Analysis of SITA policies

Eitan Bachmat

Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, 84105. ebachmat@cs.bgu.ac.il *,tel. 972-8-6477858 , fax 972-8-6477650*

Hagit Sarfati

Department of Industrial Engineering, Ben-Gurion University, Beer-Sheva, Israel, 84105.

Analysis of SITA policies

Eitan Bachmat

Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, 84105. ebachmat@cs.bgu.ac.il *,tel. 972-8-6477858 , fax 972-8-6477650*

Hagit Sarfati

Department of Industrial Engineering, Ben-Gurion University, Beer-Sheva, Israel, 84105.

Abstract

We analyze the performance of Size Interval Task Assignment (SITA) policies, for multi-host assignment in a non-preemptive environment. Assuming Poisson arrivals, we provide general bounds on the average waiting time independent of the job size distribution. We establish a general duality theory for the performance of SITA policies. We provide a detailed analysis of the performance of SITA systems when the job size distribution is Bounded Pareto and the range of job sizes tends to infinity. In particular, we determine asymptotically optimal cutoff values and provide asymptotic formulas for average waiting time and slowdown. We compare the results with the Least Work Remaining policy and compute which policy is asymptotically better for any given set of parameters. In the case of inhomogeneous hosts we determine their optimal ordering.

Key words: Multiple host scheduling policies, Size interval splitting policies, Heavy-tailed distributions, queueing theory.

1. Introduction

Many installations such as web server farms and computing centers have a multitude of hosts which can serve any incoming request. In some of these installations, especially supercomputing centers, jobs are non-preemptive, meaning that, a job that has been stopped before completion, must start from scratch, [9, 29]. There has been a growing body of research regarding scheduling policies for such multi-host systems. When the job size distribution is exponential then the Least-work-remaining, LWR, policy, which assigns any job to the next available host, is optimal, [17]. However, empirical data has suggested that job size distributions which are typical of modern system workloads are *heavy-tailed*, rather than exponential, [1, 3, 6, 7, 22, 30]. Such workloads have substantial amounts of both small and large jobs and this leads to long average waiting times, when small jobs have to wait for long jobs to finish. In the case of known job sizes these considerations have led researchers, [11, 12, 29, 18], to suggest policies where each host is responsible for a certain range of job sizes. These policies employ a set of cutoff parameters, $s_1 < s_2 < \ldots < s_{h-1}$, where h is the number of hosts in the system. The first host only handles jobs which take less than s_1 time units, the second host handles jobs which take between s_1 and s_2 time units to complete and so on. Such policies are known as size interval task assignment policies (SITA policies). These policies were generalized to the case of unknown job sizes in [9].

A major obstacle in analyzing SITA policies and comparing them to other policies is that there was no known way of calculating optimal cutoffs s_i. The cutoffs in the comparisons were based on various heuristic choices, [11, 12, 10, 33, 34, 32, 5, 21]. However, there were no estimates for how far from optimal these heuristics may be.

Another issue involves the choice of a target function. The optimal cutoffs for minimizing average waiting time differ from those which minimize average slowdown. We may ask how the optimization of one target function affects the other, is one optimization choice more robust than the other?

In this paper we develop some basic insights on the performance SITA policies.

For low loads, we provide general bounds on the average waiting time, which hold for any job size distribution. Moreover, these bounds use a universal set of cutoffs which are essentially independent of the job size distribution.

We establish a general duality theory for SITA policies which vastly generalizes the symmetry which was introduced in [8]. The duality states that some performance characteristics of SITA on seemingly very different job size distributions are the same.

We also provide a very detailed analysis of the case of Bounded Pareto job size distributions, where the size of the largest job goes to infinity. A Bounded Pareto ditribution is a distribution whose density has the form $cs^{-\alpha-1}$ on some interval $I = [a, b]$, $0 < a < b$. α is the parameter of the distribution. These distributions play an important role in characterizing computer workloads and also served as test cases for the performance of SITA systems in several studies, [8, 9, 11, 14]. we compute near optimal cutoffs for minimizing average waiting time and slowdown. using these computations we compare the performance of SITA with that of LWR on Bounded Pareto distributions.

To state our results more precisely, we use the notation $f = \Theta(g)$ to indicate that the functions f and g have the same order of magnitude, namely, there is a constant $c > 0$ such that $\frac{f}{c} < g < cf$. Similarly $f = O(g)$ means that there is a constant $c > 0$ such that $f < cg$. We always assume i.i.d exponentially distributed inter-arrivals and i.i.d job sizes. Our main results are

- (Section 3) For low load, we establish performance guarantees for a SITA system with any bounded job size distribution. If p is the ratio between the largest job and the smallest job, then there are cutoffs such that the normalized average waiting time $E(N) = E(W)/E(X)$ is $O(p^{1/h})$, where h is the number of hosts.

- (Section 4) There is a duality, defined via an involution on density functions, which preserves the behavior of $E(N)$ while reversing the mapping (ordering) of hosts. This involution maps the Pareto distribution with parameter α to the Pareto distribution with parameter $2 - \alpha$. There is a second involution which preserves the order of magnitude of mean slowdown and which maps the Pareto distrubtion with parameter α to that with parameter $1 - \alpha$.

All other results assume a Bounded Pareto job size distribution.

- (Sections 6.1, 6.2, 6.4) We provide explicit formulas for asymptotically optimal cutoffs for minimizing either mean waiting time or mean slowdown and compute explicitly the resulting mean waiting time or mean slowdown. The optimal cutoffs for mean waiting time and for mean slowdown differ substantially.

- (Section 6.4) Like LWR, SITA is most effective when the total system load is less than the load which can be handled by the most powerful host in the system. In that case mean waiting time decreases very rapidly with the introduction of each additional host. If the load is higher, then, some of the h hosts are dedicated to eliminating the excessive load, leaving the system with a smaller effective number $\tilde{h}$ of hosts. Assume $h - \rho > 1$, then SITA is asymptotically better than LWR if and only if $\alpha < \frac{[h-\rho]+1}{[h-\rho]}$. In particular, if $\alpha \leq 1$, SITA is always asymptotically better, while if $\alpha > 1$, then given enough hosts, LWR will be better.

- (Sections 6.1, 6.2, 6.4, 6.7) When the job-size distribution is Bounded Pareto with parameter $\alpha \neq 1$, the mean normalized waiting time $E(N) = E(W)/E(X)$, of SITA with the optimal cutoffs has the approximate form $E(N) = \Theta(p^{b(\alpha,\tilde{h})})$, where $b(\alpha, \tilde{h})$ is exponentially small in the number of hosts. This shows that SITA policies can handle a huge job range using very few hosts. When $\alpha = 1$ the optimal waiting time has the form $\Theta(p^{1/\tilde{h}}/log^2(p))$. The optimal waiting time is always $O(p^{1/\tilde{h}})$, for any job-size distribution for which the ratio of largest to smallest job size is at most p, so $\alpha = 1$ represents a "worst case" distribution for SITA policies. $\alpha = 1$ is also a "worst case" distribution for LWR.

- (Section 6.3) When $\alpha > 1$ optimization of cutoffs with respect to average waiting time is more robust. When $\alpha < 1/2$ optimization with respect to slowdown is more robust. When $1/2 < \alpha < 1$ optimizing one target function substantially hurts the other.
- (Section 6.8) When $\alpha > 1$ the hosts of the SITA system should be ordered in decreasing power order to minimize mean waiting time. The opposite is true when $\alpha < 1$. When minimizing mean slowdown, decreasing power holds for $\alpha > 1$, increasing power when $\alpha < 1/2$. The case of $1/2 < \alpha < 1$ is more complicated. Like the previous item, duality theory is the right context for understanding this result.
- (Section 6.9) All previous heuristics for choosing the cutoffs are suboptimal, sometimes, badly suboptimal. In particular, they can lead to performance which is unboundedly worse than optimal.

1.1. Related work

Size interval task assignment policies have been shown in some cases to outperform traditional policies such as LWR. A comparison based on a trace of jobs from a supercomputer can be found in [29]. In that simulation, the SITA policy clearly out-performs the LWR policy. Other comparisons, [12, 10, 9, 14] used the Bounded Pareto family, which is typical of heavy-tailed distributions.

The study of SITA in the inhomogeneous case, where not all hosts have the same power was considered by Feng et. al, [8]. Among other issues, the authors considered the optimal assignment of hosts to the ranges of job sizes according to their power. based on experiments with several Pareto distributions, They also showed that distributions which have a certain symmetry are oblivious to the ordering of hosts. This result was the starting point for our analysis of duality, however, the basic idea goes all the way back to Riemann, [19]. The problem of host mapping (ordering) was also considered in [12] and [10], where the target function was the optimization of mean slowdown.

The choice of cutoffs was studied in several papers, leading to several heuristics. Harchol-Balter et. al, [11] and Riska et. al, [20, 5], proposed a heuristic choice of cutoffs which load balanced the SITA system. A SITA system with this particular choice of ranges was called, in [11], a SITA-E system. Three additional heuristics for choosing the cutoffs, which created non balanced systems, were proposed in [12] and [10]. These were based on the ability to solve the two host case by brute force. A SITA system with either of these heuristic choices was called a SITA-V system. There were no estimates for how far from optimal these heuristics may be.

In a series of paper [20, 33, 34, 32, 5, 21], E. Smirni and her collaborators have studied SITA like policies in different settings, including a dynamic setting in which cutoffs can be changed to adjust to changes in the workload. The basic guiding principle to determining cutoffs is taken to be load balancing. However, there are certain adjustments. In [21], the option of balancing queue length was also explored, while in [34] unbalancing is suggested in the context of auto-correlated arrivals.

Recently, a paper by Harchol-Balter, Scheller-Wolf and Young, [14], compares the asymptotic behavior of SITA policies with that of LWR for certain families of bimodal, trimodal, hyper-exponential and Bounded Pareto distributions. The paper only considers whether the performance is bounded or not. It is shown that all combinations of SITA and LWR being bounded or unbounded are possible.

2. Preliminaries

In this paper we consider multi-host assignment policies for non-preemptive systems with known job sizes. A system is non-preemptive if any job which is stopped needs to be executed from the beginning. This is a reasonable assumption for many current systems, in particular in supercomputing centers, [9]. We also assume that the sizes of jobs are i.i.d random variables. We will assume that the arrival process is always Poisson. We will denote the waiting time random variable by W. The waiting time W does not include the service time of a job.

2.1. Assignment policies

We introduce SITA policies, which are the subject of investigation in this paper. These policies have been introduced in [11]. The main feature of these policies is variance reduction at the different hosts. For more on variance reduction policies see [8, 29].

A SITA policy can be described as follows:
Consider a system with h hosts, numbered $1, \ldots, h$. Let $s_1 < s_2 < \ldots < s_{h-1}$ be a set of cutoff parameters which are fixed. We assume that job sizes are known. An incoming job of size s is dispatched to host i, such that $s_{i-1} \leq s < s_i$. Requests are serviced in each host in a first come, first served (FCFS) order. The Least-Work-Remaining assignment policy holds the jobs in a central FIFO queue and assigns each job to the next available host.

