Surprising Results on Task Assignment in Server Farms with High-Variability Workloads

Mor Harchol-Balter[*]
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA, USA
harchol@cs.cmu.edu

Alan Scheller-Wolf
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA, USA
awolf@andrew.cmu.edu

Andrew Young
Pittsburgh, PA, USA
Andrew.Richard.Young@gmail.com

ABSTRACT

This paper investigates the performance of task assignment policies for server farms, as the variability of job sizes (service demands) approaches infinity. Our results reveal that some common wisdoms regarding task assignment are flawed. The Size-Interval-Task-Assignment policy (SITA), which assigns each server a unique size range, was heretofore thought of by some as the panacea for dealing with high-variability job-size distributions. We show SITA to be inferior to the much simpler greedy policy, Least-Work-Left (LWL), for certain common job-size distributions, including many modal, hyperexponential, and Pareto distributions. We also define regimes where SITA's performance is superior, and prove simple closed-form bounds on its performance for the above-mentioned distributions.

Categories and Subject Descriptors

C.1.4 [**Processor Architectures**]: Parallel Architectures—*Distributed Architectures*; C.4 [**Performance of Systems**]: Design Studies; D.4.8 [**Operating Systems**]: Performance—*Modeling and Prediction*

General Terms

Performance,Design,Algorithms

1. INTRODUCTION

Server farms are ubiquitous, owing to their low cost (it is relatively cheap to pool together several slow servers) and their flexibility (it is easy to adjust capacity by adding and removing servers). One of the oldest and most fundamental questions arising in server farms is the question of which dispatching policy should be used for routing jobs to servers. This policy is known as the *task assignment policy*. One goal of the task assignment policy is to minimize mean response time, where response time is measured from when a job arrives until it completes.

It is well-known that empirical computer workloads such as Web file sizes, CPU process lifetimes, IP flow durations, and wireless call times have very high job-size variability, with job sizes fitting Pareto or other high-variance distributions [2, 8, 15, 26, 27]. This paper studies task assignment policies and considers the effect on response time as job size variability goes to infinity, while the mean job size stays fixed. To denote job-size variability, we use the squared coefficient of variation, $C^2 = \text{var}\,[X]/\mathbf{E}^2\,[X]$, where X represents the job size (service requirement).

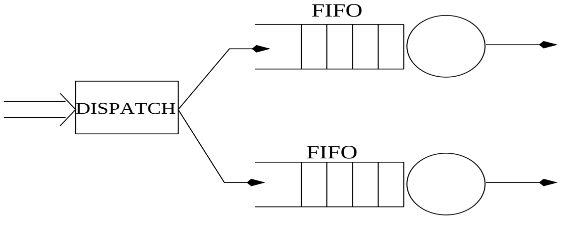


Figure 1: ***Server farm with 2 server hosts.***

Figure 1 depicts our server farm model, with $n = 2$ hosts. Jobs arrive according to a Poisson process with rate λ; the sizes of jobs are assumed to be i.i.d. from some general distribution. Each incoming job is immediately dispatched by a front-end router to one of the n server hosts. Jobs at a host are served in FCFS order, and preemption is not allowed. This model is common for supercomputing farms [13, 26], manufacturing systems [16, 5], data centers, IO systems, etc., where it is expensive to preempt jobs and thus even long jobs are typically run to completion.

For our server farm model, there are many common choices of task assignment policies. The *Round-Robin* policy assigns the first job to host 1, the second to host 2, the third to host 3, the ith to host i mod n plus 1, and so forth. The *Join-the-Shortest-Queue (JSQ)* policy assigns each incoming job to the host with the fewest *number* of jobs queued there. The *Least-Work-Left (LWL)* policy assigns each incoming job to the host with the least total work remaining. Here "work" is the sum of the remaining size of the job in service plus the sizes of all the jobs in the queue at the host. The *SITA (Size-Interval Task Assignment)* assigns a size-interval to each host, so that "short" jobs are sent to the first host, "medium-length" jobs are sent to the second host, and "long" jobs to the third, etc., where the cutoffs for differentiating size classes are chosen *optimally*, so as to minimize mean response time.

Importantly, the LWL policy is *equivalent* to the classical central FIFO queue, M/GI/n, where there are no queues at the hosts, and a free host simply takes the next job from the central queue. Specifically, under M/GI/n, jobs go to the same host as they would have under LWL and are served there at the same time as under LWL (see [13] for an inductive proof). The response times under M/GI/n and LWL are thus identical.

[*] Supported by NSF SMA/PDOS Grant CCR-0615262.

SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.
Copyright 2009 ACM 978-1-60558-511-6/09/06 ...$5.00.

While a great many papers have been written comparing the response time of different task assignment policies, e.g., [6, 7, 10, 13, 14, 20, 28, 29], all of these papers conclude (via numerical analysis, simulation, or approximation) that, for high job-size variability, the SITA policy is superior to all the other common policies above. The reason for the superiority of SITA task assignment lies in the fact that SITA allows short jobs their own "express-line," thereby giving them isolation from long jobs. Since most jobs are short jobs, the resulting mean response time is lowered.

There are several papers which specifically compare the performance of SITA to LWL [4, 7, 9, 12, 13, 14, 20, 28, 29]. All of these find that as job size variability is *increased*, SITA becomes far superior to LWL (for low C^2, SITA may be worse than LWL because not all servers are utilized; however, this behavior changes quickly as C^2 is increased).

Despite these comparisons showing that SITA outperforms LWL by orders of magnitude for high job size variability, a proof of this fact has never materialized. SITA itself is difficult to analyze, even for Poisson arrivals, because in general there is no closed-form expression for the optimal size cutoff, and hence the resulting response time. Furthermore, LWL cannot be analyzed exactly, since the M/G/n queue (equivalent to LWL) is in general only approximable. Thus, many of the existing results have used simulation to assert their claims, or have looked at phase-type job-size distributions, or heavy-traffic regimes.

In this paper, we show that the common wisdom about task assignment for high C^2 is wrong: We prove that SITA is not always superior to LWL as $C^2 \rightarrow \infty$; in fact SITA can be unboundedly worse than LWL. We show that both SITA and LWL can exhibit both convergent and divergent asymptotic behavior, depending on the load and job-size distribution. By convergent behavior, we mean that the mean response time approaches a constant as $C^2 \rightarrow \infty$ and by divergent behavior, we mean that the mean response time approaches infinity as $C^2 \rightarrow \infty$. For a server farm with n servers, system load ρ is defined as:

$$\rho = \lambda \mathbf{E}\left[X\right]$$

Note that $\rho = n$ corresponds to a fully loaded system. Some of our results require that $\rho < n - 1$. (It is known – see Section 2 – that if $\rho > n - 1$, LWL always diverges as $C^2 \rightarrow \infty$.)

Specifically, for each box in Table 1, we will illustrate an example of a class of distributions that satisfies that box. For example, looking at Box 4, we will show that there are examples of distributions where SITA diverges and LWL converges. Importantly, our examples are *not esoteric* in nature: We do not presume arcane distributions or assume very light or heavy load or a very high number of servers. To illustrate our points, it suffices to assume two server hosts only. However, in the case of the Pareto job-size distribution, we have extended our results to a general (finite) number of servers.

The distributions we use to illustrate examples in Table 1 are very common. Specifically, in Section 3, we show that the Bimodal can satisfy Boxes 2 and 4 for the 2-server system, depending on the choice of parameters, and in Section 4 we show that the Trimodal can satisfy Boxes 1 and 3. We then show in Section 5 that the traditional hyperexponential, H_2, can satisfy Boxes 2 and 4, while Section 6 shows that the 3-phase hyperexponential, H_3, can satisfy Boxes 1 and 3, again depending on the choice of parameters. The advantage of the hyperexponential job-size distribution is that it allows us to exactly analyze the performance of LWL with matrix-analytic methods [19], rather than just using bounds. Hence we can see exactly how LWL and SITA compare, including cases where they both diverge or both converge. We can cover all four boxes with either the modal distributions (Sections 3 and 4) or the Hyperexponential distributions (Sections 5 and 6). Finally, in Section 7, we consider the Bounded Pareto and Pareto job size distributions, which we find provide examples of Box 4, and also Box 3. This last result is most surprising, since SITA was specifically designed to work well under the high-variability Pareto distribution, and appears (via simulation, approximation, and numerical methods) to significantly out-perform LWL under Pareto and Bounded Pareto job size distributions. Our results show that there is however a cross-over point, at sufficiently high C^2, after which SITA diverges, while LWL might converge (Box 3) or diverge (Box 4) depending on the parameters of the Pareto. Section 2 explains the prior work in detail and, in particular, provides some explanation for *why* the above behaviors have not been observed until now.

	Convergent LWL (Section 4 & 6)	Divergent LWL (Section 3 & 5)
Convergent SITA	BOX 1	BOX 2
Divergent SITA	BOX 3	BOX 4

Table 1: We show that all four behaviors are common.

In addition to the above results, this paper also provides beautiful, simple, asymptotically-tight upper limits (as $C^2 \rightarrow \infty$) on the mean response time under SITA for the case where SITA converges. A subset of our upper bounds are shown in Table 2 below for the most common cases (e.g., H_2 with balanced branches), for server farms with two hosts. No results of this type exist in the prior literature. Table 2 illustrates these asymptotic limits for the case of the Bimodal, Trimodal and the hyperexponential distributions H_2 and H_3. As seen in Table 2, the results for the hyperexponential distributions parallel those for the modal distributions. Importantly, we see that the limiting behavior as $C^2 \rightarrow \infty$ depends only on the mean job size, $\mathbf{E}\left[X\right]$, and load ρ. Specifically, looking at the H_3 distribution, we recognize the limiting response time as that for a simple exponential job size distribution, where the mean of this limiting exponential equals the mean of the original hyperexponential.

	Convergent LWL	Divergent LWL
Convergent SITA (Modal Distributions)	(Trimodal) $\frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)}$ (for $\rho < 1$)	(Bimodal) $\frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\frac{\rho}{2})}$ (for $\rho < 2$)
Convergent SITA (Hyperexponential)	(H_3) $\frac{\mathbf{E}[X]}{1-\rho}$ (for $\rho < 1$)	(H_2) $\frac{\mathbf{E}[X]}{1-\frac{\rho}{2}}$ (for $\rho < 2$)

Table 2: Subset of typical asymptotic upper bounds on SITA mean response time proven herein.

It is reassuring to see, from Table 2, that when SITA converges, its response time can be quite good! Heretofore, there were no simple bounds on SITA's performance. Our results also indicate that when SITA diverges, LWL might converge. This too is good news since the LWL policy can be implemented as a central FIFO GI/GI/n queue, thereby obviating the need for known job sizes.

2. PRIOR WORK

The evaluation and comparison of task assignment policies is an ever-popular area of study, and there is a long list of papers on this topic. In this section, we restrict ourselves to papers that discuss either SITA or LWL, or both.

The SITA Policy

It is not clear where the idea of size-based task assignment originated, since it has been part of the common wisdom for a long time. Size-based splitting was used in the Cornell Theory Center [17], and is also mentioned in [5]. The SITA policy was formally introduced by Harchol-Balter et al. in [14], wherein it was found that, under high job-size variability (Bounded Pareto with low α), with appropriate cutoffs, mean response time under SITA is orders of magnitude lower than that under other common policies (LWL, JSQ, RANDOM, Round-Robin). A similar point was made for the TAGS algorithm (Task Assignment by Guessing Size), introduced by Harchol-Balter [13], which is similar to SITA but doesn't require knowing the size of the job. Harchol-Balter [13] finds that for job-size distributions with high variability and decreasing failure rate (again, the Bounded Pareto with low α), TAGS, like SITA, is superior to other common policies (LWL, JSQ, etc.). None of the above papers noticed that SITA could be worse than LWL under high job-size variability.

Since the introduction of the SITA policy, the SITA and TAGS algorithms have been studied in a long list of papers, all of which have touted the benefits of these algorithms under high job-size variability, but missed the fact that these policies could actually be worse than LWL under sufficiently high variability and non-heavy traffic. Thomas [29] analyzes TAGS via the Markovian process algebra PEVA and finds that TAGS performs well when job size variability is high. El-Taha and Maddah [9] analyze a variant of TAGS and prove that as $C^2 \to \infty$ this variant is superior to LWL under heavy traffic. Oida and Shinjo [20] show that SITA is superior to LWL under heavy-traffic using an integer program formulation. Ciardo et al. [7] apply SITA to web server farms, with cutoffs chosen to equalize the load, and find, via trace-driven simulation, that when job-size variability is high, the SITA policy is superior to LWL[1]. Tari et al. [28] consider a variant of SITA for heterogeneous hosts with different speeds and again find, via simulation, that SITA behaves well under high-variability job-size distributions, and Fu et al. [12] extend this result to allow jobs to be ordered by priority. Similar results are shown for a variant of TAGS by Broberg et al. [4]. Bachmat and Sarfati [1] develop a duality theory for the performance of SITA policies, allowing them to derive asymptotically-optimal cutoffs for SITA under a Bounded Pareto job-size distribution with infinite range. Feng et al. [10] prove the optimality of SITA with respect to mean response time among all policies which immediately dispatch jobs to hosts but don't know the status of the hosts (this does not include LWL). The SITA policy has received attention in many systems papers as well, e.g. [6] which discusses using SITA for web server farms or [26] which applies SITA to heavy-tailed supercomputing workloads.

