Probability in the Engineering and Informational Sciences, **24**, 2010, 219–244.
doi:10.1017/S0269964809990246

TO BALANCE OR UNBALANCE LOAD IN SIZE-INTERVAL TASK ALLOCATION

Mor Harchol-Balter
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
E-mail: harchol@cs.cmu.edu

Rein Vesilo
Department of Physics and Engineering
Macquarie University
Sydney, Australia
E-mail: rein@science.mq.edu.au

Server farms, consisting of a collection of hosts and a front-end router that dispatches incoming jobs to hosts, are now commonplace. It is well known that when job service requirements (job sizes) are highly variable, then the `Size-Interval` task assignment policy is an excellent rule for assigning jobs to hosts, since it provides isolation for short jobs by directing short jobs to one host's queue and long jobs to another host's queue. What is *not* understood is how to classify a "short" job versus a "long" job. For a long time it was believed that the size cutoff separating "short" jobs from "long" ones should be chosen to *balance* the load at the hosts in the server farm. However, recent literature has provided empirical evidence that load balancing is not always optimal for minimizing mean response time. This article provides the first *analytical criteria* for when it is preferable to *unbalance* load between two hosts using `Size-Interval` task assignment and in which direction the load should be unbalanced. Some very simple sufficient criteria are provided under which we prove that the short job host should be underloaded, and likewise for the long job host. These criteria are then used to prove that the direction of load imbalance depends on *moment index properties* related to the job size distribution. For example, under the Bounded Pareto (BP) job size distribution with parameter α and a sufficiently high upper bound (the BP is well known to be a good model of empirical computer system workloads), we show that α determines the direction of load imbalance. For $\alpha < 1$, the short job host should be underloaded; for $\alpha = 1$, load should be balanced; and for $\alpha > 1$, the long job host should be underloaded. Many other job size distributions are considered as well. We end by showing that load unbalancing can have a dramatic impact on

 0269-9648/10 $25.00

performance, reducing mean response time by an order of magnitude compared to load balancing in many common cases.

1. INTRODUCTION

Server farm architectures are ubiquitous in computing and manufacturing systems because they are inexpensive (many slow servers are less expensive than one fast one) and because they are easily scalable (it is easy to add servers and to take them away). A server farm typically consists of a front-end router that receives all incoming jobs (a.k.a., tasks) and a collection of servers (a.k.a., hosts), to which the router dispatches incoming jobs, as shown in Figure 1. The router follows a *task assignment policy*, which is a rule (a.k.a., algorithm) for assigning each incoming job to one of the hosts. A typical goal of the task assignment policy is to minimize the overall mean response time, where the *response time* of a job is the time from when it arrives until it has completed service and the *mean response time* is the average per-job response time.

In the model shown in Figure 1, the jobs assigned to each host are run in first come–first served (FCFS) order at that host. This model is consistent with service at a supercomputing center, where preempting jobs is very expensive (see Harchol-Balter [11]). It is also consistent with manufacturing centers, where jobs might represent customers or products requiring service, or any other setting where "jobs" (customers) cannot easily be preempted.

Our analysis will assume that the job sizes are independently and identically distributed according to a general bounded job size distribution and that the arrival process of jobs to the server farm is a Poisson process. We assume that the job size distribution is bounded, so that its moments are finite and, hence, mean response time is well defined. We will primarily be interested in the case in which job sizes (service requirements) are highly variable (although most of our theorems will apply to general job size distributions), in accordance with job sizes in computing environments.

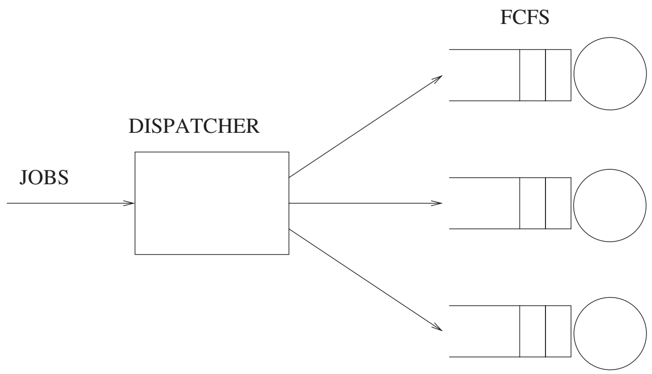

FIGURE 1. Server farm model with three hosts.

Although our results hold for general distributions, we will focus special attention on the Pareto and bounded Pareto distributions, since these are typical in computing workloads. A random variable, X, is distributed Pareto(α), where $\alpha > 0$ (it is most common that $0 < \alpha < 2$, since that is the infinite variance case), if

$$\Pr(X > t) = \left(\frac{t}{s}\right)^{-\alpha}, \qquad t \geq s.$$

(Here, s represents the smallest possible value.) The *bounded Pareto* distribution (Harchol-Balter, Crovella, and Murta [12]) has the same shape as the Pareto, except that there is a maximum upper bound on job sizes: b (for "biggest"). The Pareto and bounded Pareto distributions have been shown to well characterize Unix CPU requirements (Harchol-Balter and Downey [13]), sizes of files requested at web sites (Crovella and Bestavros [5], Crovella, Taqqu, and Bestavros [7], Riska, Smirni, and Ciardo [17]), and sizes of files in FTP transfers (Paxson and Floyd [16]).

The question of finding a good task assignment policy for server farms, so as to minimize overall mean response time, has received significant attention in the literature (Crovella, Harchol-Balter, and Murta [6], Harchol-Balter [11], Harchol-Balter [12], Schroeder and Harchol-Balter [19], Ciardo, Riska, and Smirni [4], Riska et al. [17], Tari, Broberg, Zomaya, and Baldoni [21], Ungureanu, Bradford, Katehakis, and Melamed [22]). For the context of our model, research advocates using `Size-Interval` task assignment, which assigns "short" jobs to one host, "medium" jobs to another host, "long" jobs to the third host, and so on. In Harchol-Balter [11] and Harchol-Balter et al. [12], the authors showed that when the job size distribution is highly variable, as is the case for the bounded Pareto distribution, `Size-Interval` task assignment can reduce the mean response time by orders of magnitude, compared with other common task assignment policies, like `Join-the-Shortest-Queue` (which assigns each incoming job to the host with the fewest *number* of jobs) and `Least-Work-Left` (which assigns each incoming job to the host with the least total work, where the work at a host is the sum of the sizes of jobs queued at that host). The advantage of using the `Size-Interval` policy, in the case of a highly variable job size distribution, is that it isolates short jobs from long ones, preventing short jobs from getting stuck behind long jobs and thereby greatly reducing mean response time.

Harchol-Balter et al. [12] proposed a `Size-Interval` task assignment policy called `SITA-E` with size cutoffs chosen to *equalize* ("E") load between the hosts. These equal-load cutoffs are derived in closed form for the case of the bounded Pareto job size distribution. Variations on `SITA-E` have also been advocated. The `Equiload` policy developed by Ciardo et al. [4] aimed to balance the load in different size intervals, estimating the job size distribution from previously received job sizes. The `Adaptload` policy developed by Riska et al. [17] is based on `Equiload`, but where the job size distribution is now known and the size-cutoff points dynamically change during run time. The `Least Flow-Time` algorithm developed by Tari et al. [21] uses multiple queues in the hosts for different size bands, where the front-end router forwards jobs to the "fittest" server in terms of lighter load and higher processing capacity.

Although the `Size-Interval` task assignment policy is by now well accepted, there are some important open questions. Most importantly,

> Should the *size cutoff* be chosen to balance the load between the hosts or not? And if not, then should the "small" job host be underloaded? Or, should the "large" job host be underloaded?

Although much of the literature points to choosing a size cutoff that *balances* the load between the hosts (see Harchol-Balter et al. [12], Ungureanu et al. [22], Shin and Hou [20], Cardellini, Colajanni, and Yu [3], Hwang and Jung [15]), there is also some work that suggests that it may be better *not* to balance the load (e.g. Crovella, Harchol-Balter, and Murta [6], Harchol-Balter [11], Schroeder and Harchol-Balter [19]). However, there is very little work on providing an analytical criterion for when it is preferable to balance the load and when it is preferable to unbalance the load under `Size-Interval` task assignment, even in the case of just two hosts.

Part of the difficulty is that, even in our simple model, it is very difficult to analytically determine the optimal size cutoff. The case of bounded Pareto job size distributions has been the most tractable to date. The main contributions to the analysis of the bounded Pareto distribution have been by Bachmat and Sarfati [1,2] and by Vesilo [23]. In Bachmat and Sarfati [1,2], the authors derived asymptotic expressions for the optimal size cutoffs with multiple hosts, which might be heterogeneous, both when the maximum job size increases and when the number of hosts increases. This analysis uses the concept of duality, first introduced by Feng, Misra, and Rubenstein [10], which relates results for the bounded Pareto distribution with tail index $2 - \alpha$ to results derived for the bounded Pareto distribution with tail index α. In Vesilo [23], the author derived asymptotic expressions for the proportion of load directed toward a host at the optimal load point for a two-host system, with bounded Pareto job size distribution.

This article differs from the above prior work in that we prove theorems for general job size distributions rather than restricting our attention to the bounded Pareto. For distributions other than the bounded Pareto distribution, Mathematica has been used to iteratively try out different size cutoffs (see Harchol-Balter [11]); however, this gives little insight into how size cutoffs depend on the underlying job size distribution. Certain rules of thumb for choosing the cutoff have been proposed for supercomputing workloads (Schroeder and Harchol-Balter [19]). The closest thing to an analytical approximation, in the general case, comes from Riska et al. [17], who decided on size cutoffs by approximating the job size distribution for a multiphase hyperexponential distribution. Rather than trying to determine the exact size cutoff, we instead look at the higher-level question of whether load should be balanced or unbalanced, so as to minimize mean response time.

In this article we present theorems for `Size-Interval` task assignment with two hosts and general job size distributions, stating under what conditions load should be unbalanced and whether the unbalancing should favor the small job host or the large job host. Below we provide an outline of the results in this article.

