Analysis of the Task Assignment based on Guessing Size policy

Eitan Bachmat

Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, 84105.

Josu Doncel

University of the Basque Country, UPV/EHU. Leioa, Spain. 48940.

Hagit Sarfati

Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, 84105.

Abstract

We study the Task Assignment based on Guessing Size (TAGS) policy in a parallel and homogeneous server system. The policy parameters are the number of servers h and a set of cutoffs $s_1 < s_2, \ldots, s_{h-1}$. In this policy, all the incoming jobs are routed to the first server and jobs are run up to s_1 time units. If they complete they leave the system, but jobs that do not complete after s_1 time units are killed and moved to the end of the queue of the second server, where service starts from scratch. Likewise, jobs that are executed in server i and complete service before s_i units of time, leave the system, whereas jobs that do not complete are killed and routed to the next server. We first study the stability of such system and provide a precise utilization threshold for the existence of stable parameters for a given job size distribution. We compute the threshold for several families of distributions and provide bounds for others. We show that TAGS is most stable for bounded Pareto distributions with parameter $\alpha = 1$. Besides, we provide tight bounds on the performance of the TAGS policy where the cutoffs are chosen to minimize average waiting time in the asymptotic regime where the largest job size tends to infinity for the Bounded Pareto distribution and the system load smaller than one. In this case, we show that the performance ratio between TAGS and a version called SITA which does require knowledge of job size is at most 2. We then consider more broadly the same asymptotic regime and consider a bound on the average waiting time for any distribution. This is compatible with having a conservative policy which will work well for any job size distribution. We show rather tight upper and lower bounds which again match those of SITA up to a factor of 2. These bounds are considerably lower than the corresponding bounds of competing policies such as Random or Least Work Remaining. We also show that for all these policies the bounded Pareto distribution with $\alpha = 1$ is close to being the worst possible among all distributions with the same job size range. We show using the stability results that in the same regime, if we increase the load the gap between SITA and TAGS grows dramatically and eventually, the TAGS system cannot be stabilized. The conclusion from all our analysis is that TAGS is best used as a conservative policy with minimal requirements on small systems with relatively low utilization and where the range of job sizes is large.

Keywords: Size interval routing policies, Heavy-tailed distributions, Parallel-server systems.

1. Introduction

The performance analysis of parallel-server systems, such as data centers, is a central problem in system engineering. One important performance related aspect is the choice of a routing policy. This choice can be constrained in several ways according to the input available to us and the network design of the system. In this paper, we consider an algorithm called Task Assignment based on Guessing Size

(TAGS) [14], that was designed to operate under the tightest constraints, in which we do not know the sizes of particular jobs and in which we do not know the state of the servers and so no coordination is allowed. The TAGS algorithm is based on a previous algorithm which does not require knowledge of state, but does require knowledge of the job size, the Size Interval Task assignment algorithm (SITA), [16]. It is known [12] that under the assumptions that servers are identical and arrivals are Poisson, that SITA minimizes average waiting time among stateless routing algorithms. It is also known that in the high variability regime SITA is much more effective than other stateless policies such as Random Assignment (RA), [14, 3]. In addition to stateless algorithms, there are other algorithms which do require knowledge of the state of the system. Perhaps the best known example is Least Work Left (LWL), also known as FIFO (M/G/h), where a job is assigned to the server which has the least amount of work left at the time of request arrival. Implementing LWL requires knowledge of state or a central buffer, either of which might be difficult to have. The redeeming factor is that LWL does not require knowledge of job size. The comparison between SITA and LWL depends on all system parameters such as the number of servers h, the utilization ρ, the job size distribution X and the target function, average waiting time, average slowdown and so on. Either algorithm may be better than the other depending on these circumstances.

The present paper expands on [5] where we presented an analysis of some stability and performance properties of TAGS. We will build on the main results of [5] as well as additional results, computations and comparisons to obtain a more comprehensive understanding of TAGS and its relation to RA, SITA and LWL.

In [14], the author focused mostly on the performance of TAGS with respect to the mean slowdown target function. In this paper, we will be concerned with mean (normalized) waiting time which is a more challenging target function for TAGS and SITA. We will leave a detailed analysis of mean slowdown for future work.

In particular we will be interested in a setting which we think of as a conservative approach to performance analysis. In this setting, we assume that we do not know the job size distribution and we should be prepared for any distribution within a large class. This setting resembles the worst case approach of theoretical computer science and specifically scheduling theory, but we still make stochastic assumptions about utilization, arrival processes and job size distribution. In this setting we judge the performance of an algorithm according to its guarantees on the entire family of distributions.

The particular setting that we will explore is the one where the utilization and number of servers is fixed and the family consists of distributions with very large range (ratio of largest to smallest job). When the utilization of the system is (very) low and the number of servers is small (sub-logarithmic in the range) the performance guarantee of TAGS closely resembles that of SITA, i.e., not knowing job sizes has little consequences for a conservative approach. In addition, the guarantee is far better than that of RA and even that of LWL, both of which display similarly bad guarantees. Moreover, the performance of TAGS is similar to that of SITA not only in the worst case, but for the whole family of high variability Pareto distributions. It is also interesting to note that the worst job size distributions for all algorithms are essentially the same and are derived from the Pareto distribution with parameter $\alpha = 1$. This distribution provides the biggest challenge across all algorithms and target functions and this is where TAGS and SITA shine the brightest in terms of outperforming RA and LWL.

Unfortunately, the performance of TAGS deteriorates rapidly when the utilization increases. In fact, the load that a TAGS system can handle, with any number of hosts, grows only logarithmically in the range of the job size distribution and in particular is bounded for any bounded job size distribution. This is obviously in sharp contrast with all other policies, which all can handle a utilization as large as the number of hosts. The bright spot is again distributions which are derived from the Pareto distribution with parameter $\alpha = 1$, the deterioration is slowest for such distributions. We note that performance also deteriorates if we fix the utilization and job size distribution and increase the number of hosts.

The emerging picture is that TAGS performs well when the utilization and number of servers is low and the range is very large. When the favorable conditions are met it can outperform RA by orders of magnitude. However, performance deteriorates rather rapidly for generic distributions as

these conditions are not met. The deterioration is slowest at the sweet spot, distributions which are closely related to the Pareto distribution with parameter $\alpha = 1$ and as we reach roughly 10 hosts and reasonable utilization per host and range, TAGS is outperformed by RA even at the sweet spot and is no longer effective.

The rest of the article is organized as follows. In Section 2, we put our work in the context of the existing literature. In Section 3, we describe the model of TAGS which we study in this article and, in Section 4, we analyze the stability of a system that operates under the TAGS policy. In Section 5 we provide bounds on the optimal performance of a TAGS system and in Section 6 we explore the asymptotic regime where the maximum job size tends to infinity. We present the numerical experiments we have performed in Section 7 and we give the main conclusions of our work in Section 8.

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

2. Related Work

The analytical study of how to manage a system with parallel queues has garnered much attention, see [15] for a recent book on this topic. Many existing routing policies are part of the SQ(d) family, where for each incoming job, $d \geq 2$ servers are picked uniformly at random to observe their states and the job is routed to the server with the best observed state (the least number of customers or least workload, for instance). In the extreme case we consider the state of all the servers in the system with respect to minimal load and obtain the Least Work Left policy (LWL). All these policies are obviously not stateless. It is known that these kind of systems have very good performance in many cases, [13, 21, 22, 19, 25], however the author in [26] showed that when the variability of the job size distribution is high this family of policies are not optimal. This is, in fact, the regime where TAGS outperforms the routing policies that belong to the SQ(d) family of policies [14]. Variants of TAGS, when jobs do not start from scratch, but rather are resumed in the next server are considered in [24, 23, 11, 7].

A related routing policy to TAGS is the Size Interval Task Assignment (SITA) policy. For this policy, each host serves jobs whose service demand is in a designated range [16]. Thus, the variance of the job executed in the servers decreases, which leads to a performance improvement when the number of servers increases [9]. The authors in [12] show that the SITA policy with optimal cutoffs minimizes mean response time, when the servers are non-observable and FCFS and the size of all the tasks is known. When the number of servers tends to infinity, the authors in [4, 1] show that the optimal SITA policy equalizes the loads of the servers. The author in [6] introduces a task assignment policy where the size of incoming tasks is required, but the goal is to maximize the probability of satisfying the utilization requirements of incoming tasks. The main difference of the TAGS policy with respect to the latter policy and with respect to SITA is that the TAGS policy does not require to know the size of the incoming jobs.

3. Model Description

We consider a TAGS system with h identical hosts[1]. Let $X(s)$ denote the job size distribution function associated with the job size random variable X, namely $X(s)$ is the probability that a job takes less than s time units to complete. For simplicity we will always assume that X is bounded and w.l.o.g. that the smallest job has size 1, i.e., $Pr(X < 1) = 0$ and the largest has size r, i.e., $Pr(X > r) = 0$. The range of the job size distribution is defined as the ratio between the largest and the smallest job size, which in our case is r. We note that TAGS can be described as a multi-class forward feeding network. The classes correspond to the jobs which terminate service at host i. The network is forward feeding in the sense that jobs never reenter the same server. For such

[1] As it has been already explained in [5], some of the results of this paper extend to systems with heterogeneous hosts (see Section 6C of [5] for full details).

systems it is known, [10], that stability is equivalent to the condition that each server in the network experiences a load which is less than 1. For the next section which deals with stability, these will be our only assumptions. For the other sections which deal with performance, we make the following further assumptions. We assume FCFS queues and an input stream of jobs that follows a Poisson distribution. We further assume that the size of the incoming jobs is given by a sequence of i.i.d. random variables. Let $F(s) = P(X < s)$ be the cumulative distribution function of the job size distribution, $E(X)$ its mean and $E(X^m)$ its m-th moment. We assume that F is differentiable and we write $f(s) = \frac{dF(s)}{ds}$. The load of the system is defined as $\rho = \lambda E(X)$. We have $\rho < 1$ if and only if the entire load can be handled in a stable manner with a single server.

We consider a multi-server assignment policy called TAGS. Let $s_0 = 1$ and $s_h = r$. In the policy TAGS, the servers are numbered $1, \ldots, h$ and there is a vector of $h-1$ cutoff values $\mathbf{s} = (s_1, s_2, \ldots, s_{h-1})$ verifying that $1 = s_0 < s_1 < s_2 < \cdots < s_{h-1} < s_h = r$. All incoming jobs are sent to server 1. If a job has been served before s_1 units of time in server 1, it leaves the system; otherwise, when the execution time equals s_1, it is stopped and sent to the end of the queue of server 2, where the execution starts from scratch[2]. Thus, jobs that are executed in server i have been previously executed in servers $1, 2, \ldots, i-2$ and $i-1$, respectively, $s_1, s_2, \ldots, s_{i-2}$ and s_{i-1} units of time. Besides, if a job is being processed by the ith server and its execution time is less than s_i time units, it leaves the system and, if not, it is stopped and put at the end of the queue of the next server. Jobs at the last server always run to completion.