2.1.1. Families of bounded distributions

Let X be a generic job-size random variable, which is bounded below by a positive constant. We consider the associated distribution function $X(s) = Pr(X < s)$. By changing the unit of time we can and will assume that $X(1) = 0$, which means that the smallest job has size at least 1. For any $p \geq 1$, we consider a truncated job-size random variable X_p, which is obtained by sampling X and rejecting the result if it is greater than p, repeating until a result is accepted. The associated distribution $X_p(s)$ is given by $X_p(s) = X(s)/X(p)$, for $s \leq p$ and $X_p(s) = 1$ otherwise. We say that $X_p(s)$ is the family of bounded distributions associated with $X(s)$.

2.1.2. Pareto and Bounded Pareto distributions

The job size distributions which were considered in [11] and many other studies were the Pareto distributions $B(\alpha)$, where $\alpha > 0$. The density f of $B(\alpha)$ has the form

$$f(s) = \alpha s^{-\alpha-1} \tag{1}$$

For $s \geq 1$ and zero otherwise. The bounded version $B(\alpha)_p$ has a density of the form

$$f(s) = c s^{-\alpha-1} \tag{2}$$

in the bounded range $1 \leq s \leq p$, and zero otherwise. The constant, $c > 0$, is a normalizing constant which ensures that $\int_1^p f(s)ds = 1$. A simple computation shows that the normalizing constant is

$$c = \frac{\alpha}{1 - p^{-\alpha}} \tag{3}$$

We observe that Pareto and Bounded Pareto distributions are self similar in the sense that $f(s)$ is proportional to $f(cs)$ for any $c > 0$, in the domain where both are non zero.

2.1.3. p-Asymptotics for SITA *systems*

A p-asymptotic family of SITA systems with bounded job size distributions consists of the following data:

- h - A fixed number of hosts.
- ρ - A fixed system load.
- $X(s)$ A distribution with $X(1) = 0$.
- A set of cutoff values $1 < s_1(p) < \ldots, s_{h-1}(p) < p$, parametrized by $p > 1$.

Given some p we consider a SITA system with cutoffs, $1 < s_1(p) < \ldots, s_{h-1}(p) < p$, with i.i.d job size random variables with a generic job size distribution $X_p(s)$ and a Poisson job arrival process with rate $\lambda = \rho/E(X_p)$. We let W_p be the waiting time random variable for this system and we define S_p as W_p/s, where s is the job size. Our goal is to analyze the asymptotic behavior of $E(W_p)/E(X_p)$ and $E(S_p)$ as p tends to infinity. We will be interested particularly in the case where $X(s) = B(\alpha)$. We consider the normalized quantity $N_p = W_p/E(X_p)$ and its average, $E(N) = E(W_p)/E(X_p)$, rather than $E(W_p)$, because the former is invariant under changes of the time unit.

For a single host with Poisson arrivals with rate λ the mean waiting time is given by the Pollaczek-Khinchine (P-K) formula which states that

$$E(N) = \frac{\rho}{2(1-\rho)} \frac{E(X^2)}{E(X)^2} \tag{4}$$

where $\rho = \lambda E(X)$ is the utilization.

From this we can easily see that p-asymptotic behavior of mean normalized waiting time is most interesting when $E(X^2)/E(X)^2$ is infinite, and we will take that as our working definition of a heavy-tailed distribution. In particular, when working with the Pareto distributions $B(\alpha)$ we will be interested in the range $0 < \alpha < 2$.

3. Universal cutoffs and performance guarantees for general workloads

Assume i.i.d Poisson arrivals and i.i.d job sizes with generic distribution X supported on the interval $[1, p]$ (the ratio of largest to smallet job is at most p). we provide a universal (distribution independent) bound on the average normalized waiting time in a SITA system. The bound is proved using the same set of cutoffs for any distribution, hence we get a performance guarantee when using these cutoffs

Theorem 3.1. *Let X be any job size distribution which is supported on the interval $[1, p]$ and let $\rho < 1$ be the load of a SITA system with h identical hosts. We choose $s_i(p) = p^{i/h}$, then*

$$E(N) = E(W)/E(X) \leq \frac{\rho}{2(1-\rho)} p^{1/h} \tag{5}$$

Proof: Consider host i. By the P-K formula we have $E(W_i)/E(X_i) = \rho_i E(X_i^2)/E(X_i)^2/2(1-\rho_i)$. By the choice of s_i, X_i is supported on the interval $[p^{(i-1)/h}, p^{i/h}]$. Since $E(X_i^2)/E(X_i)^2$ is time scale invariant we can rescale the time unit, dividing it by $p^{(i-1)/h}$ to obtain an equivalent distribution in the range $[1, p^{1/h}]$. Let $r = p^{1/h}$. We have a general upper bound $E(X_r^2)/E(X_r)^2 \leq r/E(X_r) \leq r$, for any random variable X_r, supported on $[1, r]$, which follows from the trivial inequality $E(X_r^2) = \int_1^r x^2 d\mu \leq \int_1^r rx d\mu = rE(X_r)$. This bound leads to the inequality

$$E(W_i)/E(X_i) \leq \frac{\rho}{2E(X_i)(1-\rho)} p^{1/h} \leq \frac{\rho}{2(1-\rho)} p^{1/h} \tag{6}$$

since $\rho_i \leq \rho$. Next, we note that $E(W)/E(X) = \sum_i \frac{p_i E(X_i)}{E(X)} (E(W_i)/E(X_i))$, where p_i is the probability that a job will be serviced by host i. This shows that $E(W)/E(X)$ is a weighted average of $E(W_i)/E(X_i)$, hence it is enough to prove the bound for each i separately. *q.e.d*

We can actually improve the bound of the theorem above by a factor of 2 with a slightly more complicated argument. the maximal mean normalized waiting for such job sizes. We claim that in fact

$$E(W_i)/E(X_i) \leq \frac{\rho}{4(1-\rho)} p^{1/h} \tag{7}$$

We sketch the argument. As in the theorem we assume that X_i is supported on the interval $[1, r]$ where $r = p^{1/h}$. Using compactness in the weak topology (or by more elementary means), it can be shown that among all distributions Y which are supported on $[1, r]$ there will be a distribution which maximizes the value of $E(Y^2)/E(Y)^2$. We observe that if a distribution Z has positive measure on an interval $[a, b]$ with $1 < a < b < r$, then it does not maximize. To obtain a distribution with a higher value we shift half the measure restricted to $[a, b]$, a distance ε to the right and the other half of the measure, a distance ε to the left. The average remains the same but $E(Y^2)$ has increased since $(x+\varepsilon)^2 + (x-\varepsilon)^2 = 2x^2 + 2\varepsilon^2 > 2x^2$. We conclude that the maximizing distribution has the form $c\delta_1 + (1-c)\delta_p$, where δ_x is the Dirac measure on x. For such a distribution $E(Z) = c + (1-c)p$, $E(Z)^2 = c^2 + c(1-c) + (1-c)^2p^2$ and $E(Z^2) = c + (1-c)p^2$. One can easily show that the optimal value of c has the form $1 - u/p$ for some bounded u. Optimizing over the possible values of u, we get $u = 1$, which leads to the result.

4. Duality for SITA systems

In this section we establish a general theory of duality (symmetry) for systems with a SITA scheduling policy. The duality depends on the target function we wish to consider. It is strongest in the case of average normalized waiting time. The duality theory generalizes a symmetry (self-duality) which was first observed in [8], proposition 6. Let $X = X_p$, be a distribution which is supported on the interval $[1, p]$, and has a density function $f = f_X$.

We let $\hat{f}(s)$ be the density defined by the following functional equation

$$\hat{f}(s) = \frac{p^3}{E(X^2)} s^{-4} f(p/s) \tag{8}$$

We will show shortly that $\hat{f}$ is indeed a density function, but we would like to present this fact as part of a more general lemma. Let $I = [a, b]$, $a > 0$, be an interval. We can define the non normalized k'th moment of f on I as $m_f^k(I) = \int_a^b s^k f(s)ds$. If $I = [1, p]$, then obviously $m_f^k(I) = E(X^k)$ and we denote it simply as m_f^k. We define the dual of the interval I to be $\hat{I} = [p/b, p/a]$. Note that $[1, p]$ is self-dual. The following lemma can essentially be traced back to B. Riemann, [19], which used a variant to prove the functional equation for his Zeta function.

Lemma 4.1. *We have the following duality*

$$m_k^{\hat{f}}(I) = \frac{p^k}{E(X^2)} m_{2-k}^f(\hat{I}) \tag{9}$$

In particular, $m_0^{\hat{f}} = \int_1^p \hat{f}(s)ds = \frac{1}{E(X^2)} m_2^f = 1$ *and so* $\hat{f}$ *is a density function.*

Proof: Letting $t = p/s$, we have

$$m_k^{\hat{f}}(\hat{I}) = \int_a^b s^k \hat{f}(s)ds = \frac{p^3}{E(X^2)} \int_a^b s^{k-4} f(p/s)ds$$

$$= -\frac{p^3}{E(X^2)} p^{k-4} p \int_{p/a}^{p/b} t^{4-k} f(t) t^{-2} dt = \frac{p^k}{E(X^2)} \int_{p/b}^{p/a} t^{2-k} f(t)dt$$

as required. *q.e.d*

Consider a SITA system with h hosts, load ρ, a job-size distribution X given by a bounded density function f, supported on the interval $[1, p]$ and cutoff parameters $1 = s_0 < s_1 < \ldots < s_{h-1} < s_h = p$. We denote the mean normalized waiting time $E(W)/E(X)$, for such a system by $E(N)(X, h, \rho, p, s_1, \ldots, s_{h-1})$. Consider $\hat{X}$, the distribution associated with the density $\hat{f}$. We define the dual cutoffs by the formula

$$\hat{s}_i = p/s_{h-i} \tag{10}$$

As a consequence of the last lemma we have the following duality statement

Theorem 4.2. *The following identity holds*

$$E(N)(X, h, \rho, p, s_1, \ldots, s_{h-1}) = E(N)(\hat{X}, h, \rho, p, \hat{s}_1, \ldots, \hat{s}_{h-1}) \tag{11}$$

Proof: let $I_i = [s_{i-1}, s_i]$ and $J_i = [\hat{s}_{i-1}, \hat{s}_i]$. Then, $\hat{I}_i = J_{h-i+1}$. The Pollaczek-khinchine formula states that the mean waiting time at host i is given by $\frac{\lambda_i}{2(1-\rho_i)}E(X_i^2)$, where λ_i is the arrival rate at the host, ρ_i is the load at the host, and $E(X_i^2)$ is the second moment for job sizes at the host. Since host i in the first `SITA` system is responsible for job sizes in I_i, the portion of jobs which arrive at host i is given by $m_0^f(I_i)$. We also have the general formula $E(X_i^j) = \frac{m_j^f(I_i)}{m_0^f(I_i)}$. We consequently have

$$\lambda_i = \lambda m_0^f(I_i) = \rho \frac{m_0^f(I_i)}{m_1^f}. \tag{12}$$

Similarly $\rho_i = \lambda_i E(X_i) = \rho \frac{m_0^f(I_i)}{m_1^f} \frac{m_1^f(I_i)}{m_0^f(I_i)} = \rho \frac{m_1^f(I_i)}{m_1^f}$. Combining we see that the average waiting time at host i is given by

$$\begin{aligned} E(W_i) &= \rho \frac{m_0^f(I_i)}{m_1^f} \frac{m_2^f(I_i)}{m_0^f(I_i)} (2(1 - \rho \frac{m_1^f(I_i)}{m_1^f}))^{-1} \\ &= \frac{\rho}{2}(1 - \rho \frac{m_1^f(I_i)}{m_1^f}))^{-1} \frac{m_2^f(I_i)}{m_1^f} \end{aligned}$$

