sensors

MDPI

Article

Size-Based Routing Policies: Non-Asymptotic Analysis and Design of Decentralized Systems †

Eitan Bachmat [1] and Josu Doncel [2,*]

[1] Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105, Israel; ebachmat@bgu.ac.il
[2] Mathematics Department, University of the Basque Country, UPV/EHU, 48940 Leioa, Spain
* Correspondence: josu.doncel@ehu.eus

† This paper is an extended version of our paper published in Bachmat, E.; Doncel, J. Non-Asymptotic Performance Analysis of Size-Based Routing Policies. In Proceedings of the 28th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Nice, France, 17–19 November 2020.

Abstract: Size-based routing policies are known to perform well when the variance of the distribution of the job size is very high. We consider two size-based policies in this paper: Task Assignment with Guessing Size (TAGS) and Size Interval Task Assignment (SITA). The latter assumes that the size of jobs is known, whereas the former does not. Recently, it has been shown by our previous work that when the ratio of the largest to shortest job tends to infinity and the system load is fixed and low, the average waiting time of SITA is, at most, two times less than that of TAGS. In this article, we first analyze the ratio between the mean waiting time of TAGS and the mean waiting time of SITA in a non-asymptotic regime, and we show that for two servers, and when the job size distribution is Bounded Pareto with parameter $\alpha = 1$, this ratio is unbounded from above. We then consider a system with an arbitrary number of servers and we compare the mean waiting time of TAGS with that of Size Interval Task Assignment with Equal load (SITA-E), which is a SITA policy where the load of all the servers are equal. We show that in the light traffic regime, the performance ratio under consideration is unbounded from above when (i) the job size distribution is Bounded Pareto with parameter $\alpha = 1$ and an arbitrary number of servers as well as (ii) for Bounded Pareto distributed job sizes with $\alpha \in (0,2)\backslash\{1\}$ and the number of servers tends to infinity. Finally, we use the result of our previous work to show how to design decentralized systems with quality of service constraints.

Keywords: parallel servers; size-based routing; queuing theory

Citation: Bachmat, E.; Doncel, J. Size-Based Routing Policies: Non-Asymptotic Analysis and Design of Decentralized Systems †. *Sensors* **2021**, *21*, 2701. https://doi.org/10.3390/s21082701

Academic Editors: Erol Gelenbe and Maria Carla Calzarossa

Received: 8 March 2021
Accepted: 8 April 2021
Published: 12 April 2021

1. Introduction

We explore the performance of a system composed of parallel servers in which the job size distribution satisfies the heavy-tailed property. This property means that the arrival of extremely large jobs occurs with a non-negligible probability and, therefore, a very small portion of the jobs account for half of the system load. For instance, jobs submitted to modern data centers follow a heavy-tailed distribution [1]. Consequently, if we model the servers of data centers as First-Come-First-Served (FCFS) queues, the presence of very long jobs might cause a substantial performance degradation as many short jobs must wait behind a large job for a long time. This problem is solved by using size-based routing policies, in which job sizes are divided into intervals. We consider that two popular size-based routing policies, the Size Interval Task Assignment (SITA), which designates an interval to each server and jobs whose size are in the interval, are executed on the corresponding server. The second size-based routing policy is called Task Assignment with Guessing Size (TAGS). Under this policy, all incoming jobs are routed to server 1 and, for each server i, if its execution time in server i time exceeds a given server-assigned threshold s_i, it is stopped and placed in the last position of the queue of server $i+1$. As SITA assumes that the size of tasks is known and routes directly to the end server, while TAGS incurs

overheads, it is clear that SITA outperforms TAGS. An important result comparing the performance of these two policies [2] states that in the asymptotic regime, where the ratio of the largest to shortest job tends to infinity and the system load is fixed and low, the average waiting time of SITA with optimal intervals is, at most, two times that of TAGS with optimal intervals.

The main contributions of this article are summarized as follows.

- We provide a non-asymptotic comparison of the mean waiting time of SITA and TAGS. We analyze a system with two servers and a job size distribution that follows the Bounded Pareto distribution with the tail parameter one. We provide a job size distribution-dependent lower bound on the ratio of average waiting times between the TAGS and SITA systems at very low rates. We compute a good estimate for the bound when the job size is an arbitrary Bounded Pareto distribution and show that unlike the case of a fixed and low load, when the load is low but not fixed, the performance ratio is unbounded. We also analyze numerically this performance ratio for bounded Pareto distributions with different tail parameters.
- For any number of servers $K \geq 2$, we compare the performance of the TAGS system with the Size Interval Task Assignment with Equal load (SITA-E) system, i.e., with a SITA policy in which the thresholds are set to load balance the servers. We analyze this system in light traffic, and we show that in this regime, asymptotically, the average waiting times of TAGS with optimal intervals and of a single server coincide. Consequently, we can use the results of [3] in which they compare the performance of SITA-E and a single server and conclude that the ratio of the average waiting time of the TAGS and SITA-E system is unbounded in a wide range of cases: (i) when the job size distribution Bounded Pareto with $\alpha = 1$ and $K \geq 2$, and (ii) when $K \to \infty$ and the job size distribution is Bounded Pareto with $\alpha \in (0,2)\backslash\{1\}$.

The stability condition for TAGS systems [2] says that the arrival rate times and the mean job size must be less than a critical load that depends on the job size distribution. Let us note that this stability condition does not depend on the number of servers present in the system. Therefore, if we increase the number of servers and the arrival rate proportionally in a TAGS system, the system will become unstable, which is not the case for other routing policies such as Bernoulli or SITA [4]. Therefore, for the design of a system with a large number of servers it is preferable to consider a decentralized system, as in [3], in which the system is divided into n load balanced and equal sized groups which can operate under either SITA or TAGS. The goal is to elucidate the maximum number of groups that can operate under TAGS so as to satisfy a quality of service constraint that states that the average waiting time of the system does not exceed a given threshold. As the performance of TAGS and SITA is very difficult to analyze in a precise manner, we focus on a suboptimal, yet reasonable solution which uses the results of [2].

A conference version of this paper appeared in [5].

The rest of the article is organized as follows. In Section 2, we present the related work. In Section 3, we describe the model we study. In Section 4, we perform the non-asymptotic comparison of the performance of TAGS and SITA for two queues, and in Section 5 we consider asymptotically an arbitrary number of queues. In Section 6, we explore how to design a decentralized system with our quality of service constraints. We discuss the results and conclude in Section 7.