The LWL Policy

The LWL policy is equivalent to the classical central-server FIFO queue, GI/GI/n as explained in Section 1. There are several key analytical papers which are concerned with the GI/GI/n under high job-size variability. None of these deal with SITA, or any task assignment policy (other than LWL). The papers most relevant to our work are those of Scheller-Wolf and Sigman [22, 24] which prove an upper bound on mean delay in a GI/GI/n system where this upper bound does not depend on any moment of service time higher than the $\frac{3}{2}$ moment, and particularly does not depend on the variance of job-size. The [24] result requires that system load ρ is less than $\lfloor n/2 \rfloor$, however [22] generalizes the result to allow for higher load, $\rho < n-1$. *This result ends up being key in our work, since we are able to show that for certain common job-size distributions (modal, H_n, Pareto, etc.), we can raise the variability unboundedly ($C^2 \to \infty$) while keeping the $\frac{3}{2}$ moment of the job-size distribution below a fixed value, hence bounding mean delay for LWL.* The converse of the [22, 24] results was presented by Scheller-Wolf and Vesilo in 2006 [25], for a large class of distributions including those in this paper. It is known that if $\rho > n-1$, then the GI/GI/n diverges as $C^2 \to \infty$ [23], hence LWL diverges too.

Whitt [30] and Foss and Korshunov [11] consider a GI/GI/2 and study the delay tail behavior when job size is subexponential. They find that for low load, the delay tail grows like the tail of the equilibrium distribution squared, whereas for high load the delay tail grows like the tail of the equilibrium distribution. These results are consistent with [24] and [25]. The M/GI/2 with heterogeneous servers has also been looked at by Boxma et al. [3], who study how high variability in the job-size distribution at one of the servers affects the other. Finally, while all of the above papers involve analytic solutions, the M/GI/n with high job-size variability has also been studied via simulation by Psounis et al. [21]. Here the authors develop an M/GI/n approximation based on two moments of the job size distribution and use that approximation to estimate the optimal number of servers.

Summary of Prior Work & Comparison with this Paper

In summary, although there have been many papers studying SITA, and quite a few comparing SITA with LWL, all have focused on the benefits of SITA over LWL for high-variability job size distributions, and *none* have noticed that SITA can be worse than LWL at high variability. By contrast, in this paper we prove that for certain common job size distributions, SITA can be worse than LWL under high variability, and in fact there are situations where SITA diverges as $C^2 \to \infty$, whereas LWL converges to a finite bound.

There are several potential reasons why these results have previously gone unnoticed. First, several of the papers comparing SITA with LWL concentrate on heavy-traffic, whereas our studies concentrate on more moderate load ($\rho < n-1$). Second, many of the papers above rely on simulation to evaluate SITA and LWL. However, simulation becomes problematic at high C^2 values. None of the above papers consider the limiting behavior of SITA as $C^2 \to \infty$. Finally, and perhaps most important, there is somewhat of a disconnect between communities like SIGMETRICS, which regularly study task assignment policies, and communities like INFORMS, which look at GI/GI/n queues. It is the *merging* of results from these two communities that inspired the idea for this paper.

3. DIVERGENT LWL VIA BIMODAL (BOXES 2 & 4)

This section will illustrate that for a class of Bimodal job-size

[1] The paper refers to SITA as EquiLoad and uses a superior variant of LWL.

distributions, mean response time under LWL diverges as $C^2 \to \infty$ (see Section 3.1), whereas mean response time under SITA may converge or diverge (see Section 3.2), depending on the parameters.

The Bimodal distribution with parameters a, b, and p is defined by the following random variable:

$$X = \begin{cases} a & \text{w.p. } p_a = p \\ b & \text{w.p. } p_b = 1-p \end{cases}$$

We further characterize the distribution as a *Q-Bimodal* by specifying a weight Q, $0 < Q < 1$, such that:

$$pa = Q\mathbf{E}[X] \quad \text{and} \quad (1-p)\,b = (1-Q)\,\mathbf{E}[X] \qquad (1)$$

We will show in Theorem 1 that a 2-server system serving a Q-Bimodal workload using an LWL policy has unbounded mean response time as $C^2 \to \infty$. In Theorem 2 we show that when $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$ mean response time under SITA is bounded from above by:

$$\mathbf{E}[T]^{\text{SITA}} \leq \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]\left(1 - 2Q\,(1-Q)\,\rho\right)}{2\left(1 - \rho + Q\,(1-Q)\,\rho^2\right)}$$

Outside this region, SITA's response time diverges as $C^2 \to \infty$. When $Q = \frac{1}{2}$ (balanced branches), $\mathbf{E}[T]^{\text{SITA}} = \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2\left(1-\frac{\rho}{2}\right)}$.

3.1 Divergent LWL

THEOREM 1. *For a 2-server system with Q-Bimodal job-size distribution under LWL,* $\mathbf{E}[T]^{LWL} \to \infty$ *as* $C^2 \to \infty$.

PROOF. Lemma 1 below guarantees that we can find a unique Bimodal distribution with parameters $a < b$ and p for any given $\mathbf{E}[X]$, $C^2 > 0$, and Q. Lemma 2 below shows that, for fixed $\mathbf{E}[X]$ and Q, as $C^2 \to \infty$, $p \to 1$, $a \to \mathbf{E}[X]$, and $b \to \infty$. Furthermore:

$$\begin{aligned} \mathbf{E}\left[X^{3/2}\right] &= pa^{3/2} + (1-p)b^{3/2} \\ &= Q\mathbf{E}[X]\sqrt{a} + (1-Q)\mathbf{E}[X] \cdot \sqrt{b} \end{aligned}$$

As $C^2 \to \infty$, we see from the above that $\mathbf{E}\left[X^{\frac{3}{2}}\right] \to \infty$, since $\sqrt{b} \to \infty$, and all the other terms are constant. Scheller-Wolf and Vesilo [25] proved that for most distributions, including all distributions in this paper, $\mathbf{E}[T]^{\text{LWL}} \to \infty$ if $\mathbf{E}\left[X^{\frac{3}{2}}\right] \to \infty$. □

LEMMA 1. *For any* $\mathbf{E}[X]$, $C^2 > 0$, *and* Q, *we can find unique parameters* $a < b$ *and* p *for a Q-Bimodal:*

$$\begin{aligned} p &= \frac{C^2 + 2Q + C\sqrt{C^2 + 4Q\,(1-Q)}}{2\,(C^2+1)} \\ a &= Q\mathbf{E}[X]\,/p \\ b &= (1-Q)\mathbf{E}[X]\,/(1-p) \end{aligned}$$

PROOF.

$$\begin{aligned} \mathbf{E}[X] &= pa + (1-p)\,b \\ \mathbf{E}\left[X^2\right] &= pa^2 + (1-p)\,b^2 = \left(C^2+1\right)\mathbf{E}^2[X] \\ C^2 + 1 &= \frac{pa^2 + (1-p)b^2}{\mathbf{E}^2[X]} = \frac{Q^2}{p} + \frac{(1-Q)^2}{1-p} \\ p &= \frac{C^2 + 2Q \pm C\sqrt{C^2 + 4Q(1-Q)}}{2(C^2+1)} \end{aligned}$$

Taking the positive root, we see that $a < b \Leftrightarrow \frac{Q\mathbf{E}[X]}{p} < \frac{(1-Q)\mathbf{E}[X]}{1-p} \Leftrightarrow \frac{p}{Q} > \frac{1-p}{1-Q}$ which we can verify from the above equation. Also, since all the terms of p are positive, $p > 0$. We verify $1 - p = \frac{C^2 + 2(1-Q) - C\sqrt{C^2+4Q(1-Q)}}{2(C^2+1)} > 0$ since $\left(C^2 + 2\,(1-Q)\right)^2 > C^2\left(C^2 + 4Q(1-Q)\right)$. Thus, $0 < p < 1$. □

LEMMA 2. *For the Bimodal with fixed* $\mathbf{E}[X]$ *and* Q, *as* $C^2 \to \infty$, $p \to 1$, $a \to Q\mathbf{E}[X]$, *and* $b \to \infty$.

PROOF. Follows immediately from Lemma 1, after dividing the numerator and denominator of p by C^2 and taking limits as $C^2 \to \infty$. □

3.2 Convergent/Divergent SITA

While we saw above that the Q-Bimodal job size distribution results in divergent mean response time for Least-Work-Left, we will now show that depending on Q, we can either get convergent or divergent behavior for SITA.

THEOREM 2. *For a 2-server system with Q-Bimodal job-size distribution with fixed* $\mathbf{E}[X]$ *and fixed* Q *where* $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$, *mean response time under SITA (for all* C^2*) is bounded from above by*

$$\mathbf{E}[T]^{SITA} \quad \leq \quad \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]\left(1 - 2Q\,(1-Q)\,\rho\right)}{2\left(1 - \rho + Q\,(1-Q)\,\rho^2\right)}$$

When $Q = \frac{1}{2}$, $\mathbf{E}[T]^{SITA} = \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2\left(1-\frac{\rho}{2}\right)}$.

When $\left|Q - \frac{1}{2}\right| \geq \frac{2-\rho}{2\rho}$, *mean response time under SITA diverges as* $C^2 \to \infty$.

PROOF. If $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$ then we can always split the jobs of size a and b into separate servers, without overloading either server, allowing the mean response time under SITA to converge. To see this, observe that:

$$\begin{aligned} \rho_a &= \lambda pa = \lambda Q\mathbf{E}[X] = Q\rho \\ \rho_b &= \lambda\,(1-p)\,b = \lambda\,(1-Q)\,\mathbf{E}[X] = (1-Q)\,\rho \end{aligned}$$

Since $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$ implies $Q < \frac{1}{\rho}$, the small-job server is not overloaded. Since the constraint on Q also implies $1 - Q < \frac{1}{\rho}$, the large job server is not overloaded either.

Note that by Lemma 1, given $\mathbf{E}[X]$ and C^2, we can always find a Q-Bimodal for any Q.

Given that we can separate the jobs with sizes a and b, the mean response time follows by conditioning and the Pollaczek-Khinchin (P-K) formula [18], where $\lambda_s = p\lambda$ and $\lambda_l = (1-p)\,\lambda$ are the arrival rates at the small-job and the large-job servers respectively:

$$\begin{aligned} \mathbf{E}[T]^{\text{SITA}} &\leq \mathbf{E}[X] + \frac{\lambda_s pa^2}{2\,(1-\lambda_s a)} + \frac{\lambda_l\,(1-p)\,b^2}{2\,(1-\lambda_l b)} \\ &= \mathbf{E}[X] + \frac{\lambda Q^2 \mathbf{E}^2[X]}{2\,(1-Q\rho)} + \frac{\lambda\,(1-Q)^2\,\mathbf{E}^2[X]}{2\,(1-(1-Q)\,\rho)} \\ &= \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]\left(1 - 2Q\,(1-Q)\,\rho\right)}{2\left(1 - \rho + Q\,(1-Q)\,\rho^2\right)} \end{aligned}$$

The above is an upper bound on SITA's performance under the optimal partition, which may not necessarily separate jobs of size a and b. When $Q = \frac{1}{2}$, $\mathbf{E}[T]^{\text{SITA}} = \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2\left(1-\frac{\rho}{2}\right)}$.

On the other hand, if $\left|Q - \frac{1}{2}\right| \geq \frac{2-\rho}{2\rho}$, it is not possible to separate the small jobs from the large jobs without overloading a server. Therefore, one server must have a mix of small and large jobs. Suppose that we fix ϵ such that $\epsilon > Q - \frac{1}{\rho} \geq 0$. Now the large-job server must run at least ϵ a-size jobs and all the b-size jobs. The contribution to mean delay from the large-job server is simply the fraction of jobs that go to the large-job server multiplied by delay at the large-job server. This comes out to:

$$\lambda\frac{\epsilon a^3 + (1-p)\,\epsilon a^2 + (1-Q)^2\,\mathbf{E}^2[X] + \epsilon a\,(1-Q)\,\mathbf{E}[X]\,b}{2\left(1-\lambda\left(\epsilon a + (1-Q)\,\mathbf{E}[X]\right)\right)}$$

The numerator of this term shows that, if $\epsilon = 0$ (when jobs can be split between the servers based on size without overloading either server), the contribution to mean delay of the large-job server is bounded. However, if $\epsilon > 0$, the presence of ϵb causes divergence in mean delay. A similar argument holds if $\epsilon > 1 - Q - \frac{1}{\rho} \geq 0$, overloading the small-job server. □

Note that $Q = \frac{1}{2}$ is always within the convergent range and provides an example of convergent SITA for all $\rho < 2$.

4. CONVERGENT LWL VIA TRIMODAL (BOXES 1 & 3)

This section will illustrate that for a class of Trimodal distributions, mean response time under LWL always converges as $C^2 \rightarrow \infty$ (Section 4.1), while that under SITA may converge or diverge (Section 4.2). The Trimodal distribution with parameters $a < b < c$, and p_a, p_b, and p_c is defined by the following random variable:

$$X = \begin{cases} a & \text{w.p. } p_a \\ b & \text{w.p. } p_b \\ c & \text{w.p. } p_c \end{cases}$$

We further specify our Trimodal distribution (which we refer to as a *k-Trimodal*) with the following relationships:

$$p_a = 1 - p_b - p_c \qquad p_b = b^{-\frac{3}{2}} \qquad p_c = b^{-3k}$$
$$c = b^{2k} \qquad k > \tfrac{1}{2}$$

The free parameters are now a, b, and k. The structure of this distribution ensures that $\mathbf{E}\left[X^{\frac{3}{2}}\right] < \infty$ for all C^2, thus guaranteeing convergence of LWL as $C^2 \rightarrow \infty$. Specifically, observe that the contribution of the medium and large jobs to $\mathbf{E}\left[X^{\frac{3}{2}}\right]$ is $p_b b^{\frac{3}{2}} + p_c c^{\frac{3}{2}} = 2$. Furthermore, we will show that, depending on the value of the parameter k compares to $\frac{3}{2}$, we can get either convergence or divergence of SITA (see Equation (3) below).

