS) queues, we give necessary conditions for the existence and uniqueness of the threshold of the SITA routing that balances the mean response times of jobs. Finally, we show that in a system with an arbitrary number of PS queues, the SITA routing that balances the mean queue lengths coincides with the SITA routing that equalizes the loads of the queues. Therefore, we show that, for this case, the thresholds are unique and we characterize them.

In the analytical part of this work, we consider a general service time distribution. To assess the quality of the SITA routing policies, we have performed numerical experiments for Bounded Pareto distributed job sizes and we have observed that the performance of the SITA routing that balances performance is worse than that of the SITA routing that equalizes the load of the servers. However, the SITA routing that balances performance performs almost optimally when the size of the smallest and the largest job is similar.

The dispatching policy we present in this paper performs the load balancing task using information about the sizes of the incoming tasks. We believe that this type of policy is interesting from a practical point of view since it removes the need for synchronization in the central queue. Besides, under these size-based routing policies, the performance of all queues is equal and, as a result, we only need to compute the performance of one queue to obtain the performance of the system. To the best of our knowledge, the analysis of size-based routing policies aiming to equalize the performance of the queues has not been performed before.

The rest of the paper is organized as follows. In Section 2 we mention related work and in Section 3 we describe the model. FCFS queues are studied in Section 4 and PS queues in Section 5. We compare the performance of our routing policy with that of other size-based load balancing schemes in Section 6. We present the conclusions of this work in Section 7.

2 RELATED WORK

How to balance the load in a system formed by a single dispatcher to a set of parallel queues has attracted the attention of researchers for many years, see, e.g., the survey [14] and the book [8]. A popular load balancing scheme is Join-the-Shortest-Queue [4, 5] where the dispatcher sends an arriving job to the queue with the least number of customers. The Power of Two [12, 13] routing policy is another important routing policy; here, for every arriving job, two servers are picked uniformly at random and Join-the-Shortest-Queue is applied to those two servers.

Size-based load balancing policies have also been widely investigated in the literature. In this type of policy, each host serves jobs whose service demand is in a designated range. Interestingly, it has been shown in [3] that if the job size distribution of arriving tasks is known and the servers are FCFS, the thresholds of the SITA policy can be chosen so as to optimize the performance of the system. This routing policy has been studied for Bounded Pareto distributed job sizes in [1, 15]. The authors in [10] compare the performance of the SITA policy with optimal thresholds with that of the Least-Work-Left policy when the variability of job service times is very high. A two-server system is considered in [11], where the authors provide conditions regarding the direction in which the load should be unbalanced in order to optimize the performance; they study the particular case of Bounded Pareto distributed job sizes.

Researchers have also been interested in studying size-based routing policies that equalize the loads of the queues. This type of routing policy has been introduced in [6, 9] and is known in the literature as SITA-E. Variations of this load balancing scheme have been studied in [2, 7].

3 MODEL DESCRIPTION

We analyze a system of K parallel queues with equal capacity and a single dispatcher. We assume that service times of incoming jobs form an i.i.d. sequence with a common distribution; X denotes a generic service time. 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_m and x_M the minimum and maximum size of the incoming jobs to the system.

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 random variable of 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 introduce

$$G(x) = \int_{x_m}^{x} yf(y)dy$$

and

$$H(x) = \int_{x_m}^{x} y^2 f(y)dy.$$

Note that $G(x)$ and $H(x)$ are increasing with x and they satisfy the following properties: $G(x_m) = H(x_m) = 0$, $G(x_M) = \mathbb{E}(X)$ and $H(x_M) = \mathbb{E}(X^2)$.

3.1 SITA Routing Policy

We now present the necessary background about SITA routing policies required for this work. In the SITA policy, there are $K+1$ thresholds that are denoted by $x_0, \ldots, x_K$ with $x_m = x_0 < x_1 < \cdots < x_{K-1} < x_K = x_M$. Jobs ranging in size from x_{i-1} to x_i are executed in queue i.