2. Related Work

Modern communication systems integrate distributed computing architectures, in which jobs are processed in parallel. Therefore, many researchers have studied how to optimally balance the load in a system with parallel servers, see in [6]. Most of the analysis assumes that the state of the servers is always known. A large class of policies which utilizes this assumption is known as the SQ(d) framework. Policies in this family operate as follows: for each incoming job, $d \geq 2$ servers are picked uniformly at random and their states are observed. The job is routed to the server whose state (number of packets or

the workload, for instance) is minimal. This family of routing policies has been studied extensively as such policies often perform well [7–11]. However, it is important to note that in [12], it was shown that when the variance of the job size distribution is very large, SQ(d) policies are not optimal. The size-based routing policies which we consider often have good performance under precisely such circumstances.

Size-based routing splits the service time distribution into intervals. By using these intervals, when the servers are FCFS, short jobs do not wait for a long time behind a long job. This is a clear advantage with respect to other routing policies that have been studied in the literature such as Bernoulli routing, for instance.

The first size-based routing policy presented in the literature was the SITA policy [13]. Under this policy, each host receives jobs whose job size is in a designated range [13]. This implies a reduction in the variability of the jobs executed in the servers. The authors of [14] provide an interesting result that shows that there exists a SITA policy (choice of intervals) that minimizes mean response time when the state of the servers cannot be observed; the servers scheduling discipline is FCFS and the size of jobs is available. The analytical expression for the optimal thresholds of a SITA policy is not known in general, even for a system with two servers. Therefore, some authors have studied alternatives such as the SITA-E policy, where the intervals are chosen to equalize the load in all the servers [3,15]. Asymptotic analysis of the optimal thresholds of a SITA policy for Bounded Pareto distributions has been carried out in [16,17] and for a large number of servers in [4]. The performance of SITA policies with two servers has been studied in [18], where the authors provide conditions under which the load should be unbalanced to minimize system performance. In [19], the authors compare the SITA policy with that of Least-Work-Left when the coefficient of variation of job sizes is large and show that the Least-Work-Left policy outperforms SITA in several scenarios.

The TAGS policy was introduced in [20] as a size-based routing policy in the setting where the sizes of incoming jobs does not need to be known. The authors of [2,21] prove that for Bounded Pareto distributions, when the ratio between the largest job size and the shortest job size tends to infinity and the system load is fixed and less than one, the ratio between the average waiting times of TAGS and SITA systems with optimal intervals is at most two. This result means that, in that regime, the penalty for not knowing the job size of incoming tasks is, at most, 2. In this work, we address a similar question but in a non-asymptotic regime.

3. Model Description

3.1. Notations

We investigate a parallel server system with K homogeneous servers. The servers are modeled as FCFS queues. Jobs arrive to the system following a Poisson distribution with rate λ. The job size is assumed to be given by a sequence of i.i.d. random variables, whose distribution we denote by X. We denote by $\mathbb{E}[X]$ the average job size and by $F(s) = \mathbb{P}[X < s]$ the cumulative distribution function of the job size distribution. We assume that $F(\cdot)$ is differentiable, and we let $f(s) = \frac{dF(s)}{ds}$ be the density function of the job size distribution. We assume that the size of the smallest job is equal to one and of the largest job is r, where $r > 1$. The system load is denoted by $\rho = \lambda\mathbb{E}[X]$.

3.2. TAGS Routing

We consider the TAGS routing policy. We assume that the servers of the system are labeled from 1 to K. Let $s_0 = 1$ and $s_K = r$. The set of parameters of the policy (intervals) is given by a vector of $K-1$ cut-off values $\mathbf{s} = (s_1, s_2, \ldots, s_{K-1})$ verifying that $x_S = s_0 < s_1 < s_2 < \cdots < s_{K-1} < s_K = x_L$.

All jobs are sent to server 1. If the job completes before s_1 time units, the job leaves the system; otherwise, the job is stopped after a runtime of s_1 and is sent to the end of the buffer of server 2, where execution starts from scratch and the process repeats. Thus, for server i, jobs that are executed in that server have already been executed in servers

$1, 2, \ldots, i-2$ and $i-1$ for $s_1, \ldots, s_{i-1}$ time units, respectively. If a job at server i completes before s_i time units, it leaves the system, and if not, it is terminated after s_i units of time and placed at the last position of the buffer of the next server. Finally, at the last server, K, jobs always run to completion.

We denote by $W^{TAGS}(\mathbf{s})$ the waiting time of jobs for TAGS routing with vector of cut-offs $\mathbf{s}$. In this work, we denote by $\mathbf{s}_T$ the optimal vector of cut-offs,

$$\mathbf{s}_T = \arg\min_{\mathbf{s}} \mathbb{E}[W^{TAGS}(\mathbf{s})].$$

3.3. SITA Routing

We consider the SITA routing policy. Let $s_0 = 1$ and $s_K = r$. Under the SITA policy, the servers are also labeled from 1 to K and, as in the TAGS policy, the set of parameters is a vector of thresholds $\mathbf{s} = (s_1, s_2, \ldots, s_{K-1})$ satisfying $x_S = s_0 < s_1 < s_2 < \cdots < s_{K-1} < s_K = x_L$. However, the SITA policy uses the job size of incoming tasks to perform the routing. Therefore, jobs ranging in size between s_{i-1} and s_i are sent directly to the ith server where they complete service.

We denote by $W^{SITA}(\mathbf{s})$ the waiting time of tasks when the system operates the SITA routing with the vector of thresholds $\mathbf{s}$. The optimal vector of the cut-offs is denoted by $\mathbf{s}^*$, i.e.,

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \mathbb{E}[W^{SITA}(\mathbf{s})].$$

We will sometimes be interested in a SITA policy whose thresholds $s_1, \ldots, s_{K-1}$ satisfy that the load of all the servers is the same, i.e.,

$$\int_{s_{i-1}}^{s_i} x f(x) dx = \int_{s_i}^{s_{i+1}} x f(x) dx,$$

for all $i = 2, \ldots, K-1$. This policy is called SITA-E, and in this work, we denote by $\mathbb{E}[W^{SITA-E}]$ the mean waiting time of incoming jobs under this policy.