Section 3 provides simple sufficient conditions for load unbalancing, described in Theorem 3.1. These *imbalance conditions* consist of two tests involving *truncated normalized moments* of the job size distribution, whereby if both tests are satisfied, then load should be unbalanced in favor of the lighter load for the short job host, and if both tests fail, then load should be unbalanced in favor of the long job host. Although Theorem 3.1 leaves some unknown territory, we show that it is very powerful in that it allows us to analyze common job size distributions. In particular, in Section 3.2 we apply Theorem 3.1 to the bounded Pareto distribution to evaluate the parameters under which load should be unbalanced and in which direction. Theorem 3.1 is also useful in helping to prove the later theorems in Section 4.

The proof of Theorem 3.1 is also interesting in its own right. It begins by forming a *low-load approximation* for the mean response time in the original problem. This low-load approximation is far simpler than the expression for mean response time in the original problem. Although assuming low load might seem limiting, we prove that the direction of imbalance for the original problem is the same as that in the low-load approximation. Hence, it suffices to derive imbalance conditions for the low-load approximation, which is the approach we follow in proving Theorem 3.1.

In Section 4 we look at an even simpler criterion for load unbalancing, based on simply evaluating the *moment index* of the *extended job size distribution*. The extended job size distribution is an infinite-support version of the original (bounded) job size distribution: Specifically, the original job size distribution, having upper bound b, is equivalent to the extended job size distribution conditioned on the extended distribution being less than b. For example, if the job size distribution is a bounded Pareto, then the corresponding extended job size distribution is the (unbounded) Pareto. Our results are stated in terms of the moment index of the extended job size distribution. For a random variable, Z, the moment index, κ_Z, is defined by (see Daley [8])

$$\kappa_Z = \sup\{r > 0 : \mathrm{E}[Z^r] < \infty\}. \qquad \textbf{(1)}$$

(For example, for an unbounded Pareto distribution with index, α, the moment index equals α.) We present two theorems:

1. (Theorem 4.1) If the first moment of the extended job size distribution is finite and the second moment is infinite, then, for job size distributions with large enough b, the large host should be underloaded.
2. (Theorem 4.2) If the moment index of the extended job size distribution satisfies $0 < \kappa_Z < 1$ and the density function of the extended distribution satisfies some additional regularity conditions (given later), then, for job size distributions with large enough b, the small host should be underloaded.

The proofs of both Theorems 4.1 and Theorem 4.2 use the earlier Theorem 3.1.

In Section 4.2 we apply Theorems 4.1 and 4.2 to fully classify a range of job size distributions that have been shown to be good empirical models of computer system workloads. Table 1 lists these distributions and their moment index. Table 2 then describes the condition under which the small job host should be underloaded for each of these distributions. It is interesting to note that the direction of load imbalance is, in many cases, simply determined by the tail parameter α and can flip as α goes from low to high.

Finally, in Section 5 we consider the detrimental effect on performance that comes from unwittingly balancing load, when one should be unbalancing load. We find that

TABLE 1. Extended Distributions with Heavy Tails

Distribution	$1 - R(t)$	r(t)	Parameters	Support	Moment Index
Pareto (I)	$(t/s)^{-\alpha}$	$\frac{\alpha}{s}(t/s)^{-\alpha-1}$	$s > 0, \alpha > 0$	(s, ∞)	α
Pareto (II)	$(1 + (t/s))^{-\alpha}$	$\frac{\alpha}{s}(1 + (t/s))^{-\alpha-1}$	$s > 0, \alpha > 0$	$(0, \infty)$	α
Burr	$\left(\frac{\kappa}{\kappa + t^\tau}\right)^\alpha$	$\frac{\alpha\tau}{\kappa}t^{\tau-1}\left(\frac{\kappa}{\kappa + t^\tau}\right)^{\alpha+1}$	$\alpha, \kappa, \tau > 0$	$(0, \infty)$	$\tau\alpha$
Log-Gamma		$\frac{\alpha^\beta}{\Gamma(\beta)}(\log t)^{\beta-1}t^{-\alpha-1}$	$\alpha, \beta > 0$	$(1, \infty)$	α
Truncated α-stable	$\Pr(\lvert Z\rvert > t)$, where Z is an α-stable random variable		$0 < \alpha < 2$	$(0, \infty)$	α
Regularly varying with index of variation $-\alpha$ (*)	$\Pr(Z > t)$, where Z is regularly varying with index of variation $-\alpha$		α	$(0, \infty)$	α

TABLE 2. Corresponding Bounded Job Size Distributions. Sufficient Conditions for Underloading the Small/Large Host

Distribution	Underload Small Host	Underload Large Host
Bounded Pareto (I)	$0 < \alpha < 1$	$1 < \alpha < 2$
Bounded Pareto (II)	$0 < \alpha < 1$	$1 < \alpha < 2$
Bounded Burr	$0 < \alpha\tau < 1$	$1 < \alpha\tau < 2$
Bounded log-Gamma	$0 < \alpha < 1$	$1 < \alpha < 2$
Bounded truncated α-stable	$0 < \alpha < 1$	$1 < \alpha < 2$
Bounded regularly varying with index of variation $-\alpha$ (*)	$0 < \alpha < 1$ and (**)	$1 < \alpha < 2$

there can be orders of magnitude difference in mean response time between balancing and unbalancing load. This underscores the importance of our problem.

2. FORMAL PROBLEM STATEMENT

Jobs arrive into the system according to a Poisson process of rate λ and job sizes are independent of each other and the arrival process. Let the service time of a generic job be represented by the random variable X. It is assumed that job sizes are defined by an absolutely continuous distribution function, $F(t)$, with continuous density, $f(t)$. It is further assumed that $f(t)$ has support[1] (s,b), where $0 \le s < b \le \infty$.

The size-cutoff value used in the system is denoted by c, with jobs smaller than c directed to the small host and jobs larger than c directed to the large host. The full ith moment of the service time (M_i), truncated ith moment ($M_i(c)$) and the normalized truncated ith moment ($m_i(c)$) are defined as follows, for $i = 1, 2$ and $s \le c \le b$:

$$M_1 = \int_s^\infty tf(t)\,dt, \qquad M_2 = \int_s^\infty t^2 f(t)\,dt,$$

$$M_1(c) = \int_s^c tf(t)\,dt, \qquad M_2(c) = \int_s^c t^2 f(t)\,dt,$$

$$m_1(c) = \frac{M_1(c)}{M_1} = \frac{1}{M_1}\int_s^c tf(t)\,dt, \qquad m_2(c) = \frac{M_2(c)}{M_2} = \frac{1}{M_2}\int_s^c t^2 f(t)\,dt. \quad \textbf{(2)}$$

Define the full and truncated loads on the small host, respectively, by

$$\rho = \lambda M_1 = \lambda \int_s^\infty tf(t)\,dt \quad \text{and} \quad \rho(c) = \lambda M_1(c) = \lambda \int_s^c tf(t)\,dt = \rho m_1(c).$$

Observe that $0 < \rho < 2$. Define the normalized truncated load on the small host by $q(c) \equiv \rho(c)/\rho$.

Because $q(c)$ and $m_2(c)$ are defined as normalized functions, they satisfy $0 \le q(c) \le 1$ and $0 \le m_2(c) \le 1$. Observe that $q(c) < \frac{1}{2}$ whenever the cutoff c is chosen such that the small host is underloaded, whereas $q(c) > \frac{1}{2}$ means that the cutoff is chosen so that the large host is underloaded.

Let T_Q be the waiting time of a job and $\mathrm{E}[T_Q]$ be the expected waiting time. By splitting the arrival stream of jobs according to size, which are independent of arrival times, the arrival streams to each of the two hosts is Poisson. Denote corresponding quantities for the small and large host by the suffixes S and L, respectively. The system then consists of two $M/G/1$ queues, with the expected waiting time in each case given by the Pollaczek–Khinchine formula:

$$\mathrm{E}[T_Q^i] = \frac{\lambda_i \mathrm{E}[X_i^2]}{2(1-\rho_i)}, \qquad i = S, L, \quad \textbf{(3)}$$

where X_i is the job size going to queue i, λ_i is the arrival rate for queue i, and ρ_i is the load on host i. The expected waiting time for the entire system is obtained by conditioning on the size of jobs:

$$\mathrm{E}[T_Q] = \Pr(X \leq c)\mathrm{E}[T_Q^S] + \Pr(X > c)\mathrm{E}[T_Q^L] = F(c)\mathrm{E}[T_Q^S] + \overline{F}(c)\mathrm{E}[T_Q^L].$$

After some manipulation, this equation can be expressed as

$$\mathrm{E}[T_Q] = \frac{M_2\lambda}{2\rho}h(c), \tag{4}$$

where

$$h(c) = \frac{m_2(c)F(c)}{a+1-q(c)} + \frac{(1-m_2(c))(1-F(c))}{a+q(c)}, \tag{5}$$

with

$$a = \frac{1-\rho}{\rho}. \tag{6}$$

Observe that $a > -\frac{1}{2}$. We will use this fact later.