To get the weighted contribution of the host to the normalized waiting time we need to multiply by $m_0^f(I_i)$, the portion of jobs arriving at host i, and to divide by $m_1^f = E(X)$. Calling the weighted contribution $E_i(N)$ we have

$$E_i(N) = \frac{\rho}{2}(1 - \rho \frac{m_1^f(I_i)}{m_1^f})^{-1} \frac{m_0^f(I_i) m_2^f(I_i)}{(m_1^f)^2}. \tag{13}$$

However, lemma 4.1 tells us that

$$\frac{m_1^f(I_i)}{m_1^f} = \frac{m_1^{\hat{f}}(J_{h-i+1})}{m_1^{\hat{f}}}$$

and that

$$\frac{m_0^f(I_i) m_2^f(I_i)}{(m_1^f)^2} = \frac{m_0^{\hat{f}}(J_{h-i+1}) m_2^{\hat{f}}(J_{h-i+1})}{(m_1^{\hat{f}})^2}$$

Which implies that $E_i(N)$ is also the contribution of the $h - i + 1$ host to the mean normalized waiting time of the dual system. *q.e.d*

As a corollary we have

Corollary 4.3. *The following equality holds*

$$E(N)(B(\alpha)_p, h, \rho, p, s_1, \ldots, s_{h-1}) = E(N)(B(2-\alpha)_p, h, \rho, p, \hat{s}_1, \ldots, \hat{s}_{h-1}) \tag{14}$$

Proof: We note that equation (13) shows that $E(N)$ can be computed using the function cf, rather than the density f, where c is any constant. This also implies that for duality it is sufficient to verify that $\hat{f}$ is proportional to $s^{-4}f(p/s)$. The function $f(s)$ is proportional to $s^{-\alpha-1}$, which implies that $s^{-4}f(p/s)$ is proportional to $s^{-4}s^{\alpha+1} = s^{\alpha-3} = s^{-(2-\alpha)-1}$ as required. *q.e.d*

We notice that in the proof we used the self similarity property of Pareto distributions which allows us to make a uniform argument for all values of p. We can repeat the argument to establish less uniform dualities between the family of Chi distributions and the family of inverse-Chi distributions, between Gamma distributions and inverse-Gamma distributions and duality within the family of Log-Normal distributions, for the latter see [8].

The involution given in equation (8) is part of a larger family of dualities. consider the involution inv_β which maps a density $f(s)$ to the density $\frac{p^{\beta-1}}{E(X^{\beta-2})}s^{-\beta}f(p/s)$. The argument of lemma 4.1 shows that $m^f_k(I)$ will satisfy

$$m^f_k(I) = \frac{p^k}{E(X^{\beta-2})}m^f_{\beta-2-k}(\hat{I}). \tag{15}$$

In addition, the computation of corollary 4.3 shows that

$$inv_\beta(B(\alpha)_p) = B(\beta-2-\alpha)_p. \tag{16}$$

Consider the expression $E_i(S)$ for the contribution to average slowdown $E(S)$ of the i'th host. A direct computation as in equation (13) shows that

$$E_i(S) = \rho(2(1-\rho\frac{m^f_1(I_i)}{m^f_1}))^{-1}\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_1 m^f_0}. \tag{17}$$

We obtain the following corollary

Corollary 4.4. *Let* $1 = s_0(p) < s_1(p) < \ldots < s_{h-1}(p) < s_h(p) = p$, *be cutoffs, then for* $\rho < 1$

$$1-\rho < \frac{E(S)(B(\alpha)_p, h, \rho, p, s_1, \ldots, s_{h-1})}{E(S)(B(1-\alpha)_p, h, \rho, p, \hat{s}_1, \ldots, \hat{s}_{h-1})} < \frac{1}{1-\rho} \tag{18}$$

Proof: Letting $\beta = 3$, we see from (15) that $\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_1 m^f_0}$ is invariant with respect to the involution inv_3. The other component of $E_i(S)$ is $\frac{1}{2(1-\rho_i)}$, is not invariant under inv_3, but rather, is invariant under inv_4. However, if $\rho < 1$, then $\frac{1}{2} < \frac{1}{2(1-\rho_i)} < \frac{1}{2(1-\rho)}$ which leads to the statement. *q.e.d*

5. Minimizing a certain class of functions

Fix $p > 1$. We consider the open simplex $\Delta_{h,p}$, consisting of possible cutoff values $1 < s_1 < s_2 < \ldots < s_{h-1} < p$. We also fix values $s_0 = 1$ and $s_h = p$. We consider functions of the general form

$$G = G(s_1, \ldots, s_{h-1}) = \sum_{i=1}^{h} f_i s_{i-1}^{-d} s_i^e \tag{19}$$

here $f_i, d, e > 0$ are fixed parameters. Functions of this form will be shown to approximate the mean normalized waiting time and the mean slowdown of SITA systems when the job sizes have a distribution $B(\alpha)_p$. Naturally, we will be interested in the cutoff values that minimize the value of G over the simplex.

We recall that a point $(s_1, \ldots, s_{h-1}) \in \Delta_{h,p}$, is called a critical point for the function G if all the partial derivatives $\frac{\partial G}{\partial s_i}$ vanish at the point. We also recall that if a function has a minimum in the open simplex it must be at a critical point. In the following section, the function $\log(x)$ will refer to $\log_{10}(x)$.

Lemma 5.1. *Let*

$$q = \frac{d}{e} \tag{20}$$

$$C_i = \log(\frac{s_{i+1}}{s_i}) \tag{21}$$

$$D_i = \frac{1}{e}(\log(\frac{e}{d}\frac{f_i}{f_{i+1}})) \tag{22}$$

and

$$\delta = -\sum_{i=1}^{h-1} \frac{q^{h-i}-1}{q^h-1} D_i \tag{23}$$

then, the unique critical point of G *is given by*

$$s_1 = p^{\frac{q-1}{q^h-1}} 10^{\delta} \tag{24}$$

and by the inductive relation

$$(\frac{s_{i+1}}{s_i}) = (\frac{e}{d}\frac{f_i}{f_{i+1}})^{\frac{1}{e}} (\frac{s_i}{s_{i-1}})^{\frac{d}{e}} \tag{25}$$

The value of the function at the critical point is given by

$$G_{crit} = \frac{1-q^{-h}}{1-q^{-1}} f_1 p^{\frac{d-e}{q^h-1}} 10^{e\delta} \tag{26}$$

This is the global minimum of G *over the simplex.*

Proof: To find the critical point we differentiate

$$\frac{\partial}{\partial s_i} G = \frac{\partial}{\partial s_i} (f_i s_{i-1}^{-d} s_i^e + f_{i+1} s_i^{-d} s_{i+1}^e) \tag{27}$$

$$= e f_i s_{i-1}^{-d} s_i^{e-1} - d f_{i+1} s_i^{-d-1} s_{i+1}^e \tag{28}$$

Setting to zero, we obtain, after a simple manipulation, the condition

$$(\frac{s_{i+1}}{s_i}) = (\frac{e}{d}\frac{f_i}{f_{i+1}})^{\frac{1}{e}} (\frac{s_i}{s_{i-1}})^{\frac{d}{e}} \tag{29}$$

Taking the logarithm of both sides, and using the definitions in the statement of the lemma, we obtain the relation

$$C_i = q C_{i-1} D_i \tag{30}$$

By induction we have

$$C_i = q^i C_0 \sum_{j=1}^{i} q^{i-j} D_j \tag{31}$$

We also have the telescopic sum

$$\sum_{i=0}^{h-1} C_i = \log(s_h/s_0) = \log(p) \tag{32}$$

From the last two equations we deduce

$$\log(p) = \sum_{i=0}^{h-1} C_i = \frac{q^h-1}{q-1} C_0 \sum_{i=1}^{h-1} \frac{q^{h-i}-1}{q-1} D_i \tag{33}$$

Defining the constant

$$\gamma = \sum_{i=1}^{h-1} \frac{q^{h-i}-1}{q-1} D_i \tag{34}$$

and exponentiating, we have

$$p = s_1^{\frac{q^h-1}{q-1}} 10^{\gamma} \tag{35}$$

or

$$s_1 = p^{\frac{q-1}{q^h-1}} 10^{\delta} \tag{36}$$

it follows from (28) that

$$\frac{e}{d} f_i s_{i-1}^{-d} s_i^{e-1} = f_{i+1} s_i^{-d-1} s_{i+1}^{e} \tag{37}$$

which after multiplication of both sides by s_i yields

$$\frac{e}{d} f_i s_{i-1}^{-d} s_i^{e} = f_{i+1} s_i^{-d} s_{i+1}^{e} \tag{38}$$

Since $\frac{e}{d} = q^{-1}$ we have by (19) that

$$G_{crit} = \frac{1-q^{-h}}{1-q^{-1}} f_1 s_1^e = \frac{1-q^{-h}}{1-q^{-1}} f_1 p^{\frac{d-e}{q^h-1}} 10^{e\delta} \tag{39}$$

as required.

Next we show that the critical point is a global minimum over the closed simplex $\bar{\Delta}_{h,p}$, which is the closure of the open simplex. Consider first the case of $h = 2$. The function G will have the form $G(s_1) = f_1 s_1^e + f_2 s_1^{-d} p^e$ and the closed simplex is given by the interval $s_1 \in [1, p]$. We need to show that the value at the critical point is less than the values of G at the endpoints 1 and p. We know that the value at the critical point is $\Theta(p^{\frac{d-e}{q^2-1}}) = \Theta((p^e)^{\frac{1}{q+1}})$. The value at the endpoint $s_1 = 1$ is given by $G(1) = f_1 + f_2 p^e = \Theta(p^e)$ and similarly the value at $s_1 = p$ is $G(p) = f_1 p^e + f_2 p^{e-d} = \Theta(p^e)$. Since $q > 0$ we see that for p large enough the value of G at the unique critical point is smaller than the values at the boundary. We conclude that the value of G at the critical point is the global minimum of G in the closed simplex (interval) $[1, p]$ when p is large enough. We claim that this property holds for all $p > 1$. Assume to the contrary that for some $p > 1$ the critical value is not minimal. The function G along with the critical point and the boundary points 1 and p all vary continuously as p changes. This means that we would have for some $p_0 > 1$, equality between the values of G at the critical point and a boundary point, either 1 or p_0. But that would entail the existence of a second critical point, which contradicts uniqueness. When $h > 2$ we notice that fixing the values of all variables s_i except one, say s_j leads to a function of the general form given in (19) with $h = 2$. A boundary point of the simplex corresponds to a set of cutoffs with $s_j = s_{j+1}$ for some $0 \le j \le h-1$. Such a point cannot produce a minimal value for G over the closed simplex since the value of s_j given all other s_i is at a boundary value for a function with $h = 2$ and hence not minimal by the preceding argument. We conclude that the minimal value is attained at an interior point of the simplex. Since there is a unique critical point in the interior of the simplex it must be the global minimum over the closed simplex as desired. *q.e.d.*

The next lemma provides a more explicit expression for G_{crit} in the simplest case where $f_i = 1$ for all i.