Remark 1 (Advantages and Disadvantages of TAGS and SITA)**.** *Before going further, it is worth explaining the major advantages and disadvantages of the presented routing policies. First, we would like to remark that both are open loop policies, i.e., they do not require information exchange between the servers and the dispatcher. This is an important advantage with respect to other routing policies such as Join the Shortest Queue, for instance. Moreover, SITA and TAGS are size-based routing policies, i.e., they both use cut-offs to determine how jobs are executed in the servers. The main advantage of SITA is the optimality result in [14]. However, SITA routing uses the size of incoming tasks to assign jobs to servers. Finally, we remark that TAGS do not use the information of the size of incoming jobs to balance the load.*

3.4. Bounded Pareto Distributions

Let $r > 1$ and $a = \frac{1}{r}$. If $1 \leq s \leq r$, then a distribution follows the Bounded Pareto distribution with parameters 1, r, and α if its density function is

$$f(s) = \frac{\alpha s^{-\alpha-1}}{(1-a^\alpha)},$$

and $f(s) = 0$ otherwise. Note that this distribution consists of the Pareto distribution with tail parameter α, but restricted to a bounded domain, i.e., $1 \leq s \leq r$. The cumulative distribution function is given by

$$F(s) = \begin{cases} 0, & s \leq 1, \\ \frac{1-(1/s)^\alpha}{1-a^\alpha}, & 1 \leq s \leq r, \\ 1, & s \geq r. \end{cases}$$

When $\alpha \neq 1$, it is easy to see that

$$\mathbb{E}[X] = \frac{\alpha}{\alpha - 1} \frac{1 - a^{\alpha - 1}}{1 - a^{\alpha}} \tag{1}$$

whereas when $\alpha = 1$

$$\mathbb{E}[X] = \frac{\ln(r)}{1 - \frac{1}{r}}. \tag{2}$$

We remark that the distributions under consideration when r is large and $0 < \alpha < 2$ are known to model well job size distributions with large variability [22]. Another interesting property of these distributions is that when $\alpha = -1$, they coincide with the uniform distribution on the interval $[1, r]$.

3.5. Application of This Theory to the Result

In this article, we are interesting in investigating the following ratio:

$$\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA}(\mathbf{s}^*)]}.$$

The analysis of this ratio can be seen as the penalty for not knowing the size of incoming tasks. The authors of [2] showed that, when r is large, this ratio is upper-bounded by 2. Therefore, they conclude that the penalty for not knowing the size of incoming jobs is, at most, 2 in the regime they consider. In this work, we also study this ratio but in a different regime. In Section 4, we consider a system with two queues, whereas in Section 5 an arbitrary number of queues.

4. Performance Comparison of TAGS and Optimal SITA with $K = 2$ Servers

We study the ratio of the mean waiting times of TAGS and SITA for Bounded Pareto distributions when r is finite. We recall that Theorem 9 in [2] says that the aforementioned performance ratio is upper-bounded by 2 when the system is at a fixed low load and r tends to infinity. In this section, we consider a system with two servers, and we show that this ratio is unbounded from above. We first provide an analysis for the case of Bounded Pareto with $\alpha = 1$. Then, we present numerical experiments that show that the ratio is also unbounded from above for Bounded Pareto-distributed jobs sizes with $\alpha \neq 1$.

4.1. Performance of the Optimal TAGS System

We consider a system with two servers operating under the TAGS policy (see Figure 1). In a system with two servers, the vector of thresholds $\mathbf{s}$ is reduced to a unique value; therefore, we denote by $\mathbb{E}[W^{TAGS}(s)]$ the mean waiting time of TAGS for the threshold s. Here, we denote by s_T the threshold value that minimizes the mean waiting time of the TAGS system, i.e.,

$$s_T = \arg\min_{s} \mathbb{E}[W^{TAGS}(s)].$$

From the definition of the TAGS system, it follows that

$$\mathbb{E}[W^{TAGS}(s_T)] = \mathbb{E}[W_1^{TAGS}(s_T)] + \left(\int_{s_T}^{r} f(x)dx\right) s_T + \left(\int_{s_T}^{r} f(x)dx\right) \mathbb{E}[W_2^{TAGS}(s_T)], \tag{3}$$

where $\mathbb{E}[W_i^{TAGS}(s_T)]$ is the mean waiting time of jobs at server i when the threshold value is s_T.

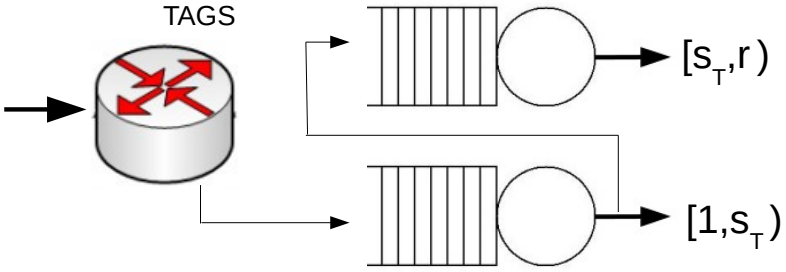

Figure 1. A system with two servers that operates under Task Assignment with Guessing Size (TAGS).

4.2. Performance of the Optimal SITA System

We now consider a system with two servers operating under the SITA policy (see Figure 2). For SITA routing, in a system with two servers, the vector of thresholds **s** is also reduced to a unique value. In this case, we denote by $\mathbb{E}[W^{SITA}(s)]$ the mean waiting time of the SITA routing system with the threshold s. We let s^* be the value of s such that the mean waiting time of the SITA system is minimized, i.e.,

$$s^* = \arg\min_s \mathbb{E}[W^{SITA}(s)].$$

The mean waiting time of jobs under the SITA policy with cutoff s^* is given by

$$\mathbb{E}[W^{SITA}(s^*)] = \left(\int_1^{s^*} f(x)dx\right)\mathbb{E}[W_1^{SITA}(s^*)] + \left(\int_{s^*}^{r} f(x)dx\right)\mathbb{E}[W_2^{SITA}(s^*)], \quad (4)$$

where $\mathbb{E}[W_i^{SITA}(s^*)]$ is the mean waiting time of jobs at server i.