The above quantities are expressed in terms of the cutoff parameter c (e.g., $h(c)$, $m_2(c)$, $F(c)$, etc.). There are times when we will wish to write these quantities instead as a function of the normalized truncated load on the small host, $q(c)$. We define $\tilde{h}(q(c)) \equiv h(c)$ and likewise for $\tilde{m}(q(c))$ and other functions. We then abuse notation a bit by simply writing $\tilde{h}(q)$ and likewise for the other functions, where the c is implicit.[2]

Using these definitions of $\tilde{F}(q)$, $\tilde{m}_2(q)$, and $\tilde{h}(q)$, Eq. (5) can be transformed into

$$\tilde{h}(q) = \frac{\tilde{m}_2(q)\tilde{F}(q)}{a+1-q} + \frac{(1-\tilde{m}_2(q))(1-\tilde{F}(q))}{a+q}. \tag{7}$$

Equation (7) is the function that we are trying to minimize in order to minimize the expected waiting time. We refer to Eq. (7) as the *original minimization problem.*

LEMMA 2.1: *The function $\tilde{h}(q)$ has a unique minimum point. We denote this point by q^* and let the cutoff value corresponding to q^* be c^*.*

PROOF: By taking derivatives of $\tilde{F}(q)$ and $\tilde{m}_2(q)$ and applying the formulas given in Appendix A, it can be shown that $\tilde{m}_2(q)$ and $\tilde{F}(q)$ are increasing functions of q, that $\tilde{F}(q)$ is concave, and that $\tilde{m}_2(q)$, $\tilde{m}_2(q)\tilde{F}(q)$, $(1-\tilde{F}(q))$, $(1-\tilde{m}_2(q))(1-\tilde{F}(q))$, $\tilde{m}_2(q)\tilde{F}(q)/(a+1-q)$, and $(1-\tilde{m}_2(q))(1-\tilde{F}(q))/(a+q)$ are all convex (we skip this straightforward proof). Using these results and the fact that $m_2(0) = F(0) = 0$ and $m_2(1) = \tilde{F}(1) = 1$, it follows that $\tilde{h}(q)$ is the sum of two convex functions, and because $f(t)$ is nonzero in its region of support, there is a unique minimum point. ■

For finite mean waiting times, the denominators in both terms on the right in Eq. (7) must be positive. Hence, for stability, we must have $a+1-q > 0$ and

$a+q>0$. This implies, if $0 \le \rho < 1$, that q must satisfy $0 \le q \le 1$, and if $1 \le \rho < 2$, that q must satisfy $1 - 1/\rho < q < 1/\rho$.

The obvious approach for finding the minimum point of $\tilde{h}(q)$ is to take its derivative and set it to zero. This gives

$$0 = \tilde{h}'(q) = \frac{\tilde{m}_2(q)\tilde{F}(q)}{(a+1-q)^2} + \frac{\tilde{m}_2'(q)\tilde{F}(q)}{a+1-q} + \frac{\tilde{m}_2(q)\tilde{F}'(q)}{a+1-q} - \frac{(1-\tilde{m}_2(q))(1-\tilde{F}(q))}{(a+q)^2} - \frac{\tilde{m}_2'(q)(1-\tilde{F}(q))}{a+q} - \frac{(1-\tilde{m}_2(q))\tilde{F}'(q)}{a+q}. \quad \textbf{(8)}$$

Unfortunately, solving Eq. (8) does not seem tractable.

3. CONDITIONS FOR UNBALANCED LOAD

The question of which host to underload can be determined by examining some simple sufficient imbalance conditions that are defined in terms of the truncated moment functions of the job size distribution and the derivatives of the truncated moment functions, all evaluated at $q = \frac{1}{2}$.

THEOREM 3.1: *If*

$$\tilde{m}_2(1/2) - (1 - \tilde{F}(1/2)) > 0, \quad \textbf{(9)}$$

$$\tilde{m}_2'(1/2) - \tilde{F}'(1/2) > 0, \quad \textbf{(10)}$$

then $q^* < \frac{1}{2}$ *and, hence, the small job host should be underloaded. This result is also true if all the inequalities are reversed, with the consequence being that* $q^* > \frac{1}{2}$, *so that the large job host is underloaded.*

In Section 3.1 we prove Theorem 3.1 and in Section 3.2 we provide an example of how the theorem can be applied when the job size distribution is Bounded Pareto.

3.1. Proof of Theorem 3.1

In order to prove Theorem 3.1, we introduce the low-load approximation, which approximates the original system for $\rho \approx 0$. In the low-load approximation, we have, from Eq. (6), $a \approx 1/\rho$, $1/(a+1-q) \approx \rho$ and $1/(a+q) \approx \rho$. Applying these approximations to Eq. (7), the equation to be minimized becomes, in the low-load approximation,

$$\tilde{h}_0(q) = \rho\left[\tilde{m}_2(q)\tilde{F}(q) + (1 - \tilde{m}_2(q))(1 - \tilde{F}(q))\right].$$

From here on, we will ignore the ρ factor (since it does not affect the minimization point) and simply define

$$\tilde{h}_0(q) \equiv \tilde{m}_2(q)\tilde{F}(q) + (1 - \tilde{m}_2(q))(1 - \tilde{F}(q))\,, \quad \textbf{(11)}$$

where the subscript 0 denotes low load. This is called the *low-load approximation* and the optimal load point for the low-load approximation is denoted by q_0^*. Observe that $\tilde{h}_0(q)$ is a convex function, as was $\tilde{h}(q)$, using the same arguments as in the proof of Lemma 2.1.

The key idea in the proof of Theorem 3.1 is that we will prove that, assuming (9) and (10),

$$q^* < \frac{1}{2} \iff q_0^* < \frac{1}{2} \quad \textbf{(12)}$$

Hence, it suffices to determine the direction of load imbalance in the low-load approximation, and that gives us the direction of load imbalance for the original problem.

PROOF OF THEOREM 3.1: We first show that the optimal load point of the low-load approximation, q_0^*, in Eq. (11) satisfies $q_0^* < 1/2$.

To prove this, differentiate Eq. (11) to get

$$\tilde{h}_0'(q) = \tilde{m}_2'(q)(2\tilde{F}(q) - 1) + \tilde{F}'(q)(2\tilde{m}_2(q) - 1).$$

Now, set $q = 1/2$ in this equation. From Appendix A, $\tilde{F}(1/2) \geq 1/2$. Furthermore, by inequality (10), $\tilde{m}_2'(1/2) > \tilde{F}'(1/2)$, and by Appendix A, $\tilde{F}'(1/2) > 0$; hence,

$$\begin{aligned}\tilde{h}_0'(1/2) &\geq \tilde{F}'(1/2)(2\tilde{F}(1/2) - 1) + \tilde{F}'(1/2)(2\tilde{m}_2(1/2) - 1)\\ &= 2\tilde{F}'(1/2)(\tilde{m}_2(1/2) - (1 - \tilde{F}(1/2))).\end{aligned} \quad \textbf{(13)}$$

By inequality (9), $\tilde{m}_2(1/2) - (1 - \tilde{F}(1/2)) > 0$. Thus, Eq. (13) implies that

$$\tilde{h}_0'(1/2) > 0. \quad \textbf{(14)}$$

Finally, since $\tilde{h}_0(q)$ is a convex function, $\tilde{h}_0'(q)$ is nondecreasing over the domain. Since $\tilde{h}_0'(q)$ is furthermore continuous, we have that the minimum value of $\tilde{h}_0(q)$ occurs at $q = q_0^* < \frac{1}{2}$.

Knowing that $q_0^* < 1/2$, we now show that the optimal load point of the original system, q^*, satisfies $q^* \leq 1/2$.

To simplify notation, define $\tilde{g}_0(q) = \tilde{m}_2(q)\tilde{F}(q)$ and $\tilde{k}_0(q) = (1 - \tilde{F}(q))(1 - \tilde{m}_2(q))$. Using these definitions, moment condition (9), $\tilde{m}_2(1/2) - (1 - \tilde{F}(1/2)) \geq 0$, can be expressed as

$$\tilde{g}_0(1/2) \geq \tilde{k}_0(1/2). \tag{15}$$

Using these definitions of $\tilde{g}_0(q)$ and $\tilde{k}_0(q)$, the low-load approximation (Eq. (11)) can be expressed as

$$\tilde{h}_0(q) = \tilde{g}_0(q) + \tilde{k}_0(q), \tag{16}$$

and the original system equation (Eq. (7)) can be expressed as

$$\tilde{h}(q) = \frac{\tilde{g}_0(q)}{a+1-q} + \frac{\tilde{k}_0(q)}{a+q}. \tag{17}$$

Differentiating this equation gives

$$\tilde{h}'(q) = \frac{\tilde{g}_0'(q)}{a+1-q} + \frac{\tilde{g}_0(q)}{(a+1-q)^2} + \frac{\tilde{k}_0'(q)}{a+q} - \frac{\tilde{k}_0(q)}{(a+q)^2}. \tag{18}$$

Now, setting $q = \frac{1}{2}$, all denominators become $a + \frac{1}{2}$, which is positive, since $a > -\frac{1}{2}$. Now since

$$\tilde{g}_0(1/2) - \tilde{k}_0(1/2) \geq 0 \quad \text{by (15)},$$
$$\tilde{g}_0'(1/2) + \tilde{k}_0'(1/2) = \tilde{h}_0'(1/2) > 0 \quad \text{by (14)},$$

we see from Eq. (18) that

$$\tilde{h}'(1/2) > 0. \tag{19}$$

Thus, it follows as in Eq. (14) (by convexity arguments) that $q^* < 1/2$. The proof with inequalities reversed is similar (hence the if and only if in relation (12). ■

3.2. Bounded Pareto Example

We now apply Theorem 3.1 to the case of the bounded Pareto job size distribution, $BP(s, b, \alpha)$. In the bounded Pareto distribution, $BP(s, b, \alpha)$, α is the tail index, s is the minimum job size, and b is the maximum job size. We will show that, for sufficiently large b, if $0 < \alpha < 1$, the small host should be underloaded, and if $1 < \alpha < 2$, the large host should be underloaded (we refer to this as the large b approximation). We will later see that these same results can be obtained from Theorems 4.2 and 4.1. To complete our analysis, we also consider the case $\alpha = 1$ and show that in this case, for sufficiently high b, the load should be equally balanced.

The fact that we require a high b value is not unrealistic. In computing workloads, b is often seven orders of magnitude higher than s (e.g., web file sizes can range from 10B to 100MB).