For a given vector of cutoffs $\mathbf{s}$, we denote by $W(\mathbf{s})$ the random variable of the waiting time of incoming jobs. For a given vector of cutoffs $\mathbf{s}$, we will be interested in analyzing the normalized mean waiting time, which is given by

$$E(\bar{W}(\mathbf{s})) = \frac{E(W(\mathbf{s}))}{E(X)},$$

where $E(W(\mathbf{s}))$ is the mean waiting time of jobs in the system.

Let $\mathbf{s}^{opt} = (s_1^{opt}, ..., s_{h-1}^{opt})$ be a vector of cutoffs that minimizes the mean waiting time of jobs among all possible cutoffs, i.e.,

$$\mathbf{s}^{opt} \in \arg\min_{\mathbf{s}} E(W(\mathbf{s})).$$

To simplify notation, we write $E(\bar{W}(\mathbf{s}^{opt})) = E(\bar{W}^*)$ for the optimal normalized mean waiting time and $E(W(\mathbf{s}^{opt})) = E(W^*)$ for the optimal mean waiting time.

Throughout the paper, we will use the notation $f \sim g$ to denote two range r dependent quantities f, g whose ratio tends to 1 as r tends to infinity. In that case we will say that f and g are asymptotically equal. If g optimizes a certain target function then we will say that f is asymptotically optimal. We also use the notation $[x]$ for the integer part of x.

Our analysis of normalized average waiting time for TAGS will rely on the approximation formulas that were introduced in [14]. The approximation assumes that the arrival process to all servers is Poisson and then uses the Pollaczek-Khinchine formula. It has been verified numerically in [5] that this is a good and conservative approximation, i.e. actual performance tends to be slightly better than the approximation, which can therefore be used safely.

The r.v. of sizes of the jobs executed in server i is denoted by X_i. The probability that a job size belongs to the interval $[s_{i-1}, s_i]$ is given by p_i. Likewise, the probability for a job to be executed in server i is denoted by $\bar{p}_i$. From the definition of the TAGS policy, it follows that $p_i = F(s_i) - F(s_{i-1})$ and $\bar{p}_i = 1 - F(s_{i-1})$. The waiting time in server i is denoted by $W_i(\mathbf{s})$ and the load in server i by ρ_i.

Consider a distribution X such that $Pr(X < 1) = 0$. We can define the restriction X_r^{res} of X to the interval $[1, r]$ by the formula $Pr(a \leq X_r^{res} \leq b) = \frac{Pr(a \leq X \leq b)}{Pr(X \leq r)}$ for all $[a, b] \subset [1, r]$ and $Pr(a \leq X \leq b) = 0$ if $a > r$. The restriction is simply the distribution induced on jobs whose size is in the interval $[1, r]$. Obviously, the range of X_r^{res} is at most r.

[2]We assume that the overhead due to jobs being reassigned to another server when their execution times exceed the cutoff value associated with the current server is negligible.

Let X be a distribution on $s \geq 1$ and consider a set of intervals $\mathbf{I}$, consisting of disjoint intervals $I_i = [s_{i,1}, s_{i,2})$, $i = 1, \ldots, k$, i.e.,

$$s_{1,1} < s_{1,2} \leq s_{2,1} < s_{2,2} \leq \ldots \leq s_{k,1} < s_{k,2}.$$

We allow $s_{k,2} = \infty$. We define the distribution $X_{\mathbf{I}}^{dis}$ obtained from X by changing the size of a job whose size is in I_i by a job of size $s_{i,1}$. This distribution coincides with X outside the union of the intervals has the value $s_{i,1}$ with probability $Pr(X \in I_i)$, since all jobs in the interval now have the smallest value in the interval, i.e., $s_{i,1}$. The discretization above corresponds to the singleton family of intervals $[r, \infty)$

As a particular instance of a discretization, consider the single interval $I_1 = [r, \infty)$. We denote this discretization by X_r^{dis}. The distribution coincides with X in the range $[1, r)$ and assigns the value r with probability $Pr(X \geq r)$.

Both the restriction X_r^{res} and the discretization X_r^{dis} are supported in the range $[1, r]$ and as $r \to \infty$ provide finer approximations of X.

We will sometimes be interested in the performance of TAGS for the Bounded Pareto distribution, that we present in the next section.

3.1. Bounded Pareto Distribution

Let $B(\alpha)$ denote the distribution whose density function for $s \geq 1$ is given by $f(s) = \alpha s^{-\alpha-1}$ and zero otherwise. The family of Bounded Pareto distributions $B_{r,\alpha}$ is the family of restrictions of $B(\alpha)$ to the intervals $[1, r]$. In more detail, let $a = 1/r$, then the density function of $B_{r,\alpha}$ has the form

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

in the range $1 \leq s \leq r$ and 0 otherwise,

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 $\alpha \neq 1$, we have

$$E(B(\alpha)) = \frac{\alpha}{\alpha - 1} \frac{1 - a^{\alpha-1}}{1 - a^\alpha}, \tag{1}$$

whereas

$$E(B(1)) = \frac{\ln(r)}{1 - a}. \tag{2}$$

We note that the Bounded Pareto distribution with $\alpha = -1$ coincides with the uniform distribution on the interval $[1, r]$.

The family of Bounded Pareto distribution with large range and $0 < \alpha < 2$ has been used in many studies as a good model for high variance job size distributions [16, 17, 8]. In particular, it was found that values of α in the range $[0.9, 1.1]$ around $\alpha = 1$ are particularly good for modeling request sizes and file sizes in internet traffic.

4. Stability Analysis

In this section, we study the stability of a system operating under the TAGS policy. We first provide a necessary and sufficient condition for the system to be stable for an arbitrary distribution and then we focus on this stability condition for several distributions.

4.1. An arbitrary distribution

Consider a TAGS system with job size distribution X, load ρ and h servers. We define $\rho_{crit}(X,h)$ to be the maximal (supremum) load that a TAGS system with job size distribution X and h servers can handle with stability. The value of $\rho_{crit}(X,h)$ is increasing in h. Consequently, we define $\rho_{crit}(X) = \lim_{h\to\infty}\rho_{crit}(X,h)$.

We note that TAGS is a multi-class forward feeding network in the sense that jobs never re-enter the same server. For such systems stability is equivalent to the condition that each server in the network experiences a load which is less than 1, [10]. The following theorem provides a formula for $\rho_{crit}(X)$.

Theorem 4.1. *Let $X(s)$ be any bounded job size distribution with range in $[1,r]$. Let*

$$M(X) = \sup_s s(1-X(s))$$

where sup *denotes the supremum, then*

$$\rho_{crit}(X) = E(X)/M(X)$$

Proof. See Appendix A □

In the above result, we present $\rho_{crit}(X)$ for a given distribution X. We now exploit the shape of the expression of $\rho_{crit}(X)$ to analyze the distribution that maximizes $\rho_{crit}(X)$ over all X in the range $[1,r]$ and we show that this distribution is $B(1)$, i.e., the Bounded Pareto distribution with $\alpha = 1$.

Theorem 4.2. *For any distribution X in the range $[1,r]$, we have $\rho_{crit}(X) \le \ln(r)+1$. The discretization of the Pareto distribution with parameter $\alpha = 1$ with respect to the interval $[r,\infty)$ achieves the bound.*

Proof. See Appendix B □

A quantity which is closely related to $\rho_{crit}(X,h)$ is $h_{min}(X,\rho)$ which is the minimal number of servers needed for a stable system with job size distribution X and load ρ, i.e., $\rho_{crit}(X, h_{min}(X,\rho)-1) \le \rho$ and $\rho_{crit}(X, h_{min}(X,\rho)) > \rho$. This number is finite if and only if $\rho < \rho_{crit}(X)$. For a given pair X,h, the value of $\rho_{crit}(X,h)$ is the maximal load ρ for which $h_{min}(X,\rho) = h$. For a given $\rho < \rho_{crit}$ and a bounded distribution X with range in $[1,r]$, the quantity $h_{min}(X,\rho)$ can be computed iteratively by the procedure we present next.

Let $s_0 = 1$. For $i \ge 1$, we assume that s_{i-1} has been computed. We let $p_i = Pr(X \ge s_{i-1})$. We obviously have $p_1 = 1$. For $s \ge s_{i-1}$ we define $X_{i,s}$ to be the restriction of X to the interval $[s_{i-1}, s]$. this means that for all t, $Pr(X_{i,s} \ge t) = \frac{Pr(t \le X \le s)}{Pr(s_{i-1} \le X \le s)}$. Let $\rho_i(s) = \frac{\rho}{E(X)}(E(X_{i,s}) + sPr(X \ge s))$. It is easy to see that $\rho_i(s)$ is a non-decreasing function of s with $\rho(s_{i-1}) = s_{i-1}Pr(X \ge s_{i-1}) < 1$, the last inequality by our assumption that $\rho < \rho_{crit}$. We let s_i be such that $\rho_i(s_i) = 1$. If no such s_i exists, i.e., $\rho_i(s) < 1$ for all $s \ge s_{i-1}$ we terminate the process and return $h_{min}(X,\rho) = i$ and $s(X,\rho) = s_{i-1}$.

We note that the s_i, $i = 1, \ldots h_{min}(X,\rho)$ cannot be used in a stable TAGS system since the utilization on server i, ρ_i will be precisely 1, and hence will not satisfy the stability condition $\rho_i < 1$. For future reference, assume, $\varepsilon > 0$ is small, we can repeat the iterative procedure constructing cutoffs s_i^ε which satisfy the stable condition $\rho_i(s_i^\varepsilon) = 1-\varepsilon$. As before, if no such s_i exists, i.e., $\rho_i(s) < 1-\varepsilon$ for all $s \ge s_{i-1}^\varepsilon$ we terminate the process and return $h_{min}^\varepsilon(X,\rho) = i$ and $s^\varepsilon(X,\rho) = s_{i-1}$. We have $\lim_{\varepsilon\to 0} s_i^\varepsilon = s_i$ and for ε small enough the process will terminate with $h_{min}^\varepsilon(X,\rho) = h_{min}(X,\rho)$.

Using the above technique, we compute in Section 7 the value of $h_{min}(X,\rho)$ for the Bounded Pareto distribution and a wide range of parameters.

4.2. The Bounded Pareto distribution

We study the critical load of Bounded Pareto distribution in this section. We first consider that $r < \infty$ and we characterize the critical load for this case in the following result.

Theorem 4.3. *Consider the Bounded Pareto job size distribution. Let $a = 1/r$ be the reciprocal to the range of the distribution.*

- *If $\alpha < 1$ and $a \leq (1-\alpha)^{1/\alpha}$ then,*

$$\rho_{crit}(X) = (1 - a^{1-\alpha})(1-\alpha)^{-1/\alpha}$$

- *If $\alpha > 1$, or, $\alpha < 1$ and $a^\alpha \geq 1 - \alpha$ then*

$$\rho_{crit}(X) = \frac{\alpha}{\alpha - 1}(\frac{1 - a^{\alpha-1}}{1 - a^\alpha})$$

- *If $\alpha = 1$ then $\rho_{crit}(X) = \frac{r}{r-1}\ln(r)$.*

Proof. For the Bounded Pareto distribution with $\alpha \neq 1$, the supremum of $s(1-F(s))$ is achieved when $s = r(1-\alpha)^\alpha$ and therefore

$$M(X) = r^{1-\alpha}\frac{(1-\alpha)^{1/\alpha}}{1-a^\alpha}\frac{\alpha}{1-\alpha}.$$

Dividing (1) by the above expression, it results that

$$\rho_{crit} = (1 - a^{1-\alpha})(1-\alpha)^{-1/\alpha}.$$

For $\alpha = 1$, since $s(1-F(s))$ is a decreasing function of s, the supremum of $s(1-F(s))$ is given when $s = 1$ and, therefore, $M(X) = 1$. As a result, we have that $\rho_{crit} = E(X)$ and using (2) the desired result follows. □

In the following result, we study the critical load for a fixed α as $r \to \infty$.