From this general definition, the following properties arise directly. On the one hand, under the SITA policy, we know that $\lambda_i = \lambda(F(x_i) - F(x_{i-1}))$. On the other hand, using conditioning, we have that

$$\mathbb{E}(X_i) = \frac{G(x_i) - G(x_{i-1})}{F(x_i) - F(x_{i-1})}$$

and

$$\mathbb{E}(X_i^2) = \frac{H(x_i) - H(x_{i-1})}{F(x_i) - F(x_{i-1})}.$$

In this work, we aim to equalize the performance of the queues and therefore we assume that the thresholds are chosen so as to achieve this goal. Let $W_i(x_{i-1}, x_i)$ be the waiting time of jobs executed by server i and $Q_i(x_{i-1}, x_i)$ the queue length at server i. Our performance measures are the mean waiting time of jobs and the mean queue length. When we investigate the former, the goal is to choose the thresholds $x_1, \ldots, x_{K-1}$ such that

$$\mathbb{E}(W_1(x_m, x_1)) = \mathbb{E}(W_2(x_1, x_2)) = \cdots = \mathbb{E}(W_K(x_{K-1}, x_M)), \tag{1}$$

whereas when we focus on the latter, the thresholds satisfy

$$\mathbb{E}(Q_1(x_m, x_1)) = \mathbb{E}(Q_2(x_1, x_2)) = \cdots = \mathbb{E}(Q_K(x_{K-1}, x_M)). \tag{2}$$

Throughout this paper, we will also be interested in other size-based policies that we now briefly present. The thresholds of the routing policy can be chosen so as to optimize the performance of the system. We call this load balancing SITA-OPT. Unfortunately, there is no closed-form expression for the performance of this routing policy for an arbitrary distribution. We also consider the SITA-E policy, where the thresholds are chosen in order to equalize the loads in the queues. For the SITA-E policy, we know that the thresholds satisfy the following condition:

$$\int_{x_m}^{x_1} xf(x)dx = \int_{x_1}^{x_2} xf(x)dx = \cdots = \int_{x_{K-1}}^{x_M} xf(x)dx. \tag{3}$$

3.2 Pros and Cons of the SITA policy with Balanced Performance

The dispatching policy we study here is a size-based policy that balances the performance of the queues. It is clear that this routing scheme is suboptimal, but we believe that it has several advantages with respect to other routing policies, as we explain below.

- *Advantages with respect to non-size based policies.* It is known that there are some routing policies that are not size-aware while their performance is optimal. The major advantage of size-based routing policies resides in the ease of implementation. Indeed, for SITA routing, no communication between the servers and the dispatcher is required.
- *Advantages with respect to the optimal SITA policy.* From [3], we know that, if the distribution of the incoming tasks is known, the thresholds of the SITA policy can be chosen in such a way that the performance of the system is optimal. An analytical expression of the performance for a system under that routing policy seems impossible to obtain even for a system formed by two queues [11]. We propose a SITA policy that is not optimal, but the performance of the system under this routing can be obtained by computing only one of the following thresholds: x_1 or x_{K-1}. To see this, note that the performance of all the queues is the same and, to obtain the performance of the first (resp. the last) queue, one only needs to know x_m and x_1 (resp. x_{K-1} and x_M); and the values of x_m and x_M are given.

4 FCFS QUEUES

In this section, we assume that the servers operate FCFS. We first study the existence of a SITA routing that balances the mean waiting times. Then, we focus on the mean queue lengths. For both cases, we show the existence of thresholds for an arbitrary number of queues, and the uniqueness of the threshold in a model with arbitrary number of queues and generally distributed job sizes.

4.1 Balancing Mean Waiting Time

We now analyze the SITA policy that balances mean waiting times of jobs when the queues are FCFS.

We first consider a system with two servers and we aim to show there exists a unique value of x such that $E(W_1(x_m, x)) =$

$E(W_2(x, x_M))$, i.e., using the well-known Pollaczek-Khinchine formula for the mean waiting time in an $M/G/1$ queue:

$$\frac{\lambda H(x)}{2(1-\lambda G(x))} = \frac{\lambda(\mathbb{E}(X^2) - H(x))}{2(1-\lambda(\mathbb{E}(X) - G(x)))}.$$

Proposition 4.1. *In a system with two FCFS queues, the threshold of the SITA policy that balances mean waiting time is unique.*

Proof. Rearranging both sides of the previous expression and simplifying, we obtain the following equivalent expression:

$$\frac{\lambda H(x)}{1-\lambda G(x)} - \frac{\lambda \mathbb{E}(X^2)}{2-\lambda \mathbb{E}(X)} = 0. \tag{4}$$

We now show that there exists a unique value of x that satisfies the previous expression. We first show that there exists at least a value of x such that (4) is satisfied. To see this, we observe that when $x \to x_m$ the LHS of (4) is negative, whereas when $x \to x_M$ it is positive.