In Theorem 3, we show that the mean response time under LWL for jobs drawn from a k-Trimodal distribution is *bounded* as $C^2 \rightarrow \infty$ for $\rho < 1$. In Theorem 4, we show that, under SITA, mean response time can converge or diverge: If the jobs are drawn from a k-Trimodal distribution with $k < \frac{3}{2}$ and $\rho < 1$ then $\mathbf{E}[T]^{\text{SITA}} = \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)}$, while if $k > \frac{3}{2}$ then mean response time is unbounded under SITA.

4.1 Convergent LWL

THEOREM 3. *For a 2-server system with k-Trimodal job-size distribution and $\rho < 1$, expected response time under LWL is bounded as $C^2 \rightarrow \infty$, and this bound is provided in Equation 2.*

PROOF. Lemma 3 below guarantees that we can find a k-Trimodal distribution with parameters a, b, c, and p_a, p_b, and p_c for any $\mathbf{E}[X]$, large enough C^2, and k. Lemma 4 below shows that, as $C^2 \rightarrow \infty$, $a \rightarrow \mathbf{E}[X]$ and $p_a \rightarrow 1$. Then

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] = p_a a^{\frac{3}{2}} + p_b b^{\frac{3}{2}} + p_c c^{\frac{3}{2}} \xrightarrow[C^2\rightarrow\infty]{} (\mathbf{E}[X])^{\frac{3}{2}} + 2$$

Since $\mathbf{E}\left[X^{\frac{3}{2}}\right] < \infty$ is bounded as $C^2 \rightarrow \infty$ and $\rho < \lfloor\frac{n}{2}\rfloor$, [24] tells us that $\mathbf{E}[T]^{\text{LWL}}$ is bounded as follows:

$$\begin{aligned}\mathbf{E}[T]^{\text{LWL}} \leq\ & \frac{\lambda^2\left(\lfloor\frac{n}{2}\rfloor^{\frac{3}{2}}\left(1+\frac{2\sqrt{3}}{9}\right)E\left[X^{\frac{3}{2}}\right]E\left[A^{\frac{3}{2}}\right] + E^2\left[X^{\frac{3}{2}}\right]\right)}{2\left(\lfloor\frac{n}{2}\rfloor - \rho\right)^2} \\ & + \frac{\lambda\rho\left(2-\frac{\rho}{\lfloor\frac{n}{2}\rfloor}\right)\sigma_A^2}{2\left(\lfloor\frac{n}{2}\rfloor-\rho\right)} + \mathbf{E}[X] \qquad (2)\end{aligned}$$

where A is a random variable representing inter-arrival times. □

LEMMA 3. *For any $\mathbf{E}[X]$, there exists C^* such that, for all $C^2 > C^*$, and $k > \frac{1}{2}$, there exists a k-Trimodal distribution with unique parameters $0 < a < b < c$ and probabilities $0 < p_a, p_b, p_c < 1$.*

PROOF. The proof demonstrates that b is monotonic in C^2 above some C^*, and so a unique value of b can be found for every such C^2, with the rest of the relationships ensuing from the definition of a k-Trimodal distribution. See Appendix for details. □

LEMMA 4. *For a k-Trimodal distribution, as $C^2 \rightarrow \infty$, we have $p_a \rightarrow 1$, $p_b \rightarrow 0$, $p_c \rightarrow 0$, $a \rightarrow \mathbf{E}[X]$, $b \rightarrow \infty$, and $c \rightarrow \infty$.*

PROOF. From the proof of Lemma 3, we know that $C^2 \rightarrow \infty$ implies $b \rightarrow \infty$. Also, from the same proof, $a \rightarrow \mathbf{E}[X]$ as $b \rightarrow \infty$. The limits for p_a, p_b, p_c, and c follow directly from the definition of the k-Trimodal as $b \rightarrow \infty$. □

4.2 Convergent/Divergent SITA

THEOREM 4. *For a 2-server system with k-Trimodal job-size distribution with mean $\mathbf{E}[X]$, parameter k, and $\rho < 1$, a SITA policy that sends all small jobs (the a's) to one server and all other jobs to the other has the following mean response time as $C^2 \rightarrow \infty$:*

$$\mathbf{E}[T]^{\textit{SITA}} \xrightarrow[C^2\rightarrow\infty]{} \begin{cases} \infty & k > \frac{3}{2} \\ \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)} + \frac{\lambda}{2} & k = \frac{3}{2} \\ \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)} & \frac{1}{2} < k < \frac{3}{2} \end{cases}$$

Any other SITA cutoff performs worse as $C^2 \rightarrow \infty$.

PROOF. We will use $\mathbf{E}\left[X_i^j\right]$ to denote the jth moment of job size on server i, ρ_i to denote the load at server i, p_i to denote the fraction of jobs assigned to server i, and W_i to denote the waiting time (delay) at server i, $i \in \{s, l\}$, where s denotes the small-job server and l denotes the large-job server.

We define q to be the fraction of medium (b) jobs sent to server s. The fraction of jobs sent to each server is given by:

$$\begin{aligned} p_s &= p_a + q \cdot p_b = 1 - (1-q)\,b^{-\frac{3}{2}} - b^{-3k} \\ p_l &= 1 - p_s = (1-q)\,b^{-\frac{3}{2}} + b^{-3k} \\ \mathbf{E}\left[X_s^j\right] &= \frac{1}{p_s}\left(a^j \cdot p_a + b^j \cdot q \cdot p_b\right) \\ \mathbf{E}\left[X_l^j\right] &= \frac{1}{p_l}\left(b^j \cdot (1-q) \cdot b^{-\frac{3}{2}} + c^j \cdot p_c\right) \\ &= \frac{1}{p_l}\left((1-q) \cdot b^{j-\frac{3}{2}} + b^{(2j-3)k}\right) \end{aligned}$$

As $C^2 \rightarrow \infty$, Lemma 3 provides that $b \rightarrow \infty$, and thus, $a \rightarrow \mathbf{E}[X]$. Many terms in $p_s \cdot \mathbf{E}[W_s]$ and $p_l \cdot \mathbf{E}[W_l]$ categorically go to zero. The remaining expressions are:

$$\begin{aligned} p_s \cdot \mathbf{E}[W_s] &\sim \frac{\lambda\left(\mathbf{E}^2[X] + qb^{\frac{1}{2}}\right)}{2(1-\rho)} \quad (\text{as } b \rightarrow \infty) \\ p_l \cdot \mathbf{E}[W_l] &\sim \frac{\lambda(1-q)\,b^{k-\frac{3}{2}}}{2} \quad (\text{as } b \rightarrow \infty) \qquad (3) \end{aligned}$$

The prior two equations demonstrate that as $C^2 \to \infty$, $p_s \cdot \mathbf{E}[W_s] + p_l \cdot \mathbf{E}[W_l] < \infty$ only when $k < \frac{3}{2}$ and $qb^{\frac{1}{2}} \to r < \infty$ (which implies $q \to 0$ and $qb^{k-\frac{3}{2}} \to 0$ since $b \to \infty$). Any system with $qb^{\frac{1}{2}} \to r > 0$ is dominated in the limit by $q = 0$. Therefore, response time is minimized when all medium jobs are assigned to server l as $b \to \infty$. In a similar manner, we can prove that response time is minimized when all small jobs are sent to s, but we have omitted the proof for lack of space.

$$\begin{aligned}\mathbf{E}[T]^{\text{SITA}} &= \mathbf{E}[X] + p_s \cdot \mathbf{E}[W_s] + p_l \cdot \mathbf{E}[W_l] \\ &\xrightarrow[C^2 \to \infty]{} \begin{cases} \infty & k > \frac{3}{2} \\ \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)} + \frac{\lambda}{2} & k = \frac{3}{2} \\ \frac{\mathbf{E}[X]}{2} + \frac{\mathbf{E}[X]}{2(1-\rho)} & \frac{1}{2} < k < \frac{3}{2} \end{cases}\end{aligned}$$

□

Theorem 4 shows that, under a SITA policy, a k-Trimodal job-size distribution with $\frac{1}{2} < k \leq \frac{3}{2}$ has finite delay as $C^2 \to \infty$, assuming no b (or c) jobs are sent to the small server. Intuitively, in this case, the small jobs receive perfect isolation, and, with k in this range, the variability in the service times on the large-job server, combined with the small probability of a large job, means that the contribution to mean response time from the large-job server vanishes. On the other hand, under a SITA policy, a Trimodal distribution with $k > \frac{3}{2}$ has unbounded delay as $C^2 \to \infty$, even if SITA gives the small jobs perfect isolation. This results from the larger variability in job sizes on the large-job server.

5. DIVERGENT LWL VIA H_2 (BOXES 2&4)

In Sections 3 and 4, we used the Q-Bimodal and k-Trimodal job size distributions to illustrate the behavior of SITA and LWL for the four boxes in Table 1. However, there are more questions to be answered, because, while we have analytic expressions for the mean response time under SITA, we only have a loose bound on LWL's mean response time. Thus in Box 1, where both LWL and SITA converge, we don't know whether LWL or SITA is superior. Likewise, for Box 4, where LWL and SITA both diverge, we don't know whether LWL or SITA diverges more quickly. To answer these remaining open questions, we turn to the hyperexponential job size distribution, which allows us to use matrix-analytic methods to evaluate the response time of LWL.

Before we can begin to evaluate performance, we first need to demonstrate that the four different behaviors shown in Table 1 can be obtained under hyperexponential job size distributions. We do so in this section and the next. Many of these arguments follow similar logic to the Q-Bimodal and k-Trimodal reasoning.

In this section, we will use a 2-phase hyperexponential distribution to illustrate the case of divergent LWL, where SITA will either converge or diverge, depending on parameters (Boxes 2 and 4). In (Section 6), we will use a 3-phase hyperexponential to illustrate the case of convergent LWL, where SITA will either converge or diverge (Boxes 1 and 3), again depending on parameters. Along the way, we will again prove beautiful asymptotic limits on the mean response time under SITA, when SITA converges.

The 2-phase Hyperexponential distribution (H_2) with parameters μ_a, μ_b, and p is defined by the following random variable:

$$X \sim \begin{cases} \text{Exp}(\mu_a) & \text{w.p. } p_a = p \\ \text{Exp}(\mu_b) & \text{w.p. } p_b = 1 - p \end{cases}$$

We again define a further parameter Q, $0 < Q < 1$, as the mean weighting, such that $\frac{p_a}{\mu_a} = Q\mathbf{E}[X]$ and $\frac{p_b}{\mu_b} = (1-Q)\mathbf{E}[X]$. If $Q = \frac{1}{2}$ then the two exponentials are of equal importance. We refer to the H_2 with additional parameter Q as the Q-H_2 distribution.

We show in Theorem 5 that a 2-server system serving job sizes from the Q-H_2 distribution using a LWL policy has unbounded expected response time as $C^2 \to \infty$. In Theorem 6 we show that, by contrast, the mean response time under SITA for the Q-H_2 might converge or diverge: when it converges, it converges to

$$\mathbf{E}[T]^{\text{SITA}} \to \frac{\mathbf{E}[X](1 - 2Q(1-Q)\rho)}{1 - \rho + Q(1-Q)\rho^2}$$

When $Q = \frac{1}{2}$, $\mathbf{E}[T]^{\text{SITA}} \to \frac{\mathbf{E}[X]}{1-\frac{\rho}{2}}$.

5.1 Divergent LWL

THEOREM 5. *For a 2-server system with Q-H_2 job-size distribution under LWL, $\mathbf{E}[T]^{\text{LWL}} \to \infty$ as $C^2 \to \infty$.*

PROOF. Lemma 5 below guarantees that we can find a Q-H_2 distribution with parameters μ_a, μ_b, and p for any $\mathbf{E}[X]$, C^2, and Q. Lemma 6 provides that $\mu_b \to 0$ as $C^2 \to \infty$. The definition of the Q-H_2 dictates that $\frac{p_a}{\mu_a} = Q\mathbf{E}[X]$ and $\frac{p_b}{\mu_b} = (1-Q)\mathbf{E}[X]$. Lemma 7 provides that:

$$\begin{aligned}\mathbf{E}\left[X^{\frac{3}{2}}\right] &= \sum_i p_i \frac{3\sqrt{\pi}}{8\mu_i^{\frac{3}{2}}} \\ &= \frac{3\sqrt{\pi}\mathbf{E}[X]}{8}\left(\frac{Q}{\mu_a^{\frac{1}{2}}} + \frac{1-Q}{\mu_b^{\frac{1}{2}}}\right) \xrightarrow[C^2 \to \infty]{} \infty\end{aligned}$$

Applying [25], $\mathbf{E}[T]^{\text{LWL}} \to \infty$ since $\mathbf{E}\left[X^{\frac{3}{2}}\right] \to \infty$. □

LEMMA 5. *For any $\mathbf{E}[X]$, $C^2 > 1$, and Q, we can find unique parameters μ_a, μ_b, and p, where $\mu_a > \mu_b$ (μ_a is the service rate for the small jobs), for a Q-H_2 distribution which satisfy:*

$$\begin{aligned}p &= \frac{C^2 + 4Q - 1 + \sqrt{(C^2-1)^2 + 8(C^2-1)Q(1-Q)}}{2(C^2+1)} \\ \mu_a &= \frac{p}{Q\mathbf{E}[X]} \\ \mu_b &= \frac{1-p}{(1-Q)\mathbf{E}[X]}\end{aligned}$$

PROOF. See Appendix. □

COROLLARY 1. *When $Q = \frac{1}{2}$,*

$$\begin{aligned}(p_a, p_b) &= \left(\frac{1 + \sqrt{\frac{C^2-1}{C^2+1}}}{2}, \frac{1 - \sqrt{\frac{C^2-1}{C^2+1}}}{2}\right) \\ (\mu_a, \mu_b) &= \left(\frac{1 + \sqrt{\frac{C^2-1}{C^2+1}}}{\mathbf{E}[X]}, \frac{1 - \sqrt{\frac{C^2-1}{C^2+1}}}{\mathbf{E}[X]}\right)\end{aligned}$$

PROOF. By substitution. □

LEMMA 6. *When $\mathbf{E}[X]$ and Q are constant, as $C^2 \to \infty$, $\mu_a \to \frac{1}{Q\mathbf{E}[X]}$, $\mu_b \to 0$, $p_a \to 1$, and $p_b \to 0$.*

PROOF. See Appendix. □

LEMMA 7. *For the hyperexponential distribution,*

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] = \frac{3\sqrt{\pi}}{8}\sum_i \frac{p_i}{\mu_i^{\frac{3}{2}}}$$

PROOF.