Corollary 4.4. *When $r \to \infty$, for the Bounded Pareto distribution with parameter α,*

- *If $\alpha < 1$ then $\rho_{crit}(X) = (1-\alpha)^{-1/\alpha}$*
- *If $\alpha = 1$ then $\rho_{crit}(X) = \infty$*
- *If $\alpha > 1$ then $\rho_{crit}(X) = \frac{\alpha}{\alpha-1}$.*

4.3. The Weibull distribution

The Weibull distribution with parameter k has a distribution function of the form

$$W_k(s) = 1 - e^{-s^k}$$

for $s \geq 0$. We will compute the critical load for this unbounded distribution. We can think of the result as the limit of the critical load for the distributions which are obtained by restricting the Weibull distribution to an interval $[\varepsilon, p]$ and letting $\varepsilon \to 0$ and $p \to \infty$. We note that after a change of time unit, this is the same considering the interval $[1, r = p/\varepsilon]$ for the scaled distribution which is also a bounded piece of a Weibull distribution. It is known that $E(W_k) = \Gamma(1 + \frac{1}{k})$ where $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$ is the Gamma function. This is easily seen by using the change of variable $t = s^k$. To compute the maximum of $s(1 - W_k(s))$ we differentiate $(s(1 - W_k(s))' = e^{-s^k}(1 - ks^{k-1})$. Setting to zero we get $s = (\frac{1}{k})^{1/k}$ and for that value of s we get $s(1 - W_k(s)) = (\frac{1}{ke})^{1/k}$, where e is the natural base. Letting $k \to \infty$, we see that $E(W_k) \to \Gamma(1) = 1$, hence the critical load tends to 1. On the other hand, if $k \to 0$, say

$k = 1/n$ with n an integer, we get $E(W_{1/n}) = \Gamma(n+1) = n!$. The maximal value of $s(1 - W_k(s))$ is $(\frac{n}{e})^n$. According to Stirling's formula, as $n \to \infty$, $\rho_{crit}(W_{1/n}) \sim \sqrt{2\pi n}$.

For the exponential distribution with rate 1, which is given by setting $k = 1$ we get that the critical load is e. The critical load is easily seen to be invariant under a change of time units, $t \to ct$ for some constant c, since both the numerator and denominator in the definition are multiplied by c. If X is a distribution we say that the distribution obtained by the change in time units is a scaling of X by the factor c. The exponential distribution with rate λ is a scaling of the exponential distribution with rate 1 by the factor λ, hence the critical load is e for all exponential distributions.

4.4. A Mixture of a set of distributions

We recall that, given a set of distributions X_i, $i = 1, \ldots, m$, and corresponding probabilities p_i, we define their mixture X by the formula $Pr(X \geq s) = \sum_{i=1}^{m} p_i Pr(X_i \geq s)$, for all $s \geq 0$. More generally we can define a degenerate mixture by having non-negative weights p_i with $\sum_i p_i \leq 1$ and using the same formula for $s > 0$. For each degenerate mixture X, we can attach a mixture $\bar{X}$ by using the probabilities $q_i = p_i / \sum_i p_i$. The degenerate mixture is retrieved from $\bar{X}$ by taking a mixture with a δ function distribution with value 0, i.e., $Pr(X \geq s) = (\sum_i p_i) Pr(\bar{X} \geq s) + (1 - \sum_i p_i) Pr(\delta_0 \geq s)$, where in general, δ_s denotes the distribution which gives the value s with probability 1. We note that $\rho_{crit}(\bar{X}) = \rho_{crit}(X)$, since both the denominator and numerator in $\rho_{crit}(X)$ are given by multiplying the corresponding quantity in $\rho_{crit}(\bar{X})$ by $\sum_i p_i$.

We have the following simple lemma about the behavior of ρ_{crit} with respect to certain mixtures.

Lemma 4.5. *If X is a (possibly degenerate) mixture of X_i, $i = 1, \ldots, m$ with corresponding non-negative weights p_i and assume that there is a number ρ such that for all $1 \leq i \leq m$, $\rho_{crit}(X_i) = \rho$, then $\rho_{crit}(X) \geq \rho$.*

Proof. From the above discussion, without loss of generality, we can replace X by $\bar{X}$ and assume that we have a true mixture. Let s be the value for which $M(X) = sPr(X \geq s)$, then $M(X) = \sum_i p_i s Pr(X_i \geq s) \leq \sum_i p_i M(X_i)$. On the other hand we have $E(X) = \sum_i p_i E(X_i)$, hence

$$\rho_{crit}(X) = E(X)/M(X) \geq \frac{\sum_i p_i E(X_i)}{\sum_i p_i M(X_i)} = \frac{\sum_i p_i \rho M(X_i)}{\sum_i p_i M(X_i)} = \rho$$

as desired. □

Given a distribution X, we can consider the family of distributions which are mixtures of scalings of X. For the exponential distribution, this family consists of the hyper-exponential distributions. We have the following corollary of the above lemma and its proof.

Corollary 4.6. *Consider a distribution X and let Y be a mixture of scalings of X, then $\rho_{crit}(Y) \geq \rho_{crit}(X)$. Moreover, if Y is a non-trivial mixture involving two different scales or more and $sPr(X \geq s)$ has a unique maximum, then $\rho_{crit}(Y) > \rho_{crit}(X)$.*

The corollary gives a lower bound of e for the critical load of all hyper-exponential distributions. We also note, that if X is bounded Pareto $B_r(\alpha)$, then we can form mixtures of scalings that will produce distributions of the form $B_q(\alpha)$, for $q > r$ and so we expect the critical load for the family of $B_r(\alpha)$ to be increasing in r and this is indeed the case.

5. Bounds on the Performance of TAGS

We consider a TAGS system, with h servers, job size distribution X, supported on the interval $[1, r]$ and utilization ρ. Let $E(\bar{W}^*_{max}(h, \rho, r))$ be the supremum of $E(\bar{W}^*(h, \rho, X))$, the normalized waiting time with optimal cutoffs, over all distributions in the range $[1, r]$, with fixed h and ρ. We first consider that the system load is such that $\rho < 1$ and then we focus on higher loads.

5.1. The case $\rho < 1$

The following two results provide an upper and lower bound on $E(\bar{W}^*_{max}(h,\rho,r))$ when $\rho < 1$. The upper bound is obtained using cutoffs that are independent of the distribution X and of ρ and in this sense, these cutoffs provide a universal performance guarantee. The bounds match up to a constant factor when h is fixed.

Theorem 5.1. *Consider the set of cutoffs $\mathbf{s}_{r,h}$ given by $s_i = r^{i/h}$, then for any $\varepsilon > 0$, there is an r_ε, such that for all $r > \varepsilon$, any TAGS system with h hosts, utilization $\rho < 1$ and any job size distribution $X = X_r$ with $Pr(X > r) = 0$ we have*

$$E(\bar{W}^*) \leq E(\bar{W}(\mathbf{s}_{r,h}) \leq (2+\varepsilon)\frac{\rho}{4(1-\rho)} r^{1/h} \tag{3}$$

Proof. Consider host i in the TAGS system. Let $E(W_i)$ denote the average waiting time at host i. Recall that p_i denotes the portion of jobs that end their service at host i and that $\bar{p}_i$ denotes the portion of jobs that pass through host i. The total average waiting time that is experienced by a job which ends service at host i is $\sum_{j=1}^{i} E(W_j)$ and hence the average waiting time is given by

$$E(W(\mathbf{s}_{r,h})) = \sum_{i=1}^{h} p_i (\sum_{j \leq i} E(W_j) = \sum_{j=1}^{h} E(W_j)(\sum_{i=j}^{h} p_i) = \sum_{j=1}^{h} \bar{p}_j E(W_j)$$

Let X_i be the job size distribution experienced by host i. The distribution X_i is a mixture of the distribution X restricted to the interval $[r^{(i-1)/h}, r^{i/h}$ and the distribution $\delta_{r^{i/h}}$ with weights $p_i/\bar{p}_i$ and $\bar{p}_{i+1}/\bar{p}_i$ respectively. Let $f(s)$ be the density of X, then we have

$$E(X_i) = \frac{\int_{s_{i-1}}^{s_i} f(s)ds + s_i \bar{p}_{i+1}}{\bar{p}_i}$$

let $E(\bar{W}_i) = E(W_i)/E(X_i)$ be the normalized waiting time of server i. By [3] we have

$$E(\bar{W}_i) \leq \frac{\rho_i}{4(1-\rho_i)} r^{1/h} \leq \frac{\rho}{4(1-\rho)} r^{1/h}$$

We write

$$E(W(s)) = \sum_{j=1}^{h} \bar{p}_j E(W_j) = \sum_{j=1}^{h} \bar{p}_j E(X_j) \frac{E(W_j)}{E(X_j)} \leq (\sum_{j=1}^{h} \bar{p}_j E(X_j)) \frac{\rho}{4(1-\rho)} r^{1/h}$$

We have

$$\sum_{j=1}^{h} \bar{p}_j E(X_j) = (\sum_{j=1}^{h} \int_{s_{j-1}}^{s_j} f(s)ds) + \sum_{j=1}^{h} s_j \bar{p}_{j+1}$$

$$= E(X) + \sum_{j=1}^{h} s_j Pr(X \geq s_j) = E(X) + \sum_{j=1}^{h} r^{j/h} Pr(X \geq r^{j/h}) = E(X) + \sum_{j=1}^{h} r^{j/h} p_j.$$

We claim that $\sum_{j=1}^{h} r^{j/h} \bar{p}_j \leq E(X) \frac{r^{1/h}}{r^{1/h}-1}$. Since $\lim_{r\to\infty} \frac{r^{1/h}}{r^{1/h}-1} = 1$, this will prove the theorem. Consider the set of intervals $\mathbf{I}$ consisting of $[1, r^{1/h}], [r^{1/h}, r^{2/h}], \ldots r^{(h-1)/h}, r$ and consider the distribution $X^{dis}_{\mathbf{I}}$. We have $Pr(X \geq r^{j/h}) = Pr(X^{dis}_{\mathbf{s}} \geq r^{j/h}) = \bar{p}_j$ for all $j = 0, \ldots, h$. In addition, since discretization never increases the moments of a distribution, we have $E(X^{dis}_{\mathbf{I}}) \leq E(X)$. We conclude that it is sufficient to show that $\sum_{j=1}^{h} r^{j/h} \bar{p}_j \leq E(X^{dis}_{\mathbf{I}}) \frac{r^{1/h}}{r^{1/h}-1}$. Since $X^{dis}_{\mathbf{I}}$ only has values in $r^{j/h}$ and $Pr(X^{dis}_{\mathbf{I}} \geq r^{j/h}) = \bar{p}_j$ by construction we have

$$E(X^{dis}_{\mathbf{I}}) = \sum_{j=0}^{h} r^{j/h} p_j \geq \sum_{j=1}^{h} r^{j/h} p_j.$$

Finally,

$$\sum_{j=1}^{h} r^{j/h}\bar{p}_j = \sum_{j=1}^{h} p_j r^{j/h}(\sum_{k=0}^{j-1} r^{-k/h}) \leq \sum_{j=1}^{h} p_j r^{j/h}\frac{r^{1/h}}{r^{1/h}-1} \leq E(X_{\mathbf{I}}^{dis}) \leq E(X)\frac{r^{1/h}}{r^{1/h}-1}$$

as required. □

In the previous result, we provide an upper bound of $E(\bar{W}^*)$. In the next result, we provide a lower bound.