The unicity follows since (4) is increasing with x because $G(x)$ and $H(x)$ are. Therefore, the desired result follows. □

This result means that there exists a unique value of x such that $\mathbb{E}(W_1(x_m, x)) = \mathbb{E}(W_2(x, x_M))$. Let $x = \phi(x_m, x_M)$ be the function such that x solves the two-server problem as in Proposition 4.1. This function is continuous and increasing with x_m and x_M.

We now aim to show that the previous result can be extended to a system with more than two queues. Hence we want to prove the uniqueness of the thresholds $x_1, \ldots, x_{K-1}$ such that (1) is satisfied for FCFS queues.

Proposition 4.2. *In a system with an arbitrary number of FCFS queues, the thresholds of the SITA policy that balance mean waiting times exist and are unique.*

Proof. Using the previous arguments, it can be shown that, for all i, if x_{i-1} and x_{i+1} are fixed, there exists a unique x_i such that $x_i = \phi(x_{i-1}, x_{i+1})$. Thus, x_i can be written as

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

We apply the previous reasoning until we write x_i as a function of only x_0, x_i and x_K, that is, $x_i = h(x_0, x_i, x_K)$, where x_0 and x_K are fixed. Hence, we have a function with one variable that satisfies Brouwer's fixed point theorem since, by the implicit function theorem, ϕ is continuous. As a result, the thresholds of the SITA routing that balance mean waiting times exist. Finally, by construction of the function ϕ, the fixed point is unique and the uniqueness of the thresholds follows. □

4.2 Balancing Mean Queue Length

We now investigate the SITA policy that balances mean queue length when the queues are FCFS. Hence, we want to show the uniqueness of the values of $x_1, \ldots, x_{K-1}$ such that

$$\begin{aligned} \frac{\lambda H(x_1)F(x_1)}{1-\lambda G(x_1)} &= \frac{\lambda(F(x_2) - F(x_1))(H(x_2) - H(x_1))}{1-\lambda(G(x_2) - G(x_1))} \\ &= \ldots = \frac{\lambda(1 - F(x_{K-1}))(\mathbb{E}(X^2) - H(x_{K-1}))}{1-\lambda(\mathbb{E}(X) - G(x_{K-1}))}. \end{aligned}$$

Given the similarity of the previous expression with that of the mean waiting time and taking into account that $F(x)$ is increasing with x and also that $F(x_m) = 0$ and $F(x_M) = 1$, the same techniques of Proposition 4.1 and Proposition 4.2 can be used to prove the existence and unicity of the SITA routing that balances mean queue lengths for FCFS queues. Therefore, in view of space limitations, we omit the proof.

Proposition 4.3. *In a system with FCFS queues, for an arbitrary number of queues, the thresholds of the SITA policy that balance mean queue lengths exist and are unique.*

In this section, we have studied the SITA routing that balances the mean waiting times of jobs and the mean queue lengths. In Section 6 we analyze the performance of both routing policies for Bounded Pareto job size distribution.

5 PS QUEUES

We now assume that the servers in each queue use processor sharing, and we aim to determine how the job sizes must be divided in intervals so as to equalize the performance of all the queues. We first focus on the mean response (sojourn) time and two queues and we give conditions for the existence and the uniqueness of the threshold. We subsequently study the SITA policy to balance the mean queue lengths for an arbitrary number of servers, and we show that balancing the mean queue lengths coincides with equalizing the loads of all servers.