For $t \in [s, b]$, the bounded Pareto distribution has density function given by

$$f(t) = \frac{\alpha t^{-\alpha-1}}{(1-(b/s)^{-\alpha})s^{-\alpha}}.$$

The density is zero elsewhere. For $s \le c \le b$, the moments and truncated moments of the bounded Pareto distribution are as follows:

$$F(c) = \frac{1-(c/s)^{-\alpha}}{1-(b/s)^{-\alpha}}, \qquad 1 - F(c) = \frac{(c/s)^{-\alpha} - (b/s)^{-\alpha}}{1-(b/s)^{-\alpha}}, \qquad \alpha > 0. \tag{20}$$

For $\alpha \neq 1$,

$$M_1 = \frac{s\alpha(1-(b/s)^{1-\alpha})}{(\alpha-1)(1-(b/s)^{-\alpha})}, \qquad q(c) = \frac{1-(c/s)^{1-\alpha}}{1-(b/s)^{1-\alpha}}. \tag{21}$$

For $\alpha = 1$,

$$M_1 = \frac{s\log(b/s)}{1-(b/s)^{-1}}, \qquad q(c) = \frac{\log(c/s)}{\log(b/s)}. \tag{22}$$

For $\alpha \neq 2$,[3]

$$M_2 = \frac{s^2\alpha(1-(b/s)^{2-\alpha})}{(\alpha-2)(1-(b/s)^{-\alpha})},$$

$$m_2(c) = \frac{1-(c/s)^{2-\alpha}}{1-(b/s)^{2-\alpha}}, \qquad 1 - m_2(c) = \frac{(c/s)^{2-\alpha} - (b/s)^{2-\alpha}}{1-(b/s)^{2-\alpha}}. \tag{23}$$

3.2.1. Large b Approximation. To derive the large b approximation, the following assumptions are made for s and b.

[A1] For $0 < \alpha < 2$, assume

$$(b/s)^{-\alpha} \ll 1, \tag{24}$$

$$(b/s)^{2-\alpha} \gg 1, \tag{25}$$

$$(b/s)^{1-\alpha} \gg 1 \quad (0 < \alpha < 1) \quad \text{and} \quad (b/s)^{1-\alpha} \ll 1 \quad (1 < \alpha < 2). \tag{26}$$

Applying assumptions **A1**, the truncated moment functions, Eqs. (20) and (23), become, for $0 < \alpha < 1$ and $1 < \alpha < 2$,

$$1 - F(c) \approx (c/s)^{-\alpha} - (b/s)^{-\alpha}, \tag{27}$$

$$q(c) \approx \begin{cases} \dfrac{(c/s)^{1-\alpha} - 1}{(b/s)^{1-\alpha}}, & 0 < \alpha < 1 \\ 1 - (c/s)^{1-\alpha}, & 1 < \alpha < 2, \end{cases} \tag{28}$$

$$m_2(c) \approx \frac{(c/s)^{2-\alpha} - 1}{(b/s)^{2-\alpha}}. \tag{29}$$

We, now, proceed to analyze different cases of α.

3.2.2. Bounded Pareto: $0 < \alpha < 1$. From Eq. (28), the large b approximation for c/s is given by

$$c/s \approx (1 + q(c)(b/s)^{1-\alpha})^{1/(1-\alpha)} = (b/s)\left((b/s)^{\alpha-1} + q(c)\right)^{1/(1-\alpha)}. \tag{30}$$

Inserting this equation into Eqs. (27) and (29) respectively, we get the following approximations for the truncated moment functions:

$$1 - \tilde{F}(q) \approx (b/s)^{-\alpha}\left((b/s)^{\alpha-1} + q\right)^{-\alpha/(1-\alpha)} - (b/s)^{-\alpha},$$
$$\tilde{m}_2(q) \approx \left((b/s)^{\alpha-1} + q\right)^{(2-\alpha)/(1-\alpha)} - (b/s)^{\alpha-2}.$$

From Appendix A we have that $\tilde{F}'(q) = M_1/c$ and $\tilde{m}'(q) = cM_1/M_2$. In addition, under assumptions **A1**, $M_1 \approx s\alpha(b/s)^{1-\alpha}/(1-\alpha)$ and $M_2 \approx s^2\alpha(b/s)^{2-\alpha}/(2-\alpha)$. These results together with Eq. (30) give

$$\tilde{F}'(q) \approx \left(\frac{\alpha}{1-\alpha}\right)(b/s)^{-\alpha}\left((b/s)^{\alpha-1} + q\right)^{-1/(1-\alpha)},$$
$$\tilde{m}_2'(q) \approx \left(\frac{2-\alpha}{1-\alpha}\right)\left((b/s)^{\alpha-1} + q\right)^{1/(1-\alpha)}.$$

The imbalance conditions evaluate to (setting $q = 1/2$)

$$\tilde{m}_2(1/2) - (1 - F(1/2)) \approx \left((b/s)^{\alpha-1} + 1/2\right)^{(2-\alpha)/(1-\alpha)} - (b/s)^{-\alpha}\left((b/s)^{\alpha-1} + 1/2\right)^{-\alpha/(1-\alpha)}, \tag{31}$$

$$\tilde{m}_2'(1/2) - F'(1/2) \approx \left(\frac{2-\alpha}{1-\alpha}\right)\left((b/s)^{\alpha-1} + 1/2\right)^{1/(1-\alpha)} - \left(\frac{\alpha}{1-\alpha}\right)(b/s)^{-\alpha}\left((b/s)^{\alpha-1} + 1/2\right)^{-1/(1-\alpha)}. \tag{32}$$

Since $(b/s)^{-\alpha} \to 0$ ($b \to \infty$), the second term on the right in both Eqs. (31) and (32) becomes zero ($b \to \infty$), whereas the first term on the right in Eqs. (31) and (32) remains bounded below by a positive value. Hence, Eqs. (31) and (32) will both be positive for b large enough. Thus, using Theorem 3.1, the small host should be underloaded in the original system, for b large enough.

3.2.3. Bounded Pareto: $1 < \alpha < 2$. From Eq. (28), the large b approximation for c/s is given by

$$c/s \approx (1 - q(c))^{1/(1-\alpha)}. \tag{33}$$

Inserting this equation into Eqs. (27) and (29) respectively, we get the following approximations for the truncated moment functions:

$$1 - \tilde{F}(q) \approx (1-q)^{\alpha/(\alpha-1)} - (b/s)^{-\alpha},$$

$$\tilde{m}_2(q) \approx \frac{(1-q)^{-(2-\alpha)/(\alpha-1)} - 1}{(b/s)^{2-\alpha}}.$$

As earlier, we have from Appendix A that $\tilde{F}'(q) = M_1/c$ and $\tilde{m}'(q) = cM_1/M_2$. In addition, under assumptions **A1**, $M_1 \approx s\alpha/(\alpha-1)$ and $M_2 \approx s^2\alpha(b/s)^{2-\alpha}/(2-\alpha)$. These results together with Eq. (33) give

$$\tilde{F}'(q) \approx \left(\frac{\alpha}{\alpha-1}\right)(1-q)^{1/(\alpha-1)},$$

$$\tilde{m}_2'(q) \approx \left(\frac{2-\alpha}{\alpha-1}\right)(b/s)^{\alpha-2}(1-q)^{-1/(\alpha-1)}.$$

The imbalance conditions evaluate to (setting $q = 1/2$)

$$\tilde{m}_2(1/2) - (1 - F(1/2)) \approx (1/2)^{-(2-\alpha)/(\alpha-1)}(b/s)^{\alpha-2} - (1/2)^{\alpha/(\alpha-1)} - (b/s)^{\alpha-2} + (b/s)^{-\alpha}, \quad \textbf{(34)}$$

$$\tilde{m}_2'(1/2) - F'(1/2) \approx \left(\frac{2-\alpha}{\alpha-1}\right)(1/2)^{-1/(\alpha-1)}(b/s)^{\alpha-2} - \left(\frac{\alpha}{\alpha-1}\right)(1/2)^{1/(\alpha-1)}. \quad \textbf{(35)}$$

Since $(b/s)^{-\alpha}, (b/s)^{\alpha-2} \to 0$ $(b \to \infty)$, the first, third, and fourth terms on the right in Eq. (34) and the first term on the right in Eq. (35) become zero $(b \to \infty)$. On the other hand, the second term on the right in both Eqs. (34) and (35) remains bounded above by a negative value. Hence, Eqs. (34) and (35) will both be negative for b large enough. Thus, using Theorem 3.1, the large host should be underloaded in the original system, for b large enough.

3.2.4. Bounded Pareto: $\alpha = 1$. We will show that if $\alpha = 1$, then the load is balanced at the optimal load point. To show this, we need to make the following additional assumptions:
[A2] For $0 < \alpha < 2$ and c in the region of the optimal point, c^*, assume

$$(b/s)^{-\alpha} \ll (c/s)^{-\alpha} \ll 1, \quad \textbf{(36)}$$

$$(b/s)^{2-\alpha} \gg (c/s)^{2-\alpha} \gg 1. \quad \textbf{(37)}$$

Assumptions A2 cannot be made a priori. We make these assumptions, simplify, and solve the minimization problem, Eq. (7), and then verify that the assumptions are correct.

Applying assumptions A2 to Eqs. (27) and (29), the truncated moment functions become

$$1 - F(c) \approx (c/s)^{-1}, \tag{38}$$

$$m_2(c) \approx \frac{c/s}{b/s}. \tag{39}$$

For $\alpha = 1$, we can compute an explicit expression for $c(q)$ directly from Eq. (22), giving $c(q)/s = (b/s)^q$. Inserting this into Eqs. (38) and (39) gives

$$1 - F(c) \approx (b/s)^{-q} \quad \text{and} \quad \tilde{m}_2(q) \approx (b/s)^{q-1}. \tag{40}$$

Hence, the minimization problem, Eq. (7), becomes

$$\tilde{h}(q) \approx \frac{(b/s)^{q-1}(1-(b/s)^{-q})}{a+1-q} + \frac{(b/s)^{-q}(1-(b/s)^{q-1})}{a+q} \equiv \tilde{g}(1-q) + \tilde{g}(q), \tag{41}$$

where $\tilde{g}(q) = (b/s)^{-q}(1-(b/s)^{-(1-q)})/(a+q)$. By symmetry, $q^* \approx 1/2$ is an optimal point. Substituting $q^* \approx 1/2$ into Eq. (40) gives the optimal cutoff point

$$c^* \approx (bs)^{1/2}. \tag{42}$$

We now verify assumptions A2. At the optimal point, $(b/s)^{-1} \ll (c/s)^{-\alpha} = ((bs)^{1/2}/s)^{-1} = (b/s)^{-1/2} \ll 1$ and $(b/s)^{2-1} = b/s \gg (c/s)^{2-\alpha} = ((bs)^{1/2}/s)^{2-1} = (b/s)^{1/2} \gg 1$. Hence, provided b is large enough, both inequalities (36) and (37) in assumptions A2 are satisfied.