Theorem 5.2. *Let $\varepsilon > 0$. There exists a distribution $X_{r,h}$ with range r, such that for any $\rho < 1$*

$$E(\bar{W}^*(h,\rho,X_{r,h})) \geq (1-\varepsilon)\frac{1}{(h+1)^2}\frac{\rho}{1-\frac{2}{h+1}\rho}r^{1/h} \tag{4}$$

Proof. For the lower bound, we look at the discretization of $B(1)$ at the set of intervals $\mathbf{I}$, given by $[1, r^{1/h}], [r^{1/h}, r^{2/h}], \ldots r^{(h-1)/h}, r$ and $[r, \infty]$. This distribution has $h+1$ values. An asymptotically optimal set of cutoffs in this case is $s_i = r^{i/h}$, $i = 1, \ldots h-1$. The distribution has $h+1$ values and a trivial calculation shows that $E(X_{r,h}) \sim h+1$. Since there are just h hosts, one of the hosts has to to span two of the values. If they are not consecutive values then the contribution of that host to the average waiting time will be at least of the order of $r^{2/h}$. We conclude that one of the hosts will span two consecutive values and the others just a single value and that is precisely what we obtain with the present cutoffs. The normalized waiting time is dominated by the last host and is asymptotically given by

$$E(\bar{W}^*) \sim \frac{1}{(h+1)^2}\frac{\rho}{1-\frac{2}{h+1}\rho}r^{1/h}$$

as required. □

5.2. The case $\rho < \rho_{crit}(X,h)$

We can generalize Theorem 5.1 to higher loads if we consider families of distributions X_r^{dis} and X_r^{res} which are constructed from an unbounded distribution X, via discretization or restriction. In that case, we can generalize to $\rho < \rho_{crit}(X,h)$.

Theorem 5.3. *Given some distribution X with $Pr(X < 1) = 0$, let X_r^{dis} be the family of discretizations of X with respect to the interval $[1,r]$ and X_r^{res} the family of restrictions with respect to the same interval. Consider a utilization $\rho < \rho_{crit}(X,h)$. Let $\bar{h} = h_{min}(X,\rho)$ and let $h_{sp} = h - \bar{h}$ be the number of spare servers. For any $\varepsilon > 0$, there is an $r(X,\rho)$ such that for all $r > r(X,\rho)$, we have*

$$E(\bar{W}_r^*) \leq (2+\varepsilon)\frac{1-\rho(X,h)+\rho}{4(\rho(X,h)-\rho)}(\frac{r}{s(X,\rho)})^{1/h_{sp}} \tag{5}$$

where $E(\bar{W}_r^*)$ denotes the normalized average waiting time in a TAGS system with h hosts, utilization ρ, job size distribution either X_r^{dis} or X_r^{res} and optimal cutoffs.

Proof. Let $[1,q]$ be an interval. It is easy to check that X_r^{dis} and X_r^{res} converge as $r \to \infty$ to X on the interval, in the sense that for any $\eta > 0$ we have for any $[a,b] \subset [1,q]$

$$\frac{1}{1+\eta} \leq \frac{Pr(X \in [a,b])}{Pr(X_r^{dis} \in [a,b])} \leq 1+\eta$$

and the same for X_r^{res} for $r \geq r_{eta}$. Also let r_η be large enough so that $\rho(X,h) - \rho(X_r^{dis},\rho) < \eta$ and the same for X_r^{res}. We also choose r_η large enough so that $h_{min}(X_r^{dis},\rho) = \bar{h}$. Choose $\delta = \delta_\eta$ small enough and r_η large enough such that $\bar{h} = h_{min}^{\delta}(X_r^{dis},\rho)$. Consider the cutoffs $s_i = s_i^{\delta}$, $i = 1, \ldots, \bar{h}$

and for $i = \bar{h}+1, \ldots, h$ we let $s_i = s_{i-1}(\frac{r}{s^{\delta}_{\bar{h}}})^{\frac{1}{h_{sp}}}$. Choose δ_η small enough and r_η large enough such that $\bar{h} = h^{\delta}_{min}(X^{dis}_r, \rho)$ and $\frac{1}{1-\eta} > \frac{s^{\delta}_{\bar{h}}}{s_{\bar{h}}} \geq 1-\eta$. By convergence we can always find such r_η and $\delta\eta$. The same is assumed to hold for X_{res_r}. The first $\bar{h}$ cutoffs are always bounded by $s(X,\rho)$, and their utilization by construction is $1-\delta$, hence for r large enough their contribution to the normalized waiting time is at most $(2+\varepsilon)\frac{1-\rho(X,h)+\rho}{4(\rho(X,h)-\rho)}(\frac{r}{s(X,\rho)})^{1/h_{sp}}\eta$. The proof now proceeds exactly as in the previous theorem, with the spare servers carrying the at most the range $[(1-\eta)s(X,\rho), r]$ and there are h_{sp} of them with utilization at most $(1+\eta)(1-\rho(X,h)+\rho)$, leading to the required result. □

We note that the same result and with the same proof holds for SITA, but in that case $h_{min}(X,\rho)$ is simply $[\rho]+1$, where $[x]$ refers to the integer part. Consequently, the spare number of servers is $h-[\rho]-1$.

6. Asymptotic analysis of TAGS

In this section, we consider that r tends to infinity. We first study this asymptotic regime for an arbitrary distribution and then we focus on the Bounded Pareto distribution.

6.1. An arbitrary distribution

We now consider the distribution $x_{r,h}$ that we defined in Theorem 5.2. The following theorem compares the performance of SITA, LWL, Random and TAGS for this distribution.

Theorem 6.1. *Let $\rho < 1$. Let $E(\bar{W}^*_T(h,\rho,X))$, $E(\bar{W}^*_S(h,\rho,X))$, be the parameter optimal normalized average waiting time of TAGS and SITA with h servers, utilization ρ and job size distribution X. Let $E(\bar{W}^*_L(h,\rho,X))$ and $E(\bar{W}^*_R(h,\rho,X))$ be the normalized average waiting time of LWL and RA systems, respectively. We have for r large enough,*

1) $E(\bar{W}^*_R(h,\rho,X_{r,1})) \sim \frac{1}{2}\frac{\rho}{2h(1-\frac{\rho}{h})}\frac{r}{2}$.

2) $E(\bar{W}^*_L(h,\rho,X_{r,1})) > (1-\varepsilon)\varepsilon\frac{(\rho/2)^h}{h!}e^{-\rho/2}\frac{r}{2}$

3) $E(\bar{W}^*_S(h,\rho,X_{r,h})) > \frac{1}{h+1}(1-\varepsilon)\frac{\rho}{2h(1-\frac{\rho}{h})}r^{1/h}$

4) $E(\bar{W}^*_T(h,\rho,X_{r,h})) > (1-\varepsilon)\frac{1}{(h+1)^2}\frac{\rho}{(1-\frac{2\rho}{h+1})}r^{1/h}$

Proof. Item 1 is simply an application of the Pollaczek-Khinchine formula when each server has utilization ρ/h. Item 4 was already shown, so we are left with items 2,3.

For LWL, consider some job x that arrives at time t. Consider the time interval $[t-(1-\varepsilon)r, t)$. We claim that if there were h large jobs of size r that arrived during that time interval, then job x will wait at least εr time. If the h large jobs all went to different servers then this is obvious, since they each have at least that amount of time to complete and if one of the h large jobs joined the queue of another of the h large jobs, thats because all the other servers had even more work left which leads to the same conclusion. Since $E(X_{r,1}) \sim 2$, we have $\lambda \sim \rho/2$. Since the probability of a large job is $1/r$, the rate of large jobs λ_r, satisfies $\lambda_r \sim \frac{\rho}{2r}$. Since arrivals are assumed to be Poisson, the number of large jobs arriving during the time interval in question is Poisson distributed with parameter $\lambda_r(1-\varepsilon)r \sim \rho(1-\varepsilon)/2$. The probability of h arrivals is asymptotically $\frac{(\rho(1-\varepsilon)/2)^h}{h!}e^{-\rho/2}$ and taking ε to zero, yields the desired bound.

For SITA, we first have to consider what cutoffs mean in the context of a discrete distribution such as $X_{r,h}$. It is natural to interpret SITA as a scheme which decides which portion of the requests of each size is handled by each host. This interpretation is stable under small perturbations of the distribution which replace the delta functions at various values by a continuous approximation. In the case of $X_{r,h}$, asymptotically, to be optimal, each server will have to be allocated to only 2 consecutive values. It is also easy to check that it is not optimal to assign a portion of a single value to a server, since handling all requests to a single value has equally negligible affect on the contribution of the server to average waiting time, and this may help lower the contribution of other servers. We conclude that in an optimal

cutoff setting, host i services a portion $c_{1,i}$ of requests with a value v and a portion $c_{2,i}$ of requests of size rv. The host utilization in this case is proportional to $c_{1,i} + c_{2,i}$, while the average normalized waiting time for the server is given by $Min(\frac{c_{1,i}}{c_{2,i}}, \frac{c_{2,i}}{c_{1,i}})r$. We see that the waiting time contribution of the host is at least that of a server for which $c_{1,i} = 1$ and $c_i = c_{2,i} = c_{1,i} + c_{2,i} - 1$. By concavity of the influence of utilization on waiting time, if we have a pair of servers with coefficients $1, c_i$ and $1, , c_j$, their joint contribution to average waiting time is at least that of two servers with identical coefficients $1, (c_i + c_j)/2$. Since there are $h + 1$ values we have $\sum_i c_{1,i} + c_{2,i} = h + 1$ and hence $\sum_i c_i = 1$ and we obtain a lower bound on the average waiting time equal to that of h hosts with coefficients $1, 1/h$ and the inequality of item 3 is obtained by applying Pollaczek-Khinchine asymptotically to such hosts. □

The condition $\rho < 1$ is essential for the favorable comparison with the other policies, since for any $\rho > 1$ and any distribution X, if we consider X_r^{res} for r such that $1+\ln(r) < \rho$, we have $\rho_{crit}(X_r^{res}) < \rho$ by Theorem 4.2. Consequently, using $[\rho]+1$ servers, we cannot compare with the other policies which can always handle a load of ρ, regardless of job size distribution X and regardless of the range.

6.2. The Bounded Pareto distribution

When $\rho < 1$, we compare the performance of TAGS and that of SITA in the following result.