5.1 Balancing Mean Response Time

We focus on the SITA routing that balances the mean response times of jobs when the servers operate under the PS discipline. We use the well-known result (cf. [8]) that the mean response time in an $M/G/1$ PS queue with mean service time $\mathbb{E}(B)$ and load ρ is given by $\mathbb{E}(S) = \frac{\mathbb{E}(B)}{1-\rho}$. Let $S_i(x_{i-1}, x_i)$ be the random variable of the response time of jobs to be executed in queue i. For a system with two queues, we now require that $\mathbb{E}(S_1(x_m, x)) = \mathbb{E}(S_2(x, x_M))$, i.e.,

$$\frac{G(x)/F(x)}{1-\lambda G(x)} = \frac{(\mathbb{E}(X) - G(x))/(1-F(x))}{1-\lambda(\mathbb{E}(X) - G(x))}. \tag{5}$$

Let $\nu(x) = \mathbb{E}[S_1] - \mathbb{E}[S_2]$. We seek to find a value $x \in [x_m, x_M]$ such that $\nu(x) = 0$. Note that if $x = x_m$ the response

time at server 1 and server 2 is equal to x_m and $\frac{\mathbb{E}[X]}{1-\lambda\mathbb{E}[X]}$ respectively, which implies $v(x) < 0$ always. Further, if the allocation is such that $[x_m, x_M)$ are routed to server 1 and x_M is routed to server 2 then the response times at server 1 and server 2 are equal to $\frac{\mathbb{E}[X]}{1-\lambda\mathbb{E}[X]}$ and x_M respectively. Therefore

$$v(x) \geq 0 \iff \lambda \geq \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}.$$

Therefore, we conclude that a threshold that balances response times exists if $\lambda \geq \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$.

The mean response times of jobs in server 1 increase with x since $G(x)/F(x) = \mathbb{E}[X|X < x]$ is the mean job size in server 1, which increases with x and $\frac{1}{1-\lambda G(x)}$ is also increasing with x. Likewise, it can be shown that the mean reponse times in server 2 decrease with x. Consequently, we have that $v(x)$ is an increasing function of x. This implies that, there exists a unique $x \in [x_m, x_M]$ such that $v(x) = 0$. Besides, from this monotonicity property it follows that if $\lambda < \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$, there does not exist a threshold that balances the response times of jobs. To see this, note that if $\lambda < \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$, $v(x)$ is negative when $x = x_m$ as well as when $x = x_M$, and, since the function $v(x)$ is monotone, it does not have a root.

If we allow the dispatcher to send jobs of exactly the same size to different servers, we say that the SITA routing is a mixed strategy. We define a mixed (or probabilistic) policy as a function π that associates with each $x \in [x_m, x_M]$ a probability measure on the set of servers. Hence, we denote by $\pi_k(x)$ the fraction of jobs of size x executed by server k. When, for all $x \in [x_m, x_M]$, there exists a server k such that $\pi_k(x) = 1$, we say the SITA routing is a pure (or deterministic) strategy.

In this section, we have proven that a pure SITA policy exists for two PS queues if and only if $\lambda \geq \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$. We now show that, when $\lambda < \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$, there exists a mixed strategy and it is unique. We observe that if the allocation is such that all $x \in [x_m, x_M]$ are routed to server 1 and nothing is routed to server 2 then it results that $\mathbb{E}[S_2] = 0$ and therefore $v(x) \geq 0$, which is the condition required for the existence of a threshold. Hence, for this case, there exists the following mixed strategy that balances mean response times: the packets of sizes $[x_m, x_M)$ are routed to server 1 with probability 1, and the packets of size x_M are routed to server 1 with probability p and to server 2 with probability $1 - p$ where $p \in (0, 1)$.

Using the previous reasoning, we have the following result.

Proposition 5.1. *In a system with two PS queues,*

- *there exists a pure SITA policy that balances mean response times if and only if* $\lambda \geq \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$,
- *if* $\lambda < \frac{1}{\mathbb{E}[X]} - \frac{1}{x_M}$, *there exists a mixed strategy that balances mean response times.*

5.2 Balancing Mean Queue Length

We now study the thresholds that balance the mean queue lengths when the queues are PS, that is, we aim to obtain the values of $x_1, \ldots, x_{K-1}$ such that

$$\frac{\lambda G(x_1)}{1-\lambda G(x_1)} = \frac{\lambda(G(x_2) - G(x_1))}{1-\lambda(G(x_2) - G(x_1))} = \ldots = \frac{\lambda(\mathbb{E}(X) - G(x_{K-1}))}{1-\lambda(\mathbb{E}(X) - G(x_{K-1}))}.$$

We write the load of queue i as $\rho_i = \lambda(G(x_i) - G(x_{i-1}))$. Hence, from the previous expression, we have that for all $i \neq j$

$$\frac{\rho_i}{1-\rho_i} = \frac{\rho_j}{1-\rho_j} \iff \rho_i(1-\rho_j) = \rho_j(1-\rho_i),$$

and from the last equation, it follows that $\rho_i = \rho_j$. As a result, the thresholds that balance the mean queue lengths also equalize the loads of the servers. Therefore, we conclude that they are known and are characterized by (3).