4. HOW THE MOMENT INDEX OF THE EXTENDED JOB SIZE DISTRIBUTION DETERMINES THE DIRECTION OF IMBALANCE

In this section we will specify broad classes of job size distributions under which the small job host should be underloaded and vice versa. To do this, we begin, in Section 4.1, by demonstrating a very simple relationship between the moment index of the extended job size distribution and the direction of load imbalance. In Section 4.2 we then apply this moment index property to a range of distributions.

4.1. Moment Index Property

All job size distributions in this article, $F(t)$, are assumed to be bounded from above by some bounding level b and from below by $s \geq 0$. Recall that it is necessary that the job size distribution be bounded from above, so that its moments are finite, allowing for a finite mean response time.

Let X be a random variable denoting the job size distribution and let $F(t)$ be its distribution function. Corresponding to any such bounded job size distribution, $F(t)$, one can define an *extended distribution* function, $R(t)$, with support (s, ∞) and density $r(t)$, as follows, by means of conditioning: Suppose Z is a random variable that has distribution function $R(t)$, then the random variable, X, is equal to Z conditioned on Z being less than or equal to b.[4] This gives $F(t) = \Pr(X \le t) = \Pr(Z \le t \mid Z \le b) = R(t)/R(b)$. Likewise, the density function of X is given by $f(t) = r(t)/R(b)$.

The results in this section require that the bounding level, b, becomes large. We show in Theorem 4.1 that when the first moment of the extended distribution is finite but the second moment is infinite, the large host should be underloaded. We show in Theorem 4.2 that when the first moment of the extended distribution is infinite and the extended density function satisfies some regularity conditions, the small host should be underloaded.

THEOREM 4.1: *Let X be a random variable denoting job size, with distribution function $F(t)$, where $s < t < b$. Let Z be a random variable denoting the extended job size. Suppose that Z satisfies* $\mathrm{E}[Z] < \infty$ *and* $\mathrm{E}[Z^2] = \infty$ *and that Z has density function $r(t)$ with support (s, ∞). Then, in a server farm with job sizes drawn from X, for b large enough, the large job host should be underloaded (i.e., $q^* > \frac{1}{2}$).*

The proof of this theorem is given in Appendix C.

Theorem 4.2 examines the case when the extended distribution function, $R(t)$, has moment index $0 < \kappa_Z < 1$ (where the random variable Z is distributed according to $R(t)$). However, unlike the case when $\mathrm{E}[Z] < \infty$, where the cutoff, $c_q(b)$, was bounded, in this case, the cutoff, $c_q(b)$, increases without bound as $b \to \infty$. To obtain bounds on $\tilde{m}_2(q, b)$ in this situation, we need to introduce into Theorem 4.2 some extra regularity conditions on $r(t)$.

We note that the condition $0 < \kappa_Z < 1$ places very few restrictions on distribution functions for which $\mathrm{E}[Z] = \infty$. The only cases excluded occur when $\kappa_Z = 0$ or 1. The case $\kappa_Z = 0$ occurs if $\mathrm{E}[Z^\delta] = \infty$, for any $\delta > 0$, and so, there are extremely heavy tails. We do not consider this case in the article. The case $\kappa_Z = 1$ and $\mathrm{E}[Z] = \infty$ occurs if $\mathrm{E}[Z^\gamma] < \infty$, for any $0 \le \gamma < 1$, but $\mathrm{E}[Z] = \infty$. We do not consider this case either in the article. The other way $\kappa_Z = 1$ can occur is if $\mathrm{E}[Z^\gamma] < \infty$ $(0 \le \gamma \le 1)$ and $\mathrm{E}[Z^\delta] > \infty$ $(\delta > 1)$, but this case is included in Theorem 4.1.

THEOREM 4.2: *Suppose the random variable Z has density function $r(t)$ with support (s, ∞) and*

(a) the moment index of Z satisfies $0 < \kappa_Z < 1$,

(b) $r(t)$ is ultimately a decreasing function of t,

(c) $r'(t)$ exists and is continuous for $t \ge s$,

(d) $\lim_{t\to\infty} -tr'(t)/r(t)$ *exists (or, equivalently,* $\lim_{t\to\infty} -t(d/dt) \log r(t)$ *exists).*

Then, for b large enough, the optimal load point in the original system optimal, q^*, is less than or equal to $1/2$.

Conditions (b), (c), and (d) are introduced to make $f(t)$ regular enough so that a lower bound on $\tilde{m}_2(1/2)$ can be found. The proof of this theorem is given in Appendix D.

4.2. Application of Moment Index Theorems to Distributions

In this subsection we consider several distributions,[5] listed in Table 1, which have been the focus of many Internet traffic studies and computer workload characterizations. The distributions we consider all have a tail index α between zero and 2 and have the property that their moment index depends on α. By applying the theorems in Section 4.1 to the extended distributions in Table 1, we can determine the direction of imbalance for the bounded versions of these distributions, which we list in Table 2, assuming that the bounding level, b, is sufficiently high.

The condition (*) for regularly varying distributions denotes that the result is limited to those distribution functions satisfying regularity conditions (b), (c), and (d) in Theorem 4.2. The truncated α-stable distribution is a special case of a regularly varying distribution and the regularity conditions (b), (c), and (d) in Theorem 4.2 are shown to be satisfied in Harchol-Balter and Vesilo [14].

5. NUMERICAL RESULTS: THE BENEFITS OF UNBALANCING LOAD

This section uses numerical studies, in the case of the bounded Pareto distribution, to compare the performance of a balanced system with one that is optimally loaded. After applying our theorems to determine the optimal direction of load imbalance, we use Newtonian search to iterate to find the optimal size-cutoff point.

In all cases, the mean, M_1, of the distribution was set to 3000 and the ratio b/s was set to the value $10^{14}/9$ (these parameter values correspond to those used in the example given in Harchol-Balter et al. [12]). The lower limit of the distribution, s, was calculated knowing the values of M_1 and b/s and using Eq. (21) for $\alpha \neq 1$ and Eq. (22) for $\alpha = 1$. The upper limit b was then computed knowing s and b/s.

Table 3 compares the optimal value of the expected waiting time, $\mathrm{E}[T_Q(q^*)]$, with the value of the expected waiting time, $\mathrm{E}[T_Q(1/2)]$, evaluated at the equally loaded point as α varies, for very light load ($\rho = 0.01$). For $\alpha = 1$, the optimal value of $\mathrm{E}[T_Q(q^*)]$ equals $\mathrm{E}[T_Q(1/2)]$, since, in this case, the optimal load point is at $q^* = 1/2$. For other values of α, $\mathrm{E}[T_Q(1/2)]$ is orders of magnitude larger than the optimal value $\mathrm{E}[T_Q(q^*)]$. For $\alpha = 1.25$, which is a typical value encountered in computing and networking systems, there is approximately a 21,060-fold increase in expected waiting time caused by balancing the load instead of optimally unbalancing the load. At the optimal load point, the load is unbalanced very heavily toward the small server, with approximately 94.4% of the load directed toward that server. As α approaches 2, the ratio $\mathrm{E}[T_Q(1/2)]/\mathrm{E}[T_Q(q^*)]$ decreases. Similar results are seen when $\alpha < 1$, except that the load is unbalanced very heavily toward the large server. In fact, we see that

Table 3. Optimal Load Point, q^*, and Comparison of Optimal Value of $\mathrm{E}[T_Q(q^*)]$ Against $\mathrm{E}[T_Q(1/2)]$, for $\rho = 0.01$ (Low Load)

α	q^*	$\mathrm{E}[T_Q(q^*)]$	$\mathrm{E}[T_Q(1/2)]$	$\mathrm{E}[T_Q(1/2)]/\mathrm{E}[T_Q(q^*)]$
0.25	0.02898	52.99	7021	132.5
0.5	0.01783	372.6	2.094e+06	5619
0.75	0.05568	9110	1.919e+08	2.106e+04
1	0.5	1.114e+05	1.114e+05	1
1.25	0.9443	9110	1.919e+08	2.106e+04
1.5	0.9822	372.6	2.094e+06	5619
1.75	0.971	52.99	7021	132.5
2	0.8634	18.71	59.22	3.165

α and $2 - \alpha$, $0 < \alpha < 1$, give identical results except that if q^* denotes the optimal load point for α, then $1 - q^*$ is the optimal load point for $2 - \alpha$.

For a medium load ($\rho = 1.0$) and a higher load ($\rho = 1.6$), similar results comparing $\mathrm{E}[T_Q(q^*)]$ to $\mathrm{E}[T_Q(1/2)]$ are observed in Tables 4 and 5, respectively. The main difference is that the penalties for not unbalancing the load become smaller as the load increases. Note, for $\rho = 1.6$, that we require that $0.375 < q < 0.625$ for a stable system.