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] = \sum_i p_i \int_0^\infty x^{\frac{3}{2}} \mu_i e^{-\mu_i x} dx$$

Evaluating the integral (integrating twice by parts and once substituting $y = \left(\frac{x}{2\mu}\right)^{\frac{1}{2}}$) gives:

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] = \frac{3\sqrt{\pi}}{8} \sum_i \frac{p_i}{\mu_i^{\frac{3}{2}}}$$

□

5.2 Convergent/Divergent SITA

We now investigate convergent and divergent SITA behavior. Theorem 6 below corresponds to Theorem 2 for the Q-Bimodal distribution, with one difference: Whereas in the Q-Bimodal distribution, any cutoff between a and b was stable, as long as neither server was overloaded by just jobs of size a or just jobs of size b, in the case of a Q-H_2 distribution, the cutoff ψ must be specified explicitly and depends on C^2.

THEOREM 6. *Given a 2-server system with Q-H_2 job-size distribution with fixed mean $\mathbf{E}[X]$ and fixed parameter Q such that $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$, there exists a ψ, which is a function of C^2, such that, as $C^2 \to \infty$, a SITA policy with cutoff ψ yields mean response time:*

$$\mathbf{E}[T]^{\textit{SITA}} \to \frac{\mathbf{E}[X]\left(1 - 2Q\left(1-Q\right)\rho\right)}{1 - \rho + Q\left(1-Q\right)\rho^2}$$

When $Q = \frac{1}{2}$, $\mathbf{E}[T]^{\textit{SITA}} \to \frac{\mathbf{E}[X]}{1-\frac{\rho}{2}}$. When $\left|Q - \frac{1}{2}\right| \geq \frac{2-\rho}{2\rho}$, $\mathbf{E}[T]^{\textit{SITA}} \to \infty$ as $C^2 \to \infty$.

PROOF. Typically, under SITA, with an H_n hyperexponential job-size distribution (with parameters $p_a, \ldots, p_n, \mu_a, \ldots, \mu_n$, $\mu_i > \mu_{i+1}$), the small-job server sees jobs drawn from every branch of the hyperexponential, as does the large-job server. However, this logic does not necessarily hold as $C^2 \to \infty$. We say that the hyperexponential job size distribution "separates in the limit" as $C^2 \to \infty$ at a cutoff $\psi\left(C^2\right)$ into small jobs and large jobs if:

1. The arrival rate at the small job server converges to $p_a\lambda$, and the mean and second moment of job sizes at the small job server converges to the mean and second moment of job sizes from an $\mathrm{Exp}(\mu_a)$.

2. The arrival rate and contribution to the mean and second moment at the large-job server of jobs drawn from the $\mathrm{Exp}(\mu_a)$ branch goes to 0.

We require convergence in the first two moments to guarantee no effect on delay in the P-K formula.

Suppose that a cutoff $\psi\left(C^2\right)$ could be identified that achieved separation in the limit for the H_2 distribution. All jobs sent to the small-job server would be drawn from $\mathrm{Exp}(\mu_a)$, and all jobs sent to the large-job server would be drawn from $\mathrm{Exp}(\mu_b)$. Thus, p_s, the proportion of jobs sent to the small server, converges to p_a, and, likewise, p_l converges to p_b. Then, since for an exponential distribution mean response time is given by $\mathbf{E}[T]^{\text{Exp}} = \frac{\mathbf{E}[X]}{1-\rho}$, mean response time for the system is:

$$\mathbf{E}[T]^{\text{SITA}} \quad \to \quad p_a \frac{\mathbf{E}[X_s]}{1-\rho_s} + p_b \frac{\mathbf{E}[X_l]}{1-\rho_l}$$

Substituting $\mathbf{E}[X_s] = \frac{1}{\mu_a} = \frac{1}{p_a} Q \mathbf{E}[X]$, $\mathbf{E}[X_l] = \frac{1}{\mu_b} = \frac{1}{p_b}\left(1-Q\right)\mathbf{E}[X]$, $\lambda_s = \lambda p_a$, $\rho_s = Q\rho$, $\lambda_l = \lambda p_b$, and $\rho_l = \left(1-Q\right)\rho$:

$$\begin{aligned} \mathbf{E}[T]^{\text{SITA}} \quad &\to \quad \frac{Q\mathbf{E}[X]}{1 - Q\rho} + \frac{\left(1-Q\right)\mathbf{E}[X]}{1 - \left(1-Q\right)\rho} \\ &\to \quad \frac{\mathbf{E}[X]\left(1 - 2Q\left(1-Q\right)\rho\right)}{1 - \rho + Q\left(1-Q\right)\rho^2} \end{aligned}$$

Substituting $Q = \frac{1}{2}$ yields $\mathbf{E}[T]^{\text{SITA}} \to \frac{\mathbf{E}[X]}{1-\frac{\rho}{2}}$.

The load on the small-job server is $\lambda Q \mathbf{E}[X]$, and the load on the large-job server is $\lambda\left(1-Q\right)\mathbf{E}[X]$. Hence, if we are to take advantage of any separability, we require $\left|Q - \frac{1}{2}\right| < \frac{2-\rho}{2\rho}$ so that neither server has load greater than or equal to 1.

It remains to show that such a $\psi\left(C^2\right)$ can be found for the H_2 distribution. Let X_a denote the first branch of the H_2 distribution and X_b the second: $X_a \sim \mathrm{Exp}(\mu_a)$ and $X_b \sim \mathrm{Exp}(\mu_b)$. In order for an H_2 to be separable in the limit as $C^2 \to \infty$, a cutoff ψ must have the property that the following six quantities go to 0, where I is an indicator random variable:

1. $p_a \mathbf{E}\left[I_{X_a > \psi}\right] = p_a e^{-\mu_a \psi}$
   L'Hôpital's rule guarantees that for any polynomial $P\left(\psi\right)$, $P\left(\psi\right) e^{-\mu_a \psi} \to 0$ as $C^2 \to \infty$ if $\psi \to \infty$ as $C^2 \to \infty$.

2. $p_a \mathbf{E}\left[X_a \cdot I_{X_a > \psi}\right] = p_a \left(\psi + \frac{1}{\mu_a}\right) e^{-\mu_a \psi}$
   Same condition as 1.

3. $p_a \mathbf{E}\left[X_a^2 \cdot I_{X_a > \psi}\right] = p_a \left(\psi^2 + \frac{2\psi}{\mu_a} + \frac{2}{\mu_a^2}\right) e^{-\mu_a \psi}$
   Same condition as 1.

4. $p_b \mathbf{E}\left[I_{X_b < \psi}\right] = p_b \left(1 - e^{-\mu_b \psi}\right)$
   Requires $p_b \to 0$ or $\mu_b \psi \to 0$ as $C^2 \to \infty$.

5. $p_b \mathbf{E}\left[X_b \cdot I_{X_b < \psi}\right] = p_b \left(\frac{1}{\mu_b} - \left(\psi + \frac{1}{\mu_b}\right) e^{-\mu_b \psi}\right)$
   Substituting $p_b = \left(1-Q\right)\mathbf{E}[X]\mu_b$ gives $\left(1-Q\right)\mathbf{E}[X]\left(1 - e^{-\mu_b\psi} - \mu_b \psi e^{-\mu_b \psi}\right)$. Thus, it suffices that $\mu_b \psi \to 0$ as $C^2 \to \infty$.

6. $p_b \mathbf{E}\left[X_b^2 \cdot I_{X_b < \psi}\right] = p_b \left(\frac{2}{\mu_b^2} - \left(\psi^2 + \frac{2\psi}{\mu_b} + \frac{2}{\mu_b^2}\right) e^{-\mu_b \psi}\right)$
   Substituting $p_b \sim \mu_b$ and the Taylor series expansion of $e^{-\mu_b \psi} = 1 - \mu_b \psi + \frac{\left(\mu_b \psi\right)^2}{2} + \sum_{i=3}^{\infty} \left(-1\right)^i \frac{\left(\mu_b\psi\right)^i}{i!}$ gives:

$$\begin{aligned} & p_b \left(\frac{2}{\mu_b^2} - \left(\psi^2 + \frac{2\psi}{\mu_b} + \frac{2}{\mu_b^2}\right) e^{-\mu_b \psi}\right) \\ \sim \quad & \left(\begin{array}{c} \frac{2}{\mu_b} - \psi^2 \mu_b - 2\psi - \frac{2}{\mu_b} + \mu_b^2 \psi^3 + 2\mu_b \psi^2 \\ +2\psi - \frac{\mu_b^3 \psi^4}{2} - \mu_b^2 \psi^3 - \mu_b \psi^2 \\ -\sum_{i=3}^{\infty} \left(-1\right)^i \frac{\mu_b^{i+1}\psi^{i+2} + 2\mu_b^i \psi^{i+1} + 2\mu_b^{i-1}\psi^i}{i!} \end{array} \right) \\ \sim \quad & -\frac{\mu_b^3 \psi^4}{2} - \sum_{i=3}^{\infty} \left(-1\right)^i \mu_b^{i-1} \psi^i \frac{\mu_b^2 \psi^2 + 2\mu_b \psi + 2}{i!} \\ = \quad & O\left(\mu_b^2 \psi^3\right) \end{aligned}$$

Thus, we require that $\psi\left(C^2\right) \to \infty$ as $C^2 \to \infty$, and, furthermore, that $\mu_b^2 \psi^3 \to 0$ (implying also that $\mu_b \psi \to 0$) as $C^2 \to \infty$. Any such ψ drives all six quantities to 0 when $C^2 \to \infty$ and thus provides separation in the limit. Now we can analyze our M/H_2/2 system as two parallel M/M/1 queues with arrival rates $p_a\lambda$ and $p_b\lambda$ and service time distributions $\mathrm{Exp}(\mu_a)$ and $\mathrm{Exp}(\mu_b)$.

Good examples of $\psi\left(C^2\right)$ with these behaviors are $\psi\left(C^2\right) = \ln C^2$, which approximates load balancing across a wide range of

C^2 when $Q = \frac{1}{2}$, and $\psi\left(C^2\right) = \sqrt{C^2}$, which approximates the optimal cutoff across a wide range of C^2 when $Q = \frac{1}{2}$. Clearly, in both cases, $\psi \to \infty$ as $C^2 \to \infty$. To see the behavior of $\mu_b^2 \psi^3$, first observe that using the Taylor expansion for $\sqrt{1+x}$ in the proof of Lemma 6 shows that $\mu_b = \frac{1}{O\left(C^2\right)}$. Now $\mu_b^2 \left(\left(C^2\right)^m\right)^3 \to 0$ for $m < \frac{2}{3}$. $\psi = \sqrt{C^2}$ meets this criterion, since $m = \frac{1}{2}$. Likewise, $\psi = \ln C^2$ provides $\mu_b^2 \left(\ln C^2\right)^3 \to 0$.

At this point, the proof is complete. While separation in the limit is an elegant concept, this proof may seem unsatisfying since it only deals with $\mathbf{E}\left[T\right]^{\text{SITA}}$ in the limit as $C^2 \to \infty$. We now present an alternative derivation that holds for all C^2, yet yields the same result in the limit, using the P-K formula directly. To aid in notation, define the following for the hyperexponential distribution:

$$\begin{aligned}
F_{u,v} &= \Pr\left\{u < X < v\right\} = \int_u^v \sum_i p_i \mu_i e^{-\mu_i x} dx \\
&= \sum_i p_i e^{-\mu_i u} - \sum_i p_i e^{-\mu_i v} \qquad (4) \\
\mathbf{E}\left[X_{u,v}\right] &= \mathbf{E}\left[X | u < x < v\right] = \frac{1}{F_{u,v}} \int_u^v x \sum_i p_i \mu_i e^{-\mu_i x} dx \\
&= \frac{1}{F_{u,v}} \sum_i p_i \left(u + \frac{1}{\mu_i}\right) e^{-\mu_i u} \\
&\quad - \frac{1}{F_{u,v}} \sum_i p_i \left(v + \frac{1}{\mu_i}\right) e^{-\mu_i v} \qquad (5)
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}\left[X_{u,v}^2\right] &= \mathbf{E}\left[X^2 | u < x < v\right] \\
&= \frac{1}{F_{u,v}} \sum_i p_i \left(u^2 + \frac{2u}{\mu_i} + \frac{2}{\mu_i^2}\right) e^{-\mu_i u} \\
&\quad - \frac{1}{F_{u,v}} \sum_i p_i \left(v^2 + \frac{2v}{\mu_i} + \frac{2}{\mu_i^2}\right) e^{-\mu_i v} \qquad (6) \\
\lambda_{u,v} &= \lambda F_{u,v} \qquad (7)
\end{aligned}$$