Proposition 5.2. *In a system with $K \geq 2$ PS queues, the thresholds of the SITA routing coincide with those of the SITA-E policy, i.e., they are given by* (3).

6 COMPARISON WITH OTHER SITA POLICIES

6.1 Analytical Comparison

We first assume that the system is in heavy traffic, i.e., the load approaches K. For this case, the load of all the queues must be close to saturation, and, in particular, the queues must be equally loaded as otherwise one queue would have load higher than one, implying instability.

As a result, in the heavy traffic regime, the thresholds of the SITA routing that balance mean waiting times of jobs and the thresholds of the SITA policy that balance mean queue lengths coincide with those of the thresholds of the SITA-E and therefore they satisfy (3).

Proposition 6.1. *For PS and FCFS queues, if the system is in heavy traffic, the thresholds of the SITA policy that balances mean waiting time of jobs and the mean queue lengths coincide with those of the SITA-E.*

We now assume that all jobs have the same size and therefore size-based routing cannot be implemented. Thus, the jobs are split in such a way that each queue receives a load equal to λ/K. As a result, the mean waiting time and mean number of customers are equal in all queues. This means that any size-based routing policy satisfies that, for constant service times, the performance of all the queues is the same:

Proposition 6.2. *For PS and FCFS queues, if service times are constant, any size-based routing policy balances the mean waiting/response times of jobs and mean queue lengths.*

Unfortunately, closed-form expressions of the performance of the SITA routing that optimizes the performance of the system are unknown even for a system with two servers. As a consequence, the comparison of the performance of our routing policy with that of the optimal SITA routing seems to be very complicated to perform for the general case. In the next section, we present the numerical experiments we have carried out to analyze the optimality of the SITA routing that balances the performance of queues.

6.2 Numerical Comparison

We study a system with two servers that operate under the FCFS discipline, aiming to compare the performance of the proposed policies with other SITA policies. We assume that the job size distribution is Bounded Pareto, that is, if $x_m \leq x \leq x_M$, then

$$f(x) = \frac{\alpha\, x_m^{\alpha}}{1-(x_m/x_M)^{\alpha}}\, x^{-\alpha-1},$$

and $f(x) = 0$ otherwise, where $\alpha > 0$. The cumulative distribution function of the job sizes is

$$F(x) = \begin{cases} 0, & x \leq x_m, \\ \frac{1-(x_m/x)^{\alpha}}{1-(x_m/x_M)^{\alpha}}, & x_m \leq x \leq x_M, \\ 1, & x \geq x_M. \end{cases}$$

The values of the thresholds for SITA-E routing for Bounded Pareto distributed job sizes are given in [9]:

$$x_j = \left(\frac{K-j}{K}\, x_m^{1-\alpha} + \frac{j}{K}\, x_M^{1-\alpha}\right)^{\frac{1}{1-\alpha}},$$

if $\alpha \neq 1$ and $x_j = x_m \left(\frac{x_M}{x_m}\right)^{\frac{j}{K}}$ if $\alpha = 1$.

In the simulations we will present in this section, we compare the performance of three policies: the dashed line represents the performance of the SITA routing that balances the performance (SITA-BAL), the dotted line represents the SITA routing that optimizes the performance (SITA-OPT) and the solid line represents the SITA routing that equalizes the load of the queues (SITA-E). In all cases, we plot the mean waiting time of jobs and the mean queue length when the maximum job size varies from 2 to 100 and the minimum job size is 1.