Theorem 6.2. *Let $\rho < 1$. For Bounded Pareto distributed job sizes with $r \to \infty$, the mean waiting time in a TAGS system with optimal cutoffs is at most two times larger than the mean waiting time of a SITA system with optimal cutoffs.*

Proof. See Appendix C. □

We know that the SITA policy requires the knowledge of the size of incoming tasks, whereas the TAGS policy does not. Therefore, a system operating under the SITA policy always outperforms a system operating under the TAGS policy. However, an important conclusion from the result of the above theorem is that, in the asymptotic regime, the penalty for not knowing job sizes is upper bounded by a factor of 2, for any value of α and any number of servers. Besides, taking into account that the vector of cutoffs that minimizes the mean waiting time also minimizes the normalized mean waiting time, we conclude that when r tends to infinity, the normalized mean waiting time in a TAGS system with optimal cutoffs is at most two times larger the normalized mean waiting time of a SITA system with optimal cutoffs.

In the following result we give an expression of the order of magnitude of the ratio of the performance of a system operating under the TAGS policy for $\rho > 1$.

Proposition 6.3. *When $\rho > 1$,*

$$\mathbb{E}[\bar{W}^*] = \Theta(r^{\frac{|2\alpha-2|}{q^{\tilde{h}}-1}}), \tag{6}$$

where $q = \frac{\alpha}{2-\alpha}$ if $\alpha > 1$, $q = \frac{2-\alpha}{\alpha}$ if $\alpha < 1$ and $q = 1$ if $\alpha = 1$ and $\tilde{h}$ is the number of spare servers.

Proof. See Appendix D. □

We observe that the order of magnitude of the performance of a system operating under the TAGS policy depends on the number of servers and the minimum number of servers to stabilize the system only through the number of spare servers.

7. Numerical experiments

7.1. The approximation equations

The mean waiting time of jobs in a system operating under the TAGS policy with Poisson arrivals has no exact analytical formula. The reason for this is that the input stream to the second server and beyond is not Poisson. We analyze the approximation suggested in [14] that consists of assuming Poisson arrivals to all servers. Indeed, under this assumption, an approximation to the average waiting

time can be computed using the Pollaczek-Khinchine equation for an M/G/1 queue. The author in [14] also suggested that the approximation will over-estimate the average waiting time since the input streams, to all but the first server, tend to be more regular than Poisson, having near constant inter-arrivals.

To analyze the true performance of TAGS and to compare it with the approximation equations, we developed a simulation of a system operating under the TAGS policy. For each run, we consider an arrival traffic of 10^8 jobs and the Bounded Pareto job size distributions with different values of α, varying from 0 to 2, and different numbers of servers. The total load of the system is $\rho = h/2$, where h is the number of servers. The smallest job size was of size 1 and the largest job size was $r = 10^4$. This value was chosen because for $\alpha = 2$, the probability of a job of size greater than s is about s^2, hence the probability for a job of size greater than 10^4 is approximately 10^{-8}. This means that for α close to 2, and 10^8 jobs in a simulation run we would not get jobs substantially greater than 10^4, therefore, there was no point in choosing a larger value for r. The values of $s_1, \ldots, s_{h-1}$ for the system were chosen to be close to optimal for minimizing average waiting time in the approximate equations.

The results of Table 1 show that the average waiting time value which is computed using the approximation equations is always close to the value computed from the simulations. Moreover, as conjectured in [14], the computed value always over-estimates the actual average waiting time. As expected, the computed values are closest to the simulation results when the number of servers is small. The computed values for $h = 2$ are essentially identical to the simulated values, except for the case $\alpha = 1.8$ where there was a 10% difference. For larger values of h the error can be as large as 20%, a value which we still consider to be very reasonable. The largest errors occur for the extreme values of α, away from the central value $\alpha = 1$, where the errors are smallest.

We have also performed experiments with the larger value of the maximum job size and we have observed that, in all the cases, the obtained results follow the pattern presented in Table 1, and therefore we can conclude that the approximate equations are fairly accurate and conservative.

7.2. Comparison with SITA-E

In Theorem 6.2, we have shown that the ratio of the performance of a system operating under the optimal TAGS policy over the performance of a system operating under the optimal SITA in the asymptotic regime is upper bounded by two, i.e,

$$\frac{\mathbb{E}[W^*]}{\min_{\mathbf{s}} \mathbb{E}[W^{SITA}(\mathbf{s})}] \leq 2.$$

We now study this performance ratio when $\rho > 1$. We have computed the mean waiting time of jobs in a system operating under the SITA-E policy (i.e, the SITA policy where the load of the servers is equalized) for the parameters considered in Table 1. We know from [18] that, in a system with two servers, the optimal SITA balances the load of the servers for the Bounded Pareto distribution with $\alpha = 1$. Therefore, the performance ratio we present in Table 2 coincides with $\frac{\mathbb{E}[W^*]}{\min_{\mathbf{s}} \mathbb{E}[W^{SITA}(\mathbf{s})}]$ for $\alpha = 1$ and $h = 2$. For the rest of the cases, since $\mathbb{E}[W^{SITA-E}] \geq \min_{\mathbf{s}} \mathbb{E}[W^{SITA}(\mathbf{s})]$, it follows that the values we show in Table 2 are lower bounds of the values of the performance ratio we are investigating in this section.

As it can be seen in Table 2, for $\alpha = 1$, the difference on the performance of TAGS and SITA increases with the number of servers. In fact, when the number of servers is 7 and 8, the performance ratio is, respectively, 60.72 and 114.15. For $\alpha \neq 1$, we have found instances where the performance ratio is extremely high. For instance, when $\alpha = 0.2$ and the number of servers is 4, the ratio $\mathbb{E}[W^*]/\mathbb{E}[W^{SITA-E}]$ is 549.21, which means that the performance of a system operating under the optimal TAGS policy is more than 549.21 times the performance of a system operating under the optimal SITA policy.

7.3. Comparison with Random Assignment

In this section we explore the performance of TAGS and we show that the improvement of TAGS upon Random assignment is largest around the value $\alpha = 1$ even at small loads. This apparent

h	α	Simulated $\mathbb{E}[W^*]$	Calculated $\mathbb{E}[W^*]$
2	0.2	1368.72	1408.87
2	0.4	577.91	583.34
2	0.6	214.06	215.64
2	0.8	79.60	79.86
2	1	33.83	34.02
2	1.2	18.07	17.99
2	1.4	11.38	11.38
2	1.6	7.83	7.88
2	1.8	5.25	5.73
3	0.2	2323.63	2657.09
3	0.4	704.26	769.79
3	0.6	185.89	197.60
3	0.8	52.70	54.32
3	1	19.89	20.35
3	1.2	11.76	12.03
3	1.4	9.74	10.12
3	1.6	9.09	10.20
3	1.8	9.59	11.14
4	0.2	11180.74	13982.31
4	0.4	1468.58	1789.23
4	0.6	255.05	287.54
4	0.8	54.47	58.84
4	1	18.82	19.72
4	1.2	12.17	13.01
4	1.4	13.28	15.15
4	1.6	22.72	28.05
4	1.8	143.89	181.70
5	0.6	484.40	597.45
5	0.8	68.04	77.26
5	1	21.04	22.81
5	1.2	15.24	17.31
5	1.4	28.93	36.34
6	0.8	98.86	117.45
6	1	25.71	29.04
6	1.2	22.46	27.19
7	0.8	173.79	221.62
7	1	33.70	39.98
7	1.2	43.60	55.88
8	1	50.17	60.72

Table 1: Comparison of actual waiting time from simulations of TAGS systems, with the estimate from the approximation formulas, which assume Poisson arrivals at all servers.

h	α	$\mathbb{E}[W^{SITA-E}]$	$\mathbb{E}[W^*]/\mathbb{E}[W^{SITA-E}]$
2	0.2	91.66	14.93
2	0.4	38.56	14.98
2	0.6	16.31	13.12
2	0.8	9.11	8.73
2	1	10.507	3.21
2	1.2	17.59	1.03
2	1.4	16.266	0.69
2	1.6	9.0235	0.86
2	1.8	4.15	1.26
3	0.2	37.618	61.76
3	0.4	16.733	42.09
3	0.6	7.33	25.36
3	0.8	3.738	14.12
3	1	3.042	6.54
3	1.2	4.88	2.41
3	1.4	6.335	1.53
3	1.6	4.557	1.99
3	1.8	2.432	3.95
4	0.2	20.358	549.21
4	0.4	9.272	158.39
4	0.6	4.195	60.79
4	0.8	2.174	25.05
4	1	1.586	12.06
4	1.2	2.209	5.51
4	1.4	3.301	4.02
4	1.6	2.806	8.09
4	1.8	1.659	86.73
5	0.6	2.715	178.42
5	0.8	1.442	47.25
5	1	1.0274	20.49
5	1.2	1.273	12
5	1.4	3.013	9.61
6	0.8	1.03	95.98
6	1	0.731	35.17
6	1.2	0.843	26.64
7	0.8	0.775	224.24
7	1	0.555	60.72
7	1.2	0.608	71.71
8	1	0.4395	114.15

Table 2: Mean waiting time of the SITA-E policy and the performance of TAGS over the performance of SITA-E for the parameters considered in Table 1.

contradiction with some performance graphs in [14] is resolved by noting that our choice of scale invariant normalization differs from the normalization employed in [14]. In [14] the value of r was fixed as 10^{10} while k was varied to insure a fixed expectation of 3000. Consequently when $\alpha = 0.1$, the range of the distribution is 56 orders of magnitude, which is more than the ratio of the age of the Universe and a single compute cycle of a modern processor. For $\alpha = 1$ the range is the much more reasonable number of about 7 orders of magnitude. This huge disparity strongly affects the reported results in [14]. As we will show the new normalization provides a much more consistent comparison between different values of α.

We have chosen to compare TAGS with Random assignment rather than LWL since they have the same system setup requirements, that is, they are both stateless. When there are more than 2 hosts in the system the problem of finding optimal values for the parameters s_i (assuming they exist) becomes difficult. We have chosen to employ a mutation based genetic algorithm to search the parameter space. The results were compared with brute force searches over the entire parameter space for systems with 2 hosts and some with 4 hosts. The comparison showed that unless ρ is very close to ρ_{crit}, the genetic algorithm finds near optimal solutions. Details about the genetic algorithm can be found at [2]. We note that most of the insights which are provided in this paper were based initially on the exploration of TAGS using the genetic algorithm. Figure 1 shows the effect of the number of hosts on the relative

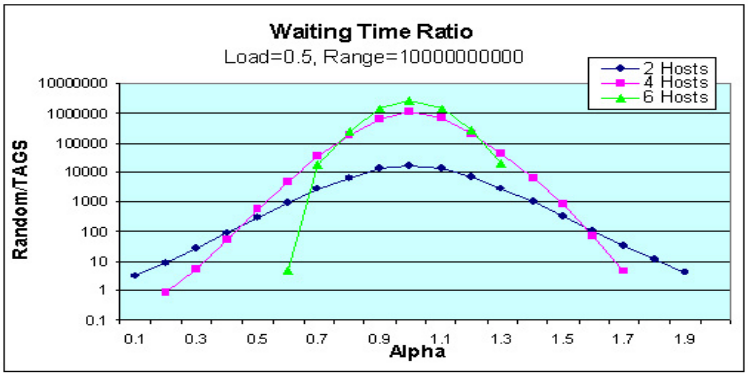

Figure 1: The effect of the number of hosts on the relative performance of TAGS.

performance of TAGS in comparison with Random assignment. The load in the experiments is fixed at 1/2 per host, hence the total load is $\rho = 1$ for 2-host systems, $\rho = 2$ for 4-host systems and $\rho = 3$ for 6-host systems. The range is fixed at the value $r = 10^{10}$. We computed for the different values of α the performance ratio between TAGS and Random assignment. As can be seen, the ratios are plotted in log scale.