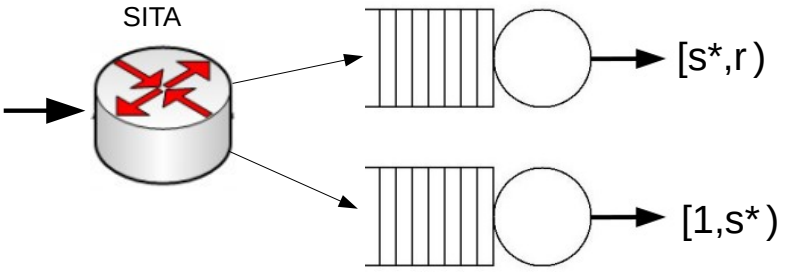

Figure 2. A system with two servers that operates under Size Interval Task Assignment (SITA).

4.3. Bounded Pareto Distribution with $\alpha = 1$

We consider the Bounded Pareto distribution with $\alpha = 1$ and compare the mean waiting time of a system with two servers operating under TAGS routing with that of a system with two servers operating under SITA routing. In the following result, we provide a lower-bound for the average waiting time of the TAGS system.

Lemma 1. *For the Bounded Pareto distribution with $\alpha = 1$, if $\lambda r < 1$,*

$$\mathbb{E}[W^{TAGS}(s_T)] > \lambda r.$$

Proof. It follows from (3) that

$$\mathbb{E}[W^{TAGS}(s_T)] \geq \mathbb{E}[W_1^{TAGS}(s_T)] + \left(\int_{s_T}^{r} f(x)dx\right)s_T.$$

For the Bounded Pareto distribution with $\alpha = 1$, we have that

$$\begin{aligned}\mathbb{E}[W_1^{TAGS}(s_T)] + \left(\int_{s_T}^{r} f(x)dx\right)s_T &= \frac{\lambda(s_T-1)}{2(1-\rho)(1-\frac{1}{r})} + \frac{1-\frac{s_T}{r}}{1-\frac{1}{r}} \\ &\geq \frac{\lambda(s_T-1)}{2(1-\frac{1}{r})} + \frac{1-\frac{s_T}{r}}{1-\frac{1}{r}} \\ &= \frac{s_T(\lambda-\frac{1}{r})+2-2\lambda}{2(1-\frac{1}{r})}.\end{aligned}$$

If $\lambda r < 1$, then $\frac{x(\lambda-\frac{1}{r})+2-2\lambda}{2(1-\frac{1}{r})}$ decreases with x. Thus,

$$\frac{s_T(\lambda-\frac{1}{r})+2-2\lambda}{2(1-\frac{1}{r})} \geq \frac{r(\lambda-\frac{1}{r})+2-2\lambda}{2(1-\frac{1}{r})} = \frac{\lambda(r-2)+1}{2(1-\frac{1}{r})} > \frac{\lambda(r-2)+\lambda r}{2(1-\frac{1}{r})} = \lambda r.$$

□

Note that, unlike previous work [2], we do not need to assume Poisson arrivals to all the servers in the above result. In the following result, we estimate, non-asymptotically, the mean waiting time of the SITA system.

Lemma 2. *For the Bounded Pareto distribution with $\alpha = 1$, when $\lambda r < 1$ and r is large,*

$$\mathbb{E}[W^{SITA}(s^*)] \leq \frac{\lambda(\sqrt{r}-1)^2}{\sqrt{r}(1-\frac{1}{r})^2}.$$

Proof. We first note that the load of each server is upper bounded by ρ. Therefore, for the Bounded Pareto distribution with $\alpha = 1$, we have that

$$\mathbb{E}[W_1^{SITA}(s^*)] \leq \frac{\lambda(s^*-1)}{2(1-\rho)(1-(1/r))}$$

and

$$\mathbb{E}[W_2^{SITA}(s^*)] \leq \frac{\lambda(r-s^*)}{2(1-\rho)(1-(1/r))},$$

where $\rho = \lambda \ln(r)$. We now note that

$$\lambda r < 1 \iff \lambda \ln(r) < \frac{\ln(r)}{r}.$$

When r is large, $\frac{\ln(r)}{r}$ tends to zero and $\rho = \lambda \ln(r)$; it follows that ρ tends to zero when $\lambda r < 1$. As a result,

$$\mathbb{E}[W_1^{SITA}(s^*)] \leq \frac{\lambda(s^*-1)}{2(1-(1/r))}$$

and

$$\mathbb{E}[W_2^{SITA}(s^*)] \leq \frac{\lambda(r-s^*)}{2(1-(1/r))}.$$

We know from [18] that for the Bounded Pareto distribution with $\alpha = 1$, s^* balances the load of both servers, and therefore

$$\int_1^{s^*} f(x)dx = \int_{s^*}^{r} f(x)dx \iff s^* = \sqrt{r}.$$

As a result,

$$\mathbb{E}[W_1^{SITA}(s^*)] \leq \frac{\lambda(\sqrt{r}-1)}{2(1-(1/r))}$$

and

$$\mathbb{E}[W_2^{SITA}(s^*)] \leq \frac{\lambda(r-\sqrt{r})}{2(1-(1/r))}.$$

Therefore, from (4), it follows that

$$\begin{aligned}
\mathbb{E}[W^{SITA}(s^*)] &\leq \left(\int_1^{s^*} f(x)dx\right)\frac{\lambda(\sqrt{r}-1)}{2(1-(1/r))} + \left(\int_{s^*}^{r} f(x)dx\right)\frac{\lambda(r-\sqrt{r})}{2(1-(1/r))} \\
&= \left(\int_1^{\sqrt{r}} f(x)dx\right)\frac{\lambda(\sqrt{r}-1)}{2(1-(1/r))} + \left(\int_{\sqrt{r}}^{r} f(x)dx\right)\frac{\lambda(r-\sqrt{r})}{2(1-(1/r))} \\
&= \frac{\lambda(1-\frac{1}{\sqrt{r}})(\sqrt{r}-1)}{2(1-(1/r))^2} + \frac{\lambda(\frac{1}{\sqrt{r}}-\frac{1}{r})(r-\sqrt{r})}{2(1-(1/r))^2} \\
&= \frac{\lambda(\frac{1}{\sqrt{r}}-\frac{1}{r})(r-\sqrt{r})}{(1-(1/r))^2} \\
&= \frac{\lambda(\sqrt{r}-1)^2}{\sqrt{r}(1-(1/r))^2}.
\end{aligned}$$

□

From the above lemmas, it follows that for $\lambda r < 1$ and large enough r,

$$\frac{\mathbb{E}[W^{TAGS}(s_T)]}{\mathbb{E}[W^{SITA}(s^*)]} \geq \frac{\lambda r}{\frac{\lambda(\sqrt{r}-1)^2}{\sqrt{r}(1-(1/r))^2}} = \frac{(\sqrt{r}+1)^2}{\sqrt{r}},$$

where the equality is obtained by simplifying the derived expression. Interestingly, the last expression does not depend on λ. We now show that it is increasing with r.