In the first set of experiments, we explore a system at low load ($\rho = 0.2$) and we set the parameter α to 1.8. We represent the performance of the three routing policies under consideration in Figure 1. In Figure 1a, we observe that the mean waiting time under the SITA-E routing is smaller than for the SITA that balances the performance. Besides, when the largest job size is close to one, the difference between the mean waiting time of the SITA routing that balances the performance and the SITA routing that optimizes the performance is very small. As it can be seen in Figure 1b, for the mean queue length, the performance of the SITA routing that balances the performance is close to that of the optimal size-based policy.

In the second set of experiments, we consider a system at medium load ($\rho = 0.9$) and we set $\alpha = 1.1$. We illustrate the performance of the three routing policies under consideration in Figure 2. In Figure 2a, we focus on the mean waiting time of jobs and we observe that the SITA-E routing performs better than the SITA routing that balances the performance and its performance coincides with the optimal performance, which is given by the SITA-OPT policy. Besides, the difference between both policies is small when the size of the smallest and largest job is small. On the other hand, we observe in Figure 2b that the SITA routing that balances the mean queue length is very close to that of SITA-E and almost optimal when the largest and the smaller job size are similar.

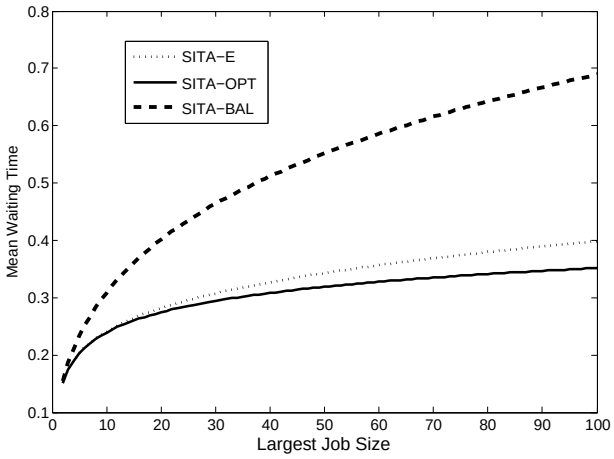
Finally, we consider a system at high load ($\rho = 1.5$) and we set $\alpha = 0.5$. We plot the performance of the three routing policies in Figure 3. We compare in Figure 3a their performance for the mean waiting time of jobs and we observe that the performance of the SITA-E routing is very close to the optimal performance and is better than that of the SITA routing that balances the mean waiting time of jobs. On the other hand, we focus on the mean queue length in Figure 3b. For this case, we observe that the performance of the SITA routing that balances the mean queue lengths is very close to that of SITA-E and the difference in the performance increases with the largest job size.

In this section, we have performed simulations to compare the performance of the SITA routing that balances the performance with that of the SITA-E routing and the SITA routing that optimizes the performance. The obtained results show that the SITA-BAL routing performs almost optimally when the size of the jobs is similar and, when the disparity of the job sizes increases, the performance of the SITA-BAL routing can be very similar to that of the SITA-E.

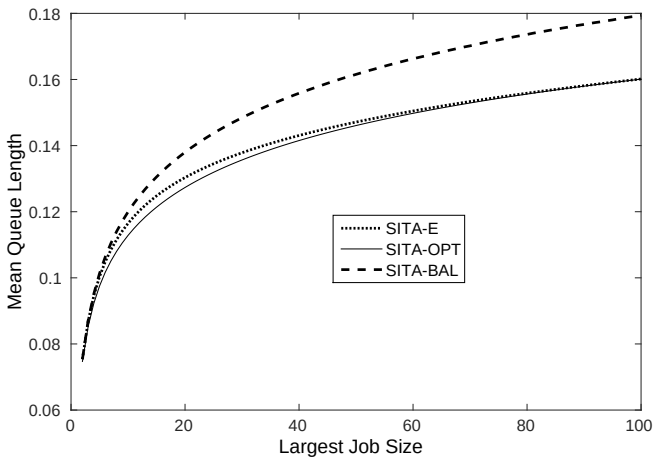
7 CONCLUSIONS

In this work, we have investigated a queueing system formed by a dispatcher that sends all the incoming Poisson traffic to K homogeneous queues. The dispatcher knows the size of each job, and carries out a size-based load balancing.