Lemma 5.2. *Let*

$$G = \sum_{i=1} s_{i-1}^{-d} s_i^e \tag{40}$$

then

$$G_{crit} = \frac{1-q^{-h}}{1-q^{-1}} q^{\frac{1}{q-1} - \frac{h}{q^h-1}} p^{\frac{d-e}{q^h-1}} \tag{41}$$

Proof: We first note that for all i

$$D_i = -\frac{1}{e}\log(q) \tag{42}$$

which leads to

$$\delta = \sum_{i=1}^{h-1} \left(\frac{q^{h-i}-1}{q^h-1}\right)\log(q) = \frac{1}{q^h-1}\left(\left(\sum_{i=1}^{h} q^{h-i}\right) - h\right)\log(q) = \left(\frac{1}{q-1} - \frac{h}{q^h-1}\right)\log(q) \tag{43}$$

from which the result follows using (39) *q.e.d*

The next result shows how the value of G_{crit} changes as the coefficients f_i vary.

Lemma 5.3. *Let $G = \sum_{i=1}^{h} f_i s_{i-1}^{-d} s_i^e$ be a function as above. Consider a set of constants $c_i > 0$, $i = 1, ..., h$. Let $\tilde{f}_i = f_i c_i$ and let $\tilde{G} = \sum_{i=1}^{h} \tilde{f}_i s_{i-1}^{-d} s_i^e$. Let*

$$\beta_i = \frac{q-1}{q^h - 1} q^{h-i} \tag{44}$$

Let

$$c = \sum_{i=1}^{h} \beta_i \log(c_i) \tag{45}$$

then

$$\tilde{G}_{crit} = G_{crit} 10^c \tag{46}$$

Proof: Using $\tilde{f}_i$ we define $\tilde{D}_i$ and $\tilde{\delta}$ as before. We compute

$$\log(\tilde{G}_{crit}) - \log(G_{crit}) = (\log(\tilde{f}_1) - \log(f_1)) + e(\tilde{\delta} - \delta)$$

Looking at the first term we have

$$\log(\tilde{f}_1) - \log(f_1) = \log(c_i)$$

To compute the second term, we consider the effect of c_i on D_i. We have

$$\tilde{D}_i = \frac{1}{e}(\log(\frac{e}{d})) + \frac{1}{e}(\log(\frac{f_i c_i}{f_{i+1} c_{i+1}})) = D_i \frac{1}{e}(\log(c_{i+1}) - \log(c_i))$$

For $i = 2, \ldots h-1$ we have $\log(c_i)$ appearing with a positive sign in $\tilde{D}_i - D_i$ and with a negative sign in $\tilde{D}_{i-1} - D_{i-1}$. The values c_1 and c_h appear only once in $\tilde{D}_1$ and $\tilde{D}_{h-1}$ respectively. Using the definition of δ we obtain

$$e(\tilde{\delta} - \delta) = -e \sum_{i=1}^{h-1} \frac{q^{h-i} - 1}{q^h - 1} (\tilde{D}_i - D_i) = -\frac{q^{h-1} - 1}{q^h - 1} \log(c_1) + \sum_{i=2}^{h} \frac{q-1}{q^h - 1} (q^{h-i}) \log(c_i)$$

$$= -\frac{q^{h-1} - 1}{q^h - 1} \log(c_1) + \sum_{i=2}^{h} \beta_i \log(c_i) \tag{47}$$

Noting that

$$\beta_1 = 1 - \frac{q^{h-1} - 1}{q^h - 1}$$

we obtain the desired result. *q.e.d*

6. Performance for Bounded Pareto job size distributions

6.1. Analysis of mean normalized waiting time when $\rho < 1$

Let us assume that the total system utilization satisfies $\rho < 1$. We consider a `SITA` system with a fixed number of hosts h and a Bounded Pareto job size distribution $B_p(\alpha)$ with α fixed. We will use the notation $f \sim g$ to denote, p dependent, quantities f, g whose ratio tends to 1 as p tends to infinity. We consider the mean normalized waiting time $E(N)$. Following corollary 4.3, when computing $E(N)$, it is enough to study the case $\alpha \geq 1$. Consider a `SITA` system with h hosts and parameters $1 = s_0(p) < s_1(p) < \ldots < s_{h-1}(p) < s_h(p) = p$, where the job size distribution is $B_p(\alpha)$, for fixed $\alpha > 1$, and the system load is fixed at $\rho < 1$. Let $E(N)(\alpha, h, \rho, p, s_1(p), \ldots, s_{h-1}(p)) = E(W_p)/E(X_p)$ be the mean normalized waiting time. We have the following asymptotic and non-asymptotic estimates.

Lemma 6.1. *Let*

$$f_1^{SITA}(\alpha,\rho) = \frac{\rho}{2(1-\rho)}\frac{(\alpha-1)^2}{(2-\alpha)\alpha} \tag{48}$$

and let

$$f_i^{SITA}(\alpha,\rho) = \frac{\rho}{2}\frac{(\alpha-1)^2}{(2-\alpha)\alpha} \tag{49}$$

for $i > 1$. Let

$$G^{SITA}(\alpha,h,\rho,p,s_1,\ldots,s_{h-1}) = \sum_{i=1}^{h} f_i^{SITA} s_{i-1}^{-\alpha} s_i^{2-\alpha} \tag{50}$$

Assume that for all $i = 1,\ldots,h$, $s_i(p)/s_{i-1}(p) \to \infty$ as $p \to \infty$, then,

$$E(N)(\alpha,h,\rho,p,s_1(p),\ldots,s_{h-1}(p)) \sim G^{SITA}(\alpha,h,\rho,p,s_1(p),\ldots,s_{h-1}(p)) \tag{51}$$

In addition, for all cutoff values $s_1 < \ldots < s_{h-1}$

$$E(N)(\alpha,h,\rho,p,s_1,\ldots,s_{h-1}) < G^{SITA}(\alpha,h,\rho,p,s_1,\ldots,s_{h-1})(1-\rho\frac{s_1^{1-\alpha}}{1-p^{1-\alpha}})^{-1}(1-p^{1-\alpha})^{-2} \tag{52}$$

Proof: As observed in the proof of corollary 4.3, we may assume that the density function of $B(\alpha)_p$ is $s^{-\alpha-1}$, since the normalizing constant does not affect the computation of $E(N)$. The formula for $E_i(N)$, the weighted contribution of host i, given in equation (13), requires the computation of $m_0^f(I_i), m_1^f(I_i), m_2^f(I_i)$ and m_1^f. We compute m_j^f

$$m_j^f = \int_1^p s^j s^{-\alpha-1} ds = \frac{1}{\alpha-j}(1-(\frac{1}{p})^{\alpha-j}) \tag{53}$$

Similarly, we compute the non-normalized moments, $m_j^f(I_i)$,

$$m_j^f(I_i) = \int_{s_{i-1}}^{s_i} s^j s^{-\alpha-1} ds = \frac{1}{\alpha-j}(s_{i-1}^{-\alpha+j} - s_i^{-\alpha+j}) \tag{54}$$

which leads to

$$\frac{m_0^f(I_i)m_2^f(I_i)}{(m_1^f)^2} = \frac{(1-\alpha)^2}{\alpha(2-\alpha)} s_{i-1}^{-\alpha} s_i^{2-\alpha}(1-(\frac{s_{i-1}}{s_i})^{\alpha})(1-(\frac{s_{i-1}}{s_i})^{2-\alpha})\frac{1}{(1-(\frac{1}{p})^{\alpha-1})^2} \tag{55}$$

We see that for any cutoff values

$$\frac{m_0^f(I_i)m_2^f(I_i)}{(m_1^f)^2} \leq \frac{(1-\alpha)^2}{\alpha(2-\alpha)} s_{i-1}^{-\alpha} s_i^{2-\alpha}(1-p^{1-\alpha})^{-2} \tag{56}$$

In addition, when $\frac{s_i(p)}{s_{i-1}(p)} \to \infty$, the last three terms of equation (55) approach 1 and we obtain

$$\frac{m_0^f(I_i)m_2^f(I_i)}{(m_1^f)^2} \sim \frac{(1-\alpha)^2}{\alpha(2-\alpha)} s_{i-1}^{-\alpha} s_i^{2-\alpha} \tag{57}$$

For the other moment expression in $E_i(N)$ we have

$$\frac{m_1^f(I_i)}{m_1^f} = (s_{i-1}^{-\alpha+1} - s_i^{-\alpha+1})(1-(\frac{1}{p})^{\alpha-1})^{-1} \tag{58}$$

For $i > 1$ this gives

$$\frac{1}{1-\rho s_{i-1}^{1-\alpha}(1-(\frac{1}{p})^{\alpha-1})^{-1}} > (1-\rho\frac{m_1^f(I_i)}{m_1^f})^{-1} \sim 1 \tag{59}$$

and for $i = 1$ we have

$$\frac{1}{1-\rho} > (1 - \rho\frac{m_1^f(I_i)}{m_1^f})^{-1} \sim 1 - \rho \tag{60}$$

Equations (57) and (59-60) combine to give equation (51), while equation (56) combines with (59-60) to give the estimate (52). *q.e.d.*

Fix α and ρ, and consider a p-asymptotic family of `SITA` systems with h hosts and generic job size distribution $B(\alpha)_p$. Let

$$s_1^{opt}(p) < \ldots < s_{h-1}^{opt}(p)$$

be a set of parameters which minimizes $E(N)$ over the simplex of all possible cutoffs $\Delta_{h,p}$. Denote the corresponding optimal mean normalized waiting time by $E(N_p)^{opt} = E(N_p)^{opt}(\alpha, h, \rho, p)$.

Lemma 6.2. *If $\rho < 1$ and $\alpha > 1$ then $\frac{s_i^{opt}(p)}{s_{i-1}^{opt}(p)} \to \infty$ as $p \to \infty$*

Proof: The proof is by induction. Consider a 2 host system and assume that there is a sequence $p(i) \to \infty$ of ranges for which $s_1^{opt}(p(i)) < c$, where c is some constant. The second host will be responsible for job sizes in the interval $[p(i), p]$ which contains the interval $[c, p]$, and so its weighted contribution to mean normalized waiting time $E_2(N_p)$ will be $\Theta(p^{2-\alpha})$, which is, in order of magnitude, more than G_{crit}^{SITA}. By the estimate (52) the order of magnitude of $E(N_p)^{opt}$ is at most that of G_{crit}^{SITA}, a contradiction. For $h > 2$ we notice that in an optimal system, each subsystem of two consecutive hosts which handles jobs in the range $[s_{i-1}, s_{i+1}]$ with a single cutoff must also be optimal. We note that the Pareto distribution is self similar (scale invariant), hence, up to scaling the workload that this subsystem observes is the same as that of a Pareto Bounded distribution with parameter α and job size in the range $[1, s_{i+1}/s_{i-1}]$. If p is large (tends to infinity), then one of the ratios s_{i+1}/s_{i-1} has to be large (at least some positive power of p), which by the 2 host case implies that both s_i/s_{i-1} and s_{i+1}/s_i are large. These relations in turn imply that the ratios s_i/s_{i-2} and s_{i+2}/s_i are large. In turn, these last two statements imply by the 2 host case that s_{i-1}/s_{i-2} and s_{i+2}/s_{i+1} are large and we continue in the same manner to show that all ratios are large (tend to infinity) as required. *q.e.d*

Theorem 6.3. *Let $q = \frac{\alpha}{2-\alpha}$. Define*

$$\tilde{s}_1(p) = p^{\frac{q-1}{q^h-1}} \left[(1-\rho)^{\frac{q^{h-1}-1}{q^h-1}} q^{\frac{1}{q-1} - \frac{h}{q^h-1}}\right]^{\frac{1}{2-\alpha}} \tag{61}$$

$$\tilde{s}_2(p) = (q(1-\rho))^{-\frac{1}{2-\alpha}} \tilde{s}_1(p)^{\frac{2}{2-\alpha}} \tag{62}$$

and

$$\tilde{s}_{i+1}(p) = q^{-\frac{1}{2-\alpha}} \tilde{s}_i(p)^{\frac{2}{2-\alpha}} \tag{63}$$

for $i \geq 2$.