Lemma 3. *The function* $\frac{(\sqrt{r}+1)^2}{\sqrt{r}}$ *is increasing with r.*

Proof. We aim to show that $\frac{(\sqrt{r}+1)^2}{\sqrt{r}}$ is increasing with r, which is true if and only if

$$\begin{gathered}
((\sqrt{r}+1)^2)'\sqrt{r} - (\sqrt{r})'(\sqrt{r}+1)^2 > 0 \iff \\
\frac{2(\sqrt{r}+1)}{2\sqrt{r}}\sqrt{r} - \frac{1}{2\sqrt{r}}(\sqrt{r}+1)^2 > 0 \iff \\
\sqrt{r}+1 - \frac{1}{2\sqrt{r}}(\sqrt{r}+1)^2 > 0 \iff \\
1 - \frac{1}{2\sqrt{r}}(\sqrt{r}+1) > 0 \iff \\
2\sqrt{r} - (\sqrt{r}+1) > 0 \iff \\
\sqrt{r} - 1 > 0
\end{gathered}$$

Therefore, the desired result follows. □

From the above results and using that $\frac{(\sqrt{r}+1)^2}{\sqrt{r}}$ tends to infinity when $r \to \infty$, it follows that when $\lambda r < 1$ and r is large, the ratio between the mean waiting time of the TAGS system and that of the SITA system is lower bounded by a function that is unbounded and the following result follows.

Theorem 1. *The ratio of the mean waiting time of the TAGS system and the mean waiting time of the SITA system is unbounded.*

4.4. Bounded Pareto with $\alpha \neq 1$

We now study the ratio between the mean waiting time of the TAGS system and that of the SITA system for Bounded Pareto distributions with $\alpha \neq 1$. We consider the evolution of this ratio when we vary r from 10 to 1000 for different values of α.

In Figure 3, we consider an arrival rate of 0.001, and we plot the ratio between the mean waiting time of the TAGS system and that of the SITA system when we vary r for several values of α larger than one. We observe that the performance ratio under study is increasing with r for the considered values of r. Therefore, this illustration shows that the result of Theorem 1 generalizes to values of α which are not equal to 1.

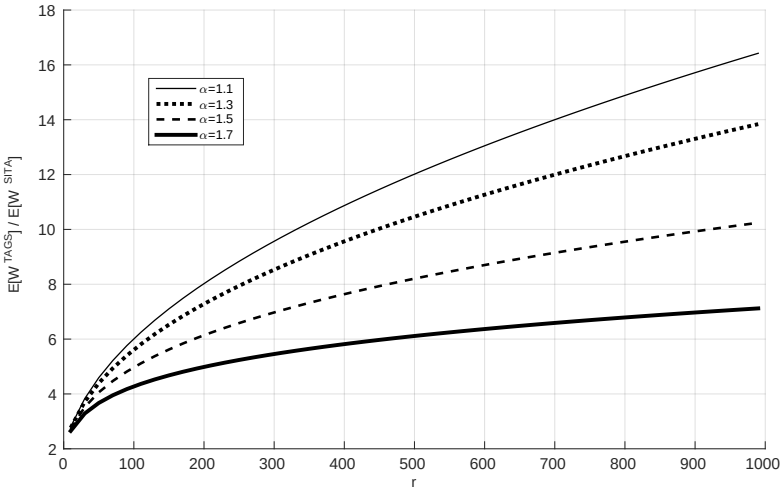

Figure 3. The ratio $\mathbb{E}[W^{TAGS}(s_T)]/\mathbb{E}[W^{SITA}(s_T)]$ as a function of r for different values of α. y-axis without units (it is the ratio of two values with the same unit) and for the x-axis the units are ms.

5. Performance Comparison of TAGS and SITA-E with $K \geq 2$ Servers

In this section, we compare the performance of TAGS and SITA-E for an arbitrary number of servers. Specifically, we study the ratio of the mean waiting time of TAGS to that of SITA-E, i.e., $\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA-E}]}$.

Let us denote by $\mathbb{E}[W^{SINGLE}(\lambda)]$ the mean waiting time of a single server with arrival rate λ. We note that the above ratio can be written as

$$\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA-E}]} = \frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SINGLE}(\lambda)]} \frac{\mathbb{E}[W^{SINGLE}(\lambda)]}{\mathbb{E}[W^{SITA-E}]}. \tag{5}$$

We now focus on the term $\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SINGLE}(\lambda)]}$. In the following result, we show that, when the arrival rate tends to zero, this ratio tends to one.

Proposition 1. *Let X be a distribution with support in $t \geq 1$. When $\lambda \to 0$, then $\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SINGLE}(\lambda)]}$ tends to one.*

Proof. A single server system is a special case of a TAGS system with degenerate cut-offs, which implies by the definition of $\mathbf{s}_T$ that $\mathbb{E}[W^{SINGLE}(\lambda)] \geq \mathbb{E}[W^{TAGS}(\mathbf{s}_T)]$. Let $s_{T,1}$ be the first component of the vector $\mathbf{s}_T$ and $p(s_{T,1})$ be the probability of jobs larger than $s_{T,1}$. The contribution to the mean waiting time waiting time of TAGS in the first server coming from jobs bigger than $s_{T,1}$ is $p(s_{T,1})s_{T,1} \geq p(s_{T,1})$, as $s_{T,1} \geq 1$. Therefore, as the waiting time of a single server goes to 0 with λ and, by the definition of $\mathbf{s}_T$, $s_{T,1}$ is optimal, we conclude that, asymptotically, as λ goes to zero, so does $p(s_{T,1})$, meaning that in the limit all jobs are

served and finish in the first server and thus the performance of the system is like that of a single server, i.e., the ratio approaches 1. □