For our 2-server system under SITA with cutoff ψ:

$$\begin{aligned}
&\mathbf{E}\left[T\right]^{\text{SITA}} \\
&= \mathbf{E}\left[X\right] \\
&\quad + F_{0,\psi} \frac{\lambda_{0,\psi} \mathbf{E}\left[X_{0,\psi}^2\right]}{2\left(1 - \lambda_{0,\psi} \mathbf{E}\left[X_{0,\psi}\right]\right)} + F_{\psi,\infty} \frac{\lambda_{\psi,\infty} \mathbf{E}\left[X_{\psi,\infty}^2\right]}{2\left(1 - \lambda_{\psi,\infty} \mathbf{E}\left[X_{\psi,\infty}\right]\right)} \\
&= \mathbf{E}\left[X\right] \\
&\quad + \frac{\lambda_{0,\psi}}{2} \frac{\frac{2p_a}{\mu_a^2} - p_a \mathbf{E}\left[X_a^2 \cdot I_{X_a > \psi}\right] + p_b \mathbf{E}\left[X_b^2 \cdot I_{X_b < \psi}\right]}{1 - \lambda\left(\frac{p_a}{\mu_a} - p_a \mathbf{E}\left[X_a \cdot I_{X_a > \psi}\right] + p_b \mathbf{E}\left[X_b \cdot I_{X_b < \psi}\right]\right)} \\
&\quad + \frac{\lambda_{\psi,\infty}}{2} \frac{p_a \mathbf{E}\left[X_a^2 \cdot I_{X_a > \psi}\right] - p_b \mathbf{E}\left[X_b^2 \cdot I_{X_b < \psi}\right] + \frac{2p_b}{\mu_b^2}}{1 - \lambda\left(p_a \mathbf{E}\left[X_a \cdot I_{X_a > \psi}\right] - p_b \mathbf{E}\left[X_b \cdot I_{X_b < \psi}\right] + \frac{p_b}{\mu_b}\right)}
\end{aligned}$$

In the limit, as $C^2 \to \infty$, we simplify using items 1-6 above:

$$\begin{aligned}
\mathbf{E}\left[T\right]^{\text{SITA}} &\to \mathbf{E}\left[X\right] + p_a \lambda \frac{\frac{p_a}{\mu_a^2}}{1 - \lambda\left(\frac{p_a}{\mu_a}\right)} + p_b \lambda \frac{\frac{p_b}{\mu_b^2}}{1 - \lambda\left(\frac{p_b}{\mu_b}\right)} \\
&\to \mathbf{E}\left[X\right] + \frac{\lambda Q^2 \mathbf{E}^2\left[X\right]}{1 - \lambda Q \mathbf{E}\left[X\right]} + \frac{\lambda\left(1 - Q\right)^2 \mathbf{E}^2\left[X\right]}{1 - \lambda\left(1 - Q\right) \mathbf{E}\left[X\right]} \\
&\to \frac{\mathbf{E}\left[X\right]\left(1 - 2Q\left(1 - Q\right)\rho\right)}{1 - \rho + Q\left(1 - Q\right)\rho^2}
\end{aligned}$$

□

Figures 2 and 3 show analytic results for SITA and LWL for a Q-H_2 job size distribution. SITA is analyzed using the closed-form expressions given above, which are functions of the cutoff ψ. We find the optimal ψ by analytically deriving $\frac{d}{d\psi}\left(\mathbf{E}\left[T\right]^{\text{SITA}}\right)$ and using Newton-Raphson to find the ψ where $\frac{d}{d\psi}\left(\mathbf{E}\left[T\right]^{\text{SITA}}\right) = 0$. The dashed line indicates the asymptotic limit for SITA as $C^2 \to \infty$, proven above, which is independent of the cutoff ψ as long as there is separation in the limit. LWL is analyzed using matrix-analytic methods. When $Q = \frac{1}{2}$, as in Figure 2, SITA converges for all loads. When $Q = 0.7$, as in Figure 3, SITA converges for low load, but diverges for higher load as separation is not possible ($Q > \frac{1}{\rho}$). LWL always diverges for a Q-H_2 distribution. The SITA analysis is precise and all SITA results that converge match our predicted values. Although the matrix-analytic method is numerical, we are guaranteed by Theorem 5 that the mean response time does indeed diverge for the Q-H_2 under LWL.

6. CONVERGENT LWL VIA H_3 (BOX 1&3)

The 3-phase Hyperexponential distribution (H_3) with parameters μ_a, μ_b, μ_c, $p_a = 1 - p_b - p_c$. p_b, and p_c is defined by the following random variable:

$$X \sim H_3 \sim \begin{cases} \text{Exp}\left(\mu_a\right) & \text{w.p. } p_a \\ \text{Exp}\left(\mu_b\right) & \text{w.p. } p_b \\ \text{Exp}\left(\mu_c\right) & \text{w.p. } p_c \end{cases}$$

We further specify our H_3 distribution (which we refer to as a k-H_3) with the following relationships:

$$p_a = 1 - p_b - p_c \qquad p_b = \mu_b^{\frac{3}{2}} \qquad p_c = \mu_b^{3k}$$
$$\mu_c = \mu_b^{2k} \qquad k > \tfrac{1}{2}$$

The free parameters are now μ_a, μ_b, and k.

In Theorem 7, we show that the mean response time under LWL for jobs drawn from a k-H_3 distribution is *bounded* as $C^2 \to \infty$. In Theorem 8, we show that, under SITA, mean response time can converge or diverge, depending on k. If $k < \frac{3}{2}$ then $\mathbf{E}\left[T\right]^{\text{SITA}} \to \frac{\mathbf{E}[X]}{1-\rho}$. If $k > \frac{3}{2}$ then mean response time is unbounded.

6.1 Convergent LWL

THEOREM 7. *For a 2-server system with k-H_3 job-size distribution, expected response time under LWL is bounded as $C^2 \to \infty$ for $\rho < 1$.*

PROOF. Lemma 8 guarantees that we can find a k-H_3 distribution with parameters μ_a, μ_b, μ_c, and p_a, p_b, and p_c for any $\mathbf{E}\left[X\right]$, large enough C^2, and k. Lemma 9 shows that, as $C^2 \to \infty$, $\frac{1}{\mu_a} \to \mathbf{E}\left[X\right]$, and $p_a \to 1$. From Lemma 7, we have

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] = \sum_i p_i \frac{3\sqrt{\pi}}{8\mu_i^{\frac{3}{2}}} \to \frac{3\sqrt{\pi}}{8}\left(\left(\mathbf{E}\left[X\right]\right)^{\frac{3}{2}} + 2\right) < \infty$$

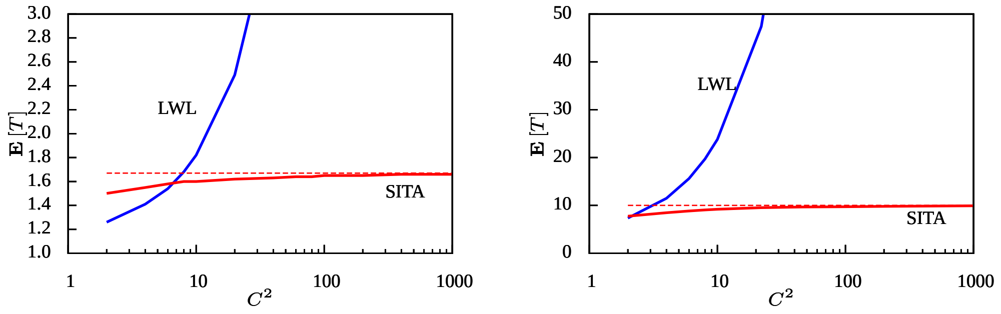
Using [24], since $\mathbf{E}\left[X^{\frac{3}{2}}\right] < \infty$, $\mathbf{E}\left[T\right]^{\text{LWL}} < \infty$ for $\rho < \lfloor \frac{n}{2} \rfloor$. □

LEMMA 8. *For any $\mathbf{E}\left[X\right]$, there exists C^* such that, for all $C^2 > C^*$, and $k > \frac{1}{2}$, there exists a k-H_3 distribution with unique parameters μ_a, μ_b, and μ_c, with $0 < \frac{1}{\mu_a} < \frac{1}{\mu_b} < \frac{1}{\mu_c}$ and probabilities $0 < p_a, p_b, p_c < 1$.*

PROOF. We proceed analogously to the k-Trimodal existence proof. See Appendix for details. □

LEMMA 9. *For the k-H_3 distribution, as $C^2 \to \infty$, $p_a \to 1$, $p_b \to 0$, $p_c \to 0$, $\frac{1}{\mu_a} \to \mathbf{E}\left[X\right]$, $\frac{1}{\mu_b} \to \infty$, and $\frac{1}{\mu_c} \to \infty$.*

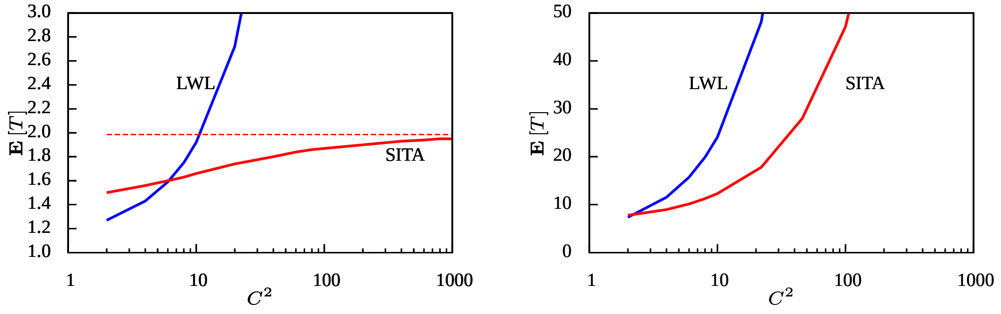
PROOF. Analogous to Theorem 4. □



(a) $\rho = 0.8$

(b) $\rho = 1.8$

Figure 2: ***Expected response time,*** $\mathbf{E}[T]$***, for SITA and LWL vs*** C^2 ***under a*** Q***-***H_2 ***distribution with*** $Q = \frac{1}{2}$ ***and (a)*** $\rho = 0.8$ ***and (b)*** $\rho = 1.8$**. The dashed line shows** $\lim_{C^2\to\infty} \mathbf{E}[T]^{\text{SITA}}$ **according to Theorem 6.**



(a) $\rho = 0.8$

(b) $\rho = 1.8$

Figure 3: ***Expected response time,*** $\mathbf{E}[T]$***, for SITA and LWL vs*** C^2 ***under a*** Q***-***H_2 ***distribution with*** $Q = 0.7$ ***and (a)*** $\rho = 0.8$ ***and (b)*** $\rho = 1.8$**. The dashed line shows** $\lim_{C^2\to\infty} \mathbf{E}[T]^{\text{SITA}}$ **according to Theorem 6.**

6.2 Convergent/Divergent SITA

We now analyze a SITA task allocation over two servers with a k-H_3 job-size distribution.

THEOREM 8. *For a 2-server system with* k-H_3 *job-size distribution, with fixed mean* $\mathbf{E}[X]$*, parameter* $k < \frac{3}{2}$*, and* $\rho < 1$*, there exists a cutoff* $\psi(C^2)$ *such that, under SITA:*

$$\mathbf{E}[T]^{SITA} \xrightarrow[C^2\to\infty]{} \begin{cases} \infty & k > \frac{3}{2} \\ \frac{\mathbf{E}[X]}{1-\rho} + \lambda & k = \frac{3}{2} \\ \frac{\mathbf{E}[X]}{1-\rho} & \frac{1}{2} < k < \frac{3}{2} \end{cases}$$

PROOF. As in Theorem 6, we seek a cutoff $\psi(C^2)$ that separates the H_3 jobs in the limit as $C^2 \to \infty$ such that the small-job server serves jobs that are $\text{Exp}(\mu_a)$ and the large-job server serves jobs that are drawn from an H_2 distribution. Let X_a denote the first branch of the H_3 distribution and X_{bc} the other two branches. Namely, $X_a \sim \text{Exp}(\mu_a)$ and $X_{bc} \sim H_2\left(\mu_b, \mu_c, \frac{p_b}{p_b+p_c}, \frac{p_c}{p_b+p_c}\right)$. In order for the k-H_3 to be separable in the limit, the cutoff ψ (which is a function of C^2) must have the property that the following six quantities go to 0:

1. $p_a\mathbf{E}[I_{X_a>\psi}] = p_a e^{-\mu_a\psi}$

2. $p_a\mathbf{E}[X_a \cdot I_{X_a>\psi}] = p_a\left(\psi + \frac{1}{\mu_a}\right)e^{-\mu_a\psi}$

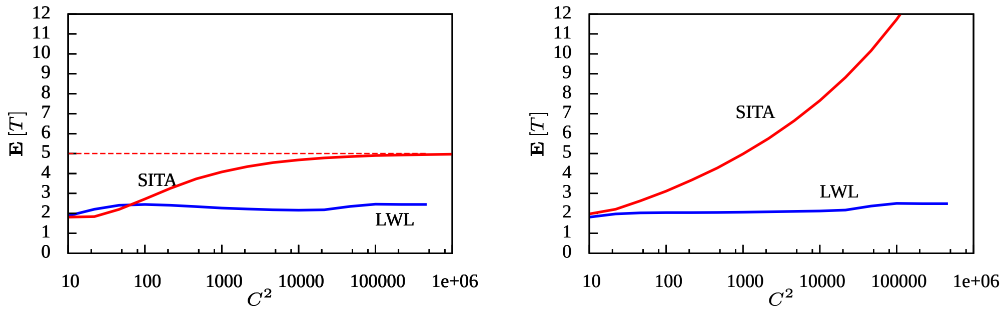
3. $p_a\mathbf{E}\left[X_a^2 \cdot I_{X_a>\psi}\right] = p_a\left(\psi^2 + \frac{2\psi}{\mu_a} + \frac{2}{\mu_a^2}\right)e^{-\mu_a\psi}$

4. $(p_b+p_c)\,\mathbf{E}[I_{X_{bc}<\psi}] = p_b\left(1-e^{-\mu_b\psi}\right) + p_c\left(1-e^{-\mu_c\psi}\right)$

5. $(p_b+p_c)\,\mathbf{E}[X_{bc}\cdot I_{X_{bc}<\psi}]$
$$= p_b\left(\frac{1}{\mu_b} - \left(\psi+\frac{1}{\mu_b}\right)e^{-\mu_b\psi}\right) + p_c\left(\frac{1}{\mu_c} - \left(\psi+\frac{1}{\mu_c}\right)e^{-\mu_c\psi}\right)$$

6. $(p_b+p_c)\,\mathbf{E}\left[X_{bc}^2\cdot I_{X_{bc}<\psi}\right]$
$$= p_b\left(\frac{2}{\mu_b^2} - \left(\psi^2+\frac{2\psi}{\mu_b}+\frac{2}{\mu_b^2}\right)e^{-\mu_b\psi}\right) + p_c\left(\frac{2}{\mu_c^2} - \left(\psi^2+\frac{2\psi}{\mu_c}+\frac{2}{\mu_c^2}\right)e^{-\mu_c\psi}\right)$$

As for the Q-H_2 distribution, as $C^2 \to \infty$, we need $\mu_a\psi \to \infty$, which follows if $\psi \to \infty$. We also need $\mu_b^2\psi^3 \to 0$, addressed shortly. Finally, we need $\mu_c^2\psi^3 \to 0$, which follows from $\mu_c = \mu_b^{2k}$ when $k > \frac{1}{2}$ and $\mu_b^2\psi^3 \to 0$. The proof of Lemma 8 shows that $\mu_b \sim (C^2)^{-\frac{1}{k}}$. Thus, if $\psi = \sqrt{C^2}$, then $\mu_b^2\psi^3 \sim (C^2)^{-\frac{2}{k}}(C^2)^{\frac{3}{2}} = (C^2)^{\frac{3}{2}-\frac{2}{k}}$. This goes to 0 when $k < \frac{4}{3}$. If $\psi = \ln C^2$, then L'Hôpital's rule gives $\mu_b^2\psi^3 \to 0$ for any $k > 0$.