If $\rho < 1$ and $\alpha > 1$ then the cutoffs $\tilde{s}_i(p)$ are asymptotically optimal in the sense that

$$E(N)(\alpha, h, \rho, p, \tilde{s}_1(p), \ldots, \tilde{s}_{h-1}(p)) \sim E(N)(\alpha, h, \rho, p, s_1(p)^{opt}, \ldots, s_{h-1}(p)^{opt})$$

Consequently

$$E(N_p)^{opt} \sim (1-q^{-h})q^{\frac{1}{q-1} - \frac{h}{q^h-1}} (\frac{1}{1-\rho})^{\frac{(q-1)q^{h-1}}{q^h-1}} \rho\frac{(\alpha-1)}{4(2-\alpha)}p^{\frac{2\alpha-2}{q^h-1}} \tag{64}$$

$$< \frac{e}{4}\frac{\alpha-1}{(2-\alpha)}\frac{\rho}{1-\rho}p^{\frac{2\alpha-2}{q^h-1}}$$

where e refers to the natural base.

Proof: By lemma 6.2 the conditions to apply lemma 6.1 on $s_i(p)^{opt}$ apply and therefore.

$$E(N)^{opt} \sim G^{SITA}(s_1(p)^{opt}, \dots, s_{h-1}(p)^{opt})$$

The function G^{SITA} is in the family of functions which are considered by lemma 5.1, with $d^{SITA} = \alpha$ and $e^{SITA} = 2 - \alpha$. We get $q^{SITA} = \frac{d}{e} = \frac{\alpha}{2-\alpha}$ which agrees with our definition. The cutoffs $\tilde{s}_i(p)$ are the values which minimize G^{SITA} over $D_{h,p}$. Consider the function $G_2 = \rho\frac{(\alpha-1)^2}{2(2-\alpha)\alpha}G_1$, where G_1 is the function given in equation (40). The function G^{SITA} is obtained from G_2 by setting $c_1 = \frac{1}{1-\rho}$ and $c_i = 1$, for $i > 1$, in lemma 5.3. The formula for G^{SITA}_{crit}, which is given on the right hand side of (64) is a direct consequence of applying lemmas 5.2 and 5.3 successively and using the equality $\frac{1}{1-q^{-1}} = \frac{\alpha}{2(\alpha-1)}$ for simplification. The last inequality comes from further simplification and noting that $q^{\frac{1}{q-1}} < e$ when $q > 1$. *q.e.d*

6.2. Analysis of mean slowdown

The analysis of the previous subsection can be applied equally well to the determination of the optimal average slowdown $E(S)^{opt}$. We will content ourselves with order of magnitude calculations, though the precise coefficients can be determined as in the case of the optimal mean normalized waiting time. From equation (17) we see that when $\rho < 1$, $E_i(S) = \Theta(\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_0 m^f_1})$. We also observe that duality with respect to inv_3 of the expression $\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_0 m^f_1}$, shows that for $\rho < 1$, $E(S)^{opt}(\alpha, h, \rho, p) = \Theta(E(S)^{opt}(1-\alpha, h, \rho, p))$. Assume, $\alpha > 1$. From equation (54) we see that $\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_0 m^f_1} = \Theta(s_{i-1}^{-(\alpha+1)} s_i^{2-\alpha})$. By lemma 5.1 this leads to

$$E(S_p)^{opt} = \Theta(p^{\frac{2\alpha-1}{\tilde{q}^h-1}}) \tag{65}$$

where $\tilde{q} = \frac{\alpha+1}{2-\alpha}$.

For $1/2 < \alpha < 1$ we have $\frac{m^f_{-1}(I_i)m^f_2(I_i)}{m^f_0 m^f_1} = \Theta(s_{i-1}^{-(\alpha+1)} s_i^{2-\alpha} p^{\alpha-1})$, because in this case $m^f_1 = \Theta(p^{1-\alpha})$, while for $\alpha > 1$, $m^f_1 = \Theta(1)$. This leads to

$$E(S_p)^{opt} = \Theta(p^{\frac{2\alpha-1}{\tilde{q}^h-1}} p^{\alpha-1}) \tag{66}$$

In particular, using (66) and duality, we obtain the following conclusion

Corollary 6.4. *Assume $\rho < 1$ and $0 < \alpha < 1$. Let $\tilde{\alpha} = \alpha$ for $\alpha > 1/2$ and $\tilde{\alpha} = 1 - \alpha$ for $\alpha < 1/2$. Let $\tilde{q} = \frac{\tilde{\alpha}+1}{2-\tilde{\alpha}}$, then,*

$$E(S)^{opt}(\alpha) = o(1)$$

for $h > log_{\tilde{q}}(\frac{2\tilde{\alpha}-1}{1-\tilde{\alpha}} + 1)$.

6.3. Relation between optimizing mean slowdown and optimizing mean normalized waiting time

We can compute the slowdown for the optimal waiting time cutoffs and vice versa. Assume first that the parameters s_i are chosen to optimize waiting time. It is easy to verify that for the optimal s_i the terms $E_i(N)$ will all have the same order of magnitude. We have $E_i(S) = \frac{m^f_1}{m^f_0}E_i(N)m^f_{-1}(I_i)/m^f_0(I_i)$. Whenever, $s_i/s_{i-1} \to \infty$, then $m^f_{-1}(I_1)/m^f_0(I_1)$ has an order of magnitude which is larger than that of $m^f_{-1}(I_i)/m^f_0(I_i)$, for $i > 1$. We conclude that only $E_1(S)$ contributes asymptotically to $E(S)$. We get that for $\alpha > 1$

$$E(S) = \Theta(s_1^{2-\alpha}) = \Theta(E(N)^{opt}) = \Theta(p^{\frac{2\alpha-2}{q^h-1}}) \tag{67}$$

For $\alpha < 1$ the same argument yields

$$E(S) = \Theta(p^{\frac{2-2\alpha}{1-q^h}\alpha-1}) \tag{68}$$

where in this case $q = \alpha/(2-\alpha) < 1$. In particular, for $\alpha < 1$, as h goes to infinity, $q^h \to 0$ and $E(S)$ approaches the order of magnitude, $p^{1-\alpha}$.

To compute the mean normalized waiting time when the slowdown is optimized, we first assume $\alpha > 1$. When we optimize slowdown all the contributions $E_i(S)$ have the same order of magnitude, and running the previous argument in reverse shows that only $E_h(N)$ will contribute. We have

$$s_{h-1} = \Theta(s_1^{\frac{\tilde{q}^{h-1}-1}{\tilde{q}-1}}) = \Theta(p^{\frac{\tilde{q}^{h-1}-1}{\tilde{q}^{h}-1}}) \tag{69}$$

Let $\tilde{r} = \frac{\tilde{q}^{h-1}-1}{\tilde{q}^h-1}$, then $\tilde{r} < 1$. We obtain

$$E(N) \sim E_h(N) = \Theta(s_{h-1}^{-\alpha} p^{2-\alpha}) = \Theta(p^{-\alpha\tilde{r}} p^{2-\alpha}) \tag{70}$$

For all $\alpha > 1$, $\tilde{r}$ tends to $\frac{1}{\tilde{q}} = \frac{2-\alpha}{\alpha+1}$, as h tends to infinity, therefore $E(N)$ tends to $p^{(2-\alpha)(1-\frac{\alpha}{\alpha+1})}$ as h goes to infinity.

For $\alpha < 1$, we similarly have $E(N) = \Theta(s_{h-1}^{-\alpha} p^{2-\alpha} p^{2\alpha-2}) = \Theta((p/s_{h-1})^{\alpha}) = \Theta(p^{(1-\tilde{r})\alpha})$. For $1/2 < \alpha < 1$, as h tends to infinity $1-\tilde{r}$ tends to $\frac{2\alpha-1}{\alpha+1}$ and so $E(N)$ tends to $p^{(2\alpha-1)(\frac{\alpha}{\alpha+1})}$. On the other hand, for $\alpha < 1/2$, $\tilde{r}$ tends to 1 as h tends to infinity, so the waiting time becomes an arbitrarily small power of p.

From these computations we conclude that for $\alpha > 1$ it is better to tune the cutoffs to provide optimal waiting time since they also lead to small slowdown, while optimizing slowdown does not lead to small waiting times.

On the contrary, for $\alpha < 1/2$ it is better to tune the system for optimal slowdown. For the range of values $1/2 < \alpha < 1$, optimizing $E(N)$ leads to relatively poor results for $E(S)$ and vice versa. This is due to the conflicting duality properties of the objective functions in that range. This shows that one needs to choose carefully the objective function for a SITA system.

6.4. Higher loads

We consider now a load $\rho > 1$. For convenience, we will assume that ρ is not an integer. We define $\tilde{h} = h - [\rho]$, where $[\rho]$ designates the integer part of ρ. Let s_{load} be such that $\lambda \int_1^{s_{load}} s^{-\alpha} ds = [\rho]$.

Theorem 6.5. *Given a load ρ, and $\alpha > 1$, the optimal average normalized waiting time satisfies*

$$E(N)^{opt}(\alpha, h, \rho, p) \sim s_{load}^{1-\alpha} E(N)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load}) \tag{71}$$

Proof: Consider a stable SITA system $S(1)$, with cutoffs s_i, such that $\rho_i < 1$. We must have $s_{[\rho]} < s_{load}$, since otherwise one of the hosts will be overloaded. We conclude that the last $\tilde{h}$ hosts will be responsible, at least, for all jobs in the range $[s_{load}, p]$. We let $S(2)$ be the subsystem of $S(1)$ which consists of the last $\tilde{h}$ hosts and let $S(3)$ be the system with the same hosts as $S(2)$, but with the subset of the traffic of $S(2)$ consisting only of jobs s with $s_{load} \leq s$. Obviously, the waiting time of any job in $S(2)$ will be greater or equal than in $S(3)$ and hence the average waiting time $E(W)_{S(2)} \geq E(W)_{S(3)}$.

We consider the average waiting time of the original system $S(1)$. Let p_{load} be the probability that a job s has size $s \geq s_{load}$. We have $E(W)_{S(1)} = \sum_{i=1}^{h} p_i E(W_i) = \sum_{i=1}^{[\rho]} p_i E(W_i) + \tilde{p} E(W)_{S(2)}$, where we consider the last $\tilde{h}$ hosts to be the single subsystem $S(2)$, with probability $\tilde{p} = 1 - \sum_{i=1}^{[\rho]} p_i$, the probability of a job of size $s \geq s_{[\rho]}$. Since $s_{load} \geq s_{[\rho]}$ we have $p_{load} \leq \tilde{p}$ and so $E(W)_{S(1)} \geq \tilde{p} E(W)_{S(2)} \geq p_{load} E(W)_{S(3)} \sim s_{load}^{-\alpha} E(W)_{S(3)}$. The job size distribution of $S(3)$ is a Bounded Pareto distribution with parameter α, restricted to the interval $[s_{load}, p]$. By scale invariance of the Pareto distribution this is the same as a time scaling by s_{load} of the Bounded Pareto distribution on the interval $[1, p/s_{load}]$. If we consider two systems with equal utilization, but with a time scaled job size distribution, then according to the P-K formula, the waiting time will also be time scaled. Consequently $E(W)_{S(3)}^{opt} = s_{load} E(W)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load})$. We also note that $E(X_p) \sim E(X_{p/s_{load}})$. Putting together these we get the asymptotic inequality

$$E(N)^{opt}(\alpha, h, \rho, p) \geq (1-\varepsilon) s_{load}^{1-\alpha} E(N)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load}) \tag{72}$$

for any given $\varepsilon > 0$ and p large enough.