From the above result, we conclude that, when $\lambda \to 0$,

$$\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA-E}]} \to \frac{\mathbb{E}[W^{SINGLE}(\lambda)]}{\mathbb{E}[W^{SITA-E}]}.$$

We now analyze the ratio $\frac{\mathbb{E}[W^{SINGLE}(\lambda)]}{\mathbb{E}[W^{SITA-E}]}$ when $\lambda \to 0$. For this case, as the mean waiting time of a single server is approximately linear in the arrival rate, it follows that

$$\frac{\mathbb{E}[W^{SINGLE}(\lambda)]}{\mathbb{E}[W^{SITA-E}]} \approx \frac{K\mathbb{E}[W^{SINGLE}(\frac{\lambda}{K})]}{\mathbb{E}[W^{SITA-E}]}.$$

The ratio $\frac{\mathbb{E}[W^{SINGLE}(\frac{\lambda}{K})]}{\mathbb{E}[W^{SITA-E}]}$ has been studied in [3]. In that work, the authors conclude that this ratio grows with the variability of the job size distribution and can be large if the variability of the job size distribution is large. For example, from Proposition 3 in [3], we conclude that for Bounded Pareto-distributed job sizes with $\alpha = 1$, the ratio $\frac{\mathbb{E}[W^{SINGLE}(\frac{\lambda}{K})]}{\mathbb{E}[W^{SITA-E}]}$ tends to infinity when the ratio between the shortest and the largest job size tends to zero, and as this result is true for any value of λ, using (5), the next result follows.

Proposition 2. *For Bounded Pareto distributed job sizes with $\alpha = 1$ and $K \geq 2$, when $\lambda \to 0$, the ratio between the mean waiting time of TAGS and SITA-E is unbounded from above.*

From this result and using that the mean waiting time of SITA-E is larger than the mean waiting time of the SITA policy with optimal thresholds, the ratio $\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA-E}]}$ is a lower bound of the ratio $\frac{\mathbb{E}[W^{TAGS}(\mathbf{s}_T)]}{\mathbb{E}[W^{SITA}(\mathbf{s}^*)]}$. Consequently, we have the following result.

Corollary 1. *For Bounded Pareto-distributed job sizes with $\alpha = 1$ and $K \geq 2$, when $\lambda \to 0$, the ratio between the mean waiting time of TAGS and the optimal SITA is unbounded from above.*

In [3], the authors also analyze the performance ratio $\frac{\mathbb{E}[W^{SINGLE}(\frac{\lambda}{K})]}{\mathbb{E}[W^{SITA-E}]}$ for Bounded Pareto distributed job sizes with tail parameter $\alpha \in (0,2)\backslash\{1\}$ and $K \to \infty$, and in Proposition 4, they show that this ratio tends to infinity when the ratio between the shortest and the largest job size tends to zero. Consequently, we obtain the following result

Proposition 3. *For Bounded Pareto distributed job sizes with $\alpha \in (0,2)\backslash\{1\}$ and $K \to \infty$, when $\lambda \to 0$, the ratio between the mean waiting time of TAGS and SITA-E is unbounded from above.*

Furthermore, the corollary below also follows immediately.

Corollary 2. *For Bounded Pareto distributed job sizes with $\alpha \in (0,2)\backslash\{1\}$ and $K \to \infty$, when $\lambda \to 0$, the ratio between the mean waiting time of TAGS and the optimal SITA is unbounded from above.*

6. Design of Decentralized Systems

We consider a decentralized system formed by n groups where, in each group, the incoming jobs are routed to K/n servers. In Figure 4, we represent an example of a decentralized system formed by 2 groups. We assume that, in each group, there is a single dispatcher that implements the SITA or TAGS policy and job sizes are Bounded Pareto-distributed.

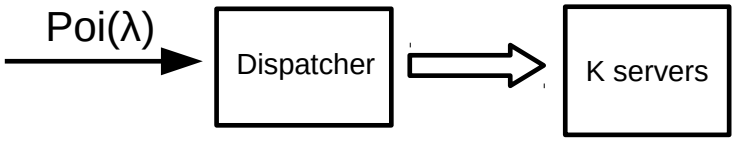

Figure 4. A decentralized system with 2 groups. Each dispatchers receives $\lambda/2$ traffic and sends it to $K/2$ servers.

In a decentralized system with n groups, when there are t groups that implement the TAGS policy and $n-t$ groups that implement SITA, the performance of the system is given by

$$P(t) = \frac{t}{n}\mathbb{E}[W^{TAGS}(\mathbf{s}_T)] + \left(1 - \frac{t}{n}\right)\mathbb{E}[W^{SITA}(\mathbf{s}^*)].$$

We assume that there is a quality of service (QoS) condition that states that the average waiting time of the system cannot exceed a given threshold d. Therefore, the system designers must decide on a policy that satisfies the condition with minimal cost (i.e., maximize the number of groups that implement the TAGS policy) and satisfies the QoS condition $P(t) \leq d$. This can be expressed as

$$\max \ t \tag{6}$$
$$\text{s.t. } P(t) \leq d. \tag{7}$$

We denote by t^* the solution to the above problem. We assume that $P(n) \leq d$; otherwise, there is no solution to this problem. Given that $P(t)$ is decreasing with t, it is clear that if $P(0) \leq d$, then $t^* = 0$, i.e., all the groups implement the SITA policy. Otherwise, using this monotonicity property, it follows that the solution of this problem is $t^* = \lfloor P^{-1}(d) \rfloor$. However, finding the analytical value of the inverse of $P()$ is difficult as an analytical expression for the mean waiting time of a system that implements the TAGS policy is not known. As a result, we focus on a suboptimal solution to this problem that uses the result of Theorem 9 in [2] to get insights into t^*.

We choose the number of groups n of the decentralized system so that the load of each group is less than one. As the arrival rate of each group is λ/n, we have that

$$\rho < 1 \iff n > \frac{\mathbb{E}[X]}{\lambda}.$$

We choose the number of groups to be the minimal value of n which is a divisor of K and is larger than $\frac{\mathbb{E}[X]}{\lambda}$.