(a) $k = 1.0$

(b) $k = 2.0$

Figure 4: *Expected response time, $\mathbf{E}[T]$, for SITA and LWL vs C^2 under a k-H_3 distribution with $\rho = 0.8$ for (a) $k = 1.0$ and (b) $k = 2.0$*. The dashed line shows $\lim_{C^2\to\infty} \mathbf{E}[T]^{\text{SITA}}$ according to Theorem 8.

Now we calculate the expected response time as the weighted average of the response time at the small-job (exponential) server plus the expected service time and P-K delay at the large-job (H_2) server. To do so, we recall Equations (4)–(7) and further define:

$$\begin{aligned} N &= p_a\mathbf{E}\left[X_a^2 \cdot I_{X_a>\psi}\right] - (p_b+p_c)\,\mathbf{E}\left[X_{bc}^2 \cdot I_{X_{bc}<\psi}\right] \\ D &= p_a\mathbf{E}\left[X_a \cdot I_{X_a>\psi}\right] - (p_b+p_c)\,\mathbf{E}\left[X_{bc} \cdot I_{X_{bc}<\psi}\right] \end{aligned}$$

Then,

$$\begin{aligned} &\mathbf{E}[W]^{\text{SITA}} \\ &= F_{0,\psi}\frac{\lambda_{0,\psi}\mathbf{E}\left[X_{0,\psi}^2\right]}{2\left(1-\lambda_{0,\psi}\mathbf{E}\left[X_{0,\psi}\right]\right)} + F_{\psi,\infty}\frac{\lambda_{\psi,\infty}\mathbf{E}\left[X_{\psi,\infty}^2\right]}{2\left(1-\lambda_{\psi,\infty}\mathbf{E}\left[X_{\psi,\infty}\right]\right)} \\ &= \frac{\lambda}{2}\frac{p_a\left(\frac{2p_a}{\mu_a^2}-N\right)}{1-\lambda\left(\frac{p_a}{\mu_a}-D\right)} + \frac{\lambda}{2}\frac{(p_b+p_c)\left(N+\frac{2p_b}{\mu_b^2}+\frac{2p_c}{\mu_c^2}\right)}{1-\lambda\left(D+\frac{p_b}{\mu_b}+\frac{p_c}{\mu_c}\right)} \end{aligned}$$

In the limit, as $C^2 \to \infty$, $N \to 0$ and $D \to 0$ using items 1-6 above. We also substitute for $\frac{p_a}{\mu_a} = \mathbf{E}[X] - \frac{p_b}{\mu_b} - \frac{p_c}{\mu_c} \to \mathbf{E}[X]$ since $\frac{p_b}{\mu_b} = \mu_b^{\frac{1}{2}} \to 0$ and $\frac{p_c}{\mu_c} = \mu_b^k \to 0$. Then:

$$\begin{aligned} &\mathbf{E}[T]^{\text{SITA}} \\ &\to \quad \mathbf{E}[X] + \frac{\lambda}{2}\frac{\mathbf{E}^2[X]}{1-\lambda\mathbf{E}[X]} + \lambda(p_b+p_c)\left(\frac{p_b}{\mu_b^2}+\frac{p_c}{\mu_c^2}\right) \\ &\to \quad \frac{\mathbf{E}[X]}{1-\rho} + \lambda\left(\mu_b + \mu_b^{\frac{3}{2}-k} + \mu_b^{3k-\frac{1}{2}} + \mu_b^{2k}\right) \\ &\xrightarrow[C^2\to\infty]{} \begin{cases} \infty & k > \frac{3}{2} \\ \frac{\mathbf{E}[X]}{1-\rho} + \lambda & k = \frac{3}{2} \\ \frac{\mathbf{E}[X]}{1-\rho} & \frac{1}{2} < k < \frac{3}{2} \end{cases} \end{aligned}$$

□

Figure 4 shows analytic results for SITA and LWL for a k-H_3 job size distribution. SITA is analyzed using the closed-form expressions given above, which are functions of the cutoff ψ. We again find the optimal ψ by Newton-Raphson. The dashed line indicates the asymptotic limit for SITA as $C^2 \to \infty$, proven above, which is independent of the cutoff ψ as long as there is separation in the limit. LWL is analyzed using matrix-analytic methods. When $k = 1$, LWL and SITA both converge provided $\rho < 1$. When $k = 2$, LWL converges provided $\rho < 1$, but SITA diverges. There is a possibility of significant error in the results for LWL for larger C^2 because of instability in the numerical solution of the matrix quadratic equation required for matrix-analytic methods. However, we are guaranteed by Theorem 7 that the mean response time does converge for the k-H_3 under LWL.

7. PARETO AND BOUNDED PARETO DISTRIBUTION (BOXES 3 & 4)

We now turn to the Pareto and Bounded Pareto distributions, which are known to well-model empirical job size distributions for a wide variety of computing applications [2, 8, 27, 15, 26].

The Bounded Pareto(k, p, α) distribution, where $0 < \alpha < 2$ and $0 < k < p$, has the following density function:

$$f(x) = \begin{cases} \frac{\alpha k^\alpha}{1-\left(\frac{k}{p}\right)^\alpha}x^{-\alpha-1} & k \le x \le p \\ 0 & \text{otherwise} \end{cases}$$

We refer to the normalizing constant as $m = \frac{\alpha k^\alpha}{1-\left(\frac{k}{p}\right)^\alpha}$. As $p \to \infty$, the Bounded Pareto distribution converges to the Pareto:

$$f(x) = \alpha k^\alpha x^{-1-\alpha} \qquad x \ge k > 0$$

For $1 < \alpha < 2$, the Pareto distribution has finite mean, but infinite variance.

We will prove that, for the Bounded Pareto and Pareto job-size distributions, the mean response time under SITA always diverges (as $C^2 \to \infty$), whereas that under LWL may converge or diverge, depending on the α-parameter of the distribution. We then extend the Pareto results to n-server systems.

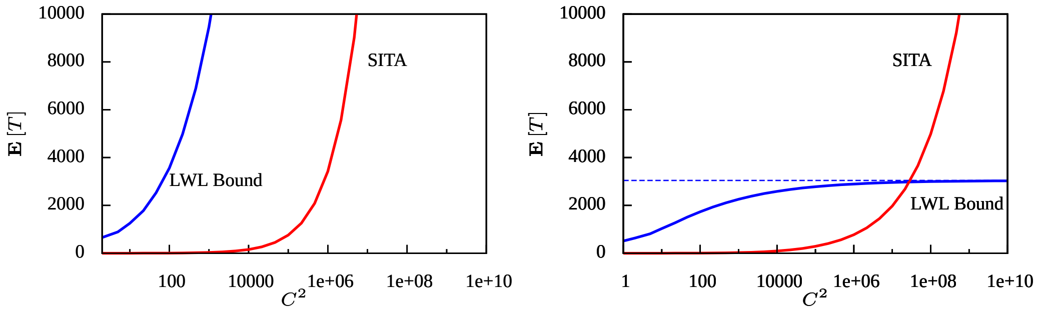
7.1 LWL

THEOREM 9. *The mean response time for a 2-server system with Bounded Pareto job-sizes under LWL is bounded if $\alpha > \frac{3}{2}$ and $\rho < 1$, regardless of C^2, including $C^2 \to \infty$. The response time is unbounded as $C^2 \to \infty$ for $\alpha \le \frac{3}{2}$ or $\rho > 1$.*

PROOF. Lemma 10 shows that, for any $\mathbf{E}[X]$, C^2 and α, there exists a Bounded Pareto(k, p, α). Lemma 11 shows that, as $C^2 \to \infty$, the Bounded Pareto converges to a Pareto$\left(\frac{\alpha-1}{\alpha}\mathbf{E}[X], \alpha\right)$ (with $p \to \infty$). The $\frac{3}{2}$ moment is given by:

$$\mathbf{E}\left[X^{\frac{3}{2}}\right] \;=\; m\int_k^p x^{\frac{1}{2}-\alpha} = \frac{m}{\frac{3}{2}-\alpha}\left(p^{\frac{3}{2}-\alpha} - k^{\frac{3}{2}-\alpha}\right)$$

The above increases with p and C^2, but is bounded as $C^2 \to \infty$. When $\rho < 1$, we can apply [24] to see that mean response time



(a) $\alpha = 1.4$

(b) $\alpha = 1.6$

Figure 5: ***Expected response time,*** $\mathbf{E}[T]$***, for SITA and LWL vs*** C^2 ***under a Bounded Pareto job size distribution with*** $\rho = 0.95$ ***and (a)*** $\alpha = 1.4$ ***and (b)*** $\alpha = 1.6$***. The dashed line shows*** $\lim_{C^2 \to \infty} \mathbf{E}[T]^{LWL}$***.***

under LWL converges. Furthermore, we see that $\mathbf{E}\left[X^{\frac{3}{2}}\right] \to \infty$ if $\alpha \leq \frac{3}{2}$. Hence, if $\alpha \leq \frac{3}{2}$, or $\rho > 1$, we see by [25] that mean response time under LWL diverges. □

COROLLARY 2. *The mean response time for a 2-server system with Pareto job sizes, $\rho < 1$, and $\frac{3}{2} < \alpha < 2$ under LWL is bounded.*

LEMMA 10. *For any $\mathbf{E}[X]$, C^2, and $\alpha > 1$, we can specify a Bounded Pareto(k, p, α).*

PROOF.

$$\begin{aligned}
\mathbf{E}[X] &= m\int_k^p x^{-\alpha} = \frac{\alpha k^\alpha}{1-\left(\frac{k}{p}\right)^\alpha}\frac{1}{1-\alpha}\left(p^{1-\alpha} - k^{1-\alpha}\right) \\
\mathbf{E}\left[X^2\right] &= m\int_k^p x^{1-\alpha} = \frac{\alpha k^\alpha}{1-\left(\frac{k}{p}\right)^\alpha}\frac{1}{2-\alpha}\left(p^{2-\alpha} - k^{2-\alpha}\right)
\end{aligned}$$

For $k = p = \mathbf{E}[X]$, applying L'Hôpital's rule to $\mathbf{E}\left[X^2\right]$ as $p \to k$ gives $\mathbf{E}\left[X^2\right] \to \mathbf{E}^2[X]$ and thus $C^2 = 0$. As $p \to \infty$, holding $\mathbf{E}[X]$ constant, $C^2 \to \infty$ but k does not. We write k as a function of p as: $k = \left(p^{1-\alpha} - \frac{1-\alpha}{m}\mathbf{E}[X]\right)^{\frac{2-\alpha}{1-\alpha}}$. Now, C^2 is a function of p, $k(p)$, and $m(p, k(p))$, all continuous for $p > 0$. Thus, there exists a value of p mapping to every value of C^2. □

LEMMA 11. *Keeping $\mathbf{E}[X]$ constant, as $C^2 \to \infty$, for the Bounded Pareto distribution, $p \to \infty$ and $k \to \frac{\alpha-1}{\alpha}\mathbf{E}[X]$ (from above for $\alpha > 1$).*

PROOF. Keeping $\mathbf{E}[X]$ constant, the only possibilities for infinite C^2 are $k = p$, which takes $m \to \infty$ and $p \to \infty$. We know from Lemma 10 that $k = p$ has $C^2 = 0$, so $C^2 \to \infty$ implies $p \to \infty$ and $k \to \frac{\alpha-1}{\alpha}\mathbf{E}[X]$. As p increases, k decreases monotonically for $\alpha > 1$: $\frac{dk}{dp} = (1-\alpha)\left(\frac{k}{p}\right)^\alpha\left(1 + \frac{\mathbf{E}[X]}{p}\right)$, and so convergence is from above. □

7.2 SITA

THEOREM 10. *The mean response time for a 2-server system with Bounded Pareto job sizes diverges under SITA as $C^2 \to \infty$.*