We observe that moving from a 2-host system to a 4-host system makes sense in an interval around the value $\alpha = 1$ but not near the extremes at 0 and 2. In fact at a load of 0.5 per host there is no stable 4-host TAGS system when $\alpha = 0.1,\ 1.8,\ 1.9,\ 2.0$. This result of the genetic algorithm was verified using the recursive procedure for determining the minimal number of hosts for a stable TAGS. Improvements upon a 2-host system were obtained only in the interval $[0.5, 1.5]$. In that interval, a 4-host system yields fantastic improvements upon 2-host systems which already display great performance in comparison to Random (and LWL) assignment. The performance of a 6-host system is better than that of a 4-host system only in the interval $[0.8, 1.2]$ and even then the improvement is small. Further experiments show that for all values of α (including 1) the performance of an 8-host system is better than the performance of a 10-host system. We see that for load per host of 1/2 the optimal group size is always small (at most 8) and in many cases is at most 4. The performance improvement over Random of the hybrid strategy will be very significant. We note though that the last conclusion does depend on the size of the range as we will see later on.

We also note that in all cases (2, 4 and 6 hosts), the best relative performance is obtained when

$\alpha \approx 1$. The experiment with two hosts was also conducted in [14]. The results of those experiments which are reported in Figure 7(a) of [14], suggest that the relative performance of TAGS is a decreasing function of α. The apparent contradiction is resolved once we recall that the normalization employed in [14] does not fix the range and is not invariant under a change of time units. As will be seen later on in Figure 3 a larger range leads to improved relative performance of TAGS, hence, smaller values of α, which have larger range in the normalization given in [14], have a distinct advantage over larger values in the assessment of the TAGS algorithm. Given a fixed range comparison, the value $\alpha = 1$ had the best relative performance in all the experiments we have conducted as can be seen from all three figures. The contrast shows how important normalization is when dealing with heavy-tailed distributions.

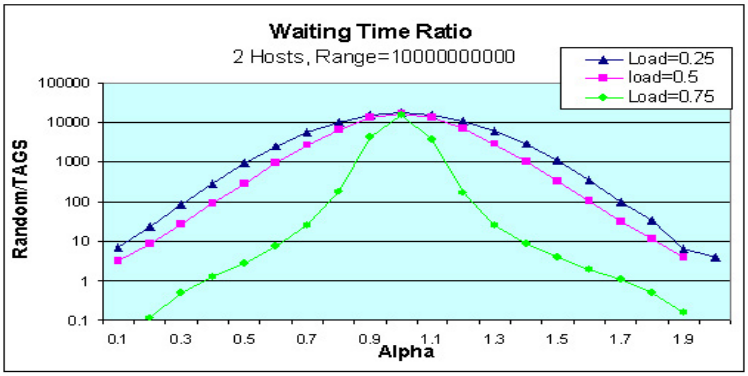

Figure 2: The effect of the load on the relative performance of TAGS.

In Figure 2 we consider the effect of the load on a 2-host system with fixed range. We see that the patterns for loads 0.25 and 0.5 per host are very similar, while for load 0.75 per host the peak of the graph is much more sharply concentrated around the value $\alpha = 1$. It is interesting to note that despite the change in shape, the size of the peak at 1 is nearly independent of the load.

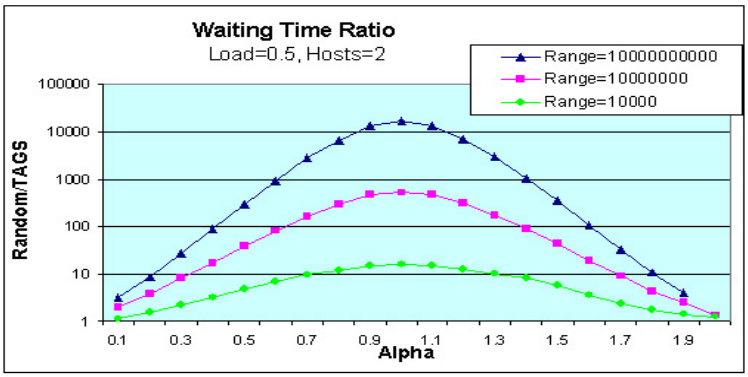

Figure 3: The effect of the range on the relative performance of TAGS.

In Figure 3 we consider the effect of the range, on a 2-host system with load 0.5 per host. We see that the range has a very strong effect on the relative performance and that TAGS does not offer substantial (order of magnitude) performance enhancements, for Bounded Pareto distributions, when the range is below 10^4 unless α is very close to 1. It is therefore important to calculate the range of the distribution and not just the parameter α when considering TAGS for handling a given distribution.

7.4. Analysis of $h_{min}(X, \rho)$

Figure 4 shows the results of computing $h_{min}(X, \rho)$ for Bounded Pareto distribution with α in the range [0.1,2] in jumps of 0.1, with the large value $r = 10^{10}$ and system loads of 1.5, 2.25, 3 and 4.5.

We observe that when the system load is $\rho = 1.5$ all values of α in the domain $[0.1, 2]$ have $h_{min}(X, 1.5) \leq 3$, i.e., 3 hosts suffice to produce a stable TAGS system. When the system load increases to $\rho = 2.25$, the α which satisfy $\alpha \geq 1.8$ do not have stable TAGS systems at all, i.e., $\rho_{crit}(X) \leq 2.25$ for these values of α. When the load further increases to $\rho = 3$ the domain of α with stable TAGS systems further decreases. We also observe that for $\alpha = 0.2$, 34 hosts are required for construction of a stable system. The reason for this large value is that the critical load $\rho_{crit}(X)$ when $\alpha = 0.2$ is very close to 3 (approximately 3.05). Finally, when the load is 4.5, there are no stable systems when $\alpha \leq 0.5$ or $\alpha \geq 1.2$. The value $\alpha = 0.6$ requires 42 hosts since the corresponding critical load is about 4.6.

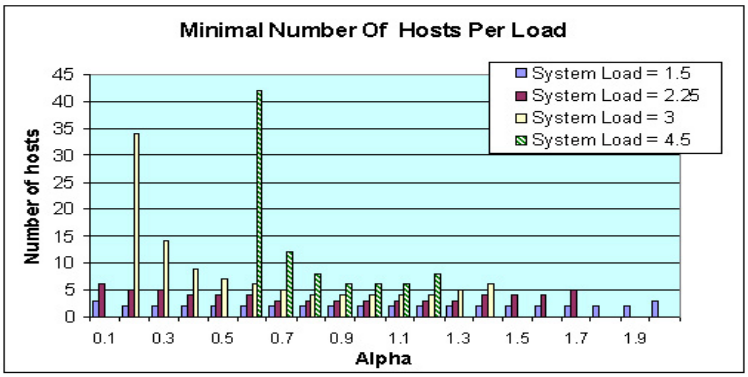

Figure 4: The effect of the number of hosts on the relative performance of TAGS.

7.5. Analysis of $\rho_{crit}(X)$ for Bounded Pareto

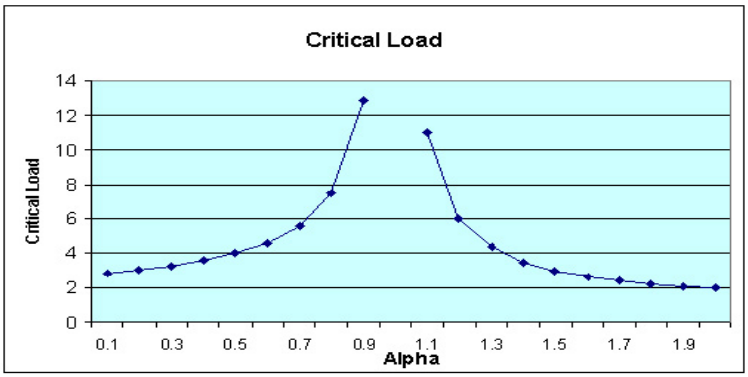

Figure 5: The critical load as r tends to infinity.

In Figure 5 we plot, the values of $\rho_{crit}(X)$ that were computed in the above corollary, for α in the domain $[0.1, 2]$, in jumps of 0.1. Asymptotically around $\alpha = 1$ it behaves like $1/|\alpha - 1|$, $\lim_{\alpha \to 0} \rho_{crit}(\alpha) = e$ and $\lim_{\alpha \to 2} \rho_{crit}(\alpha) = 2$. Outside the plotted range $\lim \alpha \to \infty \rho_{crit}(\alpha) = \lim_{\alpha \to -\infty} \rho_{crit}(\alpha) = 1$. It is interesting to note that while the figure is not symmetric around the value $\alpha = 1$, it is roughly symmetric.

8. Conclusions and future work

We have studied the stability of TAGS systems and have given an expression for the maximal load that a TAGS system can support with any number of hosts, given a job size distribution. We then computed the maximal load for several distributions such as Bounded Pareto distributions and Weibull distributions. We also provided a universal bound of $\ln(r) + 1$ on the maximal load in terms of the range r of a distribution. The bound explains why performance may deteriorate quickly as the system load increases. The analysis also shows that for a fixed range, the largest load can be supported by variants of the Bounded Pareto distribution with parameter $\alpha = 1$

We then show that at very low loads TAGS performs almost as well as SITA and that in a conservative, worst case distributional setting with large range, both provide similar performance guarantees on average waiting time which are far better than RA or even LWL.

We end by showing how TAGS and RA, which have the same minimal system requirements (no knowledge of job size or state of system) compare on Bounded Pareto distributions with waiting time as a target function. We see that the parameter $\alpha = 1$ is indeed a sweet spot for TAGS with performance that can be of order of magnitudes better than RA, but even there, as the number of hosts increases even in a modest way compared with the range, the advantage is lost. The comparison is done using the fixed range normalization which we believe is far superior to fixing the average job size and bring out the symmetry of RA and the approximate symmetry of TAGS with respect to the transformation exchanging α with $2 - \alpha$.

In future work we would like to explore more carefully the performance of both SITA and TAGS with respect to average slowdown. In terms of performance guarantees, RA and LWL still have a very weak guarantee of order of magnitude r, the same as for average normalized waiting time. For SITA and TAGS we had matching upper and lower bounds of order of magnitude $r^{1/h}$ at low loads. For average slowdown, our current upper and lower bounds do not match with the upper bound remaining of order of magnitude $r^{1/h}$ but the lower bound, [3], being the incredibly low order of magnitude $r^{\frac{1}{2^h-1}}$, the worst example, being again the Bounded Pareto with $\alpha = 1$. It would be interesting to explore if this is indeed the correct bound and compare TAGS and SITA with RA and LWL on Bounded Pareto distributions when they are properly normalized. The symmetry in this case relates α with $1 - \alpha$ and is only approximate and we should see its effect in the performance comparisons.