According to Theorem 9 in [2], when r is very large, we have that

$$2\mathbb{E}[W^{SITA}(\mathbf{s}^*)] \geq \mathbb{E}[W(\mathbf{s}_T)].$$

Thus, for r very large, we have that

$$\begin{aligned} P(t) &= \frac{t}{n}\mathbb{E}[W^{TAGS}(\mathbf{s}_T)] + \left(1 - \frac{t}{n}\right)\mathbb{E}[W^{SITA}(\mathbf{s}^*)] \\ &\leq 2\frac{t}{n}\mathbb{E}[W^{SITA}(\mathbf{s}_T)] + (1 - \frac{t}{n})\mathbb{E}[W^{SITA}(\mathbf{s}^*)] \\ &= \left(1 + \frac{t}{n}\right)\mathbb{E}[W^{SITA}(\mathbf{s}^*)]. \end{aligned}$$

Let $\tilde{P}(t) = (1 + \frac{t}{n})\mathbb{E}[W^{SITA}(\mathbf{s}^*)]$. Therefore, an alternative optimization problem to (6) and (7) consists of considering the constraint $\tilde{P}(t) \leq d$ instead of $P(t) \leq d$, i.e.,

$$\max \ t \tag{8}$$
$$\text{s.t. } \tilde{P}(t) \leq d. \tag{9}$$

The solution of this problem is clearly the minimum between n and $\left\lfloor \bar{t}^{\,*} \right\rfloor$, where $\bar{t}^{\,*}$ is such that $\tilde{P}(t^{\,*}) = d$, i.e.,

$$\left(1 + \frac{\bar{t}^{\,*}}{n}\right)\mathbb{E}[W^{SITA}(\mathbf{s}^*)] = d \iff \bar{t}^{\,*} = n\left(\frac{d}{\mathbb{E}[W^{SITA}(\mathbf{s}^*)]} - 1\right).$$

Finally, we can obtain the expression of $\mathbb{E}[\bar{W}^{SITA}(\mathbf{s}^*)]$ for Bounded Pareto-distributed job sizes with large r and $\rho < 1$ using the formula of Theorem 6.3 in [16].

Example

We consider an example with a decentralized system in which $r = 10^4$ and $\alpha = 1.25$. For this case, we obtain that $\mathbb{E}[X] = 27.027$. Thus, in a system formed by 1000 servers and arrival rate equal to 5,

$$\frac{\lambda\mathbb{E}[X]}{n} = \frac{5 \cdot 27.027}{n} < 1 \iff n > 135.135.$$

As 200 is the minimum divisor of 1000 that verifies the condition of the last expression, we assume that the decentralized system is formed by a number of groups equal to 200. This number of groups meets the condition that establishes that the load of each server is less than one:

$$\frac{\lambda\mathbb{E}[X]}{n} = \frac{5 \cdot 27.027}{200} = 0.675 < 1.$$

Using Theorem 6.3 in [16], we get that the normalized mean waiting time of a SITA system formed by 5 servers with $r = 10^4$, $\alpha = 1.25$, and $\rho = 0.675$ is given by 0.216. Therefore, if the limit d of the model is set to 0.4, we get for our example that

$$\bar{t}^{\,*} = 200\left(\frac{0.4}{0.216} - 1\right) = 170.28.$$

This result means that if 170 groups of 5 servers implement the TAGS policy and 30 groups implement the SITA policy, we can ensure that the quality of service condition (7) is satisfied.

7. Conclusions

We have studied the performance of parallel server systems with two size-based task assignment policies: SITA and TAGS. Both policies use cut-offs to determine how jobs are executed in the servers. The main difference between them is that SITA uses the size of incoming tasks to assign jobs to servers, while TAGS does not.

The goal of this paper is to analyze the ratio between the mean waiting time of TAGS and the mean waiting time of SITA. This ratio can be seen as the penalty for not knowing the size of incoming tasks. In a previous work [2], we have shown that this ratio is upper bounded by 2 when the system load is fixed and less than one and the ratio between the largest and the shortest job tends to infinity. In this article, we extend this result in several directions.

We first consider a non-asymptotic regime, and we compare the performance of SITA and TAGS for this regime. For Bounded Pareto-distributed job sizes with tail parameter 1 and two servers, we provide a lower bound on the mean waiting time of TAGS and an upper bound of the mean waiting time of SITA, in both cases assuming that the arrival rate times the largest job size is smaller than one. From these results, we provide a lower bound of the ratio of the mean waiting time of TAGS over the mean waiting time of SITA of a system with two servers and in the considered regime. We show that this lower bound has nice properties, for example, it does not depend on the arrival rate and it is increasing with the size of the largest job. We conclude that the mean waiting time of TAGS divided by the mean waiting time of SITA is unbounded from above in a system with two servers

and when the arrival rate time of the largest job size is smaller than one. We also explore numerically this performance ratio when the tail parameter is not 1, and we observe that our theoretical findings are not limited to Bounded Pareto-distributed job sizes with tail parameter 1.

Then, we analyze a system with an arbitrary number of servers, and we study the ratio of the mean waiting time of TAGS under the mean waiting time of SITA-E. We consider the light traffic regime in which the arrival rate tends to zero. For this regime, we first show that the mean waiting time of TAGS and the mean waiting time of a single server coincide. From this result, it follows that the performance ratio under consideration is proportional to the performance ratio that the authors of [3] explored. This similitude lets us conclude that the ratio of the mean waiting time of TAGS to that of SITA-E is unbounded from above for Bounded Pareto distributed job sizes with $\alpha = 1$ and $K \geq 2$, and for Bounded Pareto distributed job sizes with $\alpha \in (0, 2) \backslash \{1\}$ and $K \to \infty$.

We would like to remark that the above results show that the parameters of the system can be set in a way that the mean waiting time of TAGS can be much worse than the mean waiting time of SITA, and therefore the penalty for not knowing the job size is unbounded from above.

In this work, we also consider decentralized systems in which the system is divided into groups, where each group handles the same amount of traffic and sends the traffic to a subset of dedicated servers. Each group can operate under SITA or TAGS. The goal is to find the maximum number of groups that can operate under TAGS policy and still satisfy a given quality of service condition. We provide a suboptimal solution to this problem using the asymptotic result in [2].