Table 4. Optimal Load Point, q^*, and Comparison of Optimal Value of $\mathrm{E}[T_Q(q^*)]$ Against $\mathrm{E}[T_Q(1/2)]$, for $\rho = 1.0$ (Medium Load)

α	q^*	$\mathrm{E}[T_Q(q^*)]$	$\mathrm{E}[T_Q(1/2)]$	$\mathrm{E}[T_Q(1/2)]/\mathrm{E}[T_Q(q^*)]$
0.25	0.09504	4.009e+04	1.397e+06	34.85
0.5	0.04497	3.948e+05	4.167e+08	1055
0.75	0.07851	4.62e+06	3.818e+10	8265
1	0.5	2.216e+07	2.216e+07	1
1.25	0.9215	4.62e+06	3.818e+10	8265
1.5	0.955	3.948e+05	4.167e+08	1055
1.75	0.905	4.009e+04	1.397e+06	34.85
2	0.7147	5487	1.178e+04	2.148

Table 5. Optimal Load Point, q^*, and Comparison of Optimal Value of $\mathrm{E}[T_Q(q^*)]$ Against $\mathrm{E}[T_Q(1/2)]$, for $\rho = 1.6$ (High Load)

α	q^*	$\mathrm{E}[T_Q(q^*)]$	$\mathrm{E}[T_Q(1/2)]$	$\mathrm{E}[T_Q(1/2)]/\mathrm{E}[T_Q(q^*)]$
0.25	0.3816	1.626e+06	5.589e+06	3.438
0.5	0.3754	3.552e+08	1.667e+09	4.692
0.75	0.3751	1.82e+10	1.527e+11	8.391
1	0.5	8.866e+07	8.866e+07	1
1.25	0.6249	1.82e+10	1.527e+11	8.391
1.5	0.6246	3.552e+08	1.667e+09	4.692
1.75	0.6184	1.626e+06	5.589e+06	3.438
2	0.5698	3.014e+04	4.714e+04	1.564

6. MORE THAN TWO HOSTS

We briefly outline how our results might be extended to more than two hosts. Suppose that there are $K > 2$ hosts in the system. In that case, there are $K + 1$ size cutoffs $c_0 \equiv s < c_1 < \cdots < c_{K-1} < c_K \equiv b$, so that a job in the interval $(c_{k-1}, c_k]$, $k = 1, \ldots, K$ is directed to host k (for completeness, a job of size s is directed toward host 1). By splitting incoming jobs into K independent streams and using the Pollaczek–Khinchine formula, Eq. (5) can be rewritten as a sum of K terms, with each term representing the weighted contribution to the mean waiting time from a different host. The normalized load on host k, q_k, is defined to be the load on host k, ρ_k, divided by the total load, ρ, that is, q_k is the fraction of load directed toward host k, with $q_k = 1/K$ representing the equally loaded case. The pair of imbalance conditions in Theorem 3.1 can then be replaced by $K - 1$ pairs of conditions. Theorem 4.1 can be modified to show that the host serving the largest jobs should be underloaded and Theorem 4.2 can be modified to show that the host serving the smallest jobs should be underloaded. Our current work is on developing this approach.

7. CONCLUSION

This article has examined load distribution in a two-host system employing size-interval task assignment. We prove very simple sufficient criteria ("imbalance conditions") for when the load should be unbalanced in favor of the short job host and for when the load should be unbalanced in favor of the long job host. These imbalance conditions depend on truncated moments of the job size distribution and derivatives thereof. We also prove a beautiful result showing that the direction of load imbalance can be determined directly by the moment index properties of the extended job size distribution. This result allowed us to immediately determine the direction of load imbalance for a wide range of bounded distributions including the bounded Pareto, Burr, log-Gamma, and bounded regularly varying distributions. Finally, we investigated the impact of *not* unbalancing the load, finding that balancing the load can result in a 100-fold increase in mean response time over load unbalancing.

Although the theorems in this article are limited to just two hosts, we have preliminary analytical results along the same lines for the case of three hosts, and we believe that we can extend these results even further. In addition, we are looking into generalizing the theorems herein to the case of heterogeneous hosts (with different speeds).

Notes

1. We define the support of the density, $f(t)$, as the region where $f(t)$ is strictly positive.
2. Here's an alternative explanation that might be preferable to some readers: Since $f(c)$ is nonzero for $s < c < b$, $q(c)$ is an increasing function of c for $s \leq c \leq b$. Hence, a unique inverse function $q^{-1}(q)$ can be defined such that if $q = q(c)$, then $c = q^{-1}(q)$. Using this inverse function, any function of c, for $s \leq c \leq b$, can be replaced by an equivalent function of the normalized load on the small host, q. In particular, define

the functions $\tilde{F}(q) = F(q^{-1}(q))$, $\tilde{m}_2(q) = m_2(q^{-1}(q))$, and $\tilde{h}(q) = h(q^{-1}(q))$. (In general, we will use the tilde symbol over a function (e.g., $\tilde{h}(q)$) to denote functions whose argument is the normalized load, q. Functions without the tilde symbol (e.g., $h(c)$) will denote functions whose argument is the cutoff, c).

3. Note: We have not included equations for M_2, $M_2(c)$, and $m_2(c)$ for the case $\alpha = 2$. We also define the function $c(q) = q^{-1}(q)$, which gives the size cutoff needed to achieve a given normalized load, q.

4. A more satisfactory description for some readers might be to consider repeated trials, Z_n, from distribution function $R(t)$ and let X equal the first value of Z_n that is less than or equal to b.

5. We give brief introductions to the α-stable and regularly varying distributions in Appendix B to help the reader who is not familiar with these distributions.

References

1. Bachmat, E. & Sarfati, H. (2008). Analysis of size interval task assignment policies. In *ACM SIGMETRICS Performance Evaluation Review*. New York: ACM, pp. 107–109.
2. Bachmat, E. & Sarfati, H. (2010). Analysis of SITA policies. *Performance Evaluation* 67(1): 102–120.
3. Cardellini, V., Colajanni, M. & Yu, P.S. (1999). Dynamic load balancing on web-server systems. *IEEE Internet Computing* 3(3): 28–39.
4. Ciardo, G., Riska, A., & Smirni, E. (2001). EQUILOAD: A load balancing policy for clustered web servers. *Performance Evaluation* 46(2–3): 101–124.
5. Crovella, M.E. & Bestavros, A. (1997). Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking* 5(6): 835–846.
6. Crovella, M., Harchol-Balter, M., & Murta, C. (1998). Task assignment in a distributed system: Improving performance by unbalancing load. In *Proceedings of ACM Sigmetrics '98 Conference on Measurement and Modeling of Computer Systems Poster Session,* Madison, WI.
7. Crovella, M.E., Taqqu, M.S., & Bestavros, A. (1998). Heavy-tailed probability distributions in the world wide web. In R. Adler, R. Feldman & M. Taqqu (eds.) *A Practical Guide to Heavy Tails*. New York: Chapman & Hall, pp. 1–23.
8. Daley, D.J. (2001). The moment index of minima. *Journal of Applied Probability* 38A: 33–36.
9. Embrechts, P., Kluppelberg, C., & Mikosch, T. (1997). *Modelling extremal events*. Berlin: Springer.
10. Feng, H., Misra, V., & Rubenstein, D. (2005). Optimal state-free, size-aware dispatching for heterogeneous $M/G/$-type systems. *Performance Evaluation* 62: 475–492.
11. Harchol-Balter, M. (2002). Task assignment with unknown duration. *Journal of the ACM* 49(2): 260–288.
12. Harchol-Balter, M., Crovella, M., & Murta, C. (1999). On choosing a task assignment policy for a distributed server system. *Journal of Parallel and Distributed Computing* 59(2): 204–228.
13. Harchol-Balter, M. & Downey, A. (1997). Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems* 15(3): 253–285.
14. Harchol-Balter, M. & Vesilo, R.A. (2009). Limit properties of probability densities characterized by moment indexes. Carnegie Mellon University Computer Science Technical Report CMU-CS-09-118. Pittsburgh, PA: Carnegie Mellon University.
15. Hwang, S. & Jung, N. (2002). Dynamic scheduling of web server cluster. In *Proceedings of the Ninth International Conference on Parallel and Distributed Systems (ICPADS'02)*, pp. 563–568.
16. Paxson, V. & Floyd, S. (1995). Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking* 3(3): 226–244.
17. Riska, A., Smirni, E., & Ciardo, G. (2000). Analytic modelling of load balancing policies for tasks with heavy-tailed distributions. In *Workshop on Software and Performance*, pp. 147–157.
18. Samorodnitsky, G. & Taqqu, M.S. (1994). *Stable non-Gaussian random processes*. New York: Chapman & Hall.
19. Schroeder, B. & Harchol-Balter, M. (2000). Evaluation of task assignment policies for supercomputer servers: The case for load unbalancing and fairness, In *Proceedings of the 9th IEEE Symposium on High Performance Distributed Computing*, pp. 211–220.

20. Shin, K.G. & Hou, C.J. (1994). Design and evaluation of effective load sharing in distributed real-time systems. *IEEE Transactions on Parallel and Distributed Systems* 5(7): 704–719.
21. Tari, Z., Broberg, J., Zomaya, A., & Baldoni, R. (2005). A least flow-time first load sharing approach for distributed server farm. *Journal of Parallel and Distributed Computing* 65(7): 832–842.
22. Ungureanu, V., Bradford, P.G., Katehakis, M., & Melamed, B. (2006). Deferred assignment scheduling in clustered web servers. *Cluster Computing* 9(1): 57–65.
23. Vesilo, R. (2008). Asymptotic analysis of load distribution for size-interval task allocation with bounded Pareto job sizes, In *Proceedings of the 14th IEEE International Conference on Parallel and Distributed Systems (ICPADS'08)*, 8–10 December, Melbourne, Australia, pp. 129–137.
24. Zolotarev, V.M. (1986). *One-dimensional stable distributions*. Translations of Mathematical Monographs Volume 65. Providence, RI: American Mathematical Society.

APPENDIX A

Derivatives of Truncated Moment Functions

Taking derivatives of $q(c)$ and $m_2(c)$ in Eq. (2) gives the following equations:

$$\frac{d}{dc}q(c) = \frac{cf(c)}{M_1} \quad \text{and} \quad \frac{d}{dc}m_2(c) = \frac{c^2 f(c)}{M_2}.$$

Since $f(c) > 0$ in the region of support (which ensures that the inverse function is uniquely defined), we have $q(c(q)) = q$. Taking derivatives of this identity, using the above results and applying the chain rule gives

$$\frac{d}{dq}\tilde{F}(q) = \frac{M_1}{c} \quad \text{and} \quad \frac{d}{dq}\tilde{m}_2(q) = \frac{cM_1}{M_2}.$$

From this, we see that as q increases, $d\tilde{m}_2(q)/dq$ increases and so $\tilde{m}_2(q)$ is a convex function of q. Since $\tilde{m}_2(0) = 0$ and $\tilde{m}_2(1) = 1$, it is straightforward to show that $\tilde{m}_2(q) \leq q$. Similarly, we can show that $\tilde{F}(q)$ is a concave function of q and $\tilde{F}(q) \geq q$.