PROOF. Under SITA, with Bounded Pareto job-sizes, the P-K delay for a given cutoff ψ is given by

$$\begin{aligned}
\mathbf{E}[W] &= F(\psi)\frac{\lambda\frac{m}{-\alpha+2}\left(\psi^{-\alpha+2} - k^{-\alpha+2}\right)}{2\left(1-\lambda\frac{m}{-\alpha+1}\left(\psi^{-\alpha+1} - k^{-\alpha+1}\right)\right)} \\
&\quad +\overline{F}(\psi)\frac{\lambda\frac{m}{-\alpha+2}\left(p^{-\alpha+2} - \psi^{-\alpha+2}\right)}{2\left(1-\lambda\frac{m}{-\alpha+1}\left(p^{-\alpha+1} - \psi^{-\alpha+1}\right)\right)}
\end{aligned}$$

Lemma 11 shows that $p \to \infty$ as $C^2 \to \infty$. Now we examine the behavior of $\psi\left(C^2\right)$ as $C^2 \to \infty$. If $\psi\left(C^2\right)$ is bounded, then the second term is unbounded as $C^2 \to \infty$. However, if $\psi\left(C^2\right) \to \infty$, then the first term is unbounded. Therefore, $\mathbf{E}[W] \to \infty$, and thus $\mathbf{E}[T] = \mathbf{E}[W] + \mathbf{E}[X] \to \infty$ as $C^2 \to \infty$. □

THEOREM 11. *The mean response time for a 2-server system with Pareto job sizes and $0 < \alpha < 2$ is unbounded under SITA.*

PROOF. Assuming ψ is finite, consider jobs larger than ψ. The arrival rate of these jobs is $\lambda_L = \lambda\overline{F}(\psi) > 0$. These jobs see an M/G/1 queue where G is the conditional distribution $[X|X > \psi]$, which is still Pareto and therefore has infinite second moment. Since these large jobs have strictly positive probability $p_L = \overline{F}(\psi) > 0$, their contribution to $\mathbf{E}[T]^{\text{SITA}}$ is $p_L\lambda_L\frac{\mathbf{E}\left[X^2|X>\psi\right]}{2(1-\lambda_L\mathbf{E}[X|X>\psi])} = \infty$. If $\psi = \infty$, the small-job server sees the full Pareto job-size distribution and has unbounded $\mathbf{E}[T]^{\text{SITA}}$. □

Figure 5 compares the mean response time for Bounded Pareto workloads under LWL (the bound from Equation (2)) vs. SITA. The SITA curve is derived using the optimal cutoff ψ (determined numerically) and the above SITA equations. The figure shows that response time is bounded for LWL for $\alpha > \frac{3}{2}$, while response time for SITA is unbounded regardless of α. In the right graph, where $\alpha > \frac{3}{2}$, as $C^2 \to \infty$, SITA diverges while LWL converges. This behavior may not have been noticed in the past, since SITA's performance only starts to deteriorate rapidly after $C^2 > 10^5$, which is hard to see in simulation. The practical implication is that SITA is superior to LWL for realistic C^2 in the Bounded Pareto. In the left graph, where $\alpha < \frac{3}{2}$, both LWL and SITA diverge.

7.3 *n* Servers with Pareto Workload

For the Pareto distribution, the results for the 2-server system immediately generalize to the n-server system (for finite n). As in the 2-server system, the mean response time under LWL is bounded, if $\rho < n - 1$ and $\frac{3}{2} < \alpha < 2$, but the mean response time under SITA is never bounded.

THEOREM 12. *The mean response time for an n-server system with Pareto job sizes, $\rho < n - 1$, and $\frac{3}{2} < \alpha < 2$ under LWL is bounded.*

PROOF. Since $\mathbf{E}\left[X^{\frac{3}{2}}\right] < \infty$, we can apply [22] directly to show that LWL is bounded. □

We define SITA for an n-server system as a policy that immediately dispatches jobs with sizes between ψ_{i-1} and ψ_i (with $\psi_0 = 0$ and $\psi_n = \infty$) to server i.

THEOREM 13. *The mean response time for a n-server system with Pareto job sizes and $0 < \alpha < 2$ is unbounded under SITA.*

PROOF. Let ψ_{i-1} be the largest finite cutoff. Then server i sees a Pareto job-size distribution. As in the proof of Theorem 11, the contribution of these jobs to $\mathbf{E}[T]^{\text{SITA}}$ ensures $\mathbf{E}[T]^{\text{SITA}} = \infty$. □

8. CONCLUSION

Finding good task assignment policies for server farms is such an old, well-studied problem that one believes that all the important questions have been answered by now. Certainly that was the belief of the authors. This paper shows that there are still many things we don't understand about task assignment, when job size variability is high. Size-interval task assignment (SITA), which provides short jobs isolation from long ones, seems so natural for high-variability job sizes that is hard to imagine that it can be inferior to a much more naive policy, like Least-Work-Left (LWL) that allows short jobs to queue behind long ones. And yet, we prove that SITA's performance can be *unboundedly* worse than that of LWL, and the performance of LWL can be remarkably good. We have shown that the comparatively poor behavior of SITA can occur in wide classes of distributions (including Modal, Hyperexponential, Bounded Pareto, and Pareto) at high C^2 over a wide range of load (except heavy traffic). For some specialized forms of these distributions, SITA performs poorly even for moderate C^2 (like 10).

This discovery begs the question: When then exactly does SITA perform well? This paper answers this question too, defining the regimes under which SITA's response time converges with respect to increasing C^2, and proving the first asymptotic upper bounds on SITA's response time for certain common distributions. Thus, the message is *not* that one should discard SITA, but rather that one should carefully consider the operating regime before presuming that SITA is the best solution. To paraphrase an old nursery rhyme:

> *When SITA is good, it is very very good,*
> *But when it is bad, it is horrid.*

9. ACKNOWLEDGEMENTS

We would like to thank R. Núñez Queija for his technical comments, which greatly improved the quality of this paper.

10. REFERENCES

[1] Eitan Bachmat and Hagit Sarfati. Analysis of size interval task assigment policies. *Performance Evaluation Review*, 36(2), 2008.

[2] Paul Barford and Mark Crovella. Generating representative Web workloads for network and server performance evaluation. In *Proceedings of the ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 151–160, July 1998.

[3] Onno Boxma, Q. Deng, and Albertus Zwart. Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers. *Queueing Systems*, 40(1):5–31, 2002.

[4] James Broberg, Zahir Tari, and Panlop Zeephongsekul. Task assignment with work-conserving migration. *Parallel Computing*, 32:808–830, 2006.

[5] John Buzacott and George Shanthikumar. *Stochastic Models in Manufacturing Systems*. Prentice Hall, 1993.

[6] Valeria Cardellini, Emiliano Casalicchio, Michele Colajanni, and Philip Yu. The state of the art in locally distributed web-server systems. Technical report, 2001.

[7] Gianfranco Ciardo, Alma Riska, and Evgenia Smirni. Equiload: a load balancing policy for clustered web servers. *Performance Evaluation*, 46:101–124, 2001.

[8] Mark Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. pages 160–169, May 1996.

[9] Muhammad El-Taha and Bacel Maddah. Allocation of service time in a multiserver system. *Management Science*, 52(4):623–637, 2006.

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

[11] Serguei Foss and Dmitry Korshunov. Heavy tails in multi-server queue. *Queueing Systems*, 52:31–48, 2006.

[12] Bin Fu, James Broberg, and Zahir Tari. Task assignment strategy for overloaded systems. In *Proceedings of the Eighth IEEE International Symposium on Computers and Communications*, 2003.

[13] Mor Harchol-Balter. Task assignment with unknown duration. *Journal of the ACM*, 49(2):260–288, March 2002.

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

[15] Mor Harchol-Balter and Allen Downey. Exploiting process lifetime distributions for dynamic load balancing. In *Proceedings of ACM SIGMETRICS*, pages 13–24, Philadelphia, PA, May 1996.

[16] Wallace Hopp and Mark Spearman. *Factory Physics*. McGraw Hill/Irwin, 2 edition, 2000.

[17] Steven Hotovy, David Schneider, and Timothy O'Donnell. Analysis of the early workload on the Cornell Theory Center IBM SP2. 1996.

[18] Leonard Kleinrock. *Queueing Systems*, volume I. Theory. John Wiley & Sons, 1975.

[19] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial Mathematics, 1987.

[20] Kazumasa Oida and Kazumasa Shinjo. Characteristics of deterministic optimal routing for a simple traffic control problem. In *Performance, Computing and Communications Conference, IPCCC*, February 1999.

[21] Konstantinos Psounis, Pablo Molinero-Fernández, Balaji Prabhakar, and Fragkiskos Papadopoulos. Systems with multiple servers under heavy-tailed workloads. *Performance Evaluation*, 52:456–474, 2005.

[22] Alan Scheller-Wolf. Further delay moment results for FIFO multiserver queues. *Queueing Systems*, 34:387–400, 2000.

[23] Alan Scheller-Wolf and Karl Sigman. Delay moments for FIFO GI/GI/s queues. *Queueing Systems*, 25:77–95, 1997.

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

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

[26] Bianca Schroeder and Mor Harchol-Balter. Evaluation of task assignment policies for supercomputing servers: The case for load unbalancing and fairness. *Cluster Computing: The journal of Networks, Software Tools, and Apps*, 7(2):151–161, April 2004.

[27] Anees Shaikh, Jennifer Rexford, and Kang Shin. Load-sensitive routing of long-lived IP flows. In *Proceedings of ACM SIGCOMM*, September 1999.

[28] Zahir Tari, James Broberg, Albert Zomaya, and Roberto Baldoni. A least flow-time first load sharing approach for distributed server farm. *Journal of Parallel and Distributed Computing*, 65:832–842, 2005.

[29] Nigel Thomas. Comparing job allocation schemes where service demand is unknown. *Journal of Computer and System Sciences*, 74:1067–1081, 2008.

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

APPENDIX

LEMMA 3. *For any $\mathbf{E}[X]$, there exists C^* such that, for all $C^2 > C^*$, and $k > \frac{1}{2}$, there exists a k-Trimodal distribution with unique parameters $0 < a < b < c$ and probabilities $0 < p_a, p_b, p_c < 1$.*

PROOF. Note that since $k > \frac{1}{2}$ and $c = b^{2k}$, $c > b$ requires $b > 1$. Setting $b > 2^{\frac{2}{3}}$ guarantees that $0 < p_a, p_b, p_c < 1$. We need to prove the existence of $0 < a < b$ satisfying the desired $\mathbf{E}[X]$ and C^2. Our first constraint for a and b comes from $\mathbf{E}[X]$:

$$\begin{aligned}\mathbf{E}[X] &= p_a \cdot a + p_b \cdot b + p_c \cdot c &(8)\\ &= a \cdot \left(1 - b^{-\frac{3}{2}} - b^{-3k}\right) + b^{-\frac{1}{2}} + b^{-k}\\ a &= \frac{\mathbf{E}[X] - b^{-\frac{1}{2}} - b^{-k}}{1 - b^{-\frac{3}{2}} - b^{-3k}} &(9)\end{aligned}$$

Since we require $a > 0$, and since the denominator of a is simply p_a which is positive, we need the numerator of (9) to be positive. This is satisfied if $b > \frac{4}{\mathbf{E}^2[X]}$. For convenience, we define $b_{lb} = \max\left\{2^{\frac{2}{3}}, \frac{4}{\mathbf{E}^2[X]}\right\}$, where we seek $b > b_{lb}$.

Our second constraint comes from C^2, substituting for a from (9) (where $\sim$ denotes asymptotic convergence):

$$\begin{aligned}C^2 &= \frac{p_a \cdot a^2 + p_b \cdot b^2 + p_c \cdot c^2}{\mathbf{E}^2[X]} - 1 &(10)\\ &= \frac{a^2 \cdot \left(1 - b^{-\frac{3}{2}} - b^{-3k}\right) + b^{\frac{1}{2}} + b^k}{\mathbf{E}^2[X]} - 1 &(11)\\ &\sim \quad \widetilde{C}^2 = \frac{b^{\frac{1}{2}} + b^k}{\mathbf{E}^2[X]} \quad (\text{as } b \to \infty) &(12)\end{aligned}$$

At this point, after substituting for a, (11) specifies C^2 as a function of b. Throughout the rest of the proof, we will write $C^2(b)$ to make explicit the dependence of C^2 on b. We will now show that for $C^2(b)$ sufficiently high, there exists a b sufficiently large, which satisfies (11). We will also show that there exists a b^* and $C^* = C^2(b^*) + 1$ such that $C^2(b)$ increases monotonically with respect to b for all $b \geq b^*$, guaranteeing a one-to-one relationship between $C^2(b)$ and b for all $b \geq b^*$. We will do this by demonstrating that the derivative $\frac{d}{db}\left[C^2(b)\right]$ is positive for all $b > b^*$. If $C^2(b) = \frac{f}{g}$, then $\frac{d}{db}\left[C^2(b)\right] = \frac{gf' - fg'}{g^2} = \frac{N}{D}$. The denominator D of $\frac{d}{db}\left[C^2(b)\right]$ is positive. We now seek a region where the numerator of the derivative, N, is positive as well:

$$\frac{N}{\mathbf{E}^2[X]} =$$

$$\left(1 - b^{-\frac{3}{2}} - b^{-3k}\right)\left(\begin{array}{l}\mathbf{E}[X]\, b^{-\frac{3}{2}} + 2k\mathbf{E}[X]\, b^{-k-1}\\ -(2k+1)\, b^{-k-\frac{3}{2}} - \left(k - \frac{3}{2}\right) b^{k-\frac{5}{2}}\\ +\left(3k - \frac{1}{2}\right) b^{-3k-\frac{1}{2}}\\ +\frac{1}{2} b^{-\frac{1}{2}} + kb^{k-1}\end{array}\right)$$

$$-\left(\frac{3}{2} b^{-\frac{5}{2}} + 3kb^{-3k-1}\right)\left(\begin{array}{l}\mathbf{E}^2[X] - 2\mathbf{E}[X]\, b^{-\frac{1}{2}}\\ -2\mathbf{E}[X]\, b^{-k} + 2b^{-k-\frac{1}{2}}\\ -b^{k-\frac{3}{2}} - b^{-3k+\frac{1}{2}} + b^{\frac{1}{2}} + b^k\end{array}\right)$$