To get the reverse asymptotic inequality, we construct near optimal cutoffs. fix any $\varepsilon > 0$. For any s, define $\rho(s)$ to be the the load of jobs greater than s, given by $\rho(s) = \rho(\int_s^p s^{-\alpha} ds)/(\int_1^p s^{-\alpha} ds)$. For example, $\rho(s_{load}) = [\rho]$. Similarly let $p(s)$ be the portion of jobs of size at least s, given by, $p(s) = (\int_s^p s^{-\alpha-1} ds)/(\int_1^p s^{-\alpha-1} ds)$. We choose $\tilde{s}$ such that $\tilde{s} < s_{load}$, $\rho(\tilde{s}) < \tilde{h}$ and such that

$$\tilde{s}^{1-\alpha} p(\tilde{s}) E(W)^{opt}(\alpha, \tilde{h}, \rho(\tilde{s}), p/\tilde{s}) < (1 + \frac{\varepsilon}{2}) s_{load}^{1-\alpha} p(s_{load}) E(W)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load})$$

for all p large enough. By the continuity of $\rho(s)$, $s^{1-\alpha}$ and $p(s)$ and the continuity of $E(W)^{opt}$ with respect to the size of the largest job such a choice is possible. We set $s_{[\rho]}$ to be $\tilde{s}$ and fix $s_1, \ldots, s_{[\rho]-1}$ so that the first $[\rho]$ hosts have equal load. Since $\tilde{s} < s_{load}$ we know the hosts are stable and hence they provide a fixed and finite contribution to $E(W)$. The other cutoffs are chosen to be optimal in the range $[\tilde{s}, p]$. Since $E(W)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load})$ tends to infinity with p and the first $[\rho]$ cutoffs are fixed we have for p large enough

$$\sum_{i=1}^{[\rho]} p_i E(W_i) < \frac{\varepsilon}{2} s_{load}^{1-\alpha} p(s_{load}) E(W)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load})$$

Putting the last two inequalities together yields

$$E(N)^{opt}(\alpha, h, \rho, p) \le (1 + \varepsilon) s_{load}^{1-\alpha} E(N)^{opt}(\alpha, \tilde{h}, \rho - [\rho], p/s_{load}) \tag{73}$$

which together with equation (72) yields the desired result. *q.e.d*

6.5. Comparison with LWR

To compare the performance of SITA on Bounded Pareto distributions with that of LWR we need some order of magnitude estimates on the performance of LWR with the same workload. There is a great deal of literature on the performance of LWR in the context of heavy-tailed job size distributions. Putting together results from [24, 25, 26, 27, 31] and in particular from [28] we can conclude that if X is a Bounded Pareto distribution, with p large

$$Pr(W^{LWR} > s) = \Theta((1 - X_e(s))^{\tilde{h}}) \tag{74}$$

The left hand side is the probability that the waiting time in a LWR queue is at least s. In the right hand side X_e is the stationary-excessor (or residual-lifetime) distribution associated with X. It is given by $X_e(s) = (\int_1^s tf(t)dt)/E(X)$. The effective number of hosts $\tilde{h}$ is defined as before to be $\tilde{h} = [h - \rho] + 1$. Let $E(N)^{LWR}(h, \rho, \alpha, p)$ be the average normalized waiting time of a LWR system. Using the above estimate a straightforward computation shows that as $p \to \infty$ we have

1) $E(N)^{LWR}(\tilde{h}, \rho, \alpha, p) = O(1)$, if $\alpha \ge \frac{\tilde{h}+1}{\tilde{h}}$,

2) $E(N)^{LWR}(\tilde{h}, \rho, \alpha, p) = \Theta(p^{\tilde{h}(1-\alpha)+1})$, if $1 < \alpha < \frac{\tilde{h}+1}{\tilde{h}}$,

3) $E(N)^{LWR}(\tilde{h}, \rho, \alpha, p) = \Theta(p^{\alpha})$, if $\alpha < 1$.

When $\tilde{h} = 1$, both SITA and LWR have the same order of magnitude performance as a single host. It is hard to say which is better since the estimates for LWR are rather poor.

We assume that $\tilde{h} > 1$. From the first item we see that for $\alpha > \frac{\tilde{h}+1}{\tilde{h}}$, LWR is an asymptotically better policy than SITA. From the third item we see that for $\alpha \le 1$, SITA is asymptotically better than LWR. We conclude that for a fixed load ρ, given enough hosts, SITA will be asymptotically better than LWR if and only if $\alpha \le 1$.

More precisely, from the second item and our analysis of SITA given in equations (64) and (71), The crossover point $1 < \alpha_{\tilde{h}} < \frac{\tilde{h}+1}{\tilde{h}}$, such that LWR is asymptotically better if and only if $\alpha \ge \alpha_{\tilde{h}}$, satisfies

$$(-\tilde{h}\alpha_{\tilde{h}} + \tilde{h} + 1) F_{\tilde{h}}(\alpha_{\tilde{h}}) = (2 - \alpha_{\tilde{h}})$$

where F_k is the degree $k - 1$ (or less) polynomial $F_k(\alpha) = \sum_{m=0}^{k-1} \alpha^m (2 - \alpha)^{k-1-m}$. When $\tilde{h} = 2$, the polynomial F_1 degenerates to the constant function with value 2 and a short computation yields $\alpha_2 = \sqrt{2}$, namely, for $\alpha < \sqrt{2}$, SITA is asymptotically better. The results for average slowdown are almost identical, with negligible changes in the crossover points.

6.6. Heterogeneous hosts

We consider the case where not all hosts are equally powerful. We assume that there is a given reference host with normalized computing power of 1 unit. Following [8, 10] and others, we will assume that any job exhibits the same speed up/slowdown in processing time when moved to another host. We denote the uniform speed up/slowdown of jobs on a host in comparison with the reference host by v. We assume that the speedup values v_j, $j = 1, \ldots, h$, of hosts in the system are numbered so that they form a decreasing sequence $v_1 \geq v_2 \geq \ldots \geq v_h$. We will assume that $E(X)$, is computed w.r.t the reference host. When hosts are not equally powerful, the parameter set for a SITA system includes in addition to the cutoffs s_i, a permutation π on $1, \ldots, h$. The permutation designates that the i'th host in the system, the host which is responsible for jobs in the range $[s_{i-1}, s_i]$ has power $v_{\pi(i)}$. Given a permutation π we can define its reverse permutation π^{rev} by the formula $\pi^{rev}(i) = \pi(h - i + 1)$. We say that a system has low load if $v_1 > \rho$.

Theorem 6.6. *If $\alpha > 1$, then the identity permutation, id, provides the optimal ordering of hosts. The same holds for $\alpha < 1$ and the permutation id^{rev}.*

Proof: We first consider the low load case. We recall the mean waiting time formula for equally powerful hosts with normalized strength 1, $E(W_p)/E(X_p) \sim \sum_{i=1}^{h} f_i^{SITA} s_{i-1}^{-\alpha} s_i^{2-\alpha}$. Given a permutation π and a strength vector $\bar{v} = (v_1, \ldots v_h)$, we denote the new waiting time by W^π. We let $v_i^\pi = v_{\pi(i)}$. If $v_1^\pi < \rho$, then the cutoff s_1 is necessarily bounded and hence the optimal average normalized waiting time will have the order of magnitude corresponding to a system with less than h hosts. On the other hand, if $v_1^\pi > \rho$, then as our computations show, the optimal first cutoff $s_1(p)$ will go to infinity with p (there is no stability constraint for the first host) and consequently, the load of all other hosts will tend to 0 and stability will not be an issue when considering optimal cutoffs. this will lead to a value of $E(N^{opt})$ which has the same order of magnitude as that of a system of h homogeneous hosts with $\rho < 1$. We conclude that in the optimal ordering we have $v_1^\pi > \rho$. For such permutations we have

$$E(W_p^\pi)/E(X_p) \sim \sum_{i=1}^{h} f_i^\pi s_{i-1}^{-\alpha} s_i^{2-\alpha}$$

where the new coefficients f_i^π are given by $f_i^\pi = f_i/v_i^\pi$ for $i > 1$, while for $i = 1$ we have

$$f_1^\pi = \frac{f_1}{v_1^\pi} \frac{1-\rho}{1-\rho/v_1^\pi}$$

This is because with respect to the refernce host, the utilization of the i'th host which is originally ρ_i becomes ρ_i/v_i^π, while $E(X_i^2)/E(X_i^2)$ which also figures in the P-K formula $E(N)_i = \frac{\rho_i}{2(1-\rho_i)} \frac{E(X_i^2)}{E(X_i)^2}$, remains the same. For $i > 1$, since $\alpha > 1$, we have $\rho_i \to 0$, so dividing ρ_i by v_i^π has no asymptotic effect on the denominator term $2(1 - \rho_i)$ and the formulas follow. We can now apply lemma 5.3 with

$$c_1 = (1-\rho) \frac{1}{v_1^\pi} \frac{1}{1-\rho/v_i^\pi}$$

and

$$c_i = \frac{1}{v_i^\pi}$$

for $i > 1$. We have

$$\log(E(W_p^\pi)) - \log(E(W_p))$$

$$\sim \frac{q-1}{q^h-1}(q^{h-1})(\log(1-\rho) + \frac{q-1}{q^h-1}(q^{h-1})\log(\frac{1}{1-\rho/v_1^\pi}) + \sum_{i=1}^{h} \beta_i \log(\frac{1}{v_i^\pi})$$

The first term does not depend on v_i. The second term is minimized if $\pi(1) = 1$, since that choice maximizes v_1^π. The third term is minimized when the ordering of v_i^π opposes that of β_i. The β_i

form a descending sequence which means that $\frac{1}{v_i^\pi}$ should form an increasing sequence, which leads us to conclude that the sequence v_i^π should be decreasing, or that π should be the identity. Since the choices for the second and third term are consistent we are done. The result holds for $\alpha < 1$ by noting that the duality between α and $2 - \alpha$ still holds if for each permutation π we consider π^{rev} to be its dual.

For higher loads we assume that $\rho < \sum_{i=1}^{h} v_i$, otherwise the system cannot handle a load of ρ. We consider sub-divisions of the hosts $1, \dots, h$ into two complementary sets A and B. We will then be interested in orderings in which the hosts in A are ordered before the hosts in B. If the hosts in A can handle enough load so that the strongest host in B is stronger than the remaining load, we will be in the light load situation which was analyzed above. More formally, we consider the following definition. Let $\rho_A = \sum_{i \in A} v_i$ be the total load that hosts in A can handle. Let $b = Min(B)$ be the minimal index in B. By our convention that power is a decreasing function of the host index, v_b will be the power of the most powerful host in B. We say that a sub-division is *valid* if $0 < \rho - \rho_A < v_b$. We let $\tilde{h} = max|B|$ where the maximum is taken over all valid sub-divisions and $|B|$ denotes the size of the set B. As argued in the low load case, the order of magnitude of $E(N)^{opt}$ in a sub-division will be the same as that of a homogeneous system with $|B|$ hosts, if and only if the sub-division is valid.