References

[1] Anselmi J. and J. Doncel, Asymptotically Optimal Size-Interval Task Assignments, *IEEE Transactions on Parallel and Distributed Systems*, 2019

[2] Sarfati H., Bachmat E. and S. Kedem-Yemini , Parameter setting and exploration of TAGS using a genetic algorithm. *Proceedings of the IEEE symposium on Computational Intelligence in Scheduling, CISched 2007*, 279-285.

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

[4] Bachmat E. and A. Natanzon, Analysis of SITA queues with many servers and spacetime geometry. *SIGMETRICS Performance Evaluation Review*, 40(3), 92-94, 2012.

[5] Bachmat E., J. Doncel, and H. Sarfati, Performance and Stability Analysis of the Task Assignment Based on Guessing Size Routing Policy, *27th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 1-13, 2019.

[6] Bestavros A., Load profiling: a methodology for scheduling real-time tasks in a distributed system, *Proceedings of 17th International Conference on Distributed Computing Systems*, 449-456, 1997.

[7] Broberg J., Z. Tari, P. Zeephongsekul, Task assignment with work-conserving migration, *Parallel Computing*, 32, 80830, 2006.

[8] Crovella M.E., M. Taqqu and A. Bestavros, Heavy-tailed probability distributions in the world wide web. In *A practical guide to heavy tails*, Chapman and Hall, New York, Chapter 1, 1-23, 1998.

[9] Doncel J., Aalto S. and U. Ayesta, Performance Degradation in Parallel-Server Systems, `IEEE/ACM Transactions on Networking`, 27(2), 875-888, 2019.

[10] Down D., S.P. Meyn, Stability of acyclic multiclass queueing networks. *IEEE Trans. Automat. Control*, 40, 91619, 1995.

[11] El-Taha M., B. Maddah, Allocation of service time in a multiserver system. *Management Science*, 52(4), 623, 2006.

[12] Feng H., V. Misra and D. Rubenstein, Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems, *Performance evaluation*, 62, 475-492, 2005.

[13] Foley R., D. McDonald et al, Join the shortest queue: stability and exact asymptotics, *The Annals of Applied Probability*, 11(3), 569-607, 2001.

[14] Harchol-Balter M., Task assignment with unknown duration, *Journal of the ACM*, 49(2), 260-288, 2002.

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

[16] Harchol-Balter M., M. Crovella. C. Murta, On choosing a task assignment policy for a distributed server system, *IEEE Journal of parallel and distributed computing*, Vol. 59, 204-228, 1999.

[17] Harchol-Balter M., A. Scheller-Wolf, A. Young, Surprising Results on Task Assignment in Server Farms with High-Variability Workloads, *Proceedings of ACM SIGMETRICS 2009 Conference on Measurement and Modeling of Computer Systems*, 2009.

[18] M. Harchol-Balter and R. Vesilo, To Balance or Unbalance Load in Size-Interval Task Allocation, *Probability in the Engineering and Informational Sciences* , vol. 24, 219-244, 2010.

[19] Richa A., M. Mitzenmacher and R. Sitaraman, The power of two random choices: A survey of techniques and results, *Combinatorial Optimization*, 9, 255-304, 2001.

[20] Riska A., W. Sun, E. Smirni, G. Ciardo, AdaptLoad: effective balancing in clustered web servers under transient load conditions, *Proceedings of the 22nd International Conference on Distributed Computing Systems, (ICDCS)*, 104-112, 2002.

[21] Scheller-Wolf A., Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues: Why s Slow Servers are Better than One Fast Server for Heavy-Tailed Systems *Operations Research*, 51, 748-758, 2003.

[22] Scheller-Wolf A., R. Vesilo, Structural Interpretation and Derivation of Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues *Queueing Systems* 54, 221-232, 2007.

[23] Tari Z., J. Broberg, A. Zomaya, R. Baldoni, A least flow-time first load sharing approach for distributed server farm, *Journal of Parallel and Distributed Computing*, 65, 83242, 2005.

[24] Thomas N.,Comparing job allocation schemes where service demand is unknown, *Journal of Computer and System Sciences*, 74, 1067081, 2008.

[25] W. Winston, Optimality of the shortest line discipline, `Journal of Applied Probability`, 14(1), 181-189, 1977.

[26] W. Whitt, Deciding Which Queue to Join: Some Counterexamples, `Operations Research`, 34(1).

Appendix A. Proof of Theorem 4.1

We aim to show that the TAGS system is stable if and only if $\rho < \rho_{crit}(X)$, where $\rho_{crit}(X)$ is as defined above. We first assume
$\rho > E(X)/M(X)$ and choose an s such that $\rho > E(X)/(s(1-X(s)))$. Let $s_{i-1} \leq s \leq s_i$ be the range of job sizes whose service completes at host i. All jobs of size at least s will pass through the i'th host and the host will spend at least s time units on each such job. The rate of jobs of size at least s is $\lambda(1-X(s))$, therefore the system must satisfy $\lambda s(1-X(s)) \leq 1$ in order to be stable. By definition $\rho = \lambda E(X)$ so we get $\rho s(1-X(s))/E(X) < 1$ a contradiction.

Conversely, assume that $\rho < E(X)/M(X)$ or equivalently $\rho M(X)/E(X) < 1$. All jobs of size at least s_{i-1} pass through host i. We know that s_i is an upper bound on the time spent by host i on any such job. Therefore, the utilization on the host i is bounded from above by

$$\lambda(1-X(s_{i-1}))s_i \leq \lambda M(X)s_i/s_{i-1}$$

$$= \rho M(X)/E(X)(s_i/s_{i-1}).$$

Fix any $q > 1$ such that $q\rho M(X)/E(X) < 1$ and let $s_i = q^i$, then by the above argument we have that host i will have a utilization below 1. We see that $h = [\log_q(r)] + 1$ hosts will suffice.

Appendix B. Proof of Theorem 4.2

Let $X(s)$ be the commulative distribution function of a distribution in the range $[1, r]$, i.e., $X(s) = Pr(X < s)$. We claim that for any $\varepsilon > 0$, $X(s)$ may be approximated by a distribution function $Y(s)$ with a continuous density function, of the same range $[r]$ such that $|\rho_{crit}(X) - \rho_{crit}(Y)| < \varepsilon$. To see this, decompose the range interval $[1, r]$ into n equal sub-intervals, with endpoints $1 = x_0, x_1, ..., x_n = r$. Consider the distribution Y_n which linearly extrapolates $X(s)$ between its endpoint values on each sub-interval $[x_i, x_{i+1}]$. Since X and Y_n are both monotone non decreasing functions they are Riemann integrable and it is obvious from the definition that $\lim_n\ E(Y_n) = E(X)$. We claim that $\lim_n\ M(Y_n) = M(X)$ as well. Indeed, For any $s \in [x_i, x_{i+1}]$ we have

$$s(1-X(s)) \leq s(1-X(x_i))$$

$$= (s-s_i)(1-X(x_i)) + x_i(1-X(x_i))$$

$$\leq \frac{r}{n} + x_i(1-Y_n(s_i)) \leq \frac{r}{n} + M(Y_n)$$

hence $M(X) \leq \frac{r}{n} + M(Y_n)$. Applying the same argument to Y_n instead of X and noting that $x_i(1-Y_n(x_i)) = x_i(1-X(x_i)) \leq M(X)$ we see that $M(Y_n) \leq \frac{r}{n} + M(X)$ and both inequalities together yield the claim for n large enough.

Given the above approximation it is enough to prove the bound for continuous distributions. Let X be continuous. We claim that we can find a distribution Y such that:
1) $\rho_{crit}(X) \leq \rho_{crit}(Y)$
2) Y is also supported on $[1, r]$.
3) $M(Y) = 1$.

We always have $M(X) \geq 1(1-X(1)) = 1$. If $M(X) = 1$ there is nothing to prove, hence we assume that $M(X) > 1$. Let $\tilde{s} > 1$ be such that $\tilde{s}(1-X(\tilde{s})) = M(X)$, $\tilde{s}$ exists by continuity of X. Consider

the distribution $\tilde{Y}$ which is supported on the interval $[M(X), r]$ which is defined as follows, for $s \geq \tilde{s}$, $\tilde{Y}(s) = X(s)$ and for $M(X) \leq s \leq \tilde{s}$, $1 - \tilde{Y}(s) = M(X)/s$

By construction $M(X) = M(\tilde{Y})$. Also by the definition of $M(X)$ and the construction of $\tilde{Y}$ we have for all s, $1 - \tilde{Y}(s) \geq 1 - X(s)$. Consequently

$$E(X) = \int_{s=0}^{r} 1 - X(s)\, ds \leq \int_{s=0}^{r} 1 - \tilde{Y}(s)\, ds = E(\tilde{Y})$$

We conclude that $\rho_{crit}(X) \leq \rho_{crit}(\tilde{Y})$. Finally we define Y to be a rescaling of $\tilde{Y}$ by the formula $Y(s) = \tilde{Y}(M(X)s)$. Rescaling does not change ρ_{crit}, the support of Y is in $[1, r/M(X)]$, which is contained in $[1, r]$, and $M(Y) = 1$.

Following the above argument it is sufficient to prove the bound for continuous distributions which satisfy $M(X) = 1$. In this case we have $\rho_{crit}(X) = E(X)$, hence our goal is to bound $E(X)$. Since $M(X) = 1$ we have for any s, $s(1 - X(s)) \leq 1$ or $1 - X(s) \leq 1/s$. Consequently,

$$E(X) = \int_{0}^{r} 1 - X(s)\, ds =$$

$$\int_{0}^{1} 1\, ds + \int_{1}^{r} 1 - X(s)\, ds \leq 1 + \int_{1}^{r} 1/s\, ds = 1 + \ln(r)$$

As desired. Showing that $B(1)_r^{dis}$ achieves the bound is an easy calculation. We have $E(X) = \ln(r) + 1$ and $sPr(X \geq s) = 1$ for all s in the range.

Appendix C. Proof of Theorem 6.2

We know from [3] that, when r is large and $\min_i(s_i/s_{i-1})$ is large, the normalized mean waiting time of a system that operates under the SITA policy is given by

$$E(\bar{W}^{SITA}(\mathbf{s})) \sim \sum_{i=1}^{h} f_i^{SITA} s_{i-1}^{-\alpha} s_i^{2-\alpha}, \tag{C.1}$$

where f_i^{SITA} is given in Lemma 6.1 of [3]. We now provide an analogous result for the normalized mean waiting time of a system that operates under the TAGS policy with the same assumptions.