We expand N and group and collect its terms into $N = N_1 + N_2 + N_3 + N_4$ as follows:

1. If the term is positive and the exponent of b is ≤ -1 for all $k > \frac{1}{2}$, we group the term into N_1.
2. If the term is negative and the exponent of b is ≤ -1 for all $k > \frac{1}{2}$, we group the term into N_2.
3. If the exponent of b in the term can be either positive or negative, we group the term into N_3.
4. The one remaining term we group into N_4.

$$\begin{aligned}\frac{N_1}{\mathbf{E}^2[X]} &= \mathbf{E}[X]\, b^{-\frac{3}{2}} + 2k\mathbf{E}[X]\, b^{-k-1} + \left(3k - \frac{1}{2}\right) b^{-3k-\frac{1}{2}}\\ &\quad + (2k+1)\, b^{-k-3} + (2k+1)\, b^{-4k-\frac{3}{2}} + 3\mathbf{E}[X]\, b^{-3}\\ &\quad + 3\mathbf{E}[X]\, b^{-k-\frac{5}{2}} + \frac{3}{2} b^{-3k-2} + 6k\mathbf{E}[X]\, b^{-3k-\frac{3}{2}}\\ &\quad + 6k\mathbf{E}[X]\, b^{-4k-1} + \left(4k - \frac{3}{2}\right) b^{-2k-\frac{5}{2}} + 3kb^{-6k-\frac{1}{2}}\\ \frac{N_2}{\mathbf{E}^2[X]} &= -(2k+1)\, b^{-k-\frac{3}{2}} - \mathbf{E}[X]\, b^{-3} - 2k\mathbf{E}[X]\, b^{-k-\frac{5}{2}}\\ &\quad - \left(3k - \frac{1}{2}\right) b^{-3k-2} - 2b^{-2} - \mathbf{E}[X]\, b^{-3k-\frac{3}{2}}\\ &\quad - 2k\mathbf{E}[X]\, b^{-4k-1} - \left(3k - \frac{1}{2}\right) b^{-6k-\frac{1}{2}} - \frac{1}{2} b^{-3k-\frac{1}{2}}\\ &\quad - kb^{-2k-1} - \frac{3}{2}\mathbf{E}^2[X]\, b^{-\frac{5}{2}} - 3b^{-k-3}\\ &\quad - 3k\mathbf{E}^2[X]\, b^{-3k-1} - 6kb^{-4k-\frac{3}{2}}\\ &\quad - 3kb^{-3k-\frac{1}{2}} - 3kb^{-2k-1}\\ \frac{N_3}{\mathbf{E}^2[X]} &= kb^{k-1} + kb^{k-4} - 2kb^{k-\frac{5}{2}}\\ \frac{N_4}{\mathbf{E}^2[X]} &= \frac{1}{2} b^{-\frac{1}{2}}\end{aligned}$$

We now derive a lower bound for N, denoted by N^*. If we can guarantee $N^* > 0$ then clearly $N > 0$.

- $N_1 > 0$, so we omit those terms from N^*.
- Since $b > 1$, we replace the exponent of b in the terms of N_2 with -1, which makes N_2 more negative and N^* smaller. Define the modified term as N_2^*:

  $$\frac{N_2^*}{\mathbf{E}^2[X]} = -\left(\begin{array}{l}21k + \frac{11}{2} + (2+4k)\,\mathbf{E}[X]\\ +\left(3k + \frac{3}{2}\right)\mathbf{E}^2[X]\end{array}\right) b^{-1}$$

- $N_3 > 0$, so we omit those terms from N^*. To see this, observe that:

  $$\begin{aligned}b^{k-1} + b^{k-4} &> 2b^{k-\frac{5}{2}}\\ b^{\frac{3}{2}} + b^{-\frac{3}{2}} &> 2\end{aligned}$$

  Since $b > 2^{\frac{2}{3}}$, this is evidently true.
- N_4 is unchanged in N^*.

$$\frac{N^*}{\mathbf{E}^2[X]} = N_2^* b^{-1} + \frac{1}{2} b^{-\frac{1}{2}}$$

To ensure $N^* > 0$, we require that

$$\begin{aligned}\frac{1}{2} b^{-\frac{1}{2}} &> N_2^* b^{-1}\\ b &> \left(2N_2^*\right)^2\\ b^* &= \max\left\{\mathbf{E}[X], b_{lb}, \left(2N_2^*\right)^2\right\}\end{aligned}$$

Observe that $\mathbf{E}[X] > a$ for $k > \frac{1}{2}$ because a is the smallest value that the distribution takes on and because $c > b > b^* > \mathbf{E}[X] > a$. Hence $b > a$.

Thus, for all $b > b^*$, $C^2(b)$ increases monotonically with b. Since $0 < \widetilde{C}^2(b) - C^2(b) < 1$ for $b > 1$ and $\widetilde{C}^2(b)$ is monotonic in b for $b > 1$, no $b < b^*$ has $C^2(b) > C^*$. Thus, for any $C^2 > C^* = C^2(b^*) + 1$, a unique value of $b > a$ exists. □

LEMMA 5. *For any $\mathbf{E}[X]$, $C^2 > 1$, and Q, we can find unique parameters μ_a, μ_b, and p, where $\mu_a > \mu_b$ (μ_a is the service rate for the small jobs), for a Q-H_2 distribution which satisfy:*

$$\begin{aligned} p &= \frac{C^2 + 4Q - 1 + \sqrt{\left(C^2-1\right)^2 + 8\left(C^2-1\right)Q\left(1-Q\right)}}{2\left(C^2+1\right)} \\ \mu_a &= \frac{p}{Q\mathbf{E}[X]} \\ \mu_b &= \frac{1-p}{(1-Q)\mathbf{E}[X]} \end{aligned}$$

PROOF.

$$\begin{aligned} \mathbf{E}[X] &= \frac{p_a}{\mu_a} + \frac{p_b}{\mu_b} = \frac{p}{\mu_a} + \frac{1-p}{\mu_b} \\ \mathbf{E}\left[X^2\right] &= \frac{2p_a}{\mu_a^2} + \frac{2p_b}{\mu_b^2} = \frac{2p}{\mu_a^2} + \frac{2(1-p)}{\mu_b^2} = \left(C^2+1\right)\mathbf{E}^2[X] \end{aligned}$$

From the definition of Q-H_2, we have $\frac{p}{\mu_a} = Q\mathbf{E}[X]$ and $\frac{1-p}{\mu_b} = (1-Q)\mathbf{E}[X]$.

$$\begin{aligned} C^2 + 1 &= \frac{2Q^2}{p} + \frac{2(1-Q)^2}{1-p} \\ 0 &= p^2\left(C^2+1\right) - p\left(C^2 - 1 + 4Q\right) + 2Q^2 \\ p &= \frac{C^2 + 4Q - 1 \pm \sqrt{\left(C^2-1\right)^2 + 8\left(C^2-1\right)Q\left(1-Q\right)}}{2\left(C^2+1\right)} \end{aligned}$$

As in Lemma 1, but with $C^2 > 1$, we take the positive root so that $\mu_a > \mu_b$ and verify $0 < p < 1$. □

LEMMA 6. *When $\mathbf{E}[X]$ and Q are constant, as $C^2 \to \infty$, $\mu_a \to \frac{1}{Q\mathbf{E}[X]}$, $\mu_b \to 0$, $p_a \to 1$, and $p_b \to 0$.*

PROOF. From applying L'Hôpital's rule to the equation for p in Lemma 5, $p_a = p \to 1$ as $C^2 \to \infty$. Thus, $p_b = 1 - p \to 0$, $\mu_a = \frac{p_a}{Q\mathbf{E}[X]} \to \frac{1}{Q\mathbf{E}[X]}$, and $\mu_b = \frac{p_b}{(1-Q)\mathbf{E}[X]} \to 0$ as $C^2 \to \infty$. We will also use the following transformation of μ_b in Theorem 6:

$$\begin{aligned} \mu_b &= \frac{C^2 + 4(1-Q) - 1 - \sqrt{\left(C^2-1\right)^2 + 8\left(C^2-1\right)Q\left(1-Q\right)}}{2(1-Q)\left(C^2+1\right)\mathbf{E}[X]} \\ &= \frac{1 - \sqrt{1 + \frac{8\left(C^2-1\right)Q(1-Q) - 4C^2}{\left(C^2+1\right)^2}}}{2(1-Q)\mathbf{E}[X]} \\ &\quad + \frac{2(1-Q) - 1}{(1-Q)\left(C^2+1\right)\mathbf{E}[X]} \end{aligned} \tag{13}$$

And the rest follows. Note the correspondence between the proof of Lemma 6 and Lemma 2, where $\frac{1}{\mu_a}$ in Lemma 6 corresponds to a in Lemma 2, and $\frac{1}{\mu_b}$ in Lemma 6 corresponds to b in Lemma 2. □

LEMMA 8. *For any $\mathbf{E}[X]$, there exists C^* such that, for all $C^2 > C^*$, and $k > \frac{1}{2}$, there exists a k-H_3 distribution with unique parameters μ_a, μ_b, and μ_c, with $0 < \frac{1}{\mu_a} < \frac{1}{\mu_b} < \frac{1}{\mu_c}$ and probabilities $0 < p_a, p_b, p_c < 1$.*

PROOF. We proceed analogously to the k-Trimodal existence proof. Shortly, we will see that the k-H_3 case is almost identical to that case.

Note that since $k > \frac{1}{2}$ and $\frac{1}{\mu_c} = \left(\frac{1}{\mu_b}\right)^{2k}$, setting $\frac{1}{\mu_b} > 1$ guarantees $\frac{1}{\mu_c} > \frac{1}{\mu_b}$. We will start by setting $\frac{1}{\mu_b} > 2^{\frac{2}{3}}$, which guarantees that $0 < p_a, p_b, p_c < 1$. What remains is to prove the existence of $0 < \frac{1}{\mu_a} < \frac{1}{\mu_b}$ satisfying the desired $\mathbf{E}[X]$ and C^2. Our first constraint for μ_a and μ_b comes from $\mathbf{E}[X]$:

$$\mathbf{E}[X] = p_a \cdot \frac{1}{\mu_a} + p_b \cdot \frac{1}{\mu_b} + p_c \cdot \frac{1}{\mu_c} \tag{14}$$

$$\frac{1}{\mu_a} = \frac{\mathbf{E}[X] - \left(\frac{1}{\mu_b}\right)^{-\frac{1}{2}} - \left(\frac{1}{\mu_b}\right)^{-k}}{1 - \left(\frac{1}{\mu_b}\right)^{-\frac{3}{2}} - \left(\frac{1}{\mu_b}\right)^{-3k}} \tag{15}$$

Since we require $\frac{1}{\mu_a} > 0$, and since the denominator of $\frac{1}{\mu_a}$ is simply p_a which is positive, we need the numerator of (15) to be positive. This is satisfied if $\frac{1}{\mu_b} > \frac{4}{\mathbf{E}^2[X]}$. For convenience, we define

$$\left(\frac{1}{\mu_b}\right)_{lb} = \max\left\{2^{\frac{2}{3}}, \frac{4}{\mathbf{E}^2[X]}\right\}$$

where we seek $\frac{1}{\mu_b} > \left(\frac{1}{\mu_b}\right)_{lb}$.

Our second constraint comes from C^2, substituting for $\frac{1}{\mu_a}$ from (15):

$$C^2 = \frac{p_a \cdot \frac{2}{\mu_a^2} + p_b \cdot \frac{2}{\mu_b^2} + p_c \cdot \frac{2}{\mu_c^2}}{\mathbf{E}^2[X]} - 1 \tag{16}$$

$$\sim 2\frac{\left(\frac{1}{\mu_b}\right)^{\frac{1}{2}} + \left(\frac{1}{\mu_b}\right)^{k}}{\mathbf{E}^2[X]} + 1 \left(\text{as } \left(\frac{1}{\mu_b}\right) \to \infty\right) \tag{17}$$

At this point Equation (16), after straightforward substitution, gives us an equation specifying $\left(\frac{1}{\mu_b}\right)$. Note that this equation, however, is a near replica of equation (10), except for a factor of 2 on one of the terms, and b replaced with $\frac{1}{\mu_b}$. The factor of 2 in the derivation of C^* and $\left(\frac{1}{\mu_b}\right)^*$ do not affect the sign of the derivative of $C^2(\mu_b)$, thus the derivative will again be positive, provided that $\frac{1}{\mu_b} > \left(\frac{1}{\mu_b}\right)^*$, given by

$$\left(\frac{1}{\mu_b}\right)^* = \max\left\{\mathbf{E}[X], \left(\frac{1}{\mu_b}\right)_{lb}, \left(2N_2^*\right)^2\right\}$$

Observe that $\mathbf{E}[X] > \frac{1}{\mu_a}$ for $k > \frac{1}{2}$ because $\frac{1}{\mu_a}$ is the smallest mean of the component exponential distribution, and $p_a < 1$, and because $\frac{1}{\mu_c} > \frac{1}{\mu_b} > \left(\frac{1}{\mu_b}\right)^* > \mathbf{E}[X] > \frac{1}{\mu_a}$. Hence $\frac{1}{\mu_b} > \frac{1}{\mu_a}$.

Thus, for all $\frac{1}{\mu_b} > \left(\frac{1}{\mu_b}\right)^*$, $C^2(\mu_b)$ increases monotonically with $\frac{1}{\mu_b}$, and for any $C^2 > C^* = C^2(\mu_b^*)$, a unique value of $\frac{1}{\mu_b} > \frac{1}{\mu_a}$ exists. □