We say that a sub-division into A and B is *good* if $\tilde{h} = |B|$. Among all valid sub-divisions, the good ones are those which can lead to the smallest order of magnitude for $E(N)^{opt}$. Let j be the smallest index such that $\rho - \sum_{i=1}^{j} v_i < v_{j+1}$. Let H be the set of hosts with index i such that $i \leq j$. We claim that H and its complement, given by indices i such that $i > j$ form a good sub-division. To see this, assume that A, B form a good sub-division. Assume that $a \in A$ and $b \in B$ are indices such that $a > b$ or correspondingly $v_b > v_a$. We may assume w.l.o.g that $b = Min(B)$. Consider the sub-division $\tilde{A}, \tilde{B}$ which is obtained by moving a to B and b to A. Since A and B formed a good sub-division, we have $v_b > \rho - \rho_A = \rho - (\rho_{\tilde{A}} - v_b + v_a) = \rho - \rho_{\tilde{A}} + v_b - v_a$ which implies that $v_a > \rho - \rho_{\tilde{A}}$, but $a \in \tilde{B}$ so the sub-division $\tilde{A}, \tilde{B}$ is also good. We may repeat this process until all the indices in A are smaller than all the indices in B, which means that A is a prefix of the set of indices.

The set H is by definition the only prefix set which can be good since it maximizes the size of its complement. Assume that A, B form a valid sub-division. We choose any ordering of the hosts in A. We load each host in turn to nearly full utilization. This will cover jobs in the size interval $[1, s_A - \varepsilon]$, where s_A is such that $\lambda \int_1^{s_A} sf(s)ds = \rho_A$. By the definition of a valid sub-division the hosts in B are in the low load case, once ε is small enough. We know from theorem 6.6 that for the best performance, the hosts in B should be ordered in decreasing power order. Since the ordering in A does not affect the asymptotics of performance, we also choose decreasing power order in A. The performance of such a system will have order of magnitude $\Theta(p^{(2\alpha-2)/(q^{|B|}-1)})$. We see that good sub-divisions will lead to the best order of magnitude performance.

We want to show that among good sub-divisions, H leads to the best performance. For this we return to our change from the sub-division A, B to $\tilde{A}, \tilde{B}$. All elements in $A \cap \tilde{A} = A - a$ are used in both cases to lower load in the same fashion and so have the same effect on system performance, we may therefore simply assume that the set is empty. Similarly, load does not affect the performance coefficients f_i^π, $i > 1$, since the first host, b in this case absorbs all the load. We conclude that it is enough to consider a two host system with hosts a and b such that $v_b > v_a$ and $\rho < v_a + v_b$. If we let a be the first host then c_1 in theorem 6.6 is $\frac{1}{v_b}\frac{1}{1-\frac{\rho - v_a}{v_b}}$, while if we let b be the first host the coefficient is $\frac{1}{v_a}\frac{1}{1-\frac{\rho - v_b}{v_a}}$, however both expressions equal to $\frac{1}{v_a + v_b - \rho}$. The difference between the two scenarios, is that when a is the first host, the range for the second host is p/s_{v_a}, while if b is the first host the range is p/s_{v_b} which is smaller, hence we get better performance. We conclude that the sub-division $\tilde{A}, \tilde{B}$ is preferable, which leads us to conclude that H is the best sub-division, which leads us to ordering all the hosts in decreasing power order. *q.e.d*

6.7. The case of $\alpha = 1$

We compute the asymptotic mean normalized waiting time when $\alpha = 1$ and $\rho < 1$. The only difference from the computations for $\alpha > 1$ is in the formula for the first non normalized moment $m_1^f(I_i) = \ln(s_i) - \ln(s_{i-1})$. We have

$$\frac{m_0^f(I_i) m_2^f(I_i)}{(m_1^f)^2} \sim (\frac{1}{\ln^2(p)} s_{i-1}^{-1} s_i) \tag{75}$$

Letting $s_i = p^{i/h}$ we see that the optimal performance satisfies $E(N^{opt}) = O(p^{1/h})$. On the other hand, expression (75) also shows that for any $\varepsilon > 0$ and any i we have $s_i^{opt}(p)/s_{i-1}^{opt}(p) = O(p^{\frac{1}{h}+\varepsilon})$. This constraint together with the constraint $\Pi_{i=1}^h s_i/s_{i-1} = p$ implies that $\ln(s_i^{opt}(p)) \sim \frac{i}{h} ln(p)$, which in turn leads for the optimal s_i to the estimate $\frac{\rho}{2}(1 - \rho \frac{m_1^f(I_i)}{m_1^f})^{-1} \sim \frac{\rho}{2(1-\rho/h)}$. Using lemma 5.1 we get

$$E(N)^{opt} \sim \frac{\rho/h}{2(1-\rho/h)} \frac{p^{1/h}}{ln^2(p^{1/h})} \tag{76}$$

which is the same as the performance of a single host with job size distribution $B(p^{1/h}, 1)$ and load ρ/h.

Comparing theorem 3.1 with the result for $\alpha = 1$, shows that up to polylogarithmic factors, the performance of SITA on $B(p, 1)$ is the worst possible among all distributions which are bounded on the interval $[1, p]$. As we shall see later the same is true for the LWR policy, however, in that case $E(N)^{LWR}(B(p, 1)) = \Theta(p)$ regardless of h, hence the performance of LWR is far worse than that of SITA.

6.8. Higher moments

We can compute the asymptotics of higher moments of the slowdown and normalized waiting time, when the job size distribution is Pareto. The waiting time moments for a single host are given by Takacs' recursive formula, see [15]

$$E(W^k) = \frac{\rho}{E(X)(1-\rho)} \sum_{i=1}^{k} B_{k,i} \frac{1}{i+1} E(X^{i+1}) E(W^{k-i})$$

where $B_{k,i}$ denotes the binomial coefficient. Assuming that the job size distribution is $B(p, 1)$, and $k \geq 1$, we claim inductively that for $k + 1 > \alpha > 1$, $E(W^k) = \Theta(p^{k+1-\alpha})$, while for $\alpha < 1$, $E(W^k) = \Theta(p^k)$. For $k = 1$ the Takacs formula coincides with the P-K formula for which we have already done the computations which verify the claim. For $k > 1$, the term $E(X^{i+1})$ has order of magnitude $E(X^{i+1}) = \Theta(p^{i+1-\alpha})$, if $i+1 > \alpha$ and $\Theta(1)$ otherwise. We see that the dominant term in the sum must satisfy $i+1 > \alpha$ and we assume this inequality. Assume that $\alpha > 1$. The inductive hypothesis for $E(W^{k-i})$, $i < k$, implies that the product satisfies $E(X^{i+1})E(W^{k-i}) = \Theta(p^{k+2-2\alpha})$ when $k > i > \alpha - 1$ and $E(X^{i+1})E(W^{k-i}) = \Theta(p^{k+1-\alpha})$ when $k = i$. We see that the term with $k = i$ is dominant and we obtain the stronger result that

$$E(W^k) \sim \frac{\rho}{E(X)(i+1)(1-\rho)} E(X^{k+1})$$

A similar computation for $\alpha < 1$ shows that the result also holds in that case. We can now translate this result into a computation for SITA systems. Assuming that $s_i(p)/s_{i-1}(p)$, and following the exact same procedure that led to theorem 6.1, we can use the above asymptotics to show the following corollary which for simplicity we state for $\alpha > 1$

Corollary 6.7. *Assume $\alpha > 1$ and $\rho < 1$. Let*

$$f_1^{k,SITA}(\alpha, \rho) = \frac{\rho}{(k+1)(1-\rho)} \frac{(\alpha-1)^k}{(k+1-\alpha)\alpha} \tag{77}$$

and let

$$f_i^{k,SITA}(\alpha,\rho) = \frac{\rho}{k+1}\frac{(\alpha-1)^k}{(k+1-\alpha)\alpha} \tag{78}$$

for $i > 1$. Let

$$G^{k,SITA}(\alpha,h,\rho,p,s_1,\dots,s_{h-1}) = \sum_{i=1}^{h} f_i^{SITA} s_{i-1}^{-\alpha} s_i^{k+1-\alpha} \tag{79}$$

Assume that for all $i = 1,\dots,h$, $s_i(p)/s_{i-1}(p) \to \infty$ as $p \to \infty$, then,

$$E(N^k)(\alpha,h,\rho,p,s_1(p),\dots,s_{h-1}(p)) \sim G^{k,SITA}(\alpha,h,\rho,p,s_1(p),\dots,s_{h-1}(p)) \tag{80}$$

From the corollary and lemma 5.1 we can compute the asymptotics of $E(N^k)$ for the optimal cutoffs. We also note that from a duality standpoint it is more natural to consider another normalization of the higher moments of waiting time, namely $E(W^k)/(E(X)E(X^{k-1}))$. When $\rho < 1$, this normalization preserves order of magnitude under the involution which sends α to $k+1-\alpha$. We can also compute the asymptotics for $E(S^k) = E(W^k)E(X^{-k})$, since we know the asymptotics of both components of the product. It is interesting to note that in this case the involution mapping α to $1-\alpha$ preserves orders of magintude asymptotics for all moments at once when $\rho < 1$.

6.9. Analysis of the SITA-V heuristics

We will analyze the asymptotic behavior of the SITA-V heuristics, suggested by Harchol-balter, Crovella and Murta in [11]. We recall the three basic heuristics for determining size cutoffs:

- The front heuristic: We consider a two server system in which the first server has strength 1, while the second server has strength $h-1$. We let s_1 be the optimal cutoff point for the two server system. We continue inductively to find the other cutting points, by decreasing the number of hosts by 1 and considering the job size distribution restricted to the interval $[s_1, p]$.
- The back heuristic: This is the same as the front heuristic, but the first server has strength $h-1$ and we find s_{h-1}.
- The middle heuristic: We assume for simplicity that $h = 2^i$ is a power of two. We consider a two host system with equally powerful hosts and let the optimal cutoff be $s_{h/2}$. We continue the process inductively and independently on the ranges $[1, s_{h/2}]$ and $[s_{h/2}, p]$.

For the sake of asymptotic analysis of waiting times, the properties of the symmetry between α and $2-\alpha$ implies that it is enough to consider the front and middle heuristics. We will also content ourselves with order of magnitude calculations and skip some calculational details.