Lemma C.1. When r is large and $\min_i(s_i/s_{i-1})$,

$$E(W(\mathbf{s})) \sim \sum_{i=1}^{h} f_i s_{i-1}^{-\alpha} s_i^{2-\alpha}, \tag{C.2}$$

where for $i < h$

$$f_i = \frac{2}{\alpha} f_i^{SITA} \tag{C.3}$$

and

$$f_h = f_h^{SITA}. \tag{C.4}$$

Proof. We first show different properties that the SITA system and the TAGS system verify when they have the same cutoffs and when r is large and $\min_i(s_i/s_{i-1})$ is large:

- We compute the portion of jobs executed in server i in a SITA system, i.e., the probability of a job ranging in size between s_{i-1} and s_i, that for the Bounded Pareto distribution results

$$p_i^{SITA} = \frac{1}{1 - (\frac{1}{p})^{\alpha}} (s_{i-1}^{\alpha} - s_i^{\alpha}). \tag{C.5}$$

We also compute the portion of jobs which pass through server i in a TAGS system and it results that

$$\bar{p}_i = \frac{1}{1-(\frac{1}{p})^\alpha}(s_{i-1}^\alpha - p^\alpha) \tag{C.6}$$

By (C.5-C.6) and since we assume that s_i/s_{i-1} is large, we get $p_i^{SITA} \sim \bar{p}_i$ and $p_h^{SITA} = \bar{p}_h$. From the above expressions, it follows that, when r tends to infinity,

$$1 - \frac{p_i^{SITA}}{\bar{p}_i} \to s_{i-1}^\alpha s_i^{-\alpha}.$$

- For the arrival rate, it follows from the above reasoning that λ_i^{SITA}, which is the arrival rate to server i of the SITA system, and λ_i, which is the arrival rate of server i of the TAGS system, satisfy the following property: $\lambda_i^{SITA} \sim \lambda_i$.

- The j-th moment of the distribution of the service time of jobs in server i of the TAGS system, that is $E(X_i^j)$, satisfies that

$$E(X_i^j) = \frac{p_i^{SITA}}{\bar{p}_i} E(X_{i,SITA}^j) + (1 - \frac{p_i^{SITA}}{\bar{p}_i}) s_i^j, \tag{C.7}$$

where $E(X_{i,SITA}^j)$ is the j-th moment of the service time of jobs in server i of the corresponding SITA system. The reason for this is that, in the TAGS system, the jobs which pass through server i consist of those which do not pass onto server $i+1$ (since the job size is less than s_i) and those who do (and in this case the service time in server i is s_i).

Besides, since the distribution of jobs sizes is Bounded Pareto, it follows that, if $\alpha \neq 1$ and $j \neq 1$,

$$E(X_{i,SITA}^j) = \frac{\alpha s_{i-1}^\alpha}{1-(\frac{s_{i-1}}{s_i})^\alpha} \frac{s_{i-1}^{j-\alpha} - s_i^{j-\alpha}}{\alpha - j} \tag{C.8}$$

and if $j = 1$ and $\alpha = 1$, it follows that

$$E(X_{i,SITA}^j) = \frac{s_{i-1}}{1-(\frac{s_{i-1}}{s_i})} \ln \frac{s_i}{s_{i-1}} \tag{C.9}$$

- We now show that the load of the servers for SITA and TAGS coincides in the asympototic regime. From (C.5),(C.6), (C.7) and (C.8) and the above formula, it follows that the mean service time of jobs in server i satisfies that $E(X_i^1) \sim E(X_{i,SITA}^1)$ for $\alpha < 1$ and $i = h$ or for $\alpha > 1$ and any i, whereas for $\alpha < 1$ and $i < h$, $E(X_i^1) \sim \frac{1}{\alpha} E(X_{i,SITA}^1)$. Besides, using that (C.5),(C.6), (C.7) and (C.9), it follows that $E(X_i^1) \sim E(X_{i,SITA}^1)$ for $\alpha = 1$. Therefore, since $\lambda_i^{SITA} \sim \lambda_i$, the load of server i of the SITA system, that is ρ_i^{SITA}, and the load of server i of the TAGS system, ρ_i, satisfy that $\rho_i \sim \rho_i^{SITA}$ in the following instances: (i) $\alpha < 1$ and $i = h$, (ii) $\alpha > 1$ and any i and (iii) $\alpha = 1$. On the other hand, for $i < h$ and $\alpha < 1$, we have that $\rho_i \sim \frac{1}{\alpha}\rho_i^{SITA}$. We know from (59) of [3] that ρ_i^{SITA} tends to zero when $\alpha > 1$ for all $i > 1$ and, by the duality result of Lemma 4.1 of [3], it follows that ρ_i^{SITA} tends to zero when $\alpha < 1$ for all $i < h$. Thus, since $\rho_i \sim \frac{1}{\alpha}\rho_i^{SITA}$, ρ_i also tends to zero when $\alpha < 1$ for all $i < h$. And this implies that ρ_i and ρ_i^{SITA} coincide when r tends to infinity.

- For the second moment of the service time of jobs in server i, we have that $E(X_i^2) \sim \frac{2}{\alpha} E(X_{i,SITA}^2)$ for $i < h$ and $E(X_i^2) \sim E(X_{i,SITA}^2)$.

- Since the arrivals in both systems are Poisson, we have compute the mean waiting time of jobs in server i using the Pollaczek-Kinchine formula. Let $E(W_i^{SITA})$ be the mean waiting time of jobs in the SITA system. Using the above formulas, it follows that, for $i < h$,

$$E(W_i) \sim \frac{2}{\alpha} E(W_i^{SITA}),$$

and for $i = h$,

$$E(W_i) \sim E(W_i^{SITA}).$$

We recall that, in a TAGS system, a job that finishes service at server i spends an additional time of

$$T_i(\mathbf{s}) = \sum_{j=1}^{i-1} s_j \leq (h-1)s_{i-1},$$

being serviced at servers $1, 2, \ldots, i-1$ and that the average excess service time satisfies

$$E(T(\mathbf{s})) = \sum_{i=1}^{h} p_i T_i(\mathbf{s}) \leq (h-1)E(X)$$

or equivalently

$$\mathbb{E}(T(\mathbf{s}))/E(X) \leq h-1 \tag{C.10}$$

We have

$$E(\bar{W}(\mathbf{s})) \geq E(\bar{W}^*) \geq \min_{\mathbf{s}} E(\bar{W}^{SITA}(\mathbf{s})).$$

We know from the results of [3] that the last term of the above inequality tends to infinity when r is large. This implies that $E(\bar{W}(\mathbf{s}))$ tends to infinity when r is large, i.e., $E(W(\mathbf{s}))/E(X)$ is unbounded when r is large and by (C.10), we know that $E(T(\mathbf{s}))/E(X)$ is bounded. As a result, it follows that $\mathbb{E}(T(\mathbf{s}))$ is asymptotically negligible. Therefore, the mean waiting time of jobs for the TAGS system satisfies that

$$E(W) \sim \sum_{i=1}^{h} \bar{p}_i E(W_i), \tag{C.11}$$

whereas for the SITA system

$$E(W^{SITA}) \sim \sum_{i=1}^{h} p_i^{SITA} E(W_i^{SITA}). \tag{C.12}$$

Finally, given the asymptotic relation given above between p_i^{SITA} and $\bar{p}_i$ and between $E(W_i)$ and $E(W_i^{SITA})$, using (C.1) as well as (C.11) and (C.12), the desired result follows. □

We now provide the proof of Theorem 6.2.

Proof. Using (C.1) and (C.2), the ratio between the mean waiting time of a system that operates under the TAGS policy and of a system that operates under the SITA policy is given by

$$\frac{\min_{\mathbf{s}} E(\bar{W}^{SITA}(\mathbf{s}))}{E(\bar{W}^*)}, \tag{C.13}$$

when r is large. From the results of [3] it is known that $\arg\min_{\mathbf{s}} E(\bar{W}^{SITA}(\mathbf{s}))$ satisfies the condition that $\min_i s_i/s_{i-1} \to \infty$ and from this and the above the same will hold for TAGS. A simple scaling argument shows that the ratio is independent of r and depends only on the ratios f_i^{TAGS}/f_i^{SITA}, which as we have shown in the previous lemma it is $2/\alpha$ for $i < h$ and one for $i = h$. Therefore, we apply the result of Lemma 5.3 of [3] with $c_i = f_i^{TAGS}/f_i^{SITA}$ and it results that (C.13) is equal to $\left(\frac{2}{\alpha}\right)^{\mu}$, where $\mu = \frac{(q^{h-1}-1)q}{q^h-1}$ and $q = \frac{\alpha}{2-\alpha}$.

We still need to show that $\left(\frac{2}{\alpha}\right)^{\mu} \leq 2$. If $\alpha \geq 1$, then $\mu \leq 1$ and, therefore, it is clear that $\left(\frac{2}{\alpha}\right)^{\mu} \leq 2$. If $\alpha < 1$, then we have that $q < 1$. This implies that

$$\mu = \frac{(q^{h-1}-1)q}{q^h-1} < q = \frac{\alpha}{2-\alpha}.$$

Differentiating $(\frac{2}{\alpha})^{\frac{\alpha}{2-\alpha}}$ it is easy to verify that it is an increasing function in the interval $(0, 1]$ with value 2 at $\alpha = 1$ and the desired result follows. □

Appendix D. Proof of Proposition 6.3

We first study the case $\alpha > 1$. For this instance, we denote by i the minimal number of servers needed for a system operating under the TAGS policy to be stable. Thus, there exist i cutoffs $\tilde{s}_0, \tilde{s}_1, \ldots, \tilde{s}_{i-1}, \tilde{s}_i$ such that the load of the first i-1 servers is equal to one. The load of jobs whose size is larger than $\tilde{s}_{i-1}$ is less than one. Let $s_{i-1} < \tilde{s}_{i-1}$ such the load of jobs whose size is larger than s_{i-1} is also less than one. We choose $s_1, s_2, \ldots, s_{i-2}$ such that the load of the first $i-1$ servers is the same. Since $s_{i-1} < \tilde{s}_{i-1}$, the first $i-1$ servers are stable. The remaining load is handled by $\tilde{h} = h - i + 1$ servers and, since it is less than 1 and up to scaling the job size distribution is Bounded Pareto, from the result of Theorem 6.3, we have that the normalized mean waiting time of jobs whose size is larger than s_{i-1}, in the asymptotic regime where $r \to \infty$, is of the same order of magnitude of that of a system operating under the SITA policy, which, according to (67) of [3], is given by (6).

We now focus on the order of magnitud of the first $i-1$ servers for $\alpha > 1$. We observe that $\tilde{s}_{i-1}$ is increasing with r and, since the value of $\tilde{s}_{i-1}$ is bounded by the value $\tilde{s}_{i-1}$ for the (unbounded) Pareto distribution, which is finite. Therefore, the order of magnitude of the first $i-1$ servers is negligible in the asymptotic regime.

For $\alpha < 1$, we use a similar strategy, but we start from the last server. Hence, we define $\tilde{s}_{h-1}$ to be such that the load on the last server is precisely 1. Inductively, given $\tilde{s}_{h-j}$ we define $\tilde{s}_{h-j-1}$ to be such that the load on the $h-j$ server is equal to 1. We can proceed this way to define $\tilde{s}_{\tilde{h}}, \tilde{s}_{\tilde{h}+1}, ..., \tilde{s}_{h-1}$ such that the load of jobs whose size is in the interval $[1, \tilde{s}_{\tilde{h}}]$ is less than one. And using the same arguments as in the case $\alpha > 1$, the desired result follows.