APPENDIX B

Overview of Regularly Varying and α-Stable Distributions

A positive function, $k(t)$, is said to be regularly varying with index β, as $t \to \infty$ (we say $k(t) \in \mathcal{R}_\beta$), if, for $\lambda > 0$, $\lim_{t\to\infty} k(\lambda t)/k(t) = \lambda^\beta$. If $k(t) \in \mathcal{R}_\beta$, then it can be represented as $k(t) = L(t)t^\beta$, where $L(t)$ is a slowly varying function, defined by $\lim_{t\to\infty} L(\lambda t)/L(t) = 1$, $\lambda > 0$. Suppose X is a nonnegative random variable with distribution $R(t)$ function whose tail is regularly varying such that $1 - R(t) \in \mathcal{R}_{-\alpha}$ ($\alpha > 0$). The moments of X satisfy (see, e.g., Embrechts, Kluppelberg, and Mikosch [9, Prop A3.8(d)]) $\mathrm{E}[X]^p < \infty$ if $0 < p < \alpha$ and $\mathrm{E}[X]^p = \infty$ if $p > \alpha$. Hence, the moment index of X is α.

The density function $g(t; \alpha, \beta, \gamma, \mu)$, where $0 < \alpha \leq 2$ is the index of the distribution, $-1 \leq \beta \leq 1$ is the skewness parameter, $0 < \gamma$ is the scale parameter (or dispersion) and $-\infty < \mu < \infty$ is the location parameter, is said to be α-stable if it has the characteristic function (see Samorodnitsky and Taqqu [18])

$$\phi(x) = \int_{-\infty}^{\infty} e^{ixt} g(t;\alpha,\beta,\gamma,\mu)\,dt = \exp(ix\mu - |\gamma x^{\alpha}|(1 - i\beta\,\mathrm{sgn}(x))\Phi),$$

where $\mathrm{sgn}(x)$ is the sign of x and $\Phi = \tan(\pi\alpha/2)$, for $\alpha \neq 1$, and $\Phi = -(2/\pi)\log|t|)$, for $\alpha = 1$. The name "stable" comes from the following stability property for $g(t;\alpha,\beta,\gamma,\mu)$. If $X_i \sim g(t;\alpha,\beta,\gamma,\mu)$ for $1 \le i \le n$ and $n \ge 1$, then $\sum_{i=1}^{n} k_i(X_i - \mu)$ has density $g(t/k;\alpha,\beta,\gamma,0)/k$, where $k = (\sum_{i=1}^{n} |k_i|^{\alpha})^{1/\alpha}$. For the case $\mu = 0$, if $\alpha < 1$ and $\beta = 1$, then the support of $G(t)$ is the positive half-axis, and if $\alpha < 1$ and $\beta = -1$, then the support of $G(t)$ is the negative half-axis, otherwise, the support of $G(t)$ is the whole real line (e.g., see Zolotarev [24, Remark 4, pp. 79–80]). As a consequence, if X is a random variable with distribution function $G(t)$, there is the possibility of negative values of X, in general. It is for this reason that we use $|X|$ as the job size distribution. The moments of $|X|$ satisfy the following property (see Property 1.2.16 of Samorodnitsky and Taqqu [18]):

$$\mathrm{E}[|X|]^p < \infty, \qquad 0 < p < \alpha,$$
$$\mathrm{E}[|X|]^p = \infty, \qquad p \ge \alpha.$$

Hence, the moment index of $|X|$ is α.

APPENDIX C

Proof of Theorem 4.1

We begin by introducing some notation needed in the proof. Define $c_q(b)$ to be the size cutoff needed to give a normalized load, q, on the small host, when the maximum job size is $b < \infty$; that is, $c_q(b)$ is defined by

$$q = \frac{\int_s^{c_q(b)} tr(t)\,dt}{\int_s^b tr(t)\,dt}. \qquad \textbf{(C.1)}$$

Analogous to Eq. (C.1), define the cutoff point, $c_q(\infty)$, when the normalized load is q and the bounding level is ∞, by the following:

$$q = \frac{\int_s^{c_q(\infty)} tr(t)\,dt}{\int_s^{\infty} tr(t)\,dt}. \qquad \textbf{(C.2)}$$

Note that $\mathrm{E}[Z] < \infty$ ensures in Eq. (C.2) that $c_q(\infty)$ is finite for $0 \le q < 1$.

Define parameterized versions of the truncated moments, when the maximum job size is b, by

$$M_1(b) = \int_s^b tr(t)\,dt, \qquad M_2(b) = \int_s^b t^2 r(t)\,dt,$$

and define parameterized versions of the truncated moment functions, when the normalized load on the small host is q and the maximum job size is b, by

$$\tilde{F}(q,b) = \frac{\int_s^{c_q(b)} r(t)\,dt}{\int_s^b r(t)\,dt}, \qquad \tilde{m}_2(q,b) = \frac{\int_s^{c_q(b)} t^2 r(t)\,dt}{M_2(b)}.$$

To prove the theorem, we will use the moment conditions in the theorem statement to show, for large enough b, that the following imbalance conditions are satisfied:

$$\tilde{m}(1/2, b) - (1 - \tilde{F}(1/2, b)) < 0, \tag{C.3}$$

$$\tilde{m}_2'(1/2, b) - \tilde{F}'(1/2, b) < 0. \tag{C.4}$$

Inequalities (C.3) and (C.4) fulfill the conditions for Theorem 3.1 and so the optimal load point of the original system equation satisfies $q^* \geq 1/2$.

We first prove inequality (C.3) and then we prove inequality (C.4).

The idea behind proving inequality (C.4) is as follows. Because $\mathrm{E}[Z] < \infty$, the integral $\int_s^b tr(t)\,dt$ converges to a finite limit as b approaches infinity. We use this to show that $c_{1/2}(b)$ will be bounded above, *for all values* of $b > 0$. As a consequence, $1 - \tilde{F}(1/2, b)$ is lower bounded by a nonzero value for all $b > 0$. On the other hand, in the expression $\tilde{m}_2(1/2, b) = \int_s^{c_{1/2}(b)} t^2 r(t)\,dt / \int_s^b t^2 r(t)\,dt$, the denominator increases without bound because $\mathrm{E}[Z^2] = \infty$, but the numerator will be bounded above. Hence, $\tilde{m}_2(1/2, b)$ approaches zero as b increases. This means that for large b, we expect $\tilde{m}(1/2, b) < 1 - \tilde{F}(1/2, b)$. We now make this this intuition precise.

We will now show that $c_q(b)$ is bounded above for $0 \leq q < 1$ and for all values of $b > 0$. From Eq. (C.2) we get

$$0 = \frac{\int_s^{c_q(\infty)} tr(t)\,dt}{\int_s^\infty tr(t)\,dt} - q = \frac{1}{\int_s^\infty tr(t)\,dt}\left[\int_s^{c_q(\infty)} tr(t)\,dt - q\int_s^\infty tr(t)\,dt\right]. \tag{C.5}$$

Because $\mathrm{E}[Z] = \int_s^\infty tr(t) < \infty$, we can multiply Eq. (C.5)) by $\int_s^\infty tr(t) < \infty$ to give

$$0 = \int_s^{c_q(\infty)} tr(t)\,dt - q\int_s^\infty tr(t)\,dt.$$

Adding and subtracting $\int_s^{c_q(b)} tr(t)\,dt$ in this equation and rearranging gives

$$\begin{aligned}
0 &= \int_s^{c_q(\infty)} tr(t)\,dt - \int_s^{c_q(b)} tr(t)\,dt + \int_s^{c_q(b)} tr(t)\,dt - q\int_s^\infty tr(t)\,dt \\
&= \int_{c_q(b)}^{c_q(\infty)} tr(t)\,dt + q\left(\frac{1}{q}\int_s^{c_q(b)} tr(t)\,dt - \int_s^\infty tr(t)\,dt\right) \\
&= \int_{c_q(b)}^{c_q(\infty)} tr(t)\,dt + q\left(\frac{\int_s^b tr(t)\,dt}{\int_s^{c_q(b)} tr(t)\,dt}\int_s^{c_q(b)} tr(t)\,dt - \int_s^\infty tr(t)\,dt\right) \\
&= \int_{c_q(b)}^{c_q(\infty)} tr(t)\,dt - q\int_b^\infty tr(t)\,dt.
\end{aligned} \tag{C.6}$$

Since $\lim_{b\to\infty} \int_b^\infty r(t) = 0$, it follows from Eq. (C.6) that

$$\lim_{b\to\infty} \int_{c_q(b)}^{c_q(\infty)} tr(t)\,dt = 0. \tag{C.7}$$

We conclude from Eq. (C.7), together with $c_q(\infty) < \infty$, that $c_q(b)$ is bounded above. Suppose the contrary. In that case, $\limsup_{b\to\infty} c_q(b) = \infty$. This means there is a sequence of points

$\{b_n, n = 1, 2, \ldots\}$ such that $\lim_{n\to\infty} b_n = \infty$ for which $\lim_{n\to\infty} c_q(b_n) = \infty$. Since $r(t)$ has support (s, ∞), that must entail that $\lim_{n\to\infty} \int_{c_q(b_n)}^{c_q(\infty)} tr(t)\,dt < 0$, which contradicts Eq. (C.7).