The Size Interval Task Assignment (SITA) policy is a size-based routing policy where the service time requirements of jobs are divided in intervals and all the jobs with size ranging in a given interval are sent to the same queue. We presented a SITA policy that, instead of searching the thresholds that optimize the performance of the system, seeks the thresholds which equalize the performance (mean waiting time of jobs or mean queue lengths) of all queues. One goal of this work was thus to study the existence and uniqueness of the latter thresholds for a general distribution of the incoming job sizes. For FCFS queues, we showed the existence and uniqueness

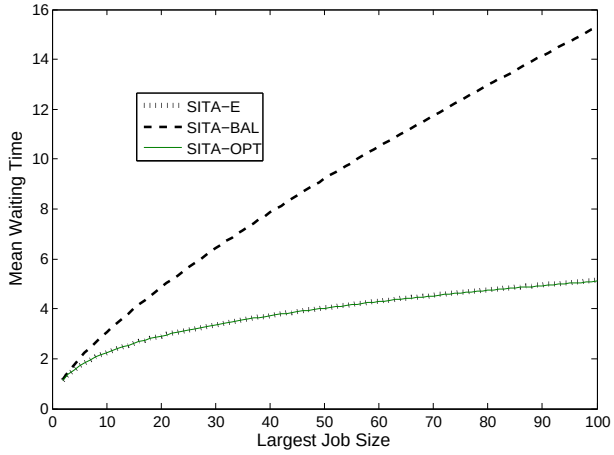


(a) Mean waiting time comparison

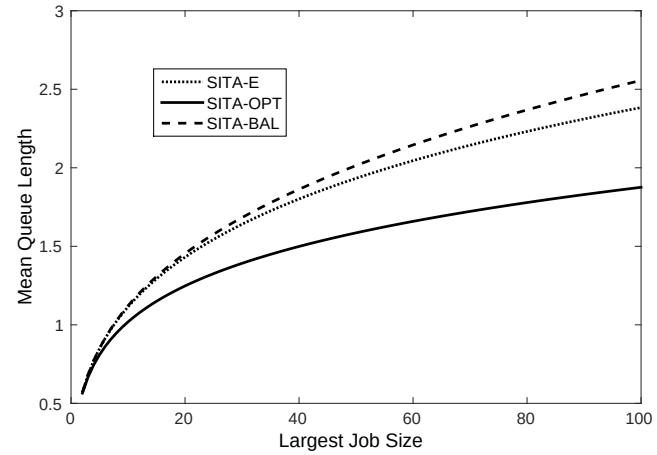


(b) Mean queue length comparison

Figure 1: Comparison of the SITA-E routing (dotted line), the SITA that optimizes the performance of the system (solid line) and the SITA routing that balances the performance (dashed line) as a function of the largest job size (x_M) when the size of the smallest job is 1, $\alpha = 1.8$ and the system load is low ($\rho = 0.2$). FCFS queue.



(a) Mean waiting time comparison.

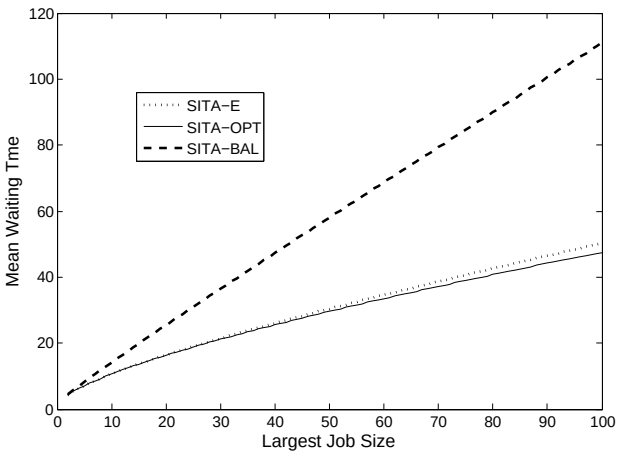


(b) Mean queue length comparison.

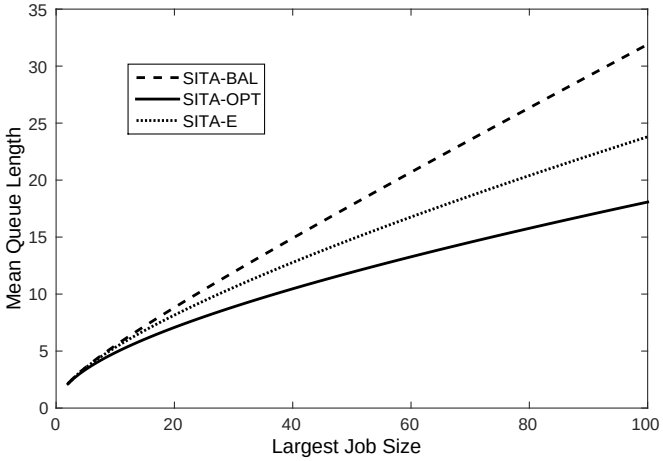
Figure 2: Comparison of the SITA-E routing (dotted line), the SITA that optimizes the performance of the system (solid line) and the SITA routing that balances the performance (dashed line) as a function of the largest job size (x_M) when the size of the smallest job is 1, $\alpha = 1.1$ and the system load is medium ($\rho = 0.9$). FCFS queue.

of these threholds in a system with an arbitrary number of queues. In a system with two PS queues, we first focused on the mean response times of jobs and we gave conditions for the existence of the threshold. Then, we concentrated on the mean queue lengths and we showed that, in a system with an arbitrary number of queues, the thresholds coincide with those that equalize the load of the system, which implies that they are unique and can be easily characterized. We have shown that the thresholds of both policies also coincide when the system is in heavy traffic.

Another goal of the study was to see how close the performance of SITA-BAL is to that of SITA-OPT. We have numerically compared the SITA dispatching policy that balances the performance of the queues with the optimal SITA routing and we have observed that, when the difference between



(a) Mean waiting time comparison.



(b) Mean queue length comparison.

Figure 3: Comparison of the SITA-E routing (dotted line), the SITA that optimizes the performance of the system (solid line) and the SITA routing that balances the performance (dashed line) as a function of the largest job size (x_M) when the size of the smallest job is 1, $\alpha = 0.5$ and the system load is high ($\rho = 1.5$). FCFS queue.

the largest and smaller job size is small, SITA-BAL performs almost as well as SITA-OPT.

The SITA routing we present here has several advantages with respect to other routing policies. In particular, there is no communication requirement between the servers and the dispatcher since the dispatcher performs the load balancing according to the size of the incoming jobs.

Several directions for further research offer themself. First, it would be interesting to consider other disciplines than FCFS or PS, and to study the existence and uniqueness of the thresholds. Another possible extension is to perform analytical, numerical and asymptotic work for a system with a large number of queues, in order to study how close to optimality SITA-BAL performs in such a system.

8 ACKNOWLEDGEMENT

The research for this paper is partly funded by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003 (Abidini, Boxma), a grant of the Belgian Government, via the IAP Bestcom Project (Boxma) and by the Basque Government through the Consolidated Research Group grant IT649-13 on "Mathematical Modeling, Simulation and Industrial Applications (M2SI)."

REFERENCES

[1] Eitan Bachmat and Hagit Sarfati. Analysis of SITA policies. *Performance Evaluation*, 67(2):102–120, 2010.

[2] Gianfranco Ciardo, Alma Riska, and Evgenia Smirni. Equiload: a load balancing policy for clustered web servers. *Performance Evaluation*, 46(2), 2001.

[3] Hanhua Feng, Vishal Misra, and Dan Rubenstein. Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems. *Performance Evaluation*, 62(1-4):475–492, October 2005.

[4] R. D. Foley and D. R. McDonald. Join the shortest queue: Stability and exact asymptotics. *Annals of Applied Probab.*, 11(3), 2001.

[5] Varun Gupta, Mor Harchol-Balter, Karl Sigman, and Ward Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

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

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

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

[9] Mor Harchol-Balter, Mark E. Crovella, and Cristina D. Murta. On choosing a task assignment policy for a distributed server system. *Journal of Parallel and Distributed Computing*, 59(2):204 – 228, 1999.

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

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

[12] Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Trans. on Parallel and Distributed Sys.*, 12(10), 2001.

[13] Andrea W Richa, M Mitzenmacher, and R Sitaraman. The power of two random choices: A survey of techniques and results. *Handbook of Randomized Computing*, 1, 2001.

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

[15] Rein 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.