The analysis presented in this article assumes several properties of Bounded Pareto distributions. A possible extension of this work would be to consider an arbitrary distributions and to compare the performance of other popular routing policies from the literature such as Join the Shortest Queue and Power of Two. Finally, we are interested in studying the application of learning techniques to parallel server systems using the approach in [23] and the interaction between the techniques presented in this work with other control methodologies such as fuzzy control or Active Disturbance Rejection Control as in [24].

Author Contributions: Conceptualization, J.D. and E.B.; Methodology, J.D. and E.B.; Software, J.D. and E.B.; Validation, J.D. and E.B.; Formal Analysis, J.D. and E.B.; Investigation, J.D. and E.B.; Resources, J.D. and E.B.; Data Curation, J.D. and E.B.; Writing–Original Draft Preparation, J.D. and E.B.; Writing–Review & Editing, J.D. and E.B.; Visualization, J.D. and E.B.; Supervision, J.D. and E.B.; Project Administration, J.D. and E.B.; Funding Acquisition, J.D. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: Josu Doncel has received funding from the Department of Education of the Basque Government through the Consolidated Research Group MATHMODE (IT1294-19), from the Marie Sklodowska-Curie grant agreement No 777778, and from the Spanish Ministry of Science and Innovation with reference PID2019-108111RB-I00 (FEDER/AEI). Eitan Bachmat's work was supported by the German Science Foundation (DFG) through the grant, Airplane Boarding, (JA 2311/3-1).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Williams, A.; Arlitt, M.; Williamson, C.; Barker, K. Web Workload Characterization: Ten Years Later. In *Web Content Delivery*; Tang, X., Xu, J., Chanson, S.T., Eds.; Springer US: Boston, MA, USA, 2005; pp. 3–21.

2. Bachmat, E.; Doncel, J.; Sarfati, H. Performance and Stability Analysis of the Task Assignment Based on Guessing Size Routing Policy. In Proceedings of the IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Rennes, France, 21–25 October 2019; pp. 1–13.
3. Doncel, J.; Aalto, S.; Ayesta, U. Performance Degradation in Parallel-Server Systems. *IEEE/ACM Trans. Netw.* **2019**, *27*, 875–888. [CrossRef]
4. Anselmi, J.; Doncel, J. Asymptotically optimal size-interval task assignments. *IEEE Trans. Parallel Distrib. Syst.* **2019**, *30*, 2422–2433. [CrossRef]
5. Bachmat, E.; Doncel, J. Non-Asymptotic Performance Analysis of Size-Based Routing Policies. In Proceedings of the 2020 28th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Nice, France, 17–19 November 2020; pp. 1–4.
6. Harchol-Balter, M. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*; Cambridge University Press: Cambridge, UK, 2013.
7. Foley, R.D.; McDonald, D.R. Join the shortest queue: Stability and exact asymptotics. *Ann. Appl. Probab.* **2001**, *11*, 569–607. [CrossRef]
8. Gupta, V.; Balter, M.H.; Sigman, K.; Whitt, W. Analysis of join-the-shortest-queue routing for web server farms. *Perform. Eval.* **2007**, *64*, 1062–1081. [CrossRef]
9. Weber, R.R. On the optimal assignment of customers to parallel servers. *J. Appl. Probab.* **1978**, *15*, 406–413. [CrossRef]
10. Richa, A.W.; Mitzenmacher, M.; Sitaraman, R. The power of two random choices: A survey of techniques and results. *Comb. Optim.* **2001**, *9*, 255–304.
11. Winston, W. Optimality of the shortest line discipline. *J. Appl. Probab.* **1977**, *14*, 181–189. [CrossRef]
12. Whitt, W. Deciding Which Queue to Join: Some Counterexamples. *Oper. Res.* **1986**, *34*, 55–62.
13. Crovella, M.E.; Harchol-Balter, M.; Murta, C.D. Task Assignment in a Distributed System (Extended Abstract): Improving Performance by Unbalancing Load. In *Proceedings of the ACM SIGMETRICS*; ACM: New York, NY, USA, 1998; pp. 268–269. [CrossRef]
14. Feng, H.; Misra, V.; Rubenstein, D. Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems. *Perform. Eval.* **2005**, *62*, 475–492. [CrossRef]
15. Harchol-Balter, M.; Crovella, M.E.; Murta, C.D. On Choosing a Task Assignment Policy for a Distributed Server System. *J. Parallel Distrib. Comput.* **1999**, *59*, 204–228. [CrossRef]
16. Bachmat, E.; Sarfati, H. Analysis of SITA policies. *Perform. Eval.* **2010**, *67*, 102–120. [CrossRef]
17. Vesilo, R. Asymptotic analysis of load distribution for size-interval task allocation with bounded Pareto job sizes. In Proceedings of the IEEE International Conference on Parallel and Distributed Systems, Melbourne, VIC, Australia, 8–10 December 2008.
18. Harchol-Balter, M.; Vesilo, R. To balance or unbalance load in size-interval task allocation. *Probab. Eng. Inform. Sci.* **2010**, *24*, 219–244. [CrossRef]
19. Harchol-Balter, M.; Scheller-Wolf, A.; Young, A.R. Surprising Results on Task Assignment in Server Farms with High-variability Workloads. In Proceedings of the SIGMETRICS, Seattle, WA, USA, 15–19 June 2009.
20. Harchol-Balter, M. Task assignment with unknown duration. In Proceedings of the 20th IEEE International Conference on Distributed Computing Systems, Taipei, Taiwan, 10–13 April 2000; pp. 214–224.
21. Bachmat, E.; Doncel, J.; Sarfati, H. Analysis of the Task Assignment based on Guessing Size policy. *Perform. Eval.* **2020**, *142*, 102122. [CrossRef]
22. Harchol-Balter, M.; Downey, A.B. Exploiting Process Lifetime Distributions for Dynamic Load Balancing. *ACM Trans. Comput. Syst.* **1997**, *15*, 253–285. [CrossRef]
23. Zhang, H.; Liu, X.; Ji, H.; Hou, Z.; Fan, L. Multi-Agent-Based Data-Driven Distributed Adaptive Cooperative Control in Urban Traffic Signal Timing. *Energies* **2019**, *12*, 1402. [CrossRef]
24. Roman, R.C.; Precup, R.E.; Petriu, E.M. Hybrid data-driven fuzzy active disturbance rejection control for tower crane systems. *Eur. J. Control.* **2021**, *58*, 373–387. [CrossRef]