Therefore, there is a fixed bound, $\hat{c}_q < \infty$ such that $c_q(b) \le \hat{c}_q$, for $b > 0$. Applying this bound to the definition of $\tilde{F}(q,b)$ gives

$$1 - \tilde{F}(q,b) = \frac{\int_{c_q(b)}^{b} r(t)\,dt}{\int_s^b r(t)\,dt} \ge \frac{\int_{\hat{c}_q}^{b} r(t)\,dt}{\int_s^b r(t)\,dt} = \frac{\int_{\hat{c}_q}^{\infty} r(t)\,dt - \int_b^{\infty} r(t)\,dt}{\int_s^b r(t)\,dt}. \qquad \textbf{(C.8)}$$

Since $\int_s^b r(t)\,dt \le 1$, it follows from Eq. (C.8) that

$$1 - \tilde{F}(q,b) \ge \int_{\hat{c}_q}^{\infty} r(t)\,dt - \int_b^{\infty} r(t)\,dt. \qquad \textbf{(C.9)}$$

Since $\int_{\hat{c}_q}^{\infty} r(t)\,dt > 0$ and $\int_b^{\infty} r(t)\,dt$ approaches zero, we conclude from Eq. (C.9) that there is a value $b_0(q)$ such that

$$1 - \tilde{F}(q,b) \ge \frac{1}{2}\int_{\hat{c}_q}^{\infty} r(t)\,dt > K_q > 0\,, \qquad b \ge b_0(q),$$

for some constant $K_q > 0$. On the other hand, applying the bound $c_q(b) \le \hat{c}_q$ to the definition of $\tilde{m}_2(q,b)$ gives

$$\tilde{m}_2(q,b) = \frac{\int_s^{c_q(b)} t^2 r(t)\,dt}{\int_s^b t^2 r(t)\,dt} < \frac{\int_s^{\hat{c}_q} t^2 r(t)\,dt}{\int_s^b t^2 r(t)\,dt}.$$

Since $\mathrm{E}[Z^2] = \infty$, it follows that $\int_s^b t^2 r(t)\,dt \to \infty$, and so we obtain

$$\tilde{m}_2(q,b) < \frac{\int_s^{\hat{c}_q} t^2 r(t)\,dt}{\int_s^b t^2 r(t)\,dt} \to 0 \qquad (b \to \infty). \qquad \textbf{(C.10)}$$

Hence, setting $q = 1/2$ in Eqs. (C.9) and (C.10), we can find a b large enough such that $\tilde{m}_2(1/2,b) - (1 - \tilde{F}(1/2,b)) < 0$.

To prove the imbalance property given in Eq. (C.4), take derivatives of $\tilde{F}(q,b)$ and $\tilde{m}_2(q,b)$ and use the results in Appendix A to give

$$\tilde{F}'(q,b) - \tilde{m}_2'(q,b) = \frac{M_1(b)}{c_q(b)}\left(1 - \frac{c_q(b)^2}{M_2(b)}\right). \qquad \textbf{(C.11)}$$

Applying the upper bound $c_q(b) \le \hat{c}_q$ gives

$$\tilde{F}'(q,b) - \tilde{m}_2'(q,b) \ge \frac{M_1(b)}{c_q(b)}\left(1 - \frac{\hat{c}_q^2}{M_2(b)}\right). \qquad \textbf{(C.12)}$$

Since $\hat{c}_q$ is a fixed value and $M_2(b) = \int_s^b t^2 r(t)\,dt$ increases to infinity, a large enough value of b can be found such that the right-hand side of Eq. (C.12) is greater than zero. Setting $q = 1/2$ proves the result.

APPENDIX D

Proof of Theorem 4.2

This appendix is devoted to proving Theorem 4.2. We require the following lemmas, which are proved in Harchol-Balter and Vesilo [14].

LEMMA D.1: *Suppose that the random variable Z has a density function $r(t)$ that is ultimately decreasing and that* $\mathrm{E}[Z] = \infty$. *Then*

$$(i) \qquad \lim_{b\to\infty} \frac{\int_s^b tr(t)\,dt}{b} = 0, \tag{D.1}$$

$$(ii) \qquad \lim_{b\to\infty} \frac{\int_s^b t^2 r(t)\,dt}{b^2} = 0. \tag{D.2}$$

LEMMA D.2: *Suppose that the random variable Z has a density function $r(t)$ that is differentiable and ultimately decreasing, that Z has moment index $0 < \kappa_Z < 1$, and that $\lim_{b\to\infty} -br'(b)/r(b)$ exists. Then*

$$(i) \qquad \lim_{b\to\infty} b^2 r(b) = \infty, \tag{D.3}$$

$$(ii) \qquad \text{if } \gamma < \kappa_Z, \text{ then } b^{\gamma+1} r(b) \quad \text{ultimately decreases monotonically to zero,} \tag{D.4}$$

$$(iii) \qquad \lim_{b\to\infty} \frac{\int_s^b tr(t)\,dt}{b^2 r(b)} = \frac{1}{(1-\kappa_Z)}, \tag{D.5}$$

$$(iv) \qquad \lim_{b\to\infty} \frac{b\int_b^\infty r(t)\,dt}{\int_s^b tr(t)\,dt} = \frac{1-\kappa_Z}{\kappa_Z}. \tag{D.6}$$

Additionally, there exists a fixed value $\hat{r} > 0$ such that

$$1 > c_{1/2}(b)/b > \hat{r} > 0 \qquad \text{for } b \text{ large enough,} \tag{D.7}$$

and the following limit applies:

$$\lim_{b\to\infty} \frac{2b\int_s^{c_{1/2}(b)} t^2 f(t)\,dt}{c_{1/2}(b)\int_s^b t^2 f(t)\,dt} = 1. \tag{D.8}$$

PROOF OF THEOREM 4.2: To prove the theorem we will show, for b large enough, that

$$\tilde{m}(1/2, b) - (1 - \tilde{F}(1/2, b)) > 0, \tag{D.9}$$

$$\tilde{m}_2'(1/2, b) - \tilde{F}'(1/2, b) > 0. \tag{D.10}$$

Inequalities (D.9) and (D.10) fulfill the conditions for Theorem 3.1 and so the optimal point of the original system equation satisfies $q^* < 1/2$. We do this by showing that $\tilde{m}_2(1/2, b)$ is lower bounded by a positive (nonzero) value.

Since $r(t)$ is a probability density with support (s,∞), we have, for b large enough, that $\int_s^b r(t)\,dt > 1/2$. Applying this inequality to the definition of $\tilde{F}(q,b)$ gives the bounds

$$1-\tilde{F}(q,b)=\frac{\int_{c_q(b)}^{b} r(t)\,dt}{\int_s^b r(t)\,dt}<\frac{\int_{c_q(b)}^{b} r(t)\,dt}{1/2}<2\int_{c_q(b)}^{\infty} r(t)\,dt. \tag{D.11}$$

The last integral in Eq. (D.11) satisfies the following identity:

$$\int_{c_q(b)}^{\infty} r(t)\,dt=\left(\frac{c_q(b)\int_{c_q(b)}^{\infty} r(t)\,dt}{\int_s^{c_q(b)} tr(t)\,dt}\right)\left(\frac{\int_s^{c_q(b)} tr(t)\,dt}{\int_s^b tr(t)\,dt}\right)\left(\frac{b}{c_q(b)}\right)\left(\frac{\int_s^b tr(t)\,dt}{b}\right).$$

Set $q=1/2$ in this identity and insert it into Eq. (D.11) to give the bound

$$1-\tilde{F}(1/2,b)<\frac{1}{2}\left(\frac{c_{1/2}(b)\int_{c_{1/2}(b)}^{\infty} r(t)\,dt}{\int_s^{c_{1/2}(b)} tr(t)\,dt}\right)\left(\frac{b}{c_{1/2}(b)}\right)\left(\frac{\int_s^b tr(t)\,dt}{b}\right).$$

Combining this bound with the definition of $\tilde{m}_2(q,b)$ gives the following bound for $\tilde{m}(1/2,b)-(1-\tilde{F}(1/2,b))$:

$$\begin{aligned}
&\tilde{m}(1/2,b)-(1-\tilde{F}(1/2,b))\\
&\quad>\frac{\int_s^{c_{1/2}(b)} t^2 f(t)\,dt}{\int_s^b t^2 f(t)\,dt}-\frac{1}{2}\left(\frac{c_{1/2}(b)\int_{c_{1/2}(b)}^{\infty} r(t)\,dt}{\int_s^{c_{1/2}(b)} tr(t)\,dt}\right)\left(\frac{b}{c_{1/2}(b)}\right)\left(\frac{\int_s^b tr(t)\,dt}{b}\right)\\
&\quad=\frac{c_{1/2}(b)}{2b}\left[\frac{2b\int_s^{c_{1/2}(b)} t^2 f(t)\,dt}{c_{1/2}(b)\int_s^b t^2 f(t)\,dt}-\left(\frac{c_{1/2}(b)\int_{c_{1/2}(b)}^{\infty} r(t)\,dt}{\int_s^{c_{1/2}(b)} tr(t)\,dt}\right)\left(\frac{b}{c_{1/2}(b)}\right)^2\left(\frac{\int_s^b tr(t)\,dt}{b}\right)\right].
\end{aligned} \tag{D.12}$$

Since $c_b(1/2)\to\infty$, we have from Eq. (D.6), that

$$\lim_{b\to\infty}\frac{c_{1/2}(b)\int_{c_{1/2}(b)}^{\infty} r(t)\,dt}{\int_s^{c_{1/2}(b)} tr(t)\,dt}=\frac{1-\kappa_Z}{\kappa_Z}. \tag{D.13}$$

Inserting Eqs. (D.7), (D.8), and (D.13) into Eq. (D.12) gives the lower bound

$$\lim_{b\to\infty}\left[\tilde{m}(1/2,b)-(1-\tilde{F}(1/2,b))\right]>\frac{\hat{r}}{2}\left[1-\left(\frac{1-\kappa_Z}{\kappa_Z}\right)\left(\frac{1}{\hat{r}}\right)^2\left(\frac{\int_s^b tr(t)\,dt}{b}\right)\right].$$

However, by Eq. (D.1), $\lim_{b\to\infty}\int_s^b tr(t)\,dt/b=0$ and so $\lim_{b\to\infty}\left[\tilde{m}(1/2,b)-(1-\tilde{F}(1/2,b))\right]>0$. Hence, for b large enough, Eq. (D.9) is satisfied.

To prove Eq. (D.9), we use Eq. (C.11). Inserting bound $c_{1/2}(b)>\hat{r}b$ into this equation gives

$$\tilde{m}_2'(1/2,b)-\tilde{F}_2'(1/2,b)>\frac{M_1(b)}{c_{1/2}(b)}\left(\frac{b^2\hat{r}^2}{M_2(b)}-1\right). \tag{D.14}$$

From Eq. (D.2), $\lim_{b\to\infty} b^2/M_2(b)=\infty$ and so, for b large enough, the right-hand side of Eq. (D.14) can be made positive, proving Eq. (D.10). ■