Theorem 6.8. *Assume $\rho < 1$, and $\alpha > 1$. Let $W_f(p), W_b(p)$ and $W_m(p)$, denote the waiting time w.r.t the front, back and middle heuristics, then*

$$W_b(p) = \Theta(W_f(p)) = \Theta(p^{(2-\alpha)^2/2}) \tag{81}$$

and

$$W_m(p) = \Omega(p^{(2-\alpha)h^{\log_2(1-\alpha/2)}}) \tag{82}$$

Sketch of proof: Since $\rho < 1$ we notice that changing the strength of a server to $h-1$ doesn't change the order of magnitude of the waiting time or the cutoff point s_1, in either the front or back heuristic. We conclude that the waiting time of the two server system will be $\Theta(p^{(2\alpha-2)/(q^2-1)}) = \Theta(p^{(2-\alpha)^2/2})$. This is also the weighted contribution of each server individually to the average waiting time. Since the heuristics don't decompose further one of the size intervals, the final performance will be commensurable to the performance after one step of the algorithm, hence the first part of the theorem. For the second part, we notice that we can write the waiting time W_2 of a two server system as $W_2(p) = \Theta(p^{(2-\alpha)^2/2}) = \Theta((p^{2-\alpha})^{1-\alpha/2}) = \Theta(W_1(p))^{1-\alpha/2}$ where W_1 denotes the waiting time of a single server system as given by the P-K formula. In the middle heuristic we iteratively repeat the process of replacing a single server by two servers, $\log_2(h)$ times. When, $\alpha > 1$, the short job server gets essentially all the requests, so its utilization is the same as that of the entire system and its shortest job is of size 1. By the self similarity of the process and the job size distribution, we conclude that the performance of the first server which is responsible for the smallest jobs will be $\Theta(W_1(p)^{(1-\alpha/2)^{\log_2(h)}})$ which leads to the second statement. *q.e.d.*

When $\rho > 1$ it may happen that the middle heuristic is no better than the other two. Consider the case where h is even and $\rho = (h-1)/2$. We have $\tilde{h} = h/2$, so the optimal waiting time will be $E(W) = \Theta(p^{(2\alpha-2)/(q^{h/2}-1)})$. On the other hand, the middle policy will consider two hosts of power $h/2 > \rho$. Therefore, it will choose $s_{h/2} = \Theta(p^{(2-\alpha)/\alpha})$. The resulting subsystem consisting of the $h/2$ hosts, responsible for the short jobs, will have to handle a load of ρ. For the subsystem we have $\tilde{h} = 1$ and we obtain the overall waiting time of a two host system as with the other two policies.

7. Conclusion and future work

We have presented an asymptotic analysis of SITA scheduling policies. Assuming Poisson arrivals, when the job size distribution is Bounded Pareto, the performance is asymptotically described by a simple formula which is amenable to analysis. The resulting analysis defines the precise parameter boundaries in which SITA outperforms LWR. The results show that the decision as to which algorithm is better is non trivial.

For more general distributions, but still assuming Poisson arrivals, SITA provides better asymptotic performance guarantees than LWR, however, that does not mean it will outperform LWR in practice, since the constants involved decrease rapidly with the number of hosts in a LWR system. Among all distributions SITA and LWR have particular difficulties in dealing with the Pareto distribution with parameter $\alpha = 1$.

There is a symmetry between the performance with workload distribution $B(\alpha)$ and performance with $B(2-\alpha)$, when we use normalized waiting time as our performance measure. When we use slowdown as a measure the symmetry exchanges α and $1-\alpha$.

In addition to the results presented in this paper, we have other results on SITA and related policies, which will be presented in forthcoming papers. We briefly overview these results.

Inspired by the methods of this paper, we have shown that under very general circumstances (Little's law applies), the cutoffs which best balance average queue lengths among the hosts, are h competitive with respect to minimizing average waiting time. This means that the average waiting time for optimal cutoffs is at most h times smaller. In many (but not all) cases better competitive bounds can be produced. This is the first general competitive rule for determining cutoffs.

We have analyzed the case of h-asymptotics in which the load and job size distribution are held fixed, while the number of hosts is increased. This regime is called in [9] the *server expansion metric*. In this regime the performance of SITA is controlled by an Euler-Lagrange equation which is very similar to that of the geodesic equation in a 2-dimensional space-time. We are still exploring the consequences of this analysis. In particular, it seems that non-uniform speed-up of hosts introduces an analogue of curvature to the model.

We have analyzed a variant of SITA known as TAGS, [9], which deals with the case of unknown job sizes. For Bounded Pareto distributions, We show that at low loads the cost of not knowing the job size is at most a factor of 2, regardless of α or the number of hosts. However, at higher loads the TAGS policy leads to serious issues of stability which we analyze completely. We also suggest a more stable variant which combines TAGS and LWR, whose performance is close to that of SITA for all loads.

Bounded Pareto distributions serve as the basic components of more complicated distributions. For example, the bimodal and trimodal distributions which are explored in [14], are obtained from

Bounded Pareto distributions via a process of discretization. We can use the methods of the present paper to provide a complete analysis of such distributions. Other variants include the double Pareto distributions, which play a role in modeling job size distributions on the web, [16]. Again, the methods which were developed in this paper can deal with such generalizations as well.

There are several interesting directions for future research which we plan to pursue. One possibility is to merge `SITA` with other known methods for improving queue performance. For instance, we may replace FCFS at each server by a better performing policy such as `Shortest job first`. As it turns out, such a change will not change the order of magnitude performance, it will only improve the constants involved. A more interesting suggestion is to combine `SITA` with cycle stealing schemes, see [13] for an example of cycle stealing. While the particular scheme presented in [13] will not lower the order of magnitude, it seems that some variants can. We hope to explore this promising direction in future work.

We also feel that there is much more to explore regarding duality as an organizing principle for understanding the behavior of `SITA`. We have seen that for Bounded Pareto distributions, the behavior of slowdown vs. waiting time optimization, and the ordering of hosts are both controlled by duality. We believe that much of the analysis can be generalized to an analysis based on the behavior of a distribution w.r.t the involutions inv_β. We also note that the involutions are strongly related to the functional equation of level 1 modular forms. In particular, some Eisenstein series, cusp forms, powers of the Theta function and powers of the Dedekind η function have nice symmetry properties w.r.t various involutions. We hope to use some insights from number theory to further explore duality.

Acknowledgments: Some of the results in this paper, are strongly related to results from the second author's masters thesis under the supervision of Prof. Gadi Rabinowitz. We would like to thank him for many valuable comments and his encouragment.

References

[1] M.F. Arlitt and C.L. Williamson, Internet web servers: Workload characterization and performance implications, *IEEE/ACM Transactions on networking*, vol. 5(5), 631-645, 1997.

[2] E. Bachmat and H. Sarfati, Analysis of Size Interval Task Assignment `SITA` policies. Extended abstract in *Performance Evaluation Reviews (PER), Special issue devoted to the proceedings of the Tenth Workshop on mathematical performance modeling and analysis (MAMA)* , Vol. 36(2), 107-109, 2008.

[3] N. Basher, A. Mahanti, C. Williamson, and M. Arlitt, A comparative analysis of web and Peer-to-Peer traffic, in *Proceedings of WWW2008*, 287-296, 2008.

[4] V. Cardellini, E. Casalicchio, M. Colajanni and P. Yu (2002), The state of the art in locally distributed web server systems, *ACM computing surveys*, vol. 34(2), 263-311.

[5] G. Ciardo, A. Riska, and E. Smirni, EQUILOAD: A load balancing policy for clustered web servers, *Performance Evaluation*, 46, 101 ? 124, 2001.

[6] M.E. Crovella and A. Bestavros, Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on networking*, Vol. 5(6), 835-846, 1997.

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

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

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

[10] M. Harchol-Balter and M. Crovella, Method and apparatus for assigning tasks in a distributed server system, U.S. patent number 6,223,205, 2001.

[11] M. Harchol-Balter, M. Crovella and 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.

[12] M. Harchol-Balter, M. Crovella and C. Murta, Task assignment in a distributed system: Improving performance by load unbalancing, Proceedings of SIGMETRICS 98, poster session, 1998. Full version published as a Boston University technical report number BUCS-TR-1997-019.

[13] M. Harchol-Balter, C. Li, T. Osogami, A. Scheller-Wolf and M. Squillante, Cycle stealing under immediate dispatch task assignment, *Proceedings of the 15th ACM symp. on parallelism in algo., SPAA 2003*, 274-285, 2003.

[14] M. Harchol-Balter, A. Scheller-Wolf, and 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. Seattle, WA, June 2009.

[15] L. Kleinrock, *Queueing systems. Volume 1: Theory*, Wiley-Interscience, 1975.

[16] M. Mitzenmacher, Dynamic models for file sizes and double Pareto distributions, *Internet mathematics*, vol 1, No. 3, pp. 305-334, 2004.

[17] R.D. Nelson and T.K. Phillips, An approximation for the mean response time for the response time of shortest queue routing, *Performance evaluation review*, vol. 7, 181-189, 1989.

[18] k. Oida and S. Saito, A packet size aware adaptive routing algorithm for parallel transmission server systems, *Journal of parallel and distributed systems*, vol. 64(1), 36-47, 2004.

[19] B. Riemann, Ueber die Anzahl der Primzahlen unter einer gegebenen Grosse (On the Number of Prime Numbers less than a Given Quantity), *Monatsberichte der Berliner Akademie*, 9 pages, November 1859. English translation by David R. Wilkins available at http://www.maths.tcd.ie/pub/HistMath/People/Riemann/Zeta/EZeta.pdf

[20] A. Riska, G. Ciardo and E. Smirni, Analytic modelling of load balancing policies for tasks with heavy-tailed distributions, *Proceedings of the workshop on software and performance 2000*, 147-157, 2000.

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

[22] D.L. Peterson and D.B. Adams, Fractal patterns in DASD I/O traffic. In *CMG Proceedings*, 1996.

[23] H. Sarfati and E. Bachmat, Performance analysis of TAGS systems, submitted.

[24] A. Scheller-Wolf, 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*, Vol. 51, 748-758, 2003.

[25] A. Scheller-Wolf, Further delay moment results for FIFO multiserver queues, *Queueing Systems*, Vol. 34, 387-400, 2000.

[26] A. Scheller-Wolf and K. Sigman, Delay moments for FIFO GI/GI/s queues, *Queueing Systems*, Vol. 25, 77-95, 1997.

[27] A. Scheller-Wolf and K. Sigman. New bounds for expected delay in FIFO GI/GI/c queues, *Queueing Systems*, 26:169-186, 1997.

[28] A. Scheller-Wolf and R. Vesilo. Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues, *Queueing Systems* 54:221-232, 2006.

[29] B. Schroeder and M. Harchol-Balter, Evaluation of task assignment policies for supercomputing servers: The case for load unbalancing and fairness, *Cluster computing*, vol. 7(2), 151-161, 2004. Preliminary version, *Proc. of the 9th IEEE Symposium on High Performance Distributed Computing (HPDC)* , 2000.

[30] A. Williams, M. Arlitt, C. Williamson, and K. Barker, Web workload characterization: Ten years after, in *Web Content Delivery, edited by X. Tang, J. Xu, S. Chanson*, Springer Verlag, 3-21, 2005.

[31] W. Whitt, The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution, *Queueing Systems*, Vol. 36, 71-87, 2000.

[32] Q. Zhang, A. Riska, W. Sun, E. Smirni, G. Ciardo, Workload-aware load balancing for clustered Web servers, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 16(3), 219-233, 2005.

[33] Q. Zhang, N. Mi, A. Riska, E. Smirni, Load Balancing for Performance Differentiation in Dual-Priority Clustered Servers, in *Proceedings of the 3rd International Conference on the Quantitative Evaluation of Systems (QEST)*, 385-394, Riverside, CA, 2006.

[34] Q. Zhang, N. Mi, A. Riska, E. Smirni, Performance-guided load (un)balancing under autocorrelated flows, *IEEE Transactions on parallel and distributed systems*, Vol. 19(5), 652-665